DOCUMENT RESUME

ED 359 575 CS 508 213

AUTHOR	Fowler, Carol A., Ed.
TITLE	Speech Research Status Report, July-December 1992.
INSTITUTION	Haskins Labs., New Haven, Conn.
REPORT NO	SR-111/112
PUB DATE	92
NOTE	344p.; For the January-June 1992 report, see ED 352 694.
PUB TYPE	Collected Works - General (020) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC14 Plus Postage.
DESCRIPTORS	Articulation (Speech); Auditory Discrimination; Communication Research; Hebrew; Language Research; Metalinguistics; *Music; *Phonology; Primary Education; Reading Comprehension; Reading Processes; *Speech Communication; Stuttering; *Word Recognition; *Written Language
IDENTIFIERS	Phonological Processing; *Speech Research

ABSTRACT

One of a series of semi-annual reports, this publication contains 25 articles which report the status and progress of studies on the nature of speech, instruments for its investigation, and practical applications. Articles are as follows: "Acoustic Shards, Perceptual Glue" (Robert E. Remez and Philip E. Rubin); "F0 Gives Voicing Information Even with Unambiguous VOTs" (Doug Whalen); "Articulatory Phonology: An Overview" (Catherine P. Browman and Louis Goldstein); "Acoustic Evidence for Gestural Overlap in Consonant Sequences" (Elizabeth C. Zsiga); "Acoustic Evidence for the Development of Gestural Coordination in the Speech of 2-Year-Olds: A Longitudinal Study" (Elizabeth Whitney Goodell and Michael Studdert-Kennedy); "Gestures, Features, and Segments in Early Child Speech" (Michael Studdert-Kennedy and Elizabeth Whitney Goodell); "An Aerodynamic Evaluation of Parkinsonian Dysarthria: Laryngeal and Supralaryngeal Manifestations" (L. Carol Gracco and others); "Effects of Alterations in Auditory Feedback and Speech Rate on Stuttering Frequency" (Joseph Kalinowsky and others); "Phonetic Recoding of Phonologically Ambiguous Printed Words" (Ram Frost and Michael Kampf); "Reading Consonants and Guessing Vowels: Visual Word Recognition in Hebrew Orthography" (Ram Frost and Shlomo Bentin); "The Reading Process Is Different for Different Orthographies: The Orthographic Depth Hypothesis" (Leonard Katz and Ram Frost); "An Examination of 'The Simple View of Reading'" (Lois G. Dreyer and Leonard Katz); "Phonological Awareness, Reading, and Reading Acquisition: A Survey and Appraisal of Current Knowledge" (Shlomo Bentin); "Morphological Analysis in Word Recognition" (Laurie B. Feldman and Darinka Andjelkovic); "Can Theories of Word Recognition Remain Stubbornly Nonphonological?" (Claudia Carello and others); "Poor Readers Are Not Easy to Fool: Comprehension of Adjectives with Exceptional Control Properties" (Paul Macaruso and others); "A Review of Daniel Reisberg (Ed.), 'Auditory Imagery'" (Bruno H. Repp); "A Review of Mari Reiss Jones and Susan Holleran (Eds.), 'Cognitive Bases of Musical Communication'" (Bruno H. Repp); "Diversity and Commonality in Music Performance; An Analysis of Timing Microstructure in Schumann's 'Traumerei'" (Bruno H. Repp); "A Review of 'Einfuhrung in die deutsche Phonetik' by Ursula Hirschfeld" (Bruno H. Repp); "Music as Motion: A Synopsis of Alexander Truslit's (1938)

'Gestaltung und Bewegung in der Musik'" (Bruno H. Repp); "Objective Performance Analysis as a Tool for the Musical Detective" (Bruno H. Repp); "Some Empirical Observations on Sound Level Properties of Recorded Piano Tunes" (Bruno H. Repp); "Probing the Cognitive Representation of Musical Time: Structural Constraints on the Perception of Timing Perturbations" (Bruno H. Repp); and "A Review of Yoh'ichi Tohkura, Eric Vatikiotis-Bateson, and Yoshinori Sagisaka (Eds.), 'Speech Perception, Production and Linguistic Structure'" (Bruno H. Repp). (RS)

Haskins Laboratories Status Report on

Speech Research

SR-111/112

JULY-DECEMBER 1992

Haskins
Laboratories
Status Report on

Speech Research

SR-111/112
JULY-DECEMBER 1992

NEW HAVEN, CONNECTICUT

Distribution Statement

Editor
Carol A. Fowler

Production
Yvonne Manning
Fawn Zefang Wang

This publication reports the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications.

Distribution of this document is unlimited. The information in this document is available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.

Correspondence concerning this report should be addressed to the Editor at the address below:

Haskins Laboratories
270 Crown Street
New Haven, Connecticut
06511-6695

Phone: (203) 865-6163 *FAX:* (203) 865-8963 *Bitnet:* HASKINS@YALEHASK
Internet: HASKINS%YALEHASK@VENUS.YCC.YALE.EDU

This Report was reproduced on recycled paper

Acknowledgment

The research reported here was made possible in part by support from the following sources:

National Institute of Child Health and Human Development

Grant HD-01994
Grant HD-21888

National Institute of Health

Biomedical Research Support Grant RR-05596

National Scierce Foundation

Grant DBS-9112198

National Institute on Deafness and Other Communication Disorders

Grant DC 00121
Grant DC 00183
Grant DC 00403
Grant DC 00016
Grant DC 00594
Grant DC 00865
Grant DC 01147
Grant DC 00044
Grant DC 00825
Grant DC 01247

Investigators

- Arthur Abramson*
- Eric Bateson*
- Fredericka Bell-Berti*
- Catherine T. Best*
- Susan Brady*
- Catherine P. Browman
- Claudia Carello*
- Franklin S. Cooper*
- Stephen Crain*
- Lois G. Dreyer*
- Alice Faber
- Laurie B. Feldman*
- Janet Fodor*
- Anne Fowler*
- Carol A. Fowler*
- Louis Goldstein*
- Carol Gracco
- Vincent Gracco
- Katherine S. Harris*
- John Hogden
- Leonard Katz*
- Rena Arens Krakow*
- Andrea G. Levitt*
- Alvin M. Liberman*
- Diane Lillo-Martin*
- Leigh Lisker*
- Anders Löfqvist
- Ignatius G. Mattingly*
- Nancy S. McGarr*
- Richard S. McGowan
- Patrick W. Nye
- Kiyoshi Oshima†
- Kenneth Pugh*
- Lawrence J. Raphael*
- Bruno H. Repp
- Hyla Rubin*
- Philip E. Rubin
- Elliot Saltzman
- Donald Shankweiler*
- Jeffrey Shaw
- Rudolph Sock‡
- Michael Studdert-Kennedy*
- Michael T. Turvey*
- Douglas Whalen

Technical Staff

- Michael D'Angelo
- Vincent Gulisano
- Donald Hailey
- Maura Herlihy
- Marion MacEachron
- Yvonne Manning
- William P. Scully
- Fawn Zefang Wang
- Edward R. Wiley

Administrative Staff

- Philip Chagnon
- Alice Dadourian
- Betty J. DeLise
- Lisa Fresa
- Joan Martinez

Students*

- Melanie Campbell
- Sandra Chiang
- Margaret Hall Dunn
- Terri Erwin
- Joseph Kalinowski
- Laura Koenig
- Betty Kollia
- Simon Levy
- Salvatore Miranda
- Maria Mody
- Weijia Ni
- Mira Peter
- Christine Romano
- Joaquin Romero
- Maria Rosetti
- Dorothy Ross
- Arlyne Russo
- Michelle Sancier
- Sonya Sheffert
- Caroline Smith
- Brenda Stone
- Mark Tiede
- Qi Wang
- Yi Xu
- Elizabeth Zsiga

*Part-time
†Visiting from University of Tokyo, Japan
‡Visiting from University of Grenoble, France

Contents

Haskins Laboratories Status Report on

Speech Research

Acoustic Shards, Perceptual Glue*

Robert E. Remez† and Philip E. Rubin

First, think of something poetic to say. Then, find the words and syntax to convey the message with clarity and wit. To give voice to the words, convert them to a series of consonants and vowels, and produce the associated gestures of articulation. Don't worry about acoustic propagation—the compliance of the atmosphere will accomplish that, conveying the signal to the listener. Your conversational partner will find acoustic attributes within the signal that distinguish consonants and vowels, will reprise the segmental series, and from it, will apprehend the words, syntax and meaning of your utterance. ("Tell me, O Muse, of the man of many ways...")

SPEECH AS PATCHWORK AND AS STREAM

The exposition of perception that derives from this setting of the speech chain emphasizes the differential value of elements within the acoustic spectrum. In fact, the speech signal is a patchwork of acoustic elements—whistle, click, buzz, hiss and hum. Given this motley assortment, it seems self-evident that perceivers of speech attend to the elemental attributes of the moment, that is, to the frequency of hiss, to the duration of hum, to the spectrum of click, thereby to determine the point of closure of the vocal tract, the placement of the constriction, the movement of the tongue; or, the consonants and vowels. Much effort has been devoted to this analysis of speech perception, and the pursuit of the distinctive acoustic cues has held the center of the field for twenty-five years. We learned that the acoustic elements are easily registered sometimes, and that auditory properties are converted to phonetic properties with little lost in translation (Massaro, 1987). In other circumstances, subtle (and not so subtle) physical properties of the acoustic signal are not represented isomorphically in perception, the departures from parity motivated phonetically (Fitch, Halwes, Erickson, & Liberman, 1980) and even lexically (Ganong, 1980). It is the objective of a substantial portion of contemporary research on speech perception to rationalize these two types of attention to specific acoustic cues in various combinations and assortments.

Rather less theoretical attention is usually paid to two key aspects of the perception of these patchwork patterns. The first can be thought of as the means by which the perceiver finds a speech signal to analyze. From the welter of diverse acoustic elements that strike the ear, the perceiver implicitly sorts acoustic components into those which belong to the talker's signal and those which are extraneous to it. This act of perceptual grouping must fuse the acoustically dissimilar components of the speech signal—clicks conjoined to whistles, and to hisses, and to buzzes, and to hums—which is quite without a convincing account in the literature on auditory perceptual organization (Remez, 1987). With the exception of speech signals, agglutination of auditory sensory elements usually occurs because of their infinitesimal if multidimensional physical similarity (Bregman, 1990). How does the perceiver even take a single speech signal uttered in a quiet room to be the product of a single source of sound? Familiar accounts of speech perception are silent on the issue of perceptual organization; and, general auditory treatments of perceptual organization only attempt to broach physically simple signals. In consequence, we are left with no explanation of this first step in the perception of speech.

We wish to express our thanks to Stefanie Berns, Louis Goldstein. Ignatius Mattingly. Lynne Nygaard. David Pisoni. and the editors for compassionate advice and fearless encouragement. This research was supported by grants from NIDCD (00308) to R. E. Remez, and from NICHHD (01994) to Haskins Laboratories.

The second aspect of speech perception that has received only glancing attention is the accommodation to indefinite variability in the spectral complexion of the acoustic elements composing natural utterances. Sources of variability are much discussed (for example, Ladefoged, 1967; Liberman & Cooper, 1972; Oshika, Zue, Weeks, Neu, & Aurbach, 1975) and are attributed to linguistic, anatomical and circumstantial factors that modulate the particular assortment of acoustic vehicles for consonants and vowels. Lacking in these discussions of the origins of variability is a specific account of the perceptual means by which this variability is tolerated, whatever its sources. To extrapolate from the familiar characterizations of pe eption, it seems that the perceiver, who is bound to attend to acoustic elements, apparently never encounters even one of them a second time. How does perception contend with this variability? We might imagine that perceivers anticipate variation through intimate familiarity with the sources driving it, though we might just as well imagine that perceivers anticipate the messages, and save ourselves the headache of explaining the parts.

To face the issues of perceptual organization and indefinite variability in the case of speech, we have been studying the perception of acoustic continuity. Our findings reveal that perceivers are acutely sensitive to coarse-grain properties of speech signals, an acuity that stands apart from proficiency in cue trading. Our initial premise is that the speech signal is not only a kind of patchwork, it is also a stream. The acoustic properties of the stream are forged by the nearly continuous gestures of the tongue, lips, jaw, velum, and larynx. In our studies of perceptual sensitivity to properties of acoustic continuity (Remez & Rubin, 1984, 1990, 1991; Remez, Rubin, Nygaard, & Howell, 1987; Remez, Rubin, Pisoni, & Carrell, 1981) we used signals consisting of three time-varying sinusoids, each of which was varied in a formant-like manner. We fabricated each sinusoidal pattern by computing the resonant center-frequencies of a natural utterance, using the analysis technique of linear prediction. The table of values produced through this analysis was used to set frequency and amplitude changes of three or four pure tones, reproducing the coarse-grain properties of the oral, nasal and fricative formants. Sinusoidal tone-complexes lack fundamental frequency, harmonic spectrum, and broadband formants (the short-term characteristics of natural speech), and therefore lack the acoustic elements on which most accounts of speech perception rest. In consequence, the time-varying properties of a sinewave pattern, specifically the coherence of the changes of the energy peaks over time, replicate natural spectral variation without also replicating the fine-grain acoustic structure typical of vocally produced sound.

Perception of Sinusoidal Signals

The perceptual effects of sinewave stimuli were easy to predict. Because the short-term spectra of three-tone signals differ drastically from natural and even synthetic speech; because no talker is capable of producing three simultaneous "whistles" with these bandwidths, in this frequency range (Busnel & Classe, 1976); and because the frequency and amplitude changes of the tones are not synchronized, the perceiver should hear three independent streams, one for each sinusoid. The perceiver should hear no phonetic qualities.

However obvious this prediction seemed, there was an equally plausible, though contrasting, prediction. Suppose the listener were able to disregard the short-term dissimilarity of a sinusoidal signal and natural speech, and could attend, instead, to the overall pattern of change of the three tones. The pattern of change of the frequency peaks resembles the resonance changes produced by a vocal tract when articulating speech. If the listener can apprehend this coherence in the time-varying properties of the nonspeech signal, then perception will resolve a phonetic message spoken by an impossible voice.

Given nonspeech stimuli whose time-varying properties are polyphonic in detail yet abstractly vocal, listeners perceived the signals in both of the ways we predicted. Those listeners who were told nothing about the stimuli heard science fiction sounds, electronic music, sirens, computer bleeps, and radio interference. Those listeners who instead were instructed to transcribe a "strangely synthesized English sentence" did exactly that—they reported the unnatural "voice" quality of the patterns, but transcribed the patterns as they would have the original natural utterances upon which we based the sinewave stimuli (Remez et al., 1981).

These studies indicate that speech perception is possible despite drastic departures from the short-term spectra of natural speech (despite absence of broadband formants, harmonic spectrum, and fundamental frequency) insofar as the time-varying properties of speech signals are preserved.

Remarkably, the listener is able to attend to the coherent time-varying properties of the acoustic pattern despite the inappropriateness of the acoustic carrier undergoing the change.

How were listeners able to understand the linguistic message borne by such anomalous and literally incoherent acoustic vehicles? Our hypothesis is that sinewave replicas preserve the spectrotemporal aspects of the speech signal that are critical in ordinary perception. The tone complexes that exhibit this typical kind of acoustic structure are only superficially unfamiliar, anomalous or incoherent. Though in detail each tone is dissimilar to the others, sufficiently so to warrant perceptual fission of each into a separate perceptual stream (see Remez, Rubin, Berns, Pardo, & Lang, 1992), the listener is also able to detect the acoustic products of a phonetically governed vocal source in the spectral changes of tone complexes. That this should occur is a kind of evidence that the sensitivity underlying phonetic perception is keyed to time-critical properties of spectral variation as much as it is to the detailed auditory effects of the elements of stimulation.

Our objective in reviewing three experiments here is to expose a bit more of what we have learned about the nature of time-varying sensitivity in speech perception. The first setting is a study that confirms the time-critical nature of the information available to perceivers in sinewave replicas. Lacking the short-term properties of speech makes these signals seem nothing like vocal sound, despite the easy impressions of consonants and vowels that they evoke. We sought to test the hypothesis directly that segmental perception depended on natural values of spectrotemporal change. In the second experiment that we review here, we showed the potential for assessing sensitivity to time-critical changes using synthetic speech. In that study, we were able to see perceptual effects of spectrotemporal variation with speechlike short-term spectra. That study promises to permit evaluation of the interaction of coarse- and fine grain effects in speech perception. Last, we describe an experiment on the issue of perceptual plasticity by which listeners come to favor time-varying properties and to ignore the anomalous and persistent short-term properties of sinewave replicas. This study ruled out a plausible counter-explanation to the one that we originally proposed.

TEMPORAL VARIATION

The first experiment tested a hypothesis critical to this explanation of the phonetic perception of tonal analogs of speech. We have claimed that phonetic perception includes the attention to coherent patterns of change in acoustic energy, and does not rely exclusively on auditory evaluation of the particular qualities of the successive, discrete acoustic elements that compose the signal (in contrast to Elman & McClelland, 1985; Massaro, 1987; Zue & Schwartz, 1980, for example). In the case of sinusoids perceived phonetically, we specifically claim that the phonetic information is conveyed by time-varying rather than momentary sensory properties. Our test manipulated the temporal coordinates of sinusoidal replicas of speech directly, in search of perceptual effects contingent on spectrotemporal attributes. Our hypothesis predicted that: 1) listeners would be able to perceive speech from signals that lacked the short-term properties of speech—our sinusoidal imitation—as long as they preserved time-variation on a natural scale; but, 2) listeners would be unable to perceive speech from signals that presented neither the short-term nor the time-varying properties of speech. Our measure here was transcription performance. We omitted short-term properties and conjointly preserved time-varying properties by imitating natural speech with three-tone signals. And, we preserved neither short-term nor time-varying properties by presenting three-tone patterns synthesized with inappropriate temporal properties.

As a control, each sinusoidal sentence was matched with a sentence produced through speech synthesis in which formant frequency and amplitude values were identical to those of the sinusoidal items. This control established a baseline for estimating the effectiveness of time-varying phonetic information by presenting both the short-term and, presumably, the time-varying structure of natural speech. The conditions in which the temporal coordinates of the synthetic speech items departed from natural rates of change permit us to test the ability of the perceiver to compensate for defective time-variation by using short-term spectral properties, which have the actual sound of speech if not its natural spectrotemporal values.[1]

Our test of the necessity and sufficiency of short-term and time-varying acoustic structure used a version of the method of limits, in which temporal variants of synthetic and sinewave sentences were presented to listeners for transcription. We used three English sentences, synthesized at a 10 ms frame rate using both the sinewave technique and a conventional serial terminal-analog speech synthesizer (Mattingly, Pollock, Levas, Scully, & Levitt, 1981). Temporal variants of these sentences were made by changing the frame rate of sinewave and speech synthesis. This yielded ten versions of each original sentence, synthesized at frame rates of 1 ms, 2 ms, 3 ms, 4 ms, 5 ms, 10 ms, 20 ms, 40 ms, 60 ms, and 80 ms. Listening sessions were blocked by synthesis type (SPEECH or SINUSOID) and by limit direction (INCREASING frame rate from the fastest rate, or DECREASING from the slowest). For example, a listener in the SINUSOID-DECREASING group heard three sequences, one for each sentence synthesized as a three-tone signal, with an initial frame rate of 80 ms, decreasing through 60 ms, 40 ms, and 20 ms, to 10 ms. (An INCREASING series started at 1 ms and increased through 2 ms, 3 ms, 4 ms, and 5 ms, to 10 ms.) Each listener participated in but a single block of trials. On each trial, a sentence was presented three times before the listener wrote a transcription in a prepared answer booklet.

The most revealing measure of transcription performance is shown in Figure 1. Here, we have marked the range of synthesis rates over which the groups performed at least half of their best transcription levels; this particular comparison was chosen to reflect the apprehension of phonetic detail from the different signals. Although synthetic speech by this generous criterion was relatively intelligible at anomalous rates of change, sinusoidal signals were only intelligible at the original rate of spectral variation, and at the most similar more rapid rate. The perceptibility of acoustic signals that lacked short-term speech properties depended here on a faithful rendering of the time-varying properties of speech signals.[2] The synthetic speech items, which did possess natural short-term properties, permitted listeners an opportunity to compensate for defective time-variation. After all, the impression of a stationary synthetic spectrum comprising a harmonic series and broadband resonances within these parametric ranges is unassailably voice-like, and we may suppose that it is no great leap for the perceiver to the ascribe a vowel or consonant segment on the basis of momentary timbres if necessary. The perception of speech sounds in sinusoidal replicas, therefore, appears in this set of tests to depend on perceptual attention to the natural attributes of spectrotemporal change. Perceptual attention to momentary stimulus properties, no closer to speech than three simultaneous pitch impressions, was inadequate to evoke phonetic attributes.

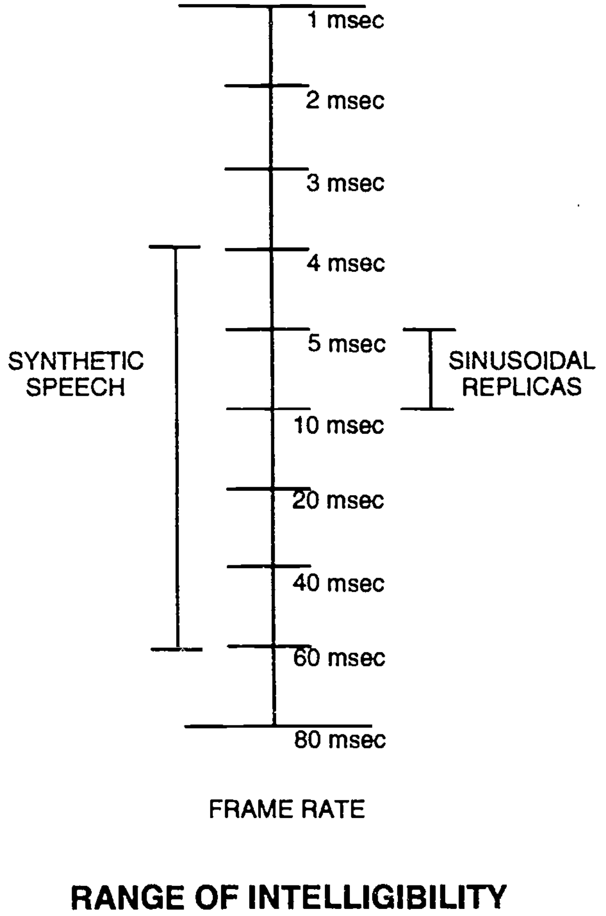


Figure 1. By the criterion that transcription performance attain 50% of best performance within synthesis type, synthetic speech was perceptible over a wide range of temporal variants of the original 10 ms/frame synthesis rate. Sinusoidal signals were not perceptible at rates that departed from the original 10 ms/frame rate.

TIME-VARYING ATTRIBUTES IN THE PERCEPTION OF SYNTHETIC SENTENCES

Although the perceptual susceptibility to phonetic properties employs standards that are expressly time-critical in nature, the prevailing approaches to perception have commonly left time out of the recipe for segmental identification, with the exception of the campaign to map the constraints of representational processes in sensory and short-term memory (for example, Liberman, Mattingly, & Turvey, 1972; Pisoni, 1973). To review a bold example of the disposition to restrict the issue of temporal properties to considerations of distinctiveness in linguistic phonetics, one recent influential model of recognition (Klatt, 1979) clearly stipulates timeless sequences of formant peaks for providing phonetic information, with only a few vowel percepts requiring definition in time-critical terms. This stands in direct contrast to the findings of our perceptual work on sinusoidal replicas of speech signals, which support a fundamental role of coherent time-variation in establishing and maintaining the speech mode of perception, however phonetic categorization occurs. The listener, according to our view, obtains phonetic information from properties of speech-like variation in the spectrum, predicated on familiarity with both linguistic structure and, implicitly, with the acoustic products of articulatory gestures.

The next step begins an uneasy adventure, due to the shortage of hypotheses about perceptual registration of spectrum variation. From the handful of experiments, including our own, that have found evidence of perceptual effects of coherent variation, no principles have yet emerged for describing the perceptual impact of variation in an acoustic signal. To press on nevertheless, we speculated that we might find a contingent effect on perception of the rate and the extent of spectrum variation. This extrapolation is based on the fact that the listener can obtain information from the patterns of variation and not only from instantaneous formant values. The experiment we performed bears some resemblance to the measurement of the modulation transfer function (Cornsweet, 1970).

Rate and Modulation Depth of Variation in Formant Frequency

The acoustic manipulation of speech that we attempted here is a distant relation of the modulation transfer function, which has been estimated in many sensory systems. Such measures characterize the responsiveness of a sensory modality to variation in energy, rather than to static properties of stimulation. Typically, such measures expose the preferential sensitivity of perceptual modalities to rates of change. In the visual case, for example, sensitivity can be described as a function of the spatial frequency and the depth of intensity variation. The analogy to our phonetic case is approximate: We hoped to describe the listener's ability to detect phonetic properties as a function of the *rate of change* and the *modulation depth* of formant frequency variation. Rate is a familiar variable in applied research, of course, especially in studies of temporal compression of speech (reviewed by Foulke & Sticht, 1969). We accomplished the rate variation conditions in the present case simply by varying the temporal value of each frame in our synthetic sentences. Initially, we produced synthetic versions of natural utterances in which the original speech rate was conserved, with a synthesis rate of 10 ms/synthesis frame. Subsequent versions were prepared with the synthesis rate set at 8 ms-, 6 ms-, and 4 ms/frame.

Modulation depth, or the extent of formant frequency variation across an utterance, is not a familiar parametric dimension of speech signals, or at least it was not familiar to us. We therefore invented a transformation of the signal that permitted us to control this structural property, which we call the *formant squash*. In our formulation, formant frequencies are represented as departures from schwa, the neutral vowel. Here, every frame of synthesis parameters specified formant frequency values as departures from nodal points of 500 Hz, 1500 Hz, and 2500 Hz for the first three formant peaks, rather than specifying the absolute formant frequencies. When we synthesized formant parameters encoded in this manner, we were able to alter the scale of variation in formant frequency by multiplying the difference between schwa and the observed values by a constant factor. We thereby controlled the modulation depth of each formant, with schwa at the shallow end, and natural variation at the other.

Consider an example of a graded series of modulated formant patterns. Figure 2 portrays the formant frequency values that were extracted from the utterance "My tee-vee has a twelve inch screen." Figure 3 shows samples of the progressive squashes that modified this signal pattern from one with natural formant variation into a pattern that incorporated only 10% of the formant frequency variation of the original natural utterance.

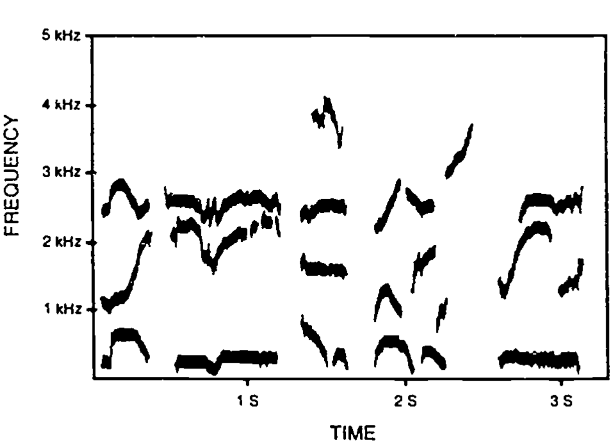


Figure 2. Formant frequency pattern for the sentence, "My tee-vee has a twelve inch screen."

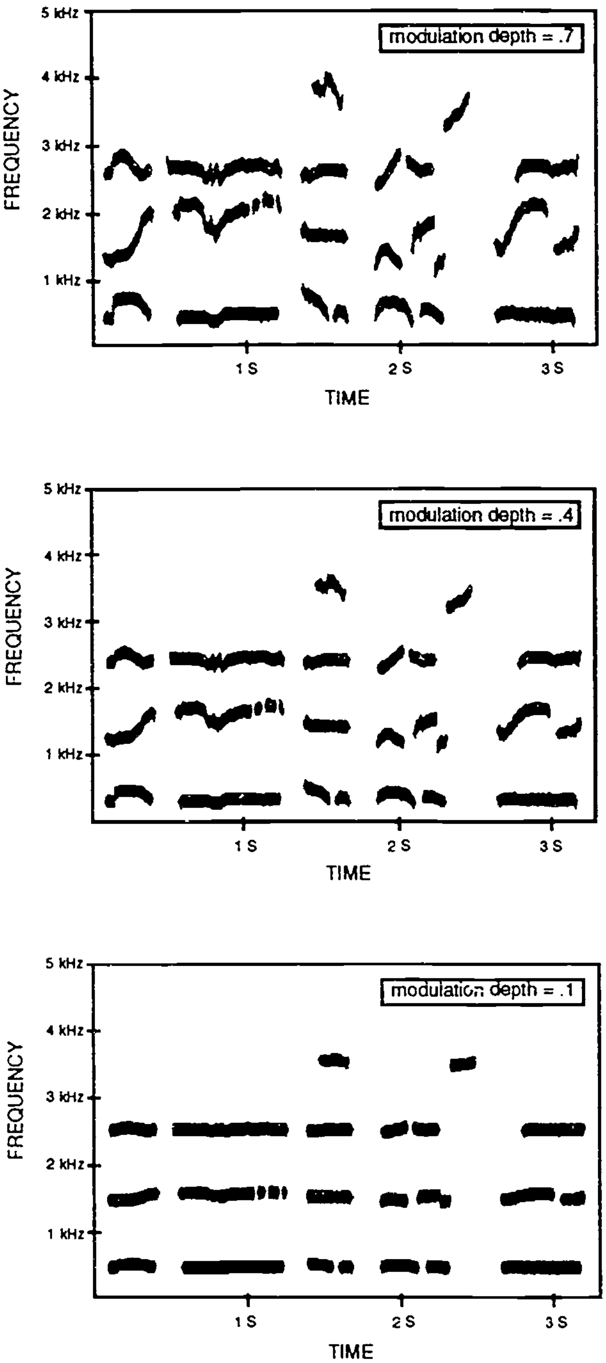


Figure 3. Three examples of formant patterns exhibiting variation in modulation depth (formant squash) are shown for the sentence, "My tee-vee..." Top panel: modulation depth = .7; Middle panel: modulation depth = .4; Bottom panel: modulation depth = .1. Nodal values are those of the neutral vowel, [ə], schwa.

The Perceptual Effects

Two synthetic sentences were prepared for this test, varying in modulation depth and in modulation rate of formant frequency. ("Kick the ball straight and follow through" was the second sentence.) At each of the four synthesis rates, ten versions were constructed for the two sentences we used, with formant squash factors of .1, .2, .3, .4, .5, .6, .7, .8, .9, and 1.0. An unchanging fundamental frequency was used to permit us to isolate the perceptual effects of frequency change in the formant pattern. Subjects in our experiments were instructed to transcribe the synthetic sentences that they heard, which were presented in order of increasing frequency variation blocked both by sentence and by synthesis rate. A single trial comprised four repetitions of the particular version of the sentence. As the subject listened to the sentences in a test session, successive trials grew progressively clearer and less schwa-like.

We scored the transcription performance for the percent of syllables correctly transcribed. Figure 4 shows the data in four functions, one for each synthesis rate, and each point represents the average performance at each degree of formant squash. The two main effects are immediately apparent. As we expected, the faster the synthesis rate and the more squashed the formant frequency variation, the poorer the transcriptions. And, there was an interaction between the effects of each, clearly visible in the close similarity of performance of the 10 ms- and 8 ms/frame conditions and the difference between those two and the 6 ms- and 4 ms/frame conditions. At 10- and 8 ms/frame, the effect of changes in modulation depth were indistinguishable, as if a phonetic constraint on the depth of spectrum variation were satisfied at either rate of modulation.

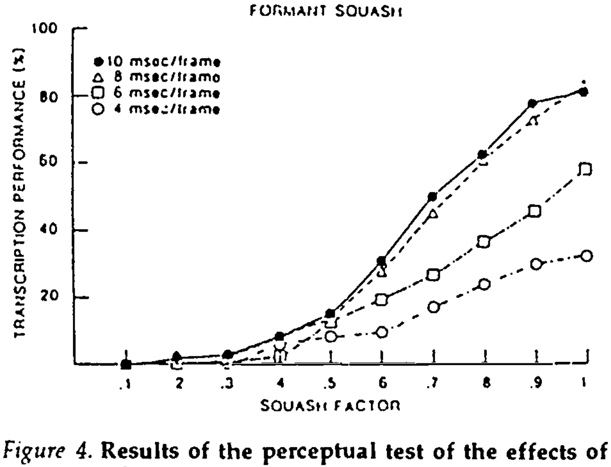


Figure 4. Results of the perceptual test of the effects of rate and frequency variation. Average percent of syllables transcribed correctly is plotted against squash factor for the four rate conditions. Subjects were blocked by sentence rate.

Implications and Extensions of the "Phonetic Modulation Transfer Function"

The effects of the combined variation of rate and extent of formant frequency change corroborate one explanation for the results of "temporal compression" experiments on speech (Foulke & Sticht, 1969). Although the comprehension of compressed speech warrants accelerating the normal process of identifying linguistic attributes from speech signals, the deterioration of performance at very rapid rates of presentation may not be due solely to overloading the perceiver's ability to translate an incoming phonetic sequence into a durably storable representation, an implicit task of speeded classification. Although listeners can obviously compensate a bit under duress of this kind, it is costly to do so, as recent studies by Whalen (1984) and Pisoni (1981) have shown. It may be due as well to the loss of acoustic structure, namely, that which is available in time-critical frequency variation. This conclusion seems justified by the differential effect of the same degree of formant squashedness, contingent on the rate with which the particular formant frequency excursions occur. Another way to indicate this is to note, again, the significance of time-varying spectra in speech perception. The phonetic value of a sequence of formant center frequencies is not invariant over temporal changes, as Klatt's model supposed. The perceiver does not identify the phonetic sequence of an utterance solely by considering the momentary spectra within the pattern. The time-course of changes in the signal appears to be critical, and incremental departures from the natural modulation depth produced small perceptual effects only within the limits of phonetic rates of change.

To develop the approach opened by this study requires several tactics. First, we need to look for the other limits—not just the progressively rapid and shallow end but slowed and exaggerated formant variations, as well, in factorial combination. We should be able to chart the limits of phonetic sensitivity for those increases from natural modulation, on the assumption that the articulations that create formant variation—the periodic opening and closing of the vocal tract, and the gymnastic advancement, retraction and elevation of the tongue—occur with characteristic periodicity, and these should be reflected in perception whatever the mechanism is. We also need to extend the tests to cases that control the classes of phonetic constituents closely, principally to differentiate two plausible alternative accounts of these effects: Whether the rate-varying formant-squashes reveal specifically time-critical perceptual sensitivity to phonetic properties, as we claim, or merely the effects of isolated miscategorizations of the few acknowledged rate-contingent segmental cues, such as may distinguish [b] from [w] in some environments (Liberman, Delattre, Gerstman, & Cooper, 1956) or [i] from [ɪ] (Ainsworth, 1972). The phonetic composition of the sentences in the present study does not favor the limited sense of time-critical. information admitted by Klatt, but only a more careful test will tell. If phonetic information is generally time-critical, then slowing the rate of formant frequency variation should have a perceptual effect similar to speeding it: Phonetic perception should depend, all other things equal, on the temporal parameters of spectrum variation no less than the particular formant frequency values.

APPARENT NATURALNESS OF SYNTHETIC AND SINEWAVE UTTERANCES

The best evidence of a perceptual susceptibility to coherent variation in the speech spectrum comes from studies with sinusoidal replicas, though there is a lingering doubt confronting us about the validity of this interpretation. Many listeners who have taken our perceptual tests told us during debriefing that sinewave signals were new to them, but that they were already familiar with synthetic speech. Though all of these listeners qualified as naive test subjects, they reported that the concept and the sound of synthetic speech was already known to them before the listening session, from movies, television and toys, and from electronic gadgets. It is reasonable to suppose that informal contact with impoverished and distorted signals of this kind affects the perception of speech. In fact, studies have shown that subjects progressively accommodate the specific acoustic phonetic correspondences employed in a commercial synthesizer, at least over a brief period (Greenspan, Nusbaum, & Pisoni, 1988). Were it possible, a direct assessment could determine whether the perceptual plasticity required in understanding a sinusoidal sentence involves familiarity with speech synthesis. Recall, the listener who hears a sinewave replica phonetically is deriving phonetic properties from the coordinated variation of the tones, while simultaneously ignoring the non-phonetic polyphonic timbre. We suppose that the ability to extract the relevant portion of immediate auditory experience, and to relegate the rest to inattention is an instance of plasticity, of which a reliable minority of our listeners is evidently incapable, or unwilling (Remez et al., 1987).

Our claim notwithstanding, the perception of sinusoidal utterances may depend on specific prior encounters with synthetic speech, as if the listener is able to exploit a special perceptual resource that has developed through familiarity with such impoverished synthetic acoustic signals. If true, this would oppose our claim that the same time-varying acoustic properties present in natural speech provide phonetic information in sinewave replicas. Consider, though, that for sinewave replicas to enjoy the same hypothetical perceptual accommodations as synthetic speech, the two kinds of signals must be treated by the perceiver as roughly the same unnatural acoustic stuff. Were we to find that the listener considers sinewave and synthetic signals to be similarly natural, we would have evidence that sinewave signals are treated as another instance of familiar synthetic speech. Accordingly, were we to find that sinewave signals differ in apparent naturalness relative to synthetic speech, such evidence would make it seem implausible that sinusoidal signals are perceived by virtue of their similarity to the synthetic speech with which the listener is familiar. To evaluate the possibility, we ran a naturalness tournament, in which natural, synthetic, sinusoidal, and amplitude-modulated sinusoidal signals were evaluated by naive listeners.

Six versions of the sentence, "My dog Bingo ran around the wall," were used in a test of relative naturalness: 1) natural speech, 2) synthetic speech, 3) sinewave replica, 4) triangle-pulsed sinewave, 5) glottal-pulsed sinewave, and 6) rectangular-notched sinewave. The natural version was spoken by one of the authors. A synthetic version of the sentence was fashioned by taking the formant tracks computed in the linear prediction analysis and using them as speech synthesis parameters, employing a cascade-type synthesizer (Mattingly et al., 1981). The values for the synthesis of fundamental frequency were derived from the instantaneous amplitude of the signal, and were not based on estimates of the fundamental of the natural utterance. The waveform of the synthetic sentence was computed and stored, for use as the synthetic version of the sentence in the naturalness test.

A sinewave version of the sentence was also constructed from the computed analysis of the natural signal, and served as the replica (condition #3) and as the main ingredient to the three other versions of the sentence used in the test. In each of those cases, a different 10 ms pulse shape was imposed throughout the digital waveform of the sinusoidal signal. Three pulse shapes were used, one in each of the versions: a triangle shape, a glottal pulse shape (Rosenberg type B; see Rosenberg, 1971), and a rectangular-shaped notch. The triangle shape was selected because it is a commonly used form of excitation in speech synthesizers; the glottal shape was used because it is a closer approximation of the glottal function of the normal adult male; the rectangle-with-notch was used to provide a truly unnatural kind of pulse.

In the naturalness tournament, each of the six versions of the sentence occurred in a paired presentation with every other version, making fifteen different pairings. On a trial, the listener was instructed to attend to each sentence and to identify the one that was more natural, by any available criterion.

To assess the relative apparent naturalness of the six acoustic versions of the sentence, we derived a naturalness index, and applied the procedure to the six different acoustic versions of the sentence. Each version occurred in five pairings, and each pairing was repeated ten times (in the randomly ordered series of pairing in the tournament). This means that the greatest number of times any single version could have been chosen over its five opponents was 50, the maximum on our index, if it had been identified on every possible occurrence as the more natural member of the trial pair. Accordingly, the minimum value was 0, denoting that a candidate version of the sentence had never been selected as the more natural utterance in any pair. Each subject, then, contributed six indexes to the group data, and the average naturalness indexes for the sentence versions are portrayed as a histogram in Figure 5.

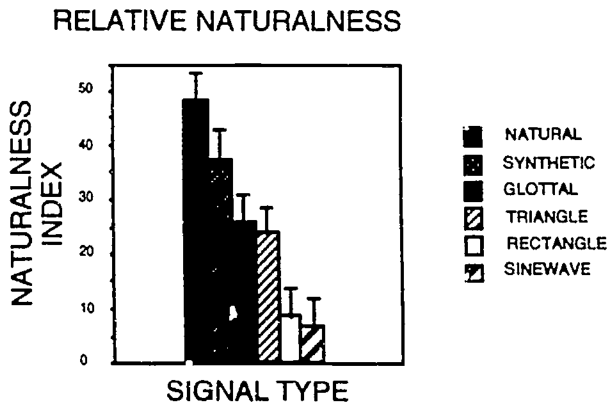


Figure 5. Group performance on a test of relative naturalness.

Subjects differentiated the six sentences by apparent naturalness, as we expected. The natural speech version proved to be the most apparently natural;[3] next was the synthetic speech version, which also differed from the four versions based on the sinewave signal. Glottal and triangle pulsed sinewaves were not found to differ from each other in apparent naturalness, though they were reported to be significantly more natural than the sinewave and rectangular notched versions. Last, no difference was found between sinewave and rectangular notched versions.

ARE TONAL ANALOGS SIMILAR TO SYNTHETIC SPEECH?

This experiment provides a clear picture of the listener's impression of the sound quality of sinusoidal replicas, by assessing naturalness of sinusoidal utterances relative to synthetic speech and to natural speech as well. Our test subjects in prior studies had mentioned that they were familiar with synthetic speech, which they impressionistically described as mechanical or electronic in sound quality (Remez et al., 1981). This raised the possibility that listeners who transcribed sinewave sentences were able to do so by first identifying sinewaves as a member of the class of familiar synthetic utterances, and then resorting to a perceptual capability that developed out of encounters with synthetic speech of various types. Had this been the case, the essential premise of our research would have been undermined, for we have hypothesized from the outset that the ability to perceive tonal replicas of utterances rests on the preservation in the sinusoidal coherence of effective time-varying acoustic attributes found in natural utterances.

Our results here show that no obvious psychoacoustic similarity can be presumed between synthetic and sinusoidal utterances, though both clearly share an apparent dissimilarity from natural speech. It is possible that listeners compensate for unnatural sounding speech by adopting a general purpose strategy accommodating anomalous timbre or peculiar sentence intonation. Nonetheless, it is implausible to suppose that the perceiver takes a sinusoidal replica to be, simply, unnatural in the familiar way, given these data. The outcome of the test shows that the acoustic properties relevant to the perceptual evaluation of naturalness place sinewave sentences well beyond the familiar unnaturalness of unmistakably synthetic speech. For this reason, we may be confident that listeners do not allocate their perceptual resources for handling sinewave utterances based on an impression of the familiar sound of synthetic speech when they take our sinewave tests.

ACOUSTIC SHARDS, PERCEPTUAL GLUE

Two problems motivated the studies that we described here, both stemming from the great diversity in the acoustic constituents of speech signals. Before perceptual analysis of a speech signal can begin, the perceiver must have a coherent signal that is fit to analyze. Though it is obvious that the perceptual organization of a speech stream occurs with the same ease that characterizes the perception of consonants and vowels, the means by which the acoustic components compose a single perceptual stream is not well understood. The perceptual unity of the components of speech incorporates dissimilar elements, periodic and aperiodic, continuous and discontinuous, simultaneous and successive, and the principal finding of our studies is that the perceiver attends to the spectrotemporal matrix in which the acoustic elements occur. This attention is specifically directed toward natural properties of variation; it is observable with synthetic speech signals that approximate natural short-term spectra; and, it occurs independent of impressions of timbre or specific experience with synthetic speech.

By taking this approach to the perceptual organization of speech signals—namely, that it occurs by virtue of susceptibility to second-order spectrotemporal properties unique to vocal articulation—we find a ready account of the perceptual accommodation to the great variability of the acoustic properties of natural speech. Because the perceiver is demonstrably sensitive to the properties of coherent variation of speech signals in addition to their elemental attributes, the theoretical possibility is hereby established for accommodating superficial acoustic novelty within familiar forms of acoustic change. Does this approach to the perceptual organization of dissimilar acoustic elements also pertain to the more familiar issues of variability and invariance described in linguistic and perceptual phonetics? The problem of acoustic variation—across many individual utterances with identical linguistic descriptions—is, at first glance, quite different from the perceptual coherence of nonstationary acoustic spectra during a tick or two of the clock. But, if there are properties of spectrotemporal variation in speech acoustics that stem specifically from the phonetic control of the articulating resonators of the supralaryngeal vocal tract, then we will not be surprised to find a congruence in the two problems of variability. The key challenge is to define the perceptual principles of spectral variation that ensure coherence of apparently dissimilar elements, permitting limitless elemental variation within the forms of change produced by phonetically governed acts of sound production.

REFERENCES

Ainsworth, W. A. (1972). Duration as a cue in the recognition of synthetic vowels. *Journal of the Acoustical Society of America, 51*, 648-651.

Bregman, A. S. (1990). *Auditory scene analysis*. Cambridge: MIT Press.

Busnel, R. G., & Classe, A. (1976). *Whistled languages*. New York: Springer Verlag.

Cornsweet, T. N. (1970). *Visual perception*. New York: Academic Press.

Elman, J. L., & McClelland, J. L. (1985). Exploiting lawful variability in the speech waveform. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 360-385). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Fitch, H. L., Halwes, T. G., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics, 27*, 343-350.

Foulke, E., & Sticht, T. G. (1969). Review of research on the intelligibility and comprehension of accelerated speech. *Psychological Review, 72*, 50-62.

Ganong, W. F., III. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 110-125.

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 421-433.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics, 7*, 279-312.

Ladefoged. P. J. (1967). *Three areas of experimental phonetics*. London: Oxford University Press.

Liberman, A. M., & Cooper, F. S. (1972). In search of the acoustic cues. In A. Valdman (Ed.), *Papers in linguistics and phonetics to the memory of Pierre Delattre* (pp. 329-338). The Hague: Mouton.

Liberman, A. M., Delattre, P. C., Gerstman, L. J., & Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology, 52*, 127-137.

Liberman, A. M., Mattingly, I. G., & Turvey, M. T. (1972). Speech codes and memory codes. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 307-334). Washington, DC: V. H. Winston.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological enquiry*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Mattingly, I. G., Pollock, S., Levas, A., Scully. W., & Levitt, A. (1981). Software synthesizer for phonetic research. *Journal of the Acoustical Society of America, 69*, S83.

Oshika, B. T., Zue, V. W., Weeks, R. V., Neu, H., & Aurbach, J. (1975). The role of phonological rules in speech understanding research. *IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-23*, 104-112.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics, 13*, 253-260.

Pisoni, D. B. (1981). Speeded classification of natural and synthetic speech in a lexical decision task. *Journal of the Acoustical Society of America, 70*, S98.

Remez, R. E. (1987). Units of organization and analysis in the perception of speech. In M. E. H. Schouten (Ed.), *Psychophysics of Speech Perception* (pp. 419-432). Dordrecht: Martinus Nijhoff.

Remez, R. E., & Rubin, P. E. (1983). The stream of speech. *Scandinavian Journal of Psychology, 24*, 63-66.

Remez, R. E., & Rubin, P. E. (1984). Perception of intonation in sinusoidal sentences. *Perception & Psychophysics*, 35, 429-440.

Remez, R. E., & Rubin, P. E. (1990). On the perception of speech from time-varying attributes: Contributions of amplitude variation. *Perception & Psychophysics, 48*, 313-325.

Remez, R. E., & Rubin, P. E. (1991). On the intonation of sinusoidal sentences: Contour and pitch height. submitted.

Remez, R. E., & Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1992). On the perceptual organization of speech. Submitted for publication.

Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 40-61.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell T. D. (1981). Speech perception without traditional speech cues. *Science, 212*, 947-950.

Rosenberg, A. E. (1971). Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America, 49*, 583-590.

Whalen, D. H. (1984). Subcategorical mismatches slow phonetic judgments. *Perception & Psychophysics, 35*, 49-64.

Zue, V. W., & Schwartz, R. M. (1980). Acoustic processing and phonetic analysis. In W. A. Lea (Ed.), *Trends in speech recognition.* (pp. 101-124). Englewood Cliffs, New Jersey: Prentice-Hall.

FOOTNOTES

[*] To appear in J. Charles-Luce, P. A. Luce & J. R. Sawusch (Eds.), *Theories in spoken language: Perception, production, and development.* Norwood, NJ: Ablex Press.

[†] Department of Psychology, Barnard College.

[1] We have described the departures from natural spectrotemporal variation that were achieved through synthesis as "defective" and "anomalous," though in the absence of a standard description of natural acoustics these designations are informal. But, the informal characterization is not implausible. The numerous phonologically and phonetically governed acoustic manifestations of variation in speech rate are hardly mimicked by the use of the frame-rate manipulation, which simply imposes a uniform temporal strain on the formant pattern.

[2] The raw transcription performance, scored for percent syllables correct, revealed an overall difference in intelligibility between the two synthesis techniques, across all conditions. In conditions in which performance was best, intelligibility of synthetic speech reached approximately 90% correct, while intelligibility of sinewave signals reached approximately 70% correct. This replicated our earlier results with sinusoidal speech, in which a substantial minority of listeners were completely unable to transcribe the signals (Remez et al., 1987). Could the differential intelligibility of synthetic speech and sinusoidal signals itself produce the range effect portrayed in Figure 1? There is small likelihood that it did, because the slopes of the linear functions fitted to the raw data differ substantially, the synthetic speech conditions with shallow slope, the sinewave conditions with steep slope.

[3] It is surprising that subjects did not uniformly —or unfailingly— select natural speech as the more natural in each of its contests. One may well wonder how subjects would have judged two instances of natural speech against each other .

F0 Gives Voicing Information even with Unambiguous VOTs*

D. H. Whalen, Arthur S. Abramson,† Leigh Lisker,‡ and Maria Mody†††

The voiced/voiceless distinction for English utterance-initial stop consonants is primarily realized as differences in the voice onset time (VOT), which is largely signalled by the time between the stop burst and the onset of voicing. The voicing of stops has also been shown to affect the vowel's F0 after release, with voiceless stops being associated with higher F0. When the VOT is ambiguous, these F0 "perturbations" have been shown to affect voicing judgments. This is to be expected of what can be considered a redundant feature, that is, that it should carry a distinction in cases where the primary feature is neutralized. However, when the voicing judgments were made as quickly as possible, an inappropriate F0 was found to slow response time even for unambiguous VOTs. This was true both of F0 contours and level F0 differences. These results reinforce the plausibility of tonogenesis, and they add further weight to the claim that listeners make full use of the signal given to them, even when overt labelling would seem to indicate otherwise.

INTRODUCTION

The voicing of utterance-initial stop consonants is perceptually determined primarily by the voice onset time (VOT), which is largely signalled by the time between release of the stop and the onset of voicing (Lisker & Abramson, 1964). This has been found to be true not only of languages such as Spanish that rely primarily on the presence or absence of voicing during closure, but also of languages such as English that, in some environments, do not voice the closure for the voiced stops and overlap the voicelessness with the release of the stop as aspiration. In addition, a falling fundamental frequency (F0) usually occurs after a voiceless stop, while a flat or rising F0 usually accompanies voiced stops (House & Fairbanks, 1953; Lehiste & Peterson, 1961; Ohde, 1984; Silverman, 1987). This has been called the F0 "perturbation" and has been found in a wide variety of languages (Hombert, 1975). In perception, perturbation effects have usually appeared only when the VOT was ambiguous, at least when the perturbation was of the same magnitude as found naturally (Abramson & Lisker, 1985; Fujimura, 1971; Whalen, Abramson, Lisker, & Mody 1990). That is, an ambiguous VOT is more likely to be heard as voiceless when the F0 is falling after the onset of voicing than if it is flat or rising.

Studies of natural productions of stops have found that most measured VOT values fall within ranges that are unambiguously interpreted (Lisker & Abramson, 1964; Shimizu, 1989). Thus it may appear that the perceptual effects of F0 on voicing, though demonstrable, are unimportant in the actual use of language. Abramson and Lisker (1985:32), for example, state that voicing judgments for certain VOT values "cannot be affected by F0." However, several studies have shown that acoustic differences that do not affect overt labelling can nonetheless affect speech processing, as shown by reaction times (Martin & Bunnell, 1981; Whalen, 1984; Whalen & Samuel, 1985; Tomiak, Mullennix, & Sawusch, 1987). These studies focused on subcategorical mismatches created by splicing segments of natural speech from one (appropriate) environment to another (an inappropriate one).

This work was supported by NIH Grant HD-01994 to Haskins Laboratories. Some of the results were presented at the Annual Meeting of the Linguistic Society of America, January, 1991, Philadelphia, PA. We would like to thank Bruno H. Repp, Carol A. Fowler, Terrance M. Nearey, and two anonymous reviewers for helpful comments.

The mismatches have involved both vowel-to-vowel and fricative/vowel coarticulation. While the effects provide clear support for the idea that subjects are sensitive to all the linguistic information given to them, it could also be argued that the mismatch might have included an abrupt change in a resonance, which might be seen as a nonlinguistic source of the delay. This might hold true even in those cases where all the stimuli had been cross-spliced, at lear˙ from one token to another (Whalen & Samuel, 1985).

The present study was designed to extend the results on mismatched cues to a situation in which there is no possibility that the coherence of the signal has been violated. We chose the influence of the F0 perturbation on identifying VOT continua, as demonstrated in previous work (Abramson & Lisker, 1985; Fujimura, 1971; Whalen et al., 1990), because in any sequence of voiceless stop followed by a voiced vowel, there must of course be a shift from a voiceless to a voiced source. Thus if there is any auditory "discontinuity" inherent in changing from a voiceless to voiced source, it is one that is normal for speech. The only manipulation we had to make was to vary the onset F0 value, a choice that should not, in itself, give rise to any auditory discontinuities. If responses are slower when the F0 information does not match that of the VOT, we can be even more confident that all the acoustic consequences of a speech gesture contribute to the perception of speech, even if the labelling fails to show it.

In addition, we wanted to assess the ability of listeners to detect these F0 differences when the VOT is unambiguous. It is the implication of certain language changes that the F0 perturbations are used perceptually: Many cases of the emergence or diversification of tone systems have been traced to the loss or realignment of a voicing distinction with a concomitant use of the perturbation's effect (Hombert, 1975; Hombert, Ohala, & Ewan, 1979; Maddieson, 1984). While the diachronic facts have not been questioned, it has remained an uneasy assumption that such small F0 differences could in fact be perceived in a natural context. The present experiments will show, at least, that these differences in F0 do affect perception, making the theory of tonogenesis that much more plausible.

I. EXPERIMENT 1

The first experiment uses reaction time to determine whether there is perceptual use of F0 for voicing judgments even when the labelling shows no effect. Such a subcategorical effect can be seen in the reaction times to stimuli with unambiguously labelled VOTs.

A. Method

1. Stimuli. Synthetic approximations to the English syllables /ba/ and /pa/ were created with the serial synthesizer designed by Ignatius G. Mattingly at Haskins Laboratories (as in Whalen et al., 1990). The vowel steady-state formants were centered at 730, 1250 and 2440 Hz with bandwidths of 100, 100, and 125 Hz. (Since this was a serial synthesizer and the bandwidths were kept constant, there was no "F1 cutback.") The formant values at the beginning of the syllable were 450, 1080 and 2300 Hz for F1, F2 and F3 respectively, and they changed linearly to reach the steady-state values after 75 ms. Vowel amplitude was level until the last 30 ms of the syllable, at which point it decreased linearly to zero. Total duration of the syllables was 250 ms.

The VOT values were 5, 10 15, 20, 25, 35 and 50 ms after the simulated release. These were obtained by turning off the voicing source (AV) and introducing aperiodic hiss (AH) for the appropriate number of synthesis frames. The F0 onset values were 98, 108, 114, 120 and 130 Hz. F0 changed linearly from these values to the steady-state F0 of 114 Hz over the first 50 ms of voicing. (Of course, with an onset of 114 Hz, the F0 was constant over the entire syllable.) These differences of onset values are similar in magnitude to those reported in the literature (e.g., Ohde, 1984). Each VOT was paired with each F0 onset, giving 35 unique stimuli.

2. Procedure. The stimuli were presented for identification as "b" or "p," with responses being made by pressing buttons labelled "b" (on the left hand side) and "p" (on the right). All subjects participated in five conditions, three unspeeded and two speeded. The first was an unspeeded condition containing five repetitions of all 35 stimuli, while the other two, which differed only in the randomization, used a subset of these consisting of the 23 stimuli which appeared in either of the two speeded conditions. One speeded condition, the F0 condition, used all F0 onset values, but only the 5, 20, and 50 ms VOTs. The other speeded condition, the VOT condition, used all VOT values but only the extreme F0 values (98 and 130 Hz). Twenty repetitions of the stimuli were randomized for the speeded conditions.

The order of conditions was as follows. First came the unspeeded condition with all the stimuli. Then came the first speeded condition, which was the F0 condition for half of the subjects and the VOT condition for the other half. After the first speeded condition came an unspeeded condition with the selected stimuli. Next, the other speeded condition was given, so that all subjects had both conditions. Finally, the unspeeded task for the selected stimuli was given one more time.

In the unspeeded conditions, subjects were to press the button when they had made their decision, limited in time only by the 2.5 s between stimuli. If unsure, they were to guess. In the speeded conditions, they were to make their response as quickly as possible, using one finger of their right (dominant) hand to press the buttons. Between trials, they rested this finger on the keyboard.

3. *Subjects.* The subjects were 12 young adults from the Yale University community who had volunteered for listening experiments. All passed an audiometric screening for both ears. They were paid for their participation.

B. Results

1. *Unspeeded Conditions.* The first unspeeded condition showed the pattern found in our earlier work (Whalen et al., 1990): The three lowest F0 values elicited about the same percentage of "b" responses collapsed over VOT (48.6, 51.0, and 52.2% for 98, 108, and 114 Hz onsets respectively), while the two higher values elicited fewer (43.6% for the 120 Hz onset and 36.4% for the 130). Figure 1 shows the effect on the judgments by giving the overall percentage of "p" responses for stimuli beginning with the stated F0 collapsed across all VOT values.

As in our earlier study, F0 did not influence judgments for unambiguous VOTs, that is, those extreme values o f VOT that received at least 80% judgments in one category (Figure 2). Although the F0 functions do not converge except in the third panel for the two extreme VOT values, there is no consistency in the ordering of the F0 functions the way there is in the ambiguous region.

The functions in Figure 1 are not monotonic, possibly due to the reduction in the number of stimuli presented. Recall that in the two unspeeded conditions with the selected stimuli, only three VOT values were used rather than seven for three of the F0 settings. This gives us somewhat less resolution in our measures, but that is acceptable for the replication of our earlier work (Whalen et al., 1990).

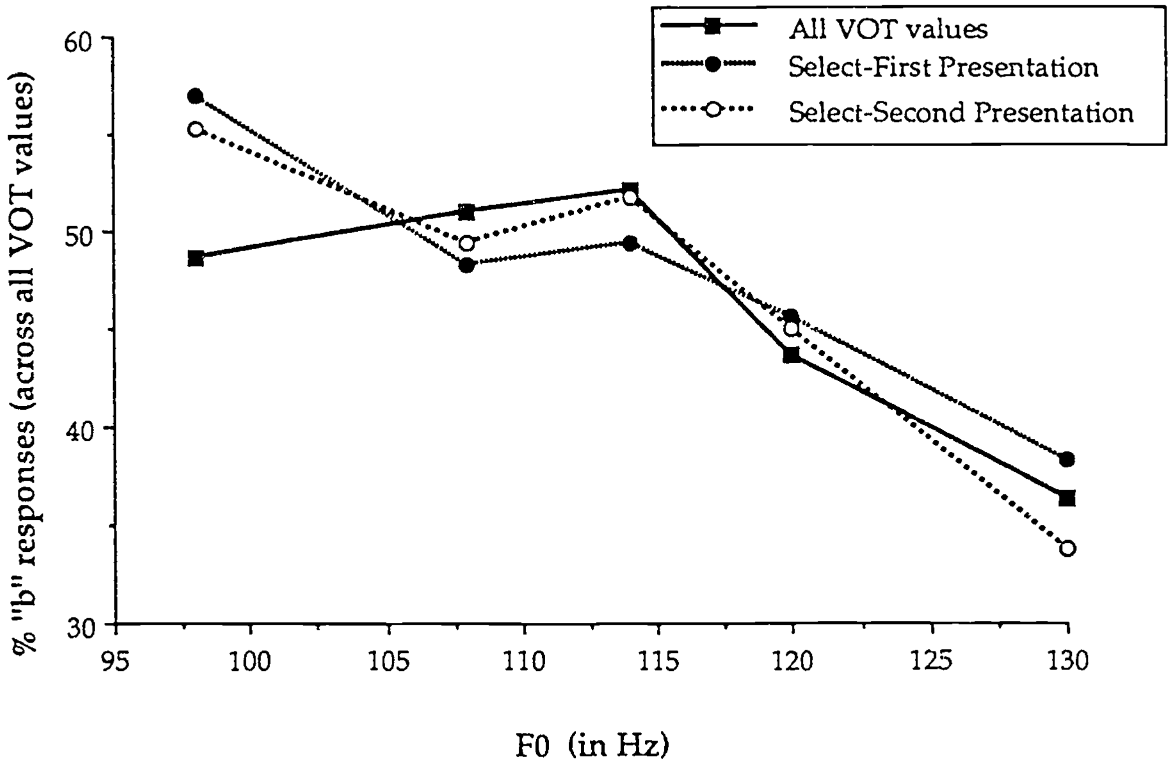


Figure 1. Responses for the 12 subjects in the three unspeeded conditions of Experiment 1, expressed as a percentage of "b" responses averaged across all VOT values.

There are two ways of looking at the selected conditions. The first is by whether the unspeeded test followed the F0 speeded condition or the VOT; the second is by the experimental order, i.e., whether the unspeeded test was the third or fifth in the experiment. The results will be considered in their experimental order, though in fact, it does not matter much since the two conditions are quite similar (Figure 1). The general result is clear, and consistent with our earlier finding: After the initial experience with these stimuli, subjects begin using F0 for voicing information in a fairly gradient fashion from the lowest F0 to the highest. The functions are not monotonic, but they are clearly different from the first condition. The only stimuli that could show a large difference are the 98 Hz stimuli, and they do show one. An analysis of variance on the proportion of "b" with the factors F0 and Condition shows a significant main effect of F0 (F(4,44) = 39.32, $p < .001$). Condition was not a significant main factor (F(2,22) 1, n.s.), but the interaction was (F(8,88) = 2.63, $p < .05$). Further analyses (separate ANOVAs for the three pairings of the conditions) showed that the first condition differed from each of the selected ones, which did not differ from each other. (A significant interaction of F0 and condition appears with the first condition and each of the selected ones (F(4,44) = 3.32 and 2.60, $p < .05$); the selected conditions did not show an interaction (F(4,44) = 1.72, n.s.).) After their initial exposure to these stimuli, the subjects make a more gradient use of F0 in their voicing judgments, as had been found in our earlier work (Whalen et al., 1990).

2. *Speeded Condition: F0.* The results for the F0 condition are shown in Figure 2. Reaction times for all twelve subjects are averaged together. "b" responses to the 50 ms VOT stimulus and "p" responses to the 5 ms VOT stimulus were excluded as being mistakes (based on the unspeeded results). These accounted for 1.6% of the responses. At the 20 ms VOT value, both responses were considered correct.

The most important result to be seen in this figure is that the extreme F0s are associated with different reaction times depending on the category label applied. The extreme F0 values are given in the thicker lines. For "b" judgments, both at the unambiguous and the ambiguous VOTs, the 98 Hz onset gave faster times than the 130 Hz onset. Conversely, the 130 Hz onset gave the faster times for the "p" judgments. The other F0 values (shown with thinner lines) tend to range in between, but their arrangement is not monotonic. There is, perhaps, not enough resolution within this rather narrow range of reaction times for a monotonic pattern to emerge with this number of repetitions.

For statistical analysis, the unambiguous items were analyzed together, and then the responses to the ambiguous stimulus. The factors were Response Category and F0. Analysis of the means and standard deviations indicated the presence of an inhomogeneity of variance, which was minimized by using a speed transform, that is 1/RT. All reported numbers are retransformed into times to make comparisons to other studies easier.

Response category was a significant main effect for the unambiguous VOTs (F(1,11) = 7.77, $p < .05$), since the "b" responses were faster by some 50 ms. F0 was not a significant main effect (F(4,44) = 2.14, $p < .10$), but the interaction of the two was (F(4,44) = 5.97, $p < .001$). The interaction shows strong evidence of a differential effect of F0 depending on which category is selected as the response. Separate analyses for each response category shows F0 to be significant in itself (F(4,44) = 3.88, $p < .05$ for the "b" responses, F(4,44) = 4.29, $p < .01$ for the "p" responses). We may therefore conclude that the appropriateness of the F0 affected decision time.

The magnitude of the reaction time difference happens to be almost that of the difference in VOTs (50 ms in reaction time versus 45 ms difference in VOT), but this is likely to be due to factors other than the VOT itself. The 50 ms VOT, though consistently heard as "p," is likely to be further from the prototypical value for /p/ than the 5 ms VOT is for /b/ (Miller & Volaitis, 1989). If so, then we would expect reaction times to be longer (Pisoni & Tash, 1974; Whalen, 1991), since less prototypical tokens are harder to identify. If our function went further along the scale, the "b" and "p" times might become equivalent. Additionally, the synthesis parameters, notably the absence of a burst and the lack of F1 attenuation in the aspirated portion, may have been more detrimental to the "p" category than the "b."

The mean values for the ambiguous items, shown in Figure 3, display the effect that we would expect, with the F0 values affecting the two judgments differently. Unfortunately, half of the subjects failed to find this VOT value ambiguous in this condition, with the consequence that their minority-category judgments are too few to give a reliable mean value. As it turns out, of the six subjects who did not find the stimuli ambiguous, three heard them primarily as "b" and three heard them mostly as "p." Therefore, two separate analyses were done, one for "b" and one for "p."

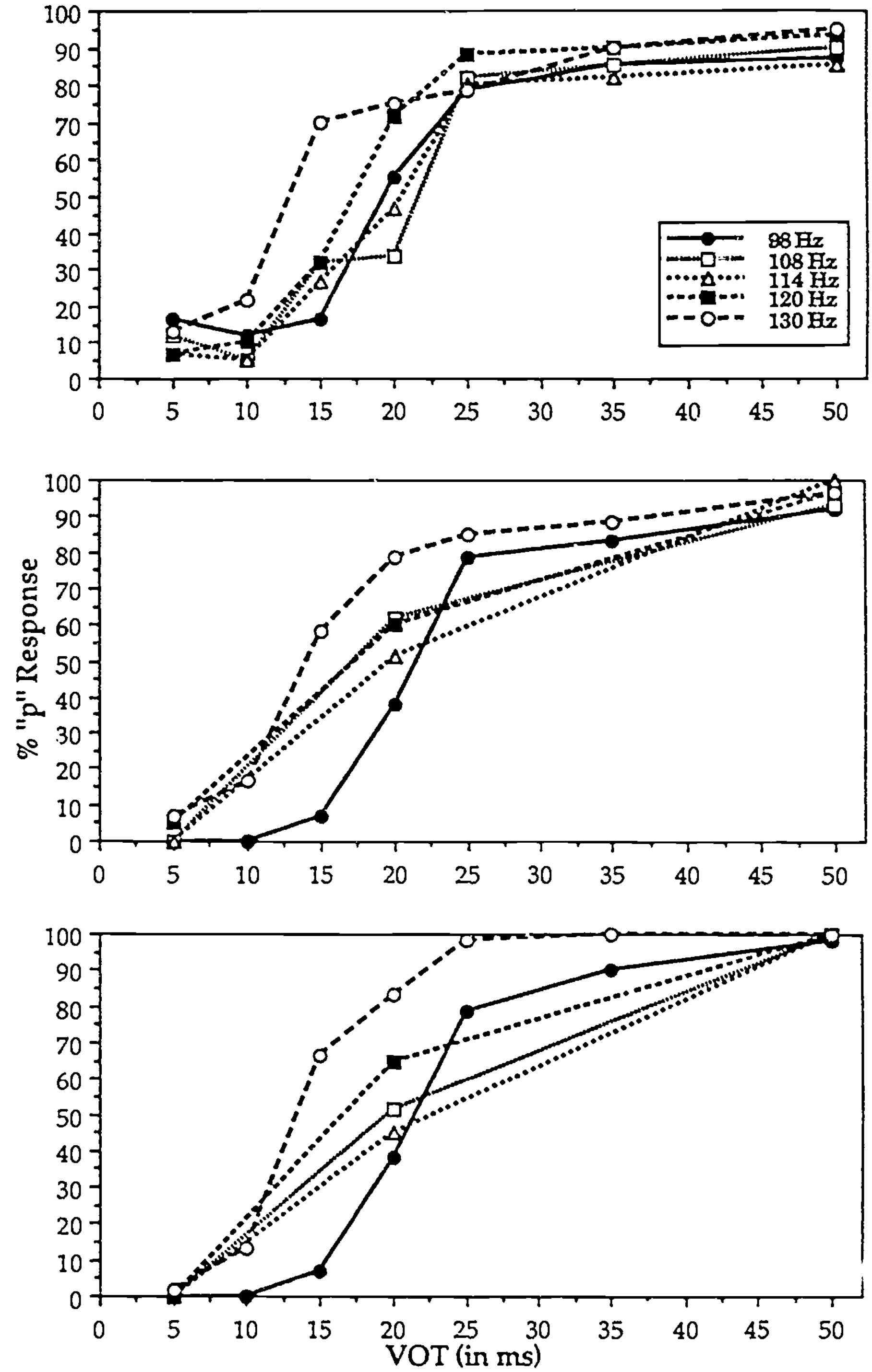


Figure 2 Responses for the 12 subjects in the three unspeeded conditions of Experiment 1, shown separately for each condition. The top panel is the first unspeeded condition. The middle panel is the first selected condition, and the bottom panel is the second selected condition.

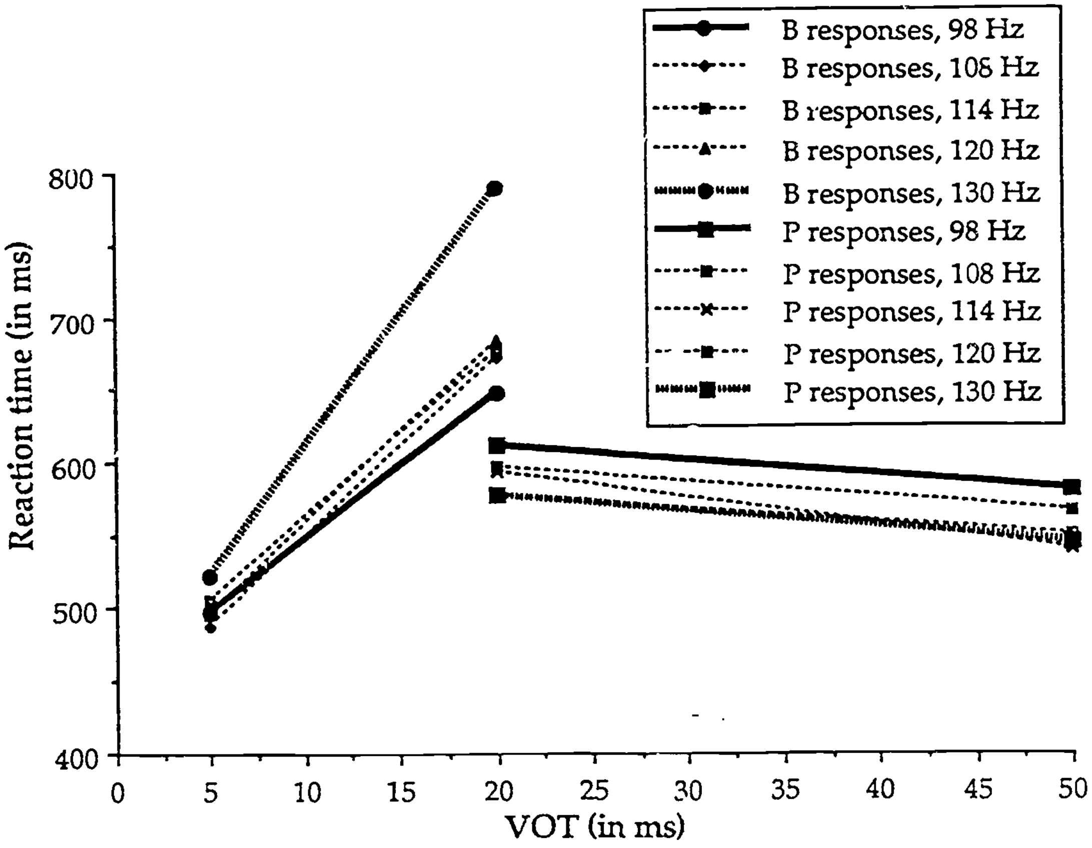


Figure 3. Reaction times for the complete range of F0 onsets, Experiment 1.

The "b" analysis included the six who heard the stimuli ambiguously plus the three who made at least 80% "b" judgments. Two of the six had no "b" responses to the 130 Hz stimulus. Those two cells were filled with the value for the most similar stimulus, namely, the 120 Hz stimulus. The "p" analysis again included the six who heard the stimuli ambiguously plus the other three subjects, those who made at least 80% "p" judgments. The means for the two sets of nine subjects are shown in Table 1. Despite the 64 ms difference, in the predicted direction, between the 98 and 130 Hz stimuli, the "b" analysis showed no effect of F0 (F(4,32) < 1, n.s.). The 85 ms difference, again in the predicted direction, for the "p" responses was a component of a significant effect (F(4,32) = 5.37, *p* < .01). Thus, though not conclusive, the evidence shows that the F0 also has an effect on reactions times to ambiguous stimuli, as well as on the labelling.

Table 1. *Reaction times for the two response categories. Each group consists of nine subjects, six being common to both.*

F0 (in Hz):	98	108	114	120	130
Means for "b"	657	686	712	683	721
Means for "p"	630	581	595	590	545

3. *Speeded Condition: VOT.* The results for the VOT condition are shown in Figure 4. As with Figure 3, the reaction time values are the means of the speeds for the 12 subjects reconverted into times. Even though the full VOT range was used, it was still the case that "b" responses above 20 ms and "p" responses below 20 ms were too few to analyze, so the two sets of functions overlap only at 20 ms. Here again, it can be seen that the 130

Hz onset slows decisions for "b" relative to the 98 Hz onset, while the reverse is true for "p" decisions. The effect appears larger at the ambiguous value for "b" but is absent for "p" at the ambiguous value.

For the statistical analysis of the unambiguous stimuli, only "b" responses to the short VOT stimuli and "p" responses to the long VOT stimuli were used. The 15, 20 and 25 ms VOTs were excluded from the analysis since they did not receive the 80% majority judgments needed to be called unambiguous The factors were Response Category, F0, and VOT. The VOT factor represented 5 and 10 for the "b" responses and 35 and 50 for the "p" responses.

The interaction of F0 and category, the most important result for the present study, was significant (F(1,11) = 6.75, $p < .05$), indicating that the effect of the same F0 differed depending on which category was being assessed. Inappropriate F0 caused an average delay of 17 ms in the identification. Response Category was also significant (F(1,11) = 5.44, $p < .05$), with "b" judgments being 52 ms faster. Neither F0 nor VOT was significant as a main effect (F(1,11) < 1, n.s.. and F(1,11) = 1.52, n.s., respectively). The interaction of Response Category and VOT was significant (F(1,11) = 6.77, $p < .05$). Times were somewhat slower the less extreme the VOT, as we would expect given previous results with stimuli that approach the ambiguous region (Pisoni and Tash, 1974; Whalen, 1991). The three-way interaction was not significant (F(1,11) = 2.34, n.s.).

C. Discussion

The unspeeded tests confirmed our previous results, showing some increase in the use of the F0 information as the subjects became more familiar with the stimuli. The speeded conditions showed that even when the VOT was unambiguous, the F0 information is taken into account.

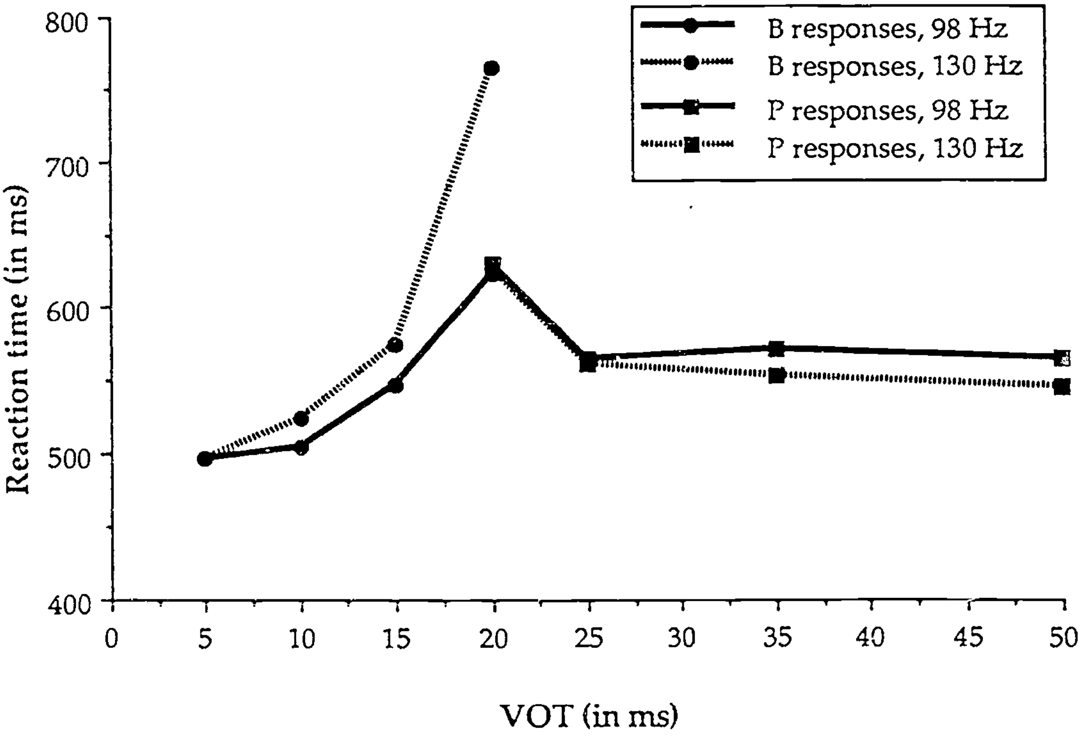


Figure 4. Reaction times for the complete range of VOT values, Experiment 1.

II. EXPERIMENT 2

The first experiment examined the dynamic perturbations at vowel onset. However, there may be differences throughout the vowel due to the stop voicing (e.g., Ohde, 1984; Whalen, 1990), though such differences are not found in every study (e.g., Löfqvist, Baer, McGarr, & Story, 1989). These level differences seem more relevant to tonogenesis than the initial perturbation since tonogenesis usually results in level tones, not contour tones. While the initial, dynamic portion of the F0 perturbation has been found in many studies, it seems that only Repp (1975) has examined level F0 differences. That study used dichotic presentation, where different syllables were presented to the two ears simultaneously and so is rather far removed from normal perception. The next experiment provides a simpler demonstration of the perceptual effectiveness of level F0 differences on voicing judgments.

A. Method

1. Stimuli. The stimuli were much like those in the first experiment, except that the F0 value at the onset was carried throughout the vowel. Thus the stimuli with an F0 of 98 Hz had 98 Hz throughout the vocalic segment, not just at the onset. As before, there were five F0 values (98, 108, 114, 120 and 130) and seven VOT values (5, 10, 15, 20, 25, 35, and 50).

2. Procedure. Two unspeeded conditions were run. In the first, five repetitions of all 35 stimuli were randomized together. In the second, only two values of F0 were used, namely, 108 and 120. These were the two values of a "b" F0 and a "p" F0 that differed by an amount closest to the 9 Hz difference found by Ohde (1984) in the midpoint of spoken vowels.

The final condition was a speeded one that used two values of F0 (108 and 120) and all seven VOT values. The instructions and equipment were the same as used in Experiment 1.

3. Subjects. Twelve subjects from the same pool used for Experiment 1 were run. Three had participated in Experiment 1. A technical problem resulted in the loss of the data for one subject, so the results of the remaining 11 will be reported.

B. Results

1. Unspeeded Conditions. Level F0s affected voicing judgments, just as F0 contours had. As can be seen in Figure 5, the percentage of "b" responses declines as the F0 increases, as predicted.

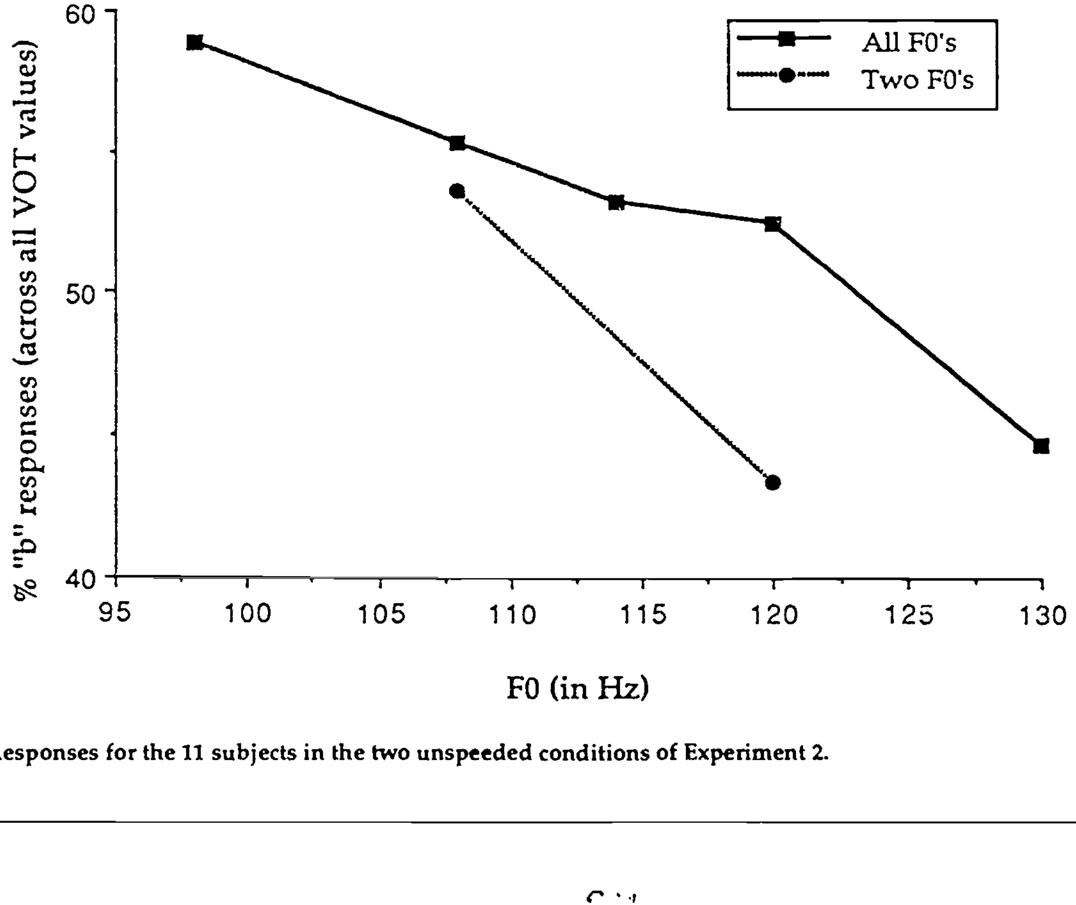


Figure 5. Responses for the 11 subjects in the two unspeeded conditions of Experiment 2.

For the statistical analysis, the total number of "b" responses across all the VOTs for each F0 was put through an analysis of variance with the single factor of F0 level. The effect of F0 on the number of "b" judgments is highly significant (F(4,40) = 7.65, $p < .001$). A Newman-Keuls post-hoc test reveals that responses to the highest F0 value are distinct from all the other F0 values, which do not differ among themselves. Still, it is clear that the F0 value of the syllable as a whole, not just the F0 perturbation, can affect the voicing judgment.

The second unspeeded condition was intended to ascertain whether subjects were able to treat the two F0 values as, in fact, separate. Since each syllable was presented as an isolated utterance, there was no immediate context, other than the experiment itself, by which to judge the relative height of the F0. So it was conceivable that the subjects would lose track of the "baseline" F0 and fail to show an effect.

As it turned out, there was a quite robust effect of the two F0 values on voicing identification, as seen in the dotted function of Figure 5. Taking the "b" responses to all of the VOT values, we find that the lower F0 elicited 53.7% "b"s, while the higher F0 elicited 43.4%. This difference was highly significant by a *t*-test (t(10) = 5.82, $p < .001$), showing that subjects were able to hear and make use of a 12 Hz difference in F0 across tokens. If anything, subjects were able to take the 120 Hz value as being at the top of the range, since it had an effect similar to the 130 Hz value in the previous condition (see Figure 5).

2. *Speeded Condition.* Before examining the response times, we need to check whether the identifications were consistent with those in the unspeeded task. If, for example, the time pressure reduced the effect of F0 on identification, we could not tell whether any reaction time difference was meaningful. In fact, the percentages were almost the same as before, with 53.2% "b"s for the 108 Hz F0, and 44.5% for the 120 Hz. This difference was also significant by a *t*-test (t(10) = 4.13, $p < .01$). Clearly, F0 retains its cue value in the speeded test.

The reaction time results are presented in Figure 6. As before, these are the means of the speeds of the responses retranslated into times. The individual boundaries for each subject varied much more than before, so that some subjects did not find the 20 ms stimulus ambiguous. Indeed, no matter where the ambiguous region was defined, there were subjects who did not, in fact, find that region ambiguous. So only the two least ambiguous VOT values of each category could be analyzed. The means of the "b" responses to the 5 and 10 ms VOTs were analyzed in one ANOVA with the factors VOT and F0, and the means of the "p" responses to the 35 and 50 ms VOTs were analyzed in another with the same factors. The "b" responses were significantly faster when the lower F0 was present (F(1,10) = 7.21, $p < .05$).

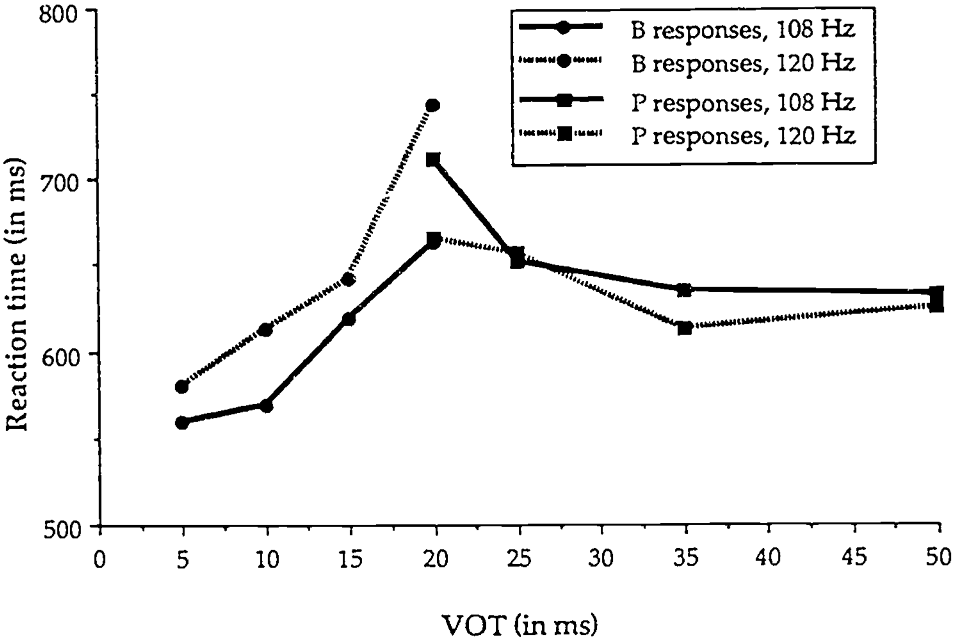


Figure 6. Reaction times for Experiment 2.

Subjects were also significantly faster on the 5 ms VOT than on the 10 (F(1,10) = 8.48, $p < .05$), as could be expected from previous work (Pisoni & Tash, 1974; Whalen, 1991). The interaction was not significant (F(1,10) = 1.15, n.s.). For the longer VOTs, the F0 effect did not reach significance (F(1,10) = 1.00, n.s.), even though the means are in the expected direction. Two subjects had large numbers of "b" responses throughout the continuum, which would contribute to making that end of the continuum less reliable than the short end. Neither the VOT effect (F(1,10) < 1, n.s.) nor the interaction (F(1,10) < 1, n.s.) was significant.

The differences in the boundary between voiced and voiceless varied so much that it was impossible to pick a single value of VOT at which all subjects had responses in both categories. In fact, the majority of the subjects (6 of the 11) did not have ambiguous cells for both keys at any one VOT. The statistical analysis of the ambiguous stimuli was therefore not attempted. Graphically, Figure 6 shows us the expected pattern of longer times for responses to stimuli with inappropriate F0 values.

So, despite the lack of significance in the longer VOTs, the shorter ones clearly show that inappropriate F0 values slow the identification of unambiguous syllables.

C. Discussion

The unspeeded tests showed that different F0 levels, even when they are only anchored by the experimental context, can be interpreted as voicing information. The speeded conditions showed that these level F0s could also be interpreted when the VOT was unambiguous, and the inappropriate values delayed identification time.

III. GENERAL DISCUSSION AND CONCLUSION

The voicing feature in English is realized as an aspiration difference (positive VOT) in utterance-initial position. Another aspect of initial voicing is that after voiceless stops, fundamental frequency (F0) falls somewhat when the voicing of the vocalic segment begins, and remains somewhat higher throughout the vowel, in contrast with voiced stops. This F0 difference has been shown to affect labelling only when the VOT was ambiguous (e.g., Abramson and Lisker, 1985). In the present experiments, listeners identifying stimuli that varied in VOT and F0 made use of the F0 information for ambiguous VOTs. This is an expected response to redundant features. The listeners also took the F0 into account with unambiguous VOTs, however, indicating that the redundant features are *always* taken into account.

Since the present experiments used synthetic stimuli to explore the listeners' behavior, there is, as always, the possibility that these results will not generalize to more natural stimuli and situations. The most clear case in the literature is the effect of lexical influence on voicing judgments (Ganong, 1980), which was found to disappear when the synthesis was improved (Burton, Baum, & Blumstein, 1989; McQueen, 1991). These concerns are greatly mitigated by the evidence of tonogenesis (see below) and by the small learning effect in the first experiment. Subjects were better able to use the F0 information within the "b" category after their initial exposure to the stimuli. It seems unlikely that an effect that depended on the strangeness of the stimuli would increase as familiarity with those stimuli increased. It would seem, rather, that subjects became more familiar with this "voice" and were able to accord it its full range of expression after the initial strangeness. This is also likely given that the two cues being dealt with in the present study are clearly phonetic, while the lexical effect mixes the phonetic with the extra-phonetic.

The mismatches inherent in these stimuli cannot be accounted for in psychophysical terms. If the stimuli included a mismatch that the ear could not resolve, we would expect there to be processing delays. However, for the vast majority of English utterance-initial stops, there must be a point at which the noise source gives way to a voiced source, whether this positive VOT is long or short and whether the category is voiced or voiceless (Lisker & Abramson, 1964). So the fact that such a change occurs in a stimulus can be neither more nor less psychophysically inappropriate for the matched F0s than for the mismatched ones. Similarly, even if a high or low F0 might be expected after a voiceless interval on purely physical terms, we need something additional to explain the F0 effect found here: high F0s slowed responses to short VOTs, but low F0s slowed responses to long VOTs. Along with the results of Whalen and Samuel (1985), we have solid evidence that the processing delays caused by these phonetic mismatches cannot be psychophysical in origin.

The results also make the theory of tonogenesis more plausible. In those cases in which the loss of a voicing distinction gives rise to new tonal categories (e.g., Abramson & Erickson, 1992;

Hombert, Ohala, & Ewan, 1979; Maddieson, 1984), we must make two assumptions. The first is that perturbations of F0 of the size found in natural productions must be perceptible. The present results, along with others, make this seem likely. Also, the learning that occurred in the present study indicates that even speakers who might not themselves depend on the F0 differences to make linguistic distinctions would be able to learn to appreciate those differences in other speakers. A second assumption is that the F0 effect of voicing must be enhanced before it can begin to be used distinctively. Otherwise, the loss of the voicing distinction would, of necessity, mean the loss of the F0 difference. This, of course, assumes that the configuration for the presence versus absence of voicing are directly responsible for the F0 perturbations. While such a view seems to hold true at present (Löfqvist et al., 1989), the number of languages that has been examined to date .s too small to reach any firm conclusions about whether the relationship is a necessary one and/or how widespread enhancement of the perturbations might be.

The present reaction time results clearly show that even when a phonologically primary feature is unambiguously specified, the perceptual system nonetheless takes a phonologically redundant feature (or purely phonetic effect) into account. This is perhaps to be expected for a system that can make use of those redundant features, but it has more often been proposed that these features are largely ignored unless the primary feature is impaired in some way. Instead, the perceptual system seems to be making use of all the information it has, even if it is phonologically redundant.

REFERENCES

Abramson, A. S., & Erickson, D. M. (1992). Tone splits and voicing shifts in Thai: Phonetic plausibility. In *Pan-Asiatic Linguistics: Proceedings of the Third International Symposium on Language and Linguistics* (Vol. 1, pp. 1-15). Bangkok: Chulalongkorn University.

Abramson, A. S., & Lisker, L. (1985). Relative power of cues: F0 shift versus voice timing. In V. A. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 25-33). New York: Academic.

Burton, M. W., Baum, S. R., & Blumstein, S. E. (1989). Lexical effects on the phonetic categorization of speech: The role of acoustic structure. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 567-575.

Fujimura, O. (1971). Remarks on stop consonants: Synthesis experiments and acoustic cues. In L. L. Hammerich, R. Jakobson, & E. Zwirner (Eds.), *Form and substance: Phonetic and linguistic papers presented to Eli Fischer-Jørgensen* (pp. 221-235). Copenhagen: Akademisk Forlag.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 110-125.

Hayes, B., & Lahiri, A. (1991). Bengali intonational phonology. *Natural Language and Linguistic Theory, 4,* 47-96.

Hombert, J. M. (1975). *Towards a theory of tonogenesis: An empirical, physiologically and perceptually-based account of the development of tonal contrasts in language.* Unpublished doctoral dissertation, University of California, Berkeley.

Hombert, J. M., Ohala, J., & Ewan, W. (1979). Phonetic explanation for the development of tones. *Language, 55.* 37-58.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America, 25,* 105-113.

Lehiste, I., & Peterson, G. E. (1961). Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America, 33,* 419-423.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20,* 385-422.

Lisker, L., & Abramson, A. S. (1971). Distinctive features and laryngeal control. *Language, 47,* 767-785.

Löfqvist, A., Baer, T., McGarr, N., & Story, R. S. (1989). The cricothyroid muscle in voicing control. *Journal of the Acoustical Society of America, 85,* 1314-1321.

Maddieson, I. (1984). The effects on F0 of a voicing distinction in sonorants and their implications for a theory of tonogenesis. *Journal of Phonetics, 12,* 9-15.

Martin, J. G., & Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects in /stri,stru/ sequences. *Journal of the Acoustical Society of America, 69,* S92.

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: stimulus quality in word-final ambiguity. *Journal of Experimental Psychology:* Human Perception and Performance, 17, 433-443.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics, 46,* 505-512.

Ohde, R. N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *Journal of the Acoustical Society of America, 75,* 224-230.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics, 15,* 285-290.

Repp, B. H. (1975). Dichotic masking of consonants by vowels. *Journal of the Acoustical Society of America, 57,* 724-735.

Shimizu, K. (1989). A cross-language study of voicing contrasts of stops. *Studia Phonologica, 23,* 1-12.

Silverman, K. (1987). *The structure and processing of fundamental frequency contours.* Unpublished doctoral dissertation, University of Cambridge.

Tomiak, G. R., Mullennix, J. W., & Sawusch, J. R. (1987). Integral processing of phonemes: Evidence for a phonetic mode of perception. *Journal of the Acoustical Society of America, 81,* 755-764.

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception and Psychophysics, 35,* 49-64.

Whalen, D. H. (1990). Coarticulation is largely planned. *Journal of Phonetics, 18,* 3-35.

Whalen, D. H. (1991). Categorical, prototypical and gradient theories of speech: Reaction time data. *Proceedings of the 12th International Congress of Phonetic Sciences* (Vol. 3, pp. 90-93). Aix-en-Provence: Universite de Provence.

Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1990). Gradient effects of fundamental frequency on stop consonant voicing judgments. *Phonetica, 47,* 36-49.

Whalen, D. H., & Samuel, A. S. (1985). Phonetic information is integrated across intervening nonlinguistic sounds. *Perception and Psychophysics, 37,* 579-587.

FOOTNOTES

[*] *Journal of the Acoustical Society of America, 93(4),* 2152-0000 (1993).

[†] Also University of Connecticut.

[‡] Also University of Pennsylvania.

[†††] Also City University of New York.

Haskins Laboratories Status Report on Speech Research
1992, SR-111/112, 23-42

Articulatory Phonology: An Overview*

Catherine P. Browman and Louis Goldstein†

INTRODUCTION

Gestures are characterizations of discrete, physically real events that unfold during the speech production process. Articulatory phonology attempts to describe lexical units in terms of these events and their interrelations, which means that gestures are basic units of contrast among lexical items as well as units of articulatory action. From our perspective, phonology is a set of relations among physically real events, a characterization of the systems and patterns that these events, the gestures, enter into. Thus, gestures are phonological events in the sense of Bird and Klein (1990).

While gestures are primitive phonological units, they do not correspond to either features or segments. Rather, they sometimes give the appearance of corresponding to features, and sometimes to segments. The issues discussed throughout are intended, among other things, to help clarify the differences among gestures, features, and segments. In addition, we will emphasize the following point throughout this paper: gestures and gestural organization can be used to capture both categorical and gradient information. Section 1 will present an overview of articulatory phonology, touching on a number of key aspects. Sections 2 and 3 will expand on examples in which a gestural analysis appears particularly fruitful. We will end in Section 4 with a discussion of how articulatory gestures provide a felicitous framework for dealing with language development.

1.1 Gestures as dynamic articulatory structures

Gestures are events that unfold during speech production and whose consequences can be observed in the movements of the speech articulators. These events consist of the formation and release of constrictions in the vocal tract. To help in explicitly modeling these events, gestures are defined in terms of *task dynamics* (Saltzman, 1986; Saltzman & Kelso, 1987; Saltzman & Munhall, 1989). Task dynamics has been used to model different kinds of coordinated multi-articulator actions, including those involved in reaching and those involved in speaking. In the case of speech, the tasks involve the formation of various constrictions relevant to the particular language being spoken. Task dynamics describes such tasks using damped second-order dynamical equations to characterize the movements; see Browman and Goldstein (1990a) and Hawkins (1992) for further discussions of the use of task dynamics to characterize speech.

One important aspect of task dynamics is that it is the motion of *tract variables* and not the motion of individual articulators that is characterized dynamically. A tract variable characterizes a dimension of vocal tract constriction; the articulators that contribute to the formation and release of this constriction are organized into a coordinative structure (Fowler, Rubin, Remez, & Turvey, 1980; Turvey, 1977). For example, the tract variable of lip aperture is affected by the action of three articulators: the upper lip, the lower lip, and the jaw. The current tract variables, and their component articulators, are displayed in Figure 1. An individual tract variable control regime is specified in terms of the set of articulators used to achieve a constriction and the values of the parameters in the dynamic equation describing its movement: target (rest position), stiffness, and damping.

Our thanks to Elliot Saltzman and Michael Studdert-Kennedy for critiquing a preliminary version of this paper, and to Lisa Zsiga and Simon Levy for help in manuscript preparation This work was supported by NSF grant BNS 882099 and NIH grants HD-01994 and DC-00121 to Haskins Laboratories.

	tract variable	articulators involved
LP	lip protrusion	upper & lower lips, jaw
LA	lip aperture	upper & lower lips, jaw
TTCL	tongue tip constrict location	tongue tip, tongue body, jaw
TTCD	tongue tip constrict degree	tongue tip, tongue body, jaw
TBCL	tongue body constrict location	tongue body, jaw
TBCD	tongue body constrict degree	tongue body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottis

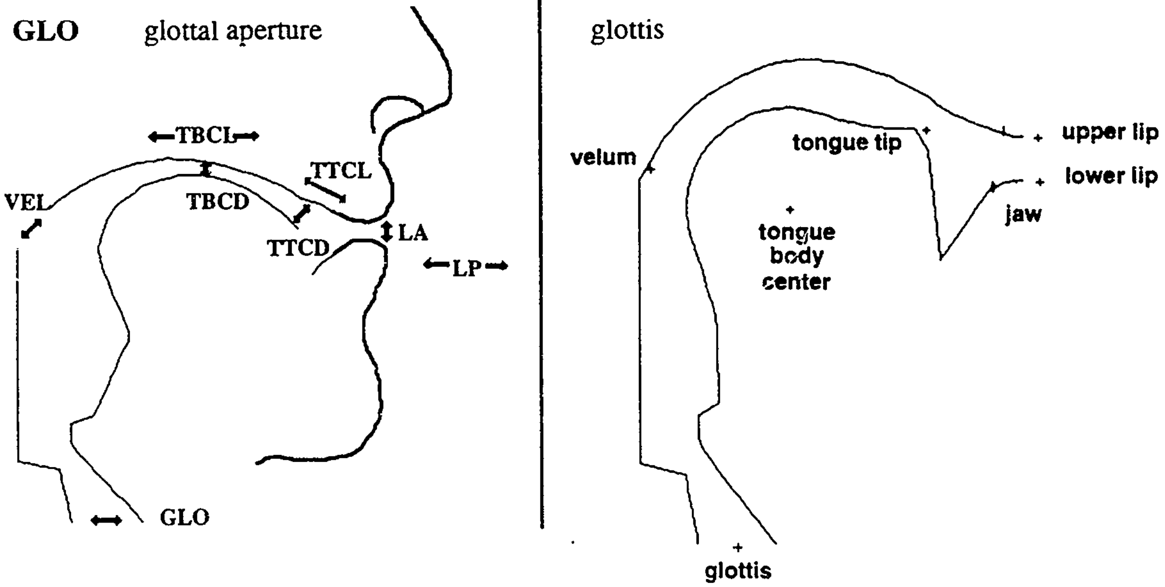


Figure 1. Tract variables and associated articulators.

These parameters provide a kind of internal structure for a control regime that underlies the spatiotemporal event in all its instances. A *gesture* in articulatory phonology is specified using a set of related tract variables. For example, in the oral tract the constriction location and degree are two dimensions of the same constriction, and therefore are considered related tract variables. In Figure 1, related tract variables contain the same first letter(s) in their names. Note that this means that each gesture is a local constriction, defined with respect to one of the five tract variable sets shown in the figure (lips, tongue tip, tongue body, velum, glottis).

Gestures can function as primitives of phonological contrast. That is, two lexical items will contrast if they differ in gestural composition. This difference can involve the presence or absence of a given gesture, parameter differences among gestures, or differences among organizations of the same gestures (discussed further in Section 1.2). This can be illustrated with the aid of displays showing the arrangement of gestural events over time. Lexical items contrast gesturally, first of all, if a given gesture is present or absent (e.g., "add" vs. "had," Figures 2a, 2b; "add" vs. "bad," Figures 2a, 2c; "bad" vs. "pad," Figures 2c, 2d; "pad" vs. "pan," Figures 2d, 2f). We assume that, in speech mode, the larynx is positioned appropriately for voicing unless otherwise instructed. Note that "had" and "bad" would typically be considered to differ from "add" by the presence of a segment, while "bad" and "pad," and "pad" and "pan," would contrast only in a single feature, voicing or nasality respectively. Gesturally, all these contrasts are conveyed by the presence or absence of a single gesture. Another kind of contrast is that in which gestures differ in their assembly, i.e., by involving different sets of articulators and tract variables, such as lip closure vs. tongue tip closure (e.g., "bad" vs. "dad," Figures 2c, 2e). All these differences are inherently categorically distinct.

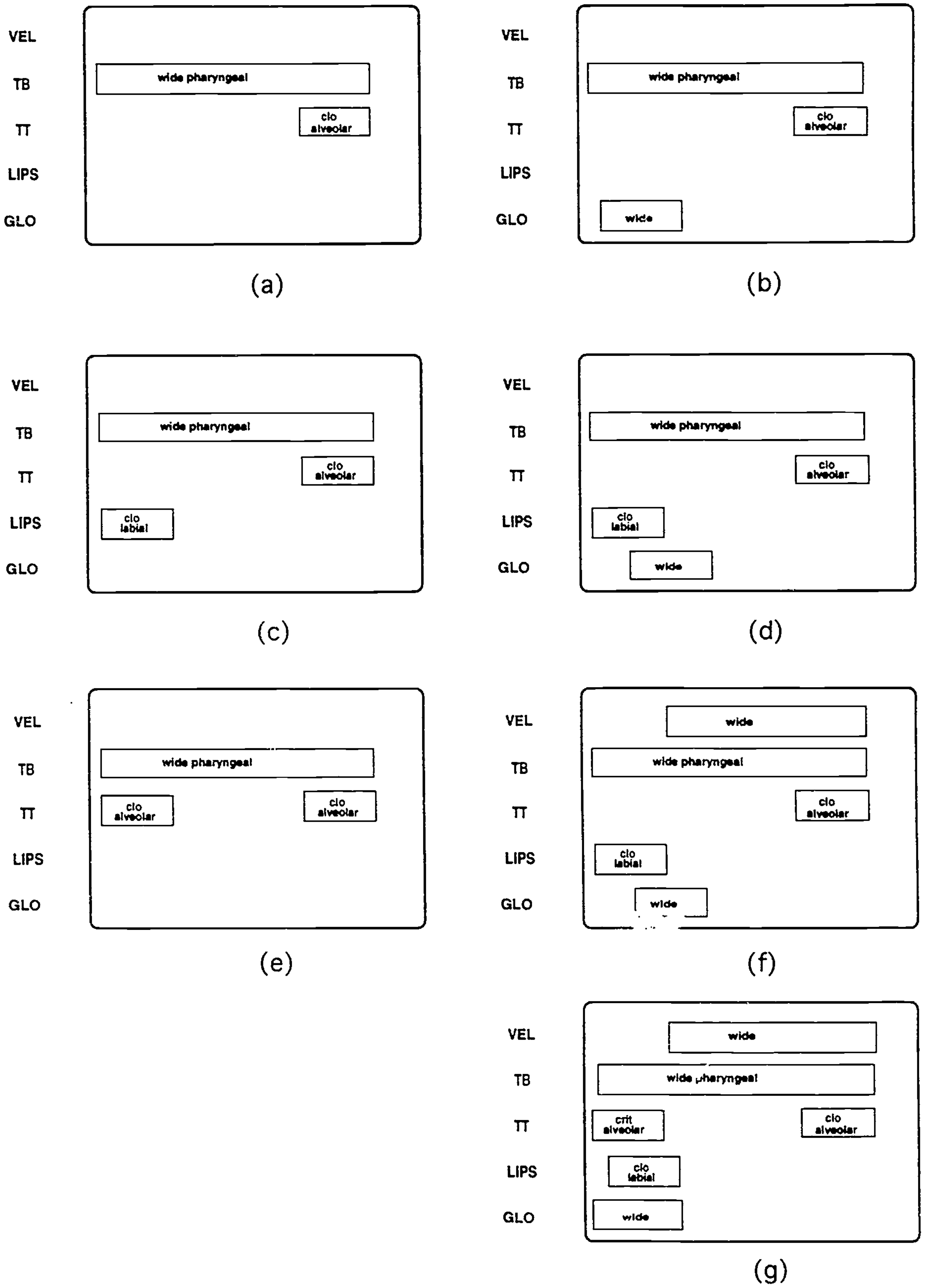


Figure 2. Schematic gestural scores. (a) "add" (b) "had" (c) "bad" (d) "pad" (e) "dad" (f) "pan" (g) "span."

Gestures can also differ parametrically, i.e., in the values of the dynamical parameters that define the spatiotemporal structure of the articulatory event, such as a target value for the tongue tip constriction degree that would lead to a complete closure vs. a critical value that would lead to the generation of turbulence (see gestures on TT tier in Figures 2g, 2e). While such differences are not inherently categorical, we have suggested (Browman & Goldstein, 1991) that distinct ranges of the possible parameter value space (for a given articulator set) will tend to be selected by a language on the basis of quantal articulatory-acoustic relations (e.g., Stevens, 1989) and/or on the basis of adaptive dispersion principles (e.g., Diehl, 1989; Lindblom & Engstrand, 1989; Lindblom, MacNeilage, & Studdert-Kennedy, 1983; Lindblom & Maddieson, 1988). In addition to target values for constriction degree, other dynamical parameters serve to distinguish gestures as well, as discussed in Browman and Goldstein (1989, 1990a): constriction location target, stiffness (possibly, vowels vs. glides), and damping (possibly, flaps vs. stops, in languages where they contrast).

Another major function of a phonological description is to represent natural classes. Since gestures are embedded in the vocal tract, the vocal tract itself acts to organize the gestures into a hierarchical articulatory geometry (Browman & Goldstein, 1989), the levels of which have been shown to represent natural classes by work in feature geometry (e.g., Sagey, 1986). The major organizational difference between this articulatory geometry and various feature geometries has been that, in the gestural approach, constriction degree (the closest gestural analog to continuancy) is low on the tree, in effect depending from the articulator node and sister to constriction location (place), whereas in feature geometries, continuancy has typically been close to the top of the feature tree. Recent work in feature geometry, however, has begun to lower the position of continuancy or its analogs such as aperture (e.g., Clements, in press). Indeed, based on generalizations about the phonological behavior of assimilations in a variety of languages, Padgett (1991) proposes that continuancy should be represented as depending from the articulator node, a proposal consistent with the gestural approach. Such a move of course supports the relevance of the gestural unit to the organization of phonological feature geometry.

For the velic and laryngeal subsystems, featural descriptions can sometimes appear very similar to gestural descriptions. Featural descriptions of the velic and laryngeal subsystems usually contain the constriction degree of the particular articulator as an inherent aspect; in these cases, they are very close to a gestural description (for example, [+nasal] corresponds to a velic opening gesture). However, even for the velic and laryngeal subsystems, there are situations in which a featural and a gestural analysis differ. For general discussions of distinctions in voicing and aspiration in the gestural framework, see Browman and Goldstein (1986) and Goldstein and Browman (1986). (This latter paper is part of an exchange with Keating, e.g., 1984, 1990, about the viability of featural and gestural accounts of various voicing phenomena). For a gestural analysis of the category of Hindi stop variously called "voiced aspirated," "breathy voiced," or "murmured," see Schiefer (1989), who compared a gestural account of these stops with a featural account in which the category is treated as a sequence of features (Dixit, 1987; also see Keating, 1990). Schiefer demonstrated that the sequential differences in these stops fall out naturally within the gestural framework, in which the breathy voice is realized with a single glottal gesture, timed comparatively late. Since gestures have an extent in time, and describe movements that change in amount of openness at different points during their realization, all the acoustic changes can be accounted for by this single glottal gesture (and its timing with respect to other gestures).

1.2 Gestural constellations: Combinations of overlapping gestures

As characterizations of physical events, gestures occur in space and over time. This has several implications. Since gestures have internal duration, they can overlap with each other; and since gestures are physical events, they are affected by physical processes occurring during the act of talking. In this section, we will focus on structure—how gestural overlap is used distinctively. Later sections will focus on process—how gestures vary in the act of talking.

The gestures that are employed in a given utterance are organized, or coordinated, into larger structures. We view the organization formed by those particular gestures as constituting the phonological structure of that utterance (or at least part of this structure). Of course, not every utterance in a language has an individual organization—there are general principles that define how classes of gestures are organized, or *phased*. These principles capture the syntagmatic aspect of a language's phonological structure, while the inventory of gestures that can participate in these organizations captures its paradigmatic aspect.

In the linguistic gestural component of the computational model currently being developed at Haskins Laboratories (see Figure 3), a first approximation of these phasing principles is used to coordinate the gestures with one another (Browman & Goldstein, 1990b). This gestural phasing results in a structure called a *gestural score*. A gestural score for the word "palm" (pronounced [p^{h}am]) can be seen in Figure 4. This representation displays the duration of the individual gestures as well as the overlap among the gestures. The horizontal extent of a given box indicates the discrete interval of time during which its particular set of values for the dynamic parameters is active. Given overlap, this means that several different gestures—sets of values—can be actively affecting the vocal tract at any particular instant in time. For example, in Figure 4, at time 50 ms, both the labial closure gesture and glottal gestures are active; by approximately time 125 ms., the labial closure gesture is no longer active but the tongue body narrow pharyngeal constriction has been activated for the vowel, so that at that point in time the glottal gesture and tongue body gesture are both active. Thus, with overlap the overall state of the vocal tract is dependent on more than one gesture. Articulatory phonology uses "tube geometry" to characterize the patterns arising from overlapping combinations of gestures. As proposed by Browman and Goldstein (1989) and further developed by Bird (1990), tube geometry represents the constriction degree effects at each level of the vocal tract (when viewed as a set of linked tubes), and in this way forms the basis for natural classes that have been defined using features such as [sonorant].

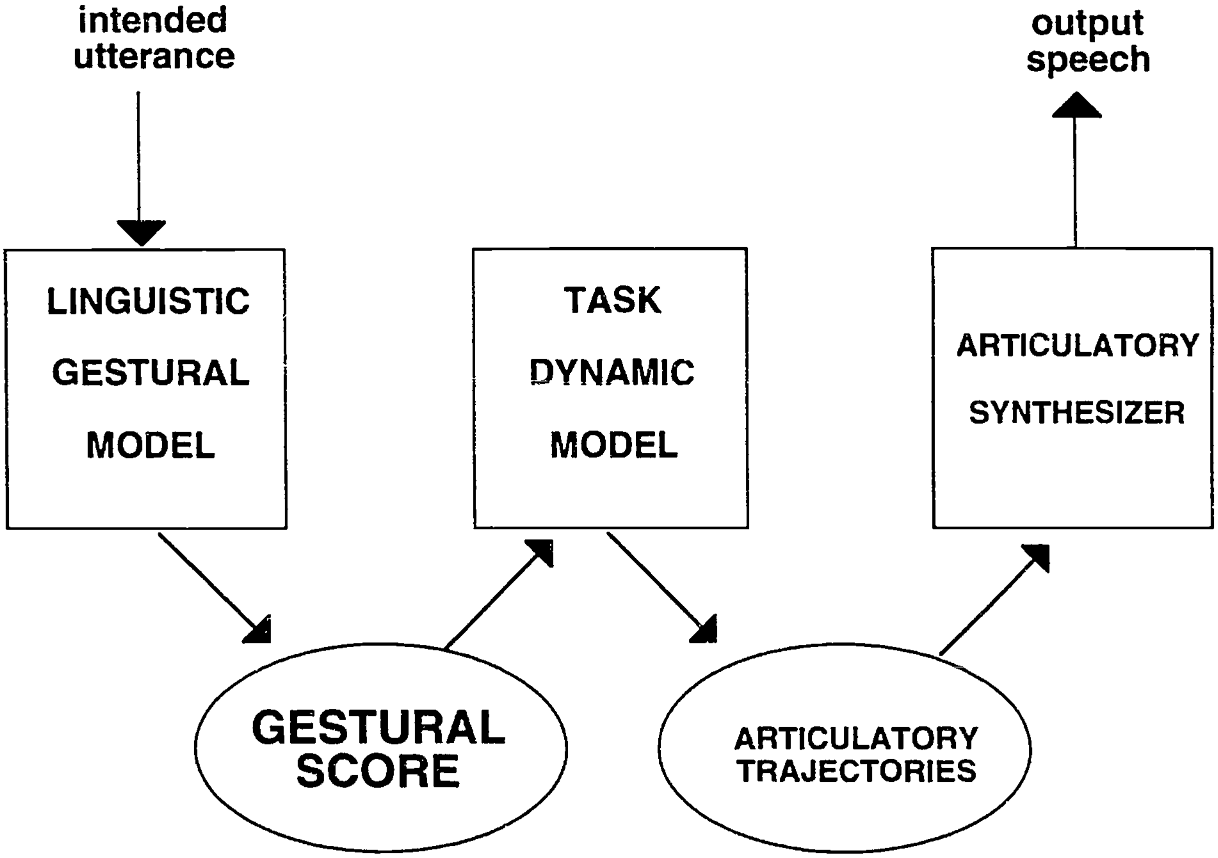


Figure 3. Gestural computational model.

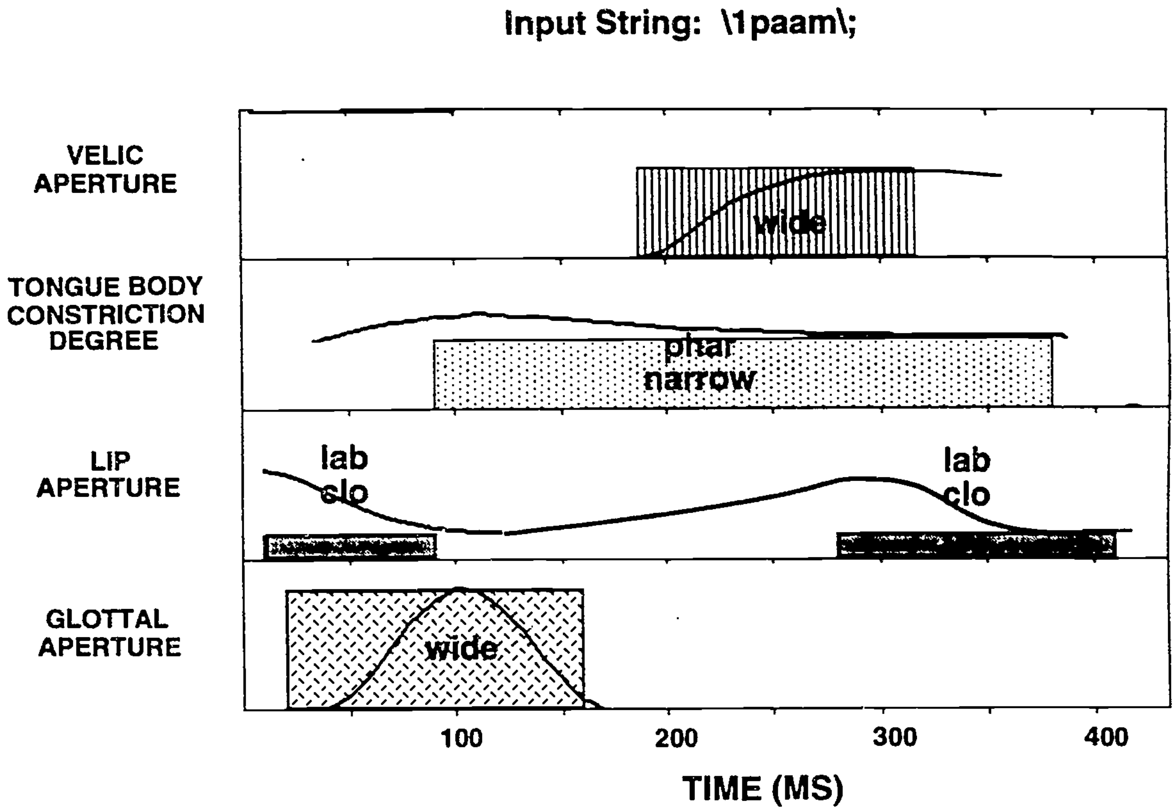


Figure 4. Gestural score for the utterance "palm" (pronounced [p^{h}am]), with boxes and tract variable motions as generated by the computational model. The input is specified in ARPAbet, so IPA /pam/ = ARPAbet /paam/. The boxes indicate gestural activation, and the curves the generated tract variable movements. Within each panel, the height of the box indicates the targeted degree of opening (aperture) for the relevant constriction: the higher the box (or curve), the greater the amount of opening.

As currently implemented in the computational model, the phasing statements coordinate pairs of gestures by specifying a particular dynamically-defined point in each gesture that is to be synchronized. A very restricted set of points is used, for consonants generally the achievement of the target or the beginning of movement away from the target, and occasionally the onset of movement towards the target. The importance of these or similar points has been noted by others. For example, Huffman (1990) suggested that closure onset and offset are among those "landmarks...[that] serve as the organizational pivots for articulatory coordination" p. 78. Krakow (1989) observed regularities in the timing of the movements of the velum and lower lip with regard to these points (to be further discussed in Section 2.2). Finally, both Kingston (1985, 1990) and Stevens (in press) have emphasized the importance of related points, but defined in the acoustic domain.

Notice (in Figure 4) that gestural scores provide an inherently underspecified representation (e.g., Browman & Goldstein, 1989), in that not every tract variable is specified at every point in time. This is most akin to the restricted underspecification argued for by Clements (1987) and Steriade (1987), among others. Notice also that gestural scores are exclusively tier-based. Hierarchical units such as syllables are currently generally represented by the mechanism of associations (phasing) among individual gestures rather than by hierarchical nodes. The only hierarchical unit for which we currently have evidence is that of the oral gestures in a (syllable-initial) consonant cluster (Browman & Goldstein, 1988). In these clusters, the oral gestures overlap only minimally rather than maximally as typically happens when gestures from different articulatory subsystems co-occur (e.g., the oral and glottal gestures in Figure 4).

Much of the richness of phonological structure, in the gestural framework, lies in the patterns of how gestures are coordinated in time with respect to one another. We have used the term *constellations* to refer to such gestural coordinations without pre-judging the correspondence between the constellations and traditional units of phonological structure (e.g., segments, syllables). Utterances comprised of the same gestures may contrast with one another in how the gestures are organized, i.e., the same gestures can form different constellations. Contrasts between nasal and prenasalized stops or between post-aspirated and pre-aspirated stops are possible examples of this kind. Considering only pair-wise combinations of gestures with a similar extent in time, Browman and Goldstein (1991) have proposed that possible contrasts in organization for these gestures are restricted to three distinct types of temporal overlap: minimal overlap, partial overlap, and complete overlap.

Gestural organization is constrained in more specific, language-dependent ways as well. For example, Browman and Goldstein (1986) proposed two organizational principles governing glottal opening-and-closing gestures occurring in word-initial onsets (for at least a subset of Germanic languages, including English): (1) that glottal peak opening is synchronized to the midpoint of any fricative gestures, and otherwise to the release of any closure gestures (following Yoshioka, Löfqvist, & Hirose, 1981) and (2) there is at most a single glottal gesture word-initially. Given these generalizations, word-initial "sp" and "p" are both presumed to have a single glottal gesture, as shown in Figures 2f and 2g (rather than the two glottal gestures for "sp" expected from a segmental analysis, see e.g., Saltzman & Munhall, 1989). The (allophonic) difference in aspiration between "sp" and "p" then follows automatically from timing principle (1) combined with the fact that gestures are events with temporal extent.

The fact that gestures are events with temporal extent can also eliminate the need for certain phonological adjacency constraints, which can often be seen to follow directly from gestural overlap. For example, much work in feature geometry (e.g., Clements, in press; McCarthy, 1988; Sagey, 1986) constrains assimilation to be the spreading of a feature to an adjacent slot, rather than the replacement of one feature by another. From the point of view of gestural overlap, many cases of "assimilation" or apparent "coarticulatory" feature-spreading follow directly from the fact that several gestures are co-occurring, either lexically or through later concatenation or sliding. (This will be discussed further in Sections 2 and 3; see also Bell-Berti & Harris, 1981, 1982; Boyce, 1990; Boyce, Krakow, Bell-Berti, & Gelfer, 1990; Fowler, 1980; Gelfer, Bell-Berti, & Harris, 1989). As these authors have also emphasized, there is no need to spread a feature, since gestures already have an inherent extent in time. A related constraint, that "total place assimilation in consonants will be restricted to immediately adjacent consonants" (Clements, in press, p. 29), also follows directly from gestural overlap. Zsiga (1993) discusses a number of cases in which overlap can account for various phonological phenomena (as well as some problem areas for a gestural account). In general, the existence of gestural overlap means that a number of phonological constraints follow automatically rather than having to be stipulated.

The general style of coordination (or phasing) between gestures may also vary from language to language. Smith (1988, 1991) has provided acoustic and articulatory evidence that temporal patterns in Italian and Japanese are affected differently by the change of an intervocalic consonant from singleton to geminate, and Dunn (1990) has found similar evidence in a comparison of Italian and Finnish. Smith found that, in Italian, no effect on the timing of the vowels was observed when consonants differed between singleton and geminate, but in Japanese, the intervowel organization was significantly altered. Such results are consistent with a gestural organization for Italian in which the vocalic gestures are directly phased with each other, and for Japanese in which vocalic gestures are phased only indirectly, by being phased with respect to the intervening consonantal gestures. In turn, such different coordination types are consistent with the characterization of Japanese as mora-timed (e.g., Han, 1962; Port, Dalby, & O'Dell, 1987) and Italian as syllable—(or possibly stress-) timed (e.g., Farnetani & Kori, 1986). The gestural account of such "rhythmic" differences as being due to a difference in direct or indirect coordination of vowels not only provides a potentially explanatory account of phonological differences, but predicts such phonetic detail as whether the vowels are shortened as intervening consonants are added (or lengthened).

2. CONTRAST AND ALLOPHONIC VARIATION

We often refer to a gestural analysis as an analysis of the "input," and more traditional analyses as analyses of the "output," where input and output refer to descriptions of the (local) articulatory gestural organization and resulting global vocal tract shape/acoustics, respectively.

Traditional segmental analyses are descriptions of the combined effects of the (overlapping) gestures in a gestural constellation, and therefore are typically descriptions of the acoustics and therefore the "output," in our terminology. Even featural descriptions often refer to attributes of segments, and are therefore often "output" descriptions. This is the source of the differences in description between the gestural approach, on the one hand[1] and segments and/or features on the other hand. An example of the descriptive differences has already been alluded to, re the voicing and aspiration issue (Browman & Goldstein, 1986; Goldstein & Browman, 1986; Keating, 1984, 1990; Schiefer, 1989). In this section, we will present a number of examples of gestural analyses of cases that have traditionally been analyzed in segmental and/or featural terms as different kinds of allophonic variation, showing that the gestural analyses capture a wider range of behavior, and do so by using general principles rather than special category-changing rules. At the same time, the underlying "input" structures capture contrast in a simple fashion.

Traditionally, the complement to contrast has been seen as identity. That is, two primitive phonological units either contrast or they are considered to be identical. Where this identity is at odds with phoneticians'/phonologists' percept of speech, this led historically to positing a single underlying phonemic (or phonological) unit, with distinct allophonic units in a more narrow phonetic representation (cf. discussion in Anderson, 1974). The same phoneme is "spelled" as categorically distinct allophones in different environments. However, when articulatory gestures are used as phonological primitives, much of the variation that was traditionally captured by a distribution of distinct allophonic units can, instead, be captured either by quantitative variation in the "input" parametric specification of a given gesture, or as a direct "output" consequence of overlap of invariant gestural units.

Generalizations. There are cases in which a gestural analysis reveals generalizations that have been missed in traditional allophonic descriptions. For example there are cases in which two very different allophonic rules (when couched in terms of segments and features) must be posited to describe what is quantitative variation in one and the same gesture in the same contexts. Further, there are cases in which particular prosodic contexts (e.g., stress and syllable positions) show a very similar influence on gestures of different types (oral and laryngeal, for example), or on their organization. We will discuss such cases below.

Relation between allophonic and other variation. There is much systematic, quantitative variation of speech gestures that has never been captured in a narrow allophonic transcription of the conventional sort, and could not be easily described in this way (e.g., differences in the magnitude and duration of stop consonant gestures in different prosodic environments—Browman & Goldstein, 1985; Kelso, V.-Bateson, Saltzman, & Kay, 1985; Munhall, Ostry, & Parush, 1985). As will be argued below, there is no principled difference between this kind of variation and the kind that has been annotated in a narrow transcription. In fact, we will examine cases in which the same parameter of variation has been treated as allophonic in some contexts and (implicitly) as quantitative in others. Moreover, as others have argued (e.g., Pierrehumbert, 1990; Sproat & Fujimura, 1989), this intermediate allophonic representation does not contribute in a useful way to the complete description of the variability. It is either unnecessary, or gets in the way of stating the generalizations. Thus, it seems that many allophonic differences are just quantitative differences that are large enough that phoneticians/phonologists have been able to notice them, and to relate them to distinctive differences found in other languages.

In this section, then, we will see that the very same syntagmatic organization will give rise to superficially different kinds of "output" variation such as "coarticulation" and allophonic differences, depending on the nature of the particular gestures in the organization (2.1, 2.2). In addition, we will see that general patterns of quantitative variation in gestural parameters can also give rise to a variety of superficially unrelated "output" consequences (2.3).

2.1 "Coarticulation" of consonants and vowels

In the phasing rules that are currently implemented in our model, oral constriction gestures are divided into two functional classes: vocalic and consonantal (Browman & Goldstein, 1990b). The distinction reflects the intrinsic differences between the two classes of gestures in their dynamical parameters. The consonantal gestures typically have a greater degree of constriction and a shorter time constant (higher stiffness) than the vocalic gestures. Syllable-sized organizations are defined by phasing (oral) consonant and vowel gestures with respect to one another. The basic relationship is that initial consonants are coordinated with vowel gesture onset, and final consonants with vowel gesture offset (the specific points being coordinated also differ in the two cases). This results in organizations in which there is substantial temporal overlap between movements associated with vowel and consonant gestures, as was seen in the gestural score of Figure 4.

When the same (invariant) consonant gesture is coproduced with different overlapping vowel gestures (e.g., in [ada] vs. [idi]), the articulator motions produced by the task dynamic model will differ, reflecting the vowel gestures' demands on the articulators that they share in common with the consonant. As discussed in Saltzman and Munhall (1989), the nature of this variation produced by the model will differ depending on whether the overlapping gestures are defined with respect to the same or distinct tract variables. In the case of distinct tract variables (e.g., TT for [d] and TB for vowels), the consonant gesture will achieve its invariantly specified tract variable (TT) target regardless of what vowel is overlapping, although the particular contribution of articulators used to achieve this target (jaw, tongue body, tongue tip) will differ depending on the vowel. Thus, the overall shape of the vocal tract produced during the tongue tip closure will differ in [ada] and [idi]. As shown in Saltzman and Munhall (1989), this difference corresponds to that seen in Öhman's (1967) X-rays. The different articulatory trajectories will produce different acoustic formant frequency transitions for the two stops, but apparently no difference in the consonant's percept (Fowler, 1980; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985).

In the case where consonants and vowels share the same (TB) tract variables (e.g., the consonant [g] as in [aga] or [igi]), the consonant and vowel gestures cannot both simultaneously achieve their targets, since they are attempting to move exactly the same structures to different positions. As a result the location (but not degree) of constriction achieved for the consonant will vary as a function of the overlapping vowel (Saltzman & Munhall, 1989). Again, this is consistent with the X-ray data of Öhman (1967). In this case, however, the difference is perceptible (at least to phoneticians), and has sometimes been represented by distinct "front" and "back" allophones.

These examples of consonant/vowel overlap illustrate two important points about gestural structures. First, they show how, as invariantly specified phonological units, gestures can give rise to context-dependent articulatory and acoustic trajectories, without having to posit any "implementation rules" for converting specific invariant (phonological) units into variable (physical) parameters. The variation follows directly from the definition of the units as parameterized task-dynamical systems, their phonological organization (pattern of overlap), and the general principles of how overlapping units blend. The same gestural structures simultaneously characterize phonological properties of the utterance (contrastive units and syntagmatic organization) and physical properties. Second, this example suggests how the very same syntagmatic structure (pattern of overlap) can yield different kinds of variation (allophonic vs. just "articulatory-acoustic"), as a function of the particular gestures involved—in particular, whether those gestures use the same or different articulator sets.

2.2 High-level units in velic and oral subsystems

Recently, the differing intergestural organization found in different (syllable) positions has been investigated in detail for two different gestural constellations in English: nasal consonants (Krakow, 1989) and /l/ (Sproat & Fujimura, submitted). Both are constellations comprising two gestures: a nasal consonant includes oral constriction and velic lowering gestures; /l/ includes tongue tip constriction and tongue body retraction gestures. Comparison of the data from these two papers reveals important similarities in how gestural organization varies as a function of position, despite differences in the traditional descriptions. For nasals, the traditional account characterizes syllable position differences by spreading the relevant feature ([nasal]) to the preceding vowel in the syllable-final case (e.g., Ladefoged, 1982), while for /l/, the position differences in certain dialects of English are handled by positing different allophones ("clear" vs. "dark," differing in the feature [back], e.g., Keating, 1985;) in initial and final position. However, as we saw with consonant-vowel overlap, this turns out to be an example in which the syntagmatic organization of the gestures is the same in these two cases, an aspect missed by the allophonic and featural descriptions.

Krakow's (1989) results show a clear difference in coordination between word-initial nasals (e.g., "see more") and word-final nasals (e.g., "seem ore"). In the word-initial case, the end of the velum lowering movement is roughly synchronous with the end of the lip closing movement. The gestures appear to be phased so that the effective achievement of their targets coincide. For the word-final case, however, the end of velum lowering occurs substantially earlier (100-350 ms) than the end of lip movement. In fact, the end of velum lowering appears to coincide with the *beginning* of the lip closing movement in this case. Syllable-position effects are similar to these word-position effects.

Sproat and Fujimura (submitted) found that the tongue body retraction (TB) and tongue tip raising (TT) movements for English /l/ also differ in their coordination as a function of word position. In word-initial position (e.g., "B. Likkovsky"), the extremum of the TB movement follows the TT extremum slightly, while in the word-final position (e.g., "Beel, equate") the TB extremum occurs substantially earlier than that for TT. Sproat and Fujimura manipulated the strength of the prosodic boundary following non-initial /l/, from none (e.g., "Beelik") to an intonation break (e.g., "Beel, equate"), and concluded that there is continuous variation in the relative timing of the two movements as a function of the boundary strength. However, examination of the relative timing data for prevocalic /l/ shows that, in general, truly word-final /l/s show the pattern with TB leading (with the magnitude of the lead affected by the strength of the following syntactic boundary), while the non-word-final cases (initial, medial, and medial before morphological boundaries) show either simultaneity or a slight lagging of TB.

There is an apparent similarity, then, in behavior of the gestures forming the constellations for nasals and /l/. Both constellations exhibit changes in relative timing as a function of word (or possibly syllable) position. In both cases, non-final position shows the gestures more nearly synchronous than in final position, and in both cases, it is the gesture with the narrower oral constriction (lip closure for the nasals, TT raising for /l/) that lags substantially in final position. In the case of the nasals, there is evidence for a specific shift in phasing: the end of velum lowering is coordinated with the end of lip closing for initials, but the beginning of the lip closing movement for finals.

It would strengthen the parallelism if evidence for such a shift also existed for /l/. Sproat and Fujimura did not examine this directly, although there is some indirect evidence in their data for such a shift. In final position, the TB gesture offset (as measured by movement extrema) precedes the TT gesture offset substantially. If the TB gesture offset were, in fact, being coordinated with TT gesture onset, as the analogy with the nasal behavior would predict, then as the TT movement increases in duration (e.g., before different boundaries), the measured offset-to-offset lag between the gestures should increase proportionally. Sproat and Fujimura measured the acoustic duration of the pre-boundary rime (which presumably is related to the acoustic duration of the /l/, and hence to the movement duration of TT); a clear correlation between this duration and the offset-to-offset lag for final /l/ can be observed in their Figure 8. This parallels a correlation between lip closure duration and offset-to-offset lag found by Krakow for the final nasals. Moreover, the points in Sproat and Fujimura's figure corresponding to non-final /l/ show TT leading, and do not appear to show any correlation between the magnitude of the TT lead and /l/ duration. This lack of correlation with duration would be expected if the offsets were being coordinated in this case, and such a lack of correlation is also found for non-final nasals.

The parallelism of nasals and /l/ reveals organizational patterns that are similar across subsystems and correlated with position in the word (or syllable). Viewing these behaviors gesturally suggests a (speculative) possible wider generalization, namely that there is a single syllable-final organizational pattern in which the wider constrictions always precede narrower constrictions (reminiscent of the sonority hierarchy; cf. also the related hypotheses of Sproat & Fujimura, submitted, and Mattingly, 1981). The same pattern would then be invoked for the (vocalic) tongue body and (consonantal) tongue tip gestures in "add," the two /l/-related tongue gestures in syllable-final /l/, and the velic and lip (or tongue) gestures in syllable-final nasals. Parallelism between the velic and oral subsystems has been noted elsewhere as well. For example, Browman (in press) showed how, if syllable-final vowel nasalization were treated as a long velic gesture, then similarities in behavior between syllable-final nasals and long oral gestures, i.e. geminates, on a gating task (Lahiri & Marslen-Wilson, 1992) could be explained.

The similarities across subsystems revealed in these studies are generalizations only in a gestural approach, and not in the more traditional analyses of these variations as being different in kind (in the nasal and /l/ example, as feature-spreading and different feature values, respectively). While the articulatory and acoustic consequences differ depending on the particular gestures involved, in a gestural approach these consequences do not need to be explicitly controlled, as they are automatic consequences of the syntagmatic organization and the particular gestures involved.

2.3 Glottal gestures: Positional (and other) variants

We have seen in previous subsections how what is traditionally described as contextual or allophonic variation can result automatically from the fact of overlap between invariant gestural units (e.g., overlap between consonants and vowels), or from differences in the characteristic patterns of overlap of gestures in syllable-initial and -final positions. In addition, some kinds of allophonic variation can be shown to result from quantitative variation in a gesture's dynamic parameters as a function of prosodic variables such as stress and position. Gestures shrink in space and in time in some contexts. This latter kind of variation is quite constrained—it scales the metric properties of a gestural event, but does not alter the composition of articulatory components out of which it is assembled.

Aspiration in English. A relevant example involves voiceless stops in English. Traditionally, these have been described as having aspirated and unaspirated allophones in different environments. Kahn (1976), for example, defines the environment that selects the aspirated allophone as "exclusively syllable-initial," with the unaspirated allophone occurring elsewhere. Kahn's rule assigns the feature [+spread glottis] in these aspirated environments, with [-spread glottis] generally being used for unaspirated allophones. This distinction is not an accurate characterization of the aspiration differences in English; nor is it either accurate or desirable to use a categorical rule to describe the aspiration of stops in English.

In many of the environments in which the output appears to be unaspirated, there is in fact a glottal opening-and-closing gesture present in the input. That is, presence or absence of aspiration in the output is generally not a discrete function of whether or not the glottis is spread, but rather is either a function of the timing of the glottal gesture with an associated oral gesture or a (gradient) function of the magnitude of the glottal gesture. The first cause of lack of aspiration in the output occurs in initial [s]-stop clusters, as mentioned in Section 1, in which lack of aspiration automatically results from the pattern of overlap among the contrastive gestures. As noted previously, English has a constraint that at most one glottal opening (spreading) gesture can occur in word-initial position. When this single gesture is associated with a fricative gesture, whether as a singleton or as a member of a sequence of oral gestures, the peak glottal opening is phased to the middle of the fricative gesture (probably its peak displacement). In the case of an [s]-stop cluster, this means that the glottis is already narrowed by the time the stop is released, which results in a "short lag" in the onset of voicing following release (VOT). This is the basis for the description of stops in such clusters as voiceless unaspirated (Lisker & Abramson, 1964).

The second cause of lack of aspiration in the output is the gradient reduction of glottal magnitude due to differences in stress and position. In analyses such as Kahn's, stress and position allophones are represented categorically. Voiceless stops are unaspirated in word-medial position before unstressed vowels (e.g., "rapid") because they are "ambisyllabic" rather than exclusively syllable-initial and therefore are represented as [-spread glottis]. However, voiceless stops are aspirated ([+spread glottis]) in the same position when before stressed vowels because they are considered to be syllable-initial. Single stops in word-initial position before either stressed or unstressed vowels are also aspirated and represented as [+spread glottis]. This categorical approach to aspiration is not supported by a recent study by Cooper (1991), who used transillumination to measure glottal aperture in four environments: initial vs. medial, before stressed and unstressed vowels.

Examining these four environments in two-syllable reiterant speech utterances (/pipip/, /titit/, and /kikik/), Cooper found, first of all, that there was a glottal spreading gesture in all four environments, contrary to the prediction that the unaspirated environment is [-spread]. Secondly, he found effects of both stress and word position on the magnitude of the glottal spreading gesture (in both space and time), with initial position and stress favoring larger gestures. Thus, the medial unstressed position showed the smallest glottal spreading gesture overall. From a gestural point of view, there is nothing special or categorically different about the medial unstressed case—it is simply the environment that shows the most gestural reduction because of the combined effect of stress and position. In an analysis such as Kahn's in which the medial unstressed case is viewed as an allophone categorically distinct from the form occurring in the other three environments, one would expect to observe qualitatively distinct laryngeal behavior in the medial unstressed case. This expectation is not borne out by Cooper's data. A weaker prediction of the categorical view is that there should be a robust interaction between stress and position factors, such that stress has a large effect medially, but little or no effect initially. This weaker prediction is also not borne out—the utterances with /t/ and /k/ generally show no interaction at all (although an interaction is observed for /p/). Cooper's own conclusion, based on additional experiments not summarized here, is that stress and word position, rather than syllable structure and aspiration category, are the relevant variables that regulate laryngeal behavior of voiceless consonants in English.

Voicelessness in final position differs from that in other positions. In final position (word or possibly syllable), the glottal spreading gesture in English is usually not observed at all (e.g., Lisker & Baer, 1984). However, the muscular activity normally associated with spreading gestures (increased activity of the posterior crico-arytenoid muscles, suppression of the interarytenoid) is found for such final stops in Lisker and Baer's data (and also in Hirose & Gay, 1972), although reduced in magnitude. This is consistent with a gestural reduction analysis: final position represents the most extreme case of reduction. However, analysis of final position is complicated by the fact that a constriction of the false (ventricular) folds is sometimes observed in this position (Fujimura & Sawashima, 1971; Manuel & Vatikiotis-Bateson, 1988). It is presumably this constriction that led Kahn to posit yet a third allophone for voiceless stops ([+constricted glottis]) in final position. Since the relation between this constriction and the muscular control of the glottis (proper) has not been explicitly investigated, it is not clear how to relate this constriction to the glottal spreading gesture.

Aspiration and "h." As reported above, positional and stress allophones of English voiceless stops result from quantitative variation in gesture magnitude (with the possible exception of the final ventricular constriction). Since the unit of reduction is the gesture, the gestural analysis predicts that similar patterns of reduction should be found, regardless of whether they have been analyzed as a segment ("h") or a feature ([+spread]). Pierrehumbert and Talkin (in press) have recently measured amount of reduction in glottal abduction for "h" in various prosodic contexts, using acoustic analysis to estimate the actual abduction. While most of their focus was on more global prosodic structure (phasal accent and intonation boundaries), they also found reduction effects due to word stress and position generally similar to those found by Cooper (1991) (although as noted above, Cooper's data shows some degree of influence of the supralaryngeal gesture on the laryngeal gesture). In a non-gestural approach, the similarity in behavior of "h" and [+spread] is not captured, since unlike aspiration in stops, the variation in "h" is not usually represented at all, even by distinct allophonic units (except where the reduction is so extreme that it is sometimes analyzed as deleted, for example in "vehicle"). In a gestural approach, however, the same reduction process gives rise to both kinds of variation.

There is also a symmetry in final position between voiceless stops and "h" in English. In final position, glottal spreading gestures are reduced to the limiting case of no observable opening. This is exactly the environment in which "h" does not occur in English. In a gestural framework, this distributional fact follows from the facts of reduction noted in voiceless stops. That is, words cannot have a contrastive glottal spreading gesture in final position, because such gestures are reduced to zero in final position, regardless of whether the glottal spreading gesture co-occurs with an oral constriction or not. (Contrast between final voiced and voiceless stops is possible only because this contrast involves other differences such as vowel length—Lisker, 1974—which can themselves be analyzed as overlap differences between consonant and vowel gestures, Fujimura, 1981). In more traditional approaches, this relationship between the distribution of "h" and the allophones of voiceless stops is not captured.

Generalizations across glottal and oral gestures. If the variation in the glottal gesture due to position and stress is in fact due to a general process, then such variation should be observed in other gestures occurring in similar environments. Similarities in the behavior of glottal and oral movements due to position and stress differences have indeed been observed.

The behavior of tongue tip movements is known to be affected by stress and position. For example, flapping of alveolar closures in English tends to occur in medial unstressed environments (Kahn, 1976), where we have seen that there is also substantial reduction in glottal spreading. If we assume that a flap is a reduced tongue tip closure gesture, reduced in time and possibly also in displacement, then the tongue tip and glottal gestures are behaving similarly. Apparent counter-examples are the medial unstressed alveolar stops that have not been considered to be flaps (e.g., in "after"). Since glottal gesture reduction applies in "after"—the "t" isn't aspirated—one would expect a reduced alveolar gesture here as well. However, these cases can be handled very nicely when input and output descriptions are properly distinguished. Although the alveolar in "after" is not considered to be a flap, it is possible that the alveolar closure *is* reduced in this context (input), but that the percept of a flap (output) depends on having an open vocal tract both before and after the reduced tongue tip movement. This analysis is related to that of Banner-Inouye (1989), who analyzes flapping in English autosegmentally as resulting from spreading of "open aperture" ([-cons]), from either side onto the timing slot associated with a coronal consonant. The phenomenon of flapping is thus analyzed by her as a short (single timing slot) open-closed-open contour that results from spreading in English. In the gestural framework, the reduction (making the movement short) would occur regardless of what other gestures are involved, but the description (or percept) of the resulting structure as a flap would depend on an open-closed-open acoustic contour (i.e., the structure in "butter" but not "after.") That is, the reduction process would always reduce the oral gesture in this environment, but the contour that is perceived as a flap would simply be one of the possible output consequences, depending on the appropriate set of gestures.

There are also potential parallels between glottal spreading and tongue tip closure gestures in final position. As we shall see in the next section, final alveolar closure gestures are subject to a variable amount of reduction in final position, including the failure to achieve any tongue tip contact. This is, of course, reminiscent of the frequent failure to see any actual glottal opening finally. When such reduced final alveolars coincide with the ventricular constriction discussed above, this produces the structure that has traditionally been described as the glottal stop [ʔ] allophone of /t/. The confluence of these events can be seen in the fibroscopic and palatographic data of Manuel and Vatikiotis-Bateson (1988).

Other oral constriction gestures also exhibit patterns of reduction similar to those exhibited by the glottal spreading and alveolar closure gestures. For example, bilabial closure gestures show effects of stress (e.g., Beckman, Edwards, & Fletcher, 1992; Kelso et al., 1985) and stress/position (Browman & Goldstein, 1985; Smith, Browman, McGowan, & Kay, submitted), similar to those shown by glottal gestures. These papers show substantial reduction of labial gestures in non-initial reduced syllables (initial reduced syllables were not examined). Thus, the reduction processes associated with stress and position in English for glottal gestures appear to be general, operating on tongue tip and labial gestures occurring in the same environments. Note again that while the variation in the dynamics of the tongue tip gesture has been represented as allophonic, the variation in the lip gesture has not been. Yet both seem to be instances of a very general reduction process, one that also operates on glottal gestures.

In addition to looking at similarities in the environments in which different kind of reduction occur, it is possible to focus on the form of the reduction itself, as observed in the dynamic properties of the gestures. Munhall et al. (1985) have demonstrated similarities in the velocity profiles of movements of the glottis and the tongue dorsum (in /k/). In addition, the quantitative changes in the kinematic properties (i.e., displacement, duration, peak velocity) for different stress conditions were shown to be similar for the tongue dorsum and glottal movements.

In summary, allophonic variation associated with prosodic variables such as position and stress has been shown, in many cases, to be a constrained quantitative and gradient variation, rather than a categorical variation. Viewing such as gradient changes within a gestural framework captures similarities in behavior across position and stress and across different featural and segmental characterizations of glottal spreading gestures, and also captures similarities in behavior across different articulatory subsystems.

3. VARIATION DURING THE ACT OF TALKING

In this section, we examine some of the consequences of using the gestural approach to analyze phonological and phonetic variation that can be attributed to processes occurring during the physical act of talking. This variation arises from two interlocking sources, one gradient and one categorical. Beginning with a contrastive canonical gestural structure, processes occurring during the act of talking will cause gradient changes that can ultimately be perceived as a categorically different gestural structure. This is due, among other things, to the fact that the acoustic (as well as articulatory) consequences of a given invariantly specified gesture will differ depending on what other gestures are concurrently active (Browman & Goldstein, 1990a, 1990b). The following examples will show how the constrained processes available in the gestural view provide a unified and explanatory view of a variety of superficially different kinds of phonetic and phonological alternations.

3.1 Speech production errors: Connected speech

One aspect of the act of talking that appears to be well handled by a gestural account is that of speech production errors. Mowrey and MacKay (1990) recorded muscle activity for [l] during experimentally induced speech errors in tongue twisters such as "Bob flew by Bligh Bay." In one session, about a third of the 150 tokens showed anomalous muscle activity, such as insertion of [l] activity in "Bob" or "Bay" and diminution of [l] activity in "flew" or "Bligh." Only five of these tokens, however, involved all-or-none behavior; most of the activity was gradient. That is, the magnitude of activity in both the inserted and "original" [l] fell on a continuum. Some of the errors were small enough so that they were not audible. The timing of the inserted activity was, however, localized and consistent. Such errors, in which the positioning (organization) is categorical but the magnitude is gradient, can be handled very naturally in a gestural framework.

Another aspect of the act of talking that is well handled in the gestural framework involves alternations that occur in connected speech. As shown in some of the data summarized below, in connected speech the patterns of gestural overlap may vary. In particular, factors associated with increased fluency (e.g., increased rate, more informal style) result in increasing the temporal overlap among gestures. Additionally, prosodic boundaries may influence the degree of overlap between neighboring gestures that belong to successive words. We have hypothesized that this kind of variation can result in changes that have traditionally been described as "fast speech" alternations of various sorts, and have presented articulatory evidence for this (Browman & Goldstein, 1990a, 1990b). However, it is important to note that such gestural sliding is endemic in talking (e.g., Hardcastle, 1985), and not limited to the cases that have been noted as alternations. Thus, this is another situation (like those discussed in Section 2) in which some, but not others, of the results of a single gradient process have been noted in phonetic transcriptions. In a gestural account, a single generalization (increase in overlap) characterizes all these cases.

Evidence for increased overlap as rate increases has been presented for consonant and vowel gestures (Engstrand, 1988; Gay, 1981) and for the laryngeal gestures for two voiceless consonants in contiguous words (Munhall & Löfquist, 1992). Hardcastle (1985) has presented evidence for variation in gestural overlap as a function of prosodic boundary strength as well as rate. Using electropalatography, he measured overlap in time between the dorsal closure for /k/ and the onset of the tip/blade contact for a following /l/. The /kl/ sequences employed included word-initial clusters and examples in which the /k/ and /l/ were separated by various boundaries (syllable, word, clause, and sentence). Sentences were read at fast and slow rates. In general, the amount of overlap was consistently greater at the fast rate than at the slow rate. The effect was observed in all phonological and syntactic contexts, but was largest at the clause and sentence boundaries. Here, slow rates often showed long "separation" intervals between the gestures (rather than overlap), while fast rates tended to show considerable overlap, often greater than that seen in the within-word or within-phrase cases. Thus, both rate and prosodic boundaries influence gestural overlap.

In this example, variation in gestural overlap did not produce changes that have been described as connected speech alternations. However, we have proposed (Browman & Goldstein, 1990b) that there are circumstances in which increased overlap would result in such alternations. One such circumstance we refer to as gestural "hiding." This occurs when gestures employing distinct tract variables (cf. Section 2.1) increase their overlap to such an extent that even though all the relevant constrictions are formed, one of them may be acoustically (and perceptually) hidden by another overlapping gesture (or gestures). X-ray evidence for this hiding analysis was provided in Browman and Goldstein (1990b). For example, two productions of the sequence "perfect memory" were analyzed, one produced as part of a word list (and thus with an intonation boundary between the two words), the other produced as part of a fluent phrase. In the fluent phrase version, the final [t] of "perfect" was not audible, and it would be conventionally analyzed as an example of alveolar stop deletion in clusters (e.g., Guy, 1980).

However, the articulator movements suggested that the alveolar closure gesture (for the [t]) still occurred in the fluent version, with much the same magnitude as in the word list version that had a clearly audible final [t]. The difference was that in the fluent version, the alveolar closure was completely overlapped by other stop gestures—the closure portion by the preceding velar closure ([k]), the release portion by the following labial closure (for the [m]). Thus, from the point of view of an articulatory phonology, all the phonetic units (gestures) were present in both versions. The difference between the list and fluent forms was due to variation in the gradient details of overlap, a process for which there is independent evidence. In other contexts, for example when a velic lowering gesture co-occurred with the hidden gesture, hiding produced apparent assimilations, rather than deletions. Thus, in the phrase "seven plus" produced at a fast rate, the final consonant of "seven" was audibly [m], but evidence for an alveolar closure was still present. Only a single gesture was hidden (the oral alveolar closure gesture) and not a segment-sized constellation of gestures. It is precisely this fact that leads to the percept of assimilation rather than deletion in this kind of example.

In analyzing casual speech alternations as resulting from gestural overlap, we were led to make the strong hypothesis (Browman & Goldstein, 1990b) that *all* examples of fluent speech alternations are due to two gradient modifications to gestural structure during the act of talking—(a) increase in overlap and (b) decrease in gesture magnitude. (The latter modification is related to the gestural modifications as a function oi prosodic structure discussed in Section 2). A typical example of magnitude reduction might be the pronunciation of the mediai (velar) consonant in "cookie" as a fricative rather than as a stop (Brown, 1977). Under this hypothesis, casual speech variation is quite constrained: all the lexical phonological units are present, though they may be decreased in magnitude and overlapped by other gestures. Gestures are never changed into other gestures, nor are gestures added.

3.2 Assimilation of final alveolars

A related hypothesis has been proposed by Nolan (in press), based on analyses of apparent assimilations of single final alveolar stops to following labial and velar stops (e.g., /t/——>[k] in "...late calls..."). Using electropalatographic contact patterns, he found that the final alveolars were present, but reduced in degree to a variable extent, in the forms that were perceived as assimilated (see also Barry, 1985; Kerswill, 1985 for examples of such "residual" tongue tip gestures). Moreover, even in cases in which no alveolar electropalatographic contact was observed, the assimilated forms were perceptually distinguishable from forms with no lexical alveolar stop gesture at all (e.g., assimilated "bed" vs. "beg"). These findings led Nolan to propose that "differences in lexical phonological form will always result in distinct articulatory gestures." From the point of view of articulatory phonology, this constraint follows quite naturally—the phonological form *is* an organization of gestural events.

Nolan's experiments on the class of final alveolar assimilations focussed on the role played by the reduction of the tongue tip gesture. In addition to reduction, however, the overlap between that gesture (reduced or not) and the following stop gesture may play a role in perceived assimilations. The role of overlap in the acoustics and perception of similar assimilations was investigated by Byrd (1990). Using the computational gestural model discussed in section 1.2, Byrd generated utterances with a continuum of overlap for each of the phrases "bad ban" and "bab dan" by systematically varying the overlap between the alveolar and bilabial closure gestures. She found an asymmetry between the perceptions of the gestures in word-final position. When the first word ended in [d], the word-final alveolar was perceived as being assimilated to the following [b] when overlap increased substantially. However, with the same amount of overlap, the word-final [b] was not assimilated, and in fact, the following word-initial [d] in such cases tended to be perceived as being assimilated to the [b]. (An asymmetry in the same direction, although less extreme, was found when subjects listened to the first word extracted). Byrd related this perceptual asymmetry in favor of the labial closure to the VC and CV formant transitions produced by synchronous (overlapping) labial and alveolar closure gestures. In general, such formant transitions were more similar to those produced by labial stops alone than those produced by alveolar stops alone. Thus, the effect of overlap tended to obscure final alveolars but not final labials. This could contribute to the tendency in English for final alveolar stops (but not final labials or velars) to assimilate to following stops (Gimson, 1962).

The simulation results of Byrd suggest that formant frequency transitions into final alveolar stops should vary as a function of the following stop (as long as they are at least partially overlapping). This hypothesized acoustic "context effect" was confirmed in an investigation of natural speech by Zsiga and Byrd (1990). They examined formant frequency transitions into the medial closure in phrases like "bad pick," "bad tick," and "bad kick" produced at different rates. The major finding was that formant transitions shifted away from those expected for an alveolar stop towards those expected for the following consonant—either a labial stop, as in "bad pick," or a velar stop, as in "bad kick." In the case of the following labial, the effects on formant transitions agreed with those observed in Byrd's simulations of "bed ban" in which the labial closure gesture overlapped the alveolar gesture—both F2 and F3 were lower at the offset of the first vowel for "bad pick" than for "bad tick." The magnitude of these effects was generally smaller than that found in Byrd's complete synchrony condition, which is consistent with the fact that final alveolar consonants in this natural speech experiment were actually perceived as such and were not assimilated to the following labials or velars. In general, perceptual assimilation should occur only when the effects of gradient overlap and reduction exceed some perceptual threshold.

A second finding of Zsiga and Byrd's was that, for utterances where the second word in the phrase began with a velar stop (e.g., "bad kick"), a systematic relation was observed between temporal and spectral properties as rate was varied. When rate variation resulted in a decrease in the total duration of the medial closure, there was also an increase in the velar effects seen in the formant transitions. This relation can be simply accounted for by assuming that these cases involve increased overlap between the tongue tip and tongue body gestures.

Finally, an ongoing experiment by A. Suprenant is explicitly testing the relative contributions of overlap and gestural magnitude to the percept of final stops. The experiment employs tokens of utterances like "MY pot puddles" collected at the X-ray microbeam facility at the University of Wisconsin. These tokens show variation in both the magnitude of tongue tip raising for the final [t] in "pot" and in the temporal overlap of that gesture and the lip closure gesture of the following word. Listeners are presented with these sentences in a speeded "detection" task. Preliminary results suggest that detection of "t" is a function both of its magnitude and amount of overlap with the following consonant.

3.3 Reduced syllable deletion

Assimilations (and deletions) of stop consonants represent only one kind of fluent speech alternation. Another example that follows directly from changes in gestural overlap is deletion of schwa in reduced syllables. For example, in a word like "beret," the vowel in the first syllable, either [əʳ] or [ɹ̩] may be apparently deleted in continuous speech, producing something transcribed as [bɹeɪ]. The tendency for deletion has been shown to be a "graded" one, dependent on a number of contextual factors (e.g., Dalby, 1984). We have demonstrated (Browman & Goldstein, 1990a) that the concomitant shift in syllabicity could be the perceptual consequence of an increase in overlap between the initial labial closure gesture and the tongue gestures for the "r." This was shown by using the computational gestural model to generate a continuum in which the degree of overlap or separation between the control regimes for the labial closure and the "r" varied in small steps. In the canonical organization for "beret," the labial and "r" gestures did not overlap at all. This meant that the labial gesture was released before the "r" was formed. This differed from the canonical organization for "bray," in which the gestures were partially overlapping (like the velar and "l" gesture in the clusters illustrated in Hardcastle, 1985). When listening to items from the continuum in a forced choice test, subjects responded with "bray" to items in which labial and "r" gestures overlapped, and "beret" to items in which they did not overlap.

Thus it is possible to view reduced syllable deletion as resulting from an increase in gestural overlap in fluent speech. This treatment is attractive for two reasons. First, it treats deletion as resulting from the same general process that gives rise to other (superficially unrelated) alternations. Second, it leaves us with the claim that all phonetic units constituting a lexical item are still present in fluent speech; only the overlap has changed, in a predictably gradient way. This seems to be a more natural treatment than one which would assume that an important structural unit (a syllable) is suddenly and completely eliminated in fluent, connected speech.

Another important aspect of this treatment of reduced syllables is the fact that the lexical difference between "bray" and "beret" was modeled only in terms of the coordination of labial closure and "r" gestures. There was no explicit tongue gesture for a schwa. This hypothesis was sufficient to generate gestural scores that produced speech with the appropriate perceptual properties, for both "bray" and "beret." In addition, the overlap of the vertical components of their articulatory trajectories was consistent with tokens of this distinction collected using the X-ray microbeam system at the University of Wisconsin (Browman & Goldstein, 1990a).

However, in another investigation of reduced syllables (Browman & Goldstein, in press), data analysis and modeling revealed that an explicit tongue gesture for a schwa *was* required in utterances of the form ['pVpəpVp], although the target of the required gesture was completely colorless in that it was the average of the tongue body positions for all full vowels for that speaker. Therefore, at the very least, development of a more complete typology of the gestural structure of reduced syllables is needed, and is currently being pursued, to evaluate the phonological and morphological conditions for schwas of various kinds, both in English and other languages. With respect to deletion processes, however, we should note that even if there is a tongue gesture associated with a particular schwa, increase in overlap between consonants on either side of it could result in hiding that gesture. Thus, even if an active schwa gesture is required in a word like "difficult," increase in overlap so that the labiodental fricative and the velar stop partially overlap could result in hiding of this gesture.

In summary, increase in overlap among gestures in fluent speech is a general gradient process that can produce apparent (perceived) discrete alternations. The examples above were describable as consonant deletions, consonant assimilations, and vowel deletions; another possible example is that of epenthetic stops in English (e.g., Anderson, 1976; Ohala, 1974), as discussed in Browman and Goldstein (1990b). However, the fact that stop epenthesis in words like "tense" is not found in some dialects of English (South African: Fourakis, 1980) raises the larger issue of variability of fluent speech alternations across dialects and languages. That is, if the process of increase in overlap is a completely general property of talking, why does it create epenthetic stops in one dialect but not another? We have suggested (Browman & Goldstein, 1989) that such dialect/language differences may arise from differences in the canonical patterns of coordination in the different languages. Two kinds of coordination differences are relevant here. First, languages may differ in the amount of canonical overlap between two gestures. For example, sequences of stops in English are canonically partially overlapping (Catford, 1977), whereas sequences in Georgian, for example, are canonically non-overlapping, i.e., are released stops (Anderson, 1974). We would expect that an amount of increase in overlap that produces hiding in English would not necessarily do so in a language such as Georgian. Second, two gestures may be directly phased with respect to one another in one language, but only indirectly phased in another language (as discussed in Section 1.2). It is possible that gestures that are directly phased will be more likely to retain their canonical organization in connected speech.

4. DEVELOPMENTAL DATA

Developmental studies show that a child's first words are stored and retrieved not as phonemes but as holistic patterns of "articulatory routines" (e.g., Ferguson & Farwell, 1975; Fry, 1966; Locke, 1983; Studdert-Kennedy, 1987; Vihman, 1991). Recent research has suggested that the basic units of these articulatory routines are discrete gestures that emerge pre-linguistically (during babbling), and which can be seen as early "gross" versions of the gestures that adults use (e.g., Browman & Goldstein, 1989; Studdert-Kennedy, 1987; Studdert-Kennedy & Goodell, in press). Further development can be viewed as differentiation (in terms of parameter values), and coordination of these basic gestures. For example, other recent studies (Fowler, Brady, & Curley, 1991; Nittrouer, Studdert-Kennedy, & McGowan, 1989) have shown that coordination into segmental-sized units (one kind of constellation) only appears gradually during the course of language acquisition, which not only supports the contention that phonemes are not present in a child's first words, but also suggests that higher-level units are formed out of smaller units during the course of language development. If so, then articulatory phonology would provide a very appropriate approach to child language, and its use would facilitate the study of language development both theoretically and methodologically, since both child and adult utterances can be described in terms of the same basic primitives of gestures.

Fowler, Brady, and Curley (1991) studied experimentally induced speech production errors in CVC utterances by children and adults, using phonetic transcriptions by trained listeners to indicate the existence of an error. The purpose of the study was to test the hypothesis that organization into phonological structures smaller than the level of the lexical item only appears gradually during the course of language-learning. Fowler et al. found that younger children were much more prone to blend features in their errors than were adults, as in the error "bam till" from the utterance "pam dill." Adults were correspondingly more likely to retain higher level organization, whether segmental or subsyllabic, tnat is to produce the error "dam pill" from the utterance "pam dill." Thus, in this experiment with single-segment onsets, onset (or segment) exchanges increased with age (4 & 5-year-olds 33%, 8-year-olds 44%, and adults 74%), while feature blends decreased (4 & 5-year-olds 33%, 8-year-olds 18%, and adults 8%).

The Fowler et al. results support the hypothesis that lexical organization intermediate between the levels of the feature (or gesture) and the word develops as part of learning the language. However, the results do not distinguish between a featural analysis and a gestural analysis. Another study, that of Studdert-Kennedy and Goodell (in press), supports the gesture as the unit out of which words are formed as the child develops language. This study focussed on another kind of "error," the differences between the child's pronunciation and the canonical adult one. The utterances of a child in transition from babble to speech (91-106 weeks) were recorded. The errors in these utterances were argued to arise either from "paradigmatic confusions among similar gestures...or from syntagmatic difficulties in coordinating the gestures that form a particular word" (p. 20).

If gestures originate as pre-linguistic units of action, and gradually develop into the units of contrast, as argued by Studdert-Kennedy (1987) and Browman and Goldstein (1989), then it is possible to see a continuity of development in language. If these gestures then serve as the primitives that are further coordinated in the language-learning process, such continuity includes higher-level phonological units as well as the fundamental contrastive units.

REFERENCES

Anderson, S. R. (1974). *The organization of phonology*. New York: Academic Press.

Anderson, S. R. (1976). Nasal consonants and the internal structure of segments. *Language, 52*, 326-344.

Banner-Inouye, S. (1989). The flap as a contour segment. *UCLA Working Papers in Phonetics, 72*, 40-81.

Barry, M. (1985). A palatographic study of connected speech processes. *Cambridge Papers in Phonetics and Experimental Linguistics, 4*.

Beckman, M. E., Edwards, J., & Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In D. Docherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II* (pp. 68-86). London: Cambridge University Press.

Bell-Berti, F., & Harris, K. S. (1981). A temporal model of speech production. *Phonetica, 38*, 9-20.

Bell-Berti, F., & Harris, K. S. (1982). Temporal patterns of coarticulation: Liprounding. *Journal of the Acoustical Society of America, 71*, 449-454.

Bird, S. (1990). *Constraint-based phonology*. Doctoral dissertation, University of Edinburgh.

Bird, S., & Klein, E. (1990). Phonological events. *Journal of Linguistics, 26*, 33-56.

Boyce, S. E. (1990). Coarticulatory organization for lip rounding in Turkish and English. *Journal of the Acoustical Society of America, 88*, 2584-2595.

Boyce, S. E., Krakow, R. A., Bell-Berti, F., & Gelfer, C. (1990). Converging sources of evidence for dissecting articulatory movements into core gestures. *Journal of Phonetics, 18*, 173-188.

Browman, C. P. (1992). Comments on Chapter 9. In D. Docherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II* (Chapter 9, pp. 257-260). London: Cambridge University Press.

Browman, C. P., & Goldstein, L. (1985). Dynamic modeling of phonetic structure. In V. A. Fromkin (Ed.), *Phonetic linguistics* (pp. 35-53). New York: Academic Press.

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3*, 219-252.

Browman, C. P., & Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica, 45*, 140-155.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*, 201-251.

Browman, C. P., & Goldstein, L. (1990a). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics, 18*, 299-320.

Browman, C. P., & Goldstein, L. (1990b). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman, *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 341-376). Cambridge: Cambridge University Press.

Browman, C. P., & Goldstein, L. (1991). Gestural structures: Distinctiveness, phonological processes, and historical change. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception* (pp. 313-338). Hillsdale NJ: Lawrence Erlbaum Associates.

Browman, C. P., & Goldstein, L. (1992). 'Targetless' schwa: An articulatory analysis. In D. Docherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II* (pp. 26-56). London: Cambridge University Press.

Brown, G. (1977). *Listening to spoken English*. London: Longman Group Ltd.

Byrd, D. (1992). Perception of assimilation in consonant clusters: A gestural model. *Phonetica, 49*, 1-24.

Catford, J. C. (1977). *Fundamental problems in phonetics*. Bloomington: Indiana University Press.

Clements, G. N. (1987). Toward a substantive theory of feature specification. *Proceedings of NELS 18, 1*, 79-89.

Clements, G. N. (in press). Place of articulation in consonants and vowels: A unified theory. In B. Laks & A. Rialland (Eds.), *L'Architecture et la géométrie des représentations phonologiques*. Paris: Editions du C.N.R.S.

Cooper, A. (1991). *An articulatory account of aspiration in English*. Unpublished doctoral dissertation, Yale University.

Dalby, J. M. (1984). *Phonetic structure of fast speech in American English*. Unpublished doctoral dissertation, Indiana University.

Diehl, R. (1989). Remarks on Steven's quantal theory of speech. *Journal of Phonetics, 17*, 71-78.

Dixit, R. P. (1987). Mechanisms for voicing and aspiration: Hindi and other languages compared. *UCLA Working Papers in Phonetics, 67*, 49-102.

Dunn, M. H. (1990). A phonetic study of syllable structure in Finnish and Italian. Paper presented at the 26th meeting of the Chicago Linguistic Society, parasession on the syllable, April.

Engstrand, O. (1988). Articulatory correlates of stress and speaking rate in Swedish VCV utterances. *Journal of the Acoustical Society of America, 83*, 1863-1875.

Farnetani, E., & Kori, S. (1986). Effects of syllable and word structure on segmental durations in spoken Italian. *Speech Communication, 5,* 17-34.

Ferguson, C. A., & Farwell, C. B. (1975). Words and sounds in early language acquisition. *Language, 51,* 419-439.

Fourakis, M. S. (1980). A phonetic study of sonorant-fricative clusters in two dialects of English. *Research Institute in Phonetics, 1,* 167-200, Indiana University:

Fowler, A., Brady, S., & Curley, S. (1991). The phoneme as an emergent structure: Evidence from speech errors. Paper presented at the April 1991 SRCD meeting.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing control. *Journal of Phonetics, 8,* 113-133.

Fowler, C. A., Rubin, P., Remez, R. E., & Turvey, M. T. (1980). Implications for speech production of a general theory of action. In B. Butterworth (Ed.), *Language production.* New York: Academic Press.

Fry, D. B. (1966). The development of the phonological system in the normal and the deaf child. In F. Smith & G. Miller (Eds.), *The genesis of language: A psycholinguistic approach* (pp. 187-206). Cambridge, MA: MIT Press.

Fujimura, O. (1981). Elementary gestures and temporal organization—What does an articulatory constraint mean? In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 101-110). Amsterdam: North-Holland.

Fujimura, O., & Sawashima, M. (1971). Consonant sequences and laryngeal control. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics, University of Tokyo, 5,* 1-6.

Gay, T. (1981). Mechanisms in the control of speech rate. *Phonetica, 38,* 148-158.

Gelfer, C. E., Bell-Berti, F., & Harris, K. S. (1989). Determining the extent of coarticulation: Effects of experimental design. *Journal of the Acoustical Society of America, 86,* 2443-2445.

Gimson, A. C. (1962). *An introduction to the pronunciation of English.* London: Edward Arnold Publishers, Ltd.

Goldstein, L., & Browman, C. P. (1986). Representation of voicing contrasts using articulatory gestures. *Journal of Phonetics, 14,* 339-342.

Guy, G. R. (1980). Variation in the group and the individual: The case of final stop deletion. In W. Labov (Ed.), *Locating language in time and space* (pp. 1-36). New York: Academic Press.

Han, M. (1962). The feature of duration in Japanese. *Study of Sounds, 10,* 65-80.

Hardcastle, W. J. (1985). Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. *Speech Communication, 4,* 247-263.

Hawkins, S. (1992). An introduction to task dynamics. In G. Docherty & D. R. Ladd (Ed.), *Papers in Laboratory Phonology II* (pp. 9-25). London: Cambridge University Press.

Hirose, H., & Gay, T. (1972). The activity of the intrinsic laryngeal muscles in voicing control: Electromyographic study. *Phonetica, 25,* 140-164.

Huffman, M. K. (1990). Implementation of nasal: Timing and articulatory landmarks. *UCLA Working Papers in Phonetics, 75*

Kahn, D. (1976). *Syllable-based generalizations in English phonology* Bloomington: University of Indiana Linguistics Club.

Keating, P. A. (1984). A phonetic and phonological representation of stop consonant voicing. *Language, 60,* 286-319.

Keating, P. A. (1985). CV phonology, experimental phonetics. and coarticulation. *UCLA Working Papers in Phonetics, 62,* 1-13.

Keating, P. A. (1990) Phonetic representations in a generative grammar. *Journal of Phonetics, 18,* 321-334.

Kelso, J. A. S., V.-Bateson, E., Saltzman, E., & Kay, B. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America, 77,* 266-280.

Kerswill, P. E. (1985). A sociophonetic study of connected speech processes in Cambridge English: An outline and some results. *Cambridge Papers in Phonetics and Experimental Linguistics, 4.*

Kingston, J. (1985). *The phonetics and phonology of the timing of oral and glottal events.* Unpublished doctoral dissertation. University of California, Berkeley.

Kingston, J. (1990). Articulatory binding: In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 406-434). Cambridge: Cambridge University Press.

Krakow, R. A. (1989). *The articulatory organization of syllables: A kinematic analysis of labial and velic gestures.* Unpublished doctoral dissertation, Yale University.

Ladefoged, P. (1982). *A course in phonetics* (2nd ed.). New York: Harcourt Brace Jovanovich.

Lahiri, A., & Marslen-Wilson, W. (1992). Lexical processing and phonological representation. In D. Docherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II* (pp. 229-254). London: Cambridge University Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431-461.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1-36.

Lindblom, B., & Engstrand, O. (1989). In what sense is speech quantal? *Journal of Phonetics, 17,* 107-121.

Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (1983). Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie, & O. Dahl (Eds.), *Explanations of linguistic universals* (pp. 181-203). Mouton: The Hague.

Lindblom, B., & Maddieson, I. (1988). Phonetic universals in consonant systems. In L. M. Hyman & C. N. Li (Eds.), *Language, speech, and mind* (pp. 62-78). London: Routledge.

Lisker, L. (1974). On time and timing in speech. In T. A. Sebeok (Ed.), *Current trends in linguistics, Vol. 12* (pp. 2387-2418). The Hague: Mouton.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20,* 385-422.

Lisker, L., & Baer, T. (1984). Laryngeal management at utterance-internal word boundary in American English. *Language and Speech, 27,* 163-171.

Locke, J. L. (1983). *Phonological acquisition and change.* New York: Academic Press.

Manuel, S. Y., & V.-Bateson, E. (1988). Oral and glottal gestures and acoustics of underlying /t/ in English. *Journal of the Acoustical Society of America, 84,* S84.

Mattingly, I. G. (1981). Phonetic representation and speech synthesis by rule. In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 415-420). Amsterdam: North-Holland.

McCarthy, J. J. (1988). Feature geometry and dependency: A review. *Phonetica, 45,* 84-108.

Mowrey, R. A., & MacKay, I. R. A (1990) Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America, 88,* 1299-1312.

Munhall, K., Löfqvist, A. (1992). Gestural aggregation in speech: laryngeal gestures. *Journal of Phonetics, 20,* 111-126

Munhall, K. G., Ostry, D. J., & Parush, A (1985). Characteristics of velocity profiles of speech movements. *Journal of Experimental Psychology. Human Perception and Performance, 11(4),* 457-474.

Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables as spoken by children and adults. *Journal of Speech and Hearing Research, 32,* 120-132.

Nolan, F. (1992). The descriptive role of segments: Evidence from assimilation. In D. Docherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II* (pp. 261-280). London: Cambridge University Press.

Ohala, J. J. (1974). Experimental historical Holland. In J. M. Anderson & C. Jones (Eds.), *Historical linguistics* (pp. 353-389). Amersterdam: North Holland.

Öhman, S. E. G. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America, 41*, 310-320.

Padgett, J. (1991). *Stricture in feature geometry*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Pierrehumbert, J. (1990). Phonological and phonetic representation. *Journal of Phonetics, 18*, 375-394.

Pierrehumbert, J., & Talkin, D. (1992). Lenition of /h/ and glottal stop. In D. R. Docherty & D. Ladd (Eds.), *Papers in Laboratory Phonology II* (pp. 90-117). London: Cambridge University Press.

Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America, 81*, 1574-1585.

Sagey, E. C. (1986). *The representation of features and relations in non-linear phonology*. Unpublished doctoral dissertation, MIT.

Saltzman, E. (1986). Task dynamic coordination of the speech articulators: A preliminary model. In H. Heuer & C. Fromm (Eds.), *Experimental Brain Research Series 15* (pp. 129-144). New York: Springer-Verlag.

Saltzman, E., & Kelso, J. A. S. (1987). Skilled actions: A task dynamic approach. *Psychological Review, 94*, 84-106.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*, 333-382.

Schiefer, L. (1989). 'Voiced aspirated' or 'breathy voiced' and the case for articulatory phonology. *Forschungsberichte des Instituts fur Phonetick und Sprachliche Kommunikation der Universitat München, 27*, 257-278.

Smith, C. (1988). A cross-linguistic contrast in consonant and vowel timing. *Journal of the Acoustical Society of America*, 86, S84.

Smith, C. (1991). The timing of vowel and consonant gestures in Italian and Japanese. Paper presented at the 12th International Congress of Phonetic Sciences, Aix-en-Provence. France. August 19-24.

Smith, C., Browman, C. P., McGowan, R., & Kay, B. (submitted). Extracting dynamic parameters from speech movement data.

Sproat, R., & Fujimura, O. (1989). Articulatory evidence for the non- categoricalness of English /l/ allophones. Paper presented at the LSA annual meeting, Washington, DC, December.

Sproat, R., & Fujimura, 0. (submitted). Allophonic variation in English/l/and its implications for phonetic implementation.

Steriade, D. (1987). Redundant values. *CLS, 23*, 339-363.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*, 3-45.

Stevens, K. N. (in press). Phonetic evidence for hierarchies of features. In P. A. Keating (Ed.), *Papers in Laboratory Phonology III*. Cambridge: Cambridge University Press.

Studdert-Kennedy, M. (1987). The phoneme as a perceptuomotor structure. In A. Allport, D. MacKay. W. Prinz, & E. Scheerer (Eds.), *Language perception and production* (pp. 67-84). London: Academic Press.

Studdert-Kennedy, M., & Goodell, E. W. (in press). Gestures, features and segments in early child speech.In B. de Gelder & J. Morais (Eds.), *Language and literacy: Comparative approaches*. Cambridge MA: MIT Press.

Turvey, M. T. (1977). Preliminaries to a theory of action with reference to vision. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing: Toward an ecological psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Vihman, M. M. (1991). Ontogeny of phonetic gestures: Speech production. In I. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception* (pp. 69-84). Hillsdale, NJ: Lawrence Erlbaum.

Yoshioka, H., Löfqvist, A., & Hirose, H. (1981). Laryngeal adjustments in the production of consonant clusters and geminates in American English. *Journal of the Acoustical Society of America, 70*, 1615-1623.

Zsiga, E. C. (1993). *Gradient rules in phonology and phonetics*. Unpublished doctoral dissertation, Yale University.

Zsiga, E. C., & Byrd, D. (1990). Acoustic evidence for gestural overlap in consonant sequences. *Journal of the Acoustical Society of America, 88*, S82.

FOOTNOTES

[*] *Phonetica, 49*, 155-180 (1992).

[†] Also Department of Linguistics, Yale University.

Haskins Laboratories Status Report on Speech Research
1992, SR-111/112, 43-62

Acoustic Evidence for Gestural Overlap in Consonant Sequences*

Elizabeth C. Zsiga†

Acoustic evidence for temporal overlap of the two closure gestures in the environment VC#CV was investigated. It was hypothesized that evidence of C2 would be found in the VC formant transitions and would increasingly dominate the transitions as rate (and by hypothesis, overlap) increased. Twenty repetitions (ten at a normal rate and ten at a rapid rate) of word pairs where the first word ended in /d/ and the second began with /p/, /t/, or /k/ were elicited in a sentence context from four subjects. F2 and F3 transitions from the midpoint of V1 to just before closure were then measured. In all environments, C2 had a clear influence on the VC formant transitions. The rate effects were less clear. For the /d#k/ environment, a significant correlation was found between more prominent velar transitions and increasing ratio of vowel duration to consonant closure duration, which may be considered a measure of increasing consonant overlap. The acoustic influence of C2 on V1 suggests considerable temporal overlap of the two closure gestures, and at least for the d#k case, increasing overlap as a function of fluency.

1. INTRODUCTION

The framework of articulatory phonology (first described in Browman & Goldstein, 1986 and elaborated on in their subsequent papers (1988, 1989, 1990a, b)) has drawn attention to the importance of understanding the patterns of gestural overlap in speech. The theory has focused in particular on the fact that there is a significant amount of overlap between sequential consonant gestures. Such overlap, discussed for example by Catford (1977), was first demonstrated instrumentally by Hardcastle and Roach (1977), using electropalatographic data from -VCCV- utterances. X-ray microbeam studies (Browman & Goldstein, 1988, 1990b) confirmed that in utterances like "perfect memory," or "seven plus" the lips begin moving toward the labial closure beginning the second word before the tongue tip closure ending the first word is released. Articulatory phonology proposes that gestural overlap increases in casual, rapid speech, and that this increased overlap may account for the apparent consonant deletions and assimilations characteristic of such speech. Assimilation of final alveolars is particularly common: for example, in the fluent pronunciation of "that boy and that girl" (two of the examples discussed by Gimson, 1962). The hypothesis is that in fast or casual speech gestural overlap may increase to the point where the labial gesture completc'v masks the preceding alveolar. The alveolar closure may still be made, but its acoustic effects will be largely hidden. As this proposal makes crucial reference to the acoustic consequences of gestural overlap, it is impo"tant to investigate those consequences. This experiment investigates whether there is acoustic evidence for the proposed patterns of gestural overlap in the formant transitions of vowels that precede consonant sequences.

A preliminary version of this paper was presented at the 120th meeting of the Acoustical Society of America, November, 1990. Dani Byrd and I collected and began the analysis of the data together, and without her, this experiment would never have been done. While taking responsibility for all the findings and interpretations in this paper, I would like to acknowledge the insightful comments of Michael Studdert-Kennedy, Bruno Repp, and Doug Whalen, and especially the profound influence and considerable assistance, at all stages of this work, of Louis Goldstein. This research was supported by NIH Grant HD-01994 to Haskins Laboratories and an NSF graduate fellowship to the author.

In the act of speaking, articulations overlap. Fowler (1980, p. 114) describes the complete lack of "temporal discreteness" in both articulatory and acoustic records: "The different kinds of gestures go on simultaneously, and thus there are no borders perpendicular to the time axis in an articulatory or acoustic record to separate one segment from another." Instrumental studies demonstrating overlap among speech gestures include Hardcastle (1985), Hardcastle and Roach (1977),and Marchal (1988) Öhman (1966), Perkell (1969),. These studies have shown, as Marchal (p. 287) puts it, that "an inescapable feature of speech production is the well-attested overlapping of speech segments." Articulatory phonology has taken the overlapping nature of speech to be basic to the formulation of phonetic and phonological generalizations. The temporal phasing, and changes in phasing, of articulatory gestures play a crucial role in this framework. The gesture, defined as "an abstract characterization of coordinated task-directed movements of articulators within the vocal tract" (Browman & Goldstein, 1989, p. 206), is the basic constituent of phonetic and phonological description. For example, a wide glottal opening gesture and a tongue tip raising gesture that approximates the alveolar ridge constitute an /s/. Through rules of temporal phasing, gestures become organized into larger contrastive units and into meaningful utterances: "The pattern of organization, or constellation, of gestures corresponding to a given utterance is embodied in a set of phasing principles...that specify the spatiotemporal coordination of the gestures" (1989, p. 211). The movements of gestures are not timed with respect to an external clock but only with respect to the internal stages of some other gesture. The effective temporal phasing between gestures may change, however, especially in fast or casual speech, with important consequences.

Many researchers (e.g., Barry, 1985, Kaisse, 1985) have described the differences between careful "canonical" pronunciation and pronunciation in "connected" fast or casual speech. Catford (1977) discusses the various consequences that arise when two consonants (especially those made at different places of articulation) become adjacent in fluent pronunciation. Word-final alveolars have been a particular focus of discussion as the consonants most likely to undergo deletion or assimilation when followed by another consonant (Avery & Rice, 1989, Byrd, 1991, Gimson, 1962, Guy, 1980, Paradis & Prunet, 1991). Articulatory phonology proposes that many of the processes of deletion, insertion, and assimilation described by these authors are due to changes in the temporal relations between gestures. There will be different consequences (hiding, revealing, or blending of gestures) depending on the gestures involved and the extent of the overlap, yet the theory proposes that "all result from two simple kinds of changes to the gestural score: (1) reduction in the magnitude of individual gestures (in both time and space) and (2) increase in overlap among gestures" (Browman & Goldstein, 1989, p. 214). An example (from Browman & Goldstein, 1990b) of gestural hiding as a result of increased overlap among gestures is the fluent pronunciation of the phrase "perfect memory," in which the final /t/ of the first word is apparently deleted. X-ray microbeam tracings of the utterance show that an alveolar closure *was* produced. It could not be heard, however, because the gesture was completely overlapped by the preceding velar and following labial stops, so that any acoustic effect was hidden by the other closures. Articulatory phonology predicts that other apparent deletions or assimilations of final consonants are the result of increased gestural overlap in fluent speech. That prediction, particularly with respect to the deletion or assimilation of final alveolar stops (as in "that boy") is examined in this paper.

There are two parts to the prediction. The first is that increased gestural overlap in such consonant sequences will result in an acoustic output consistent with the percept that the first consonant has been deleted.[1] Both the duration of closure and the characteristic vowel-to-consonant formant patterns will be affected by increased overlap. Concerning changes in duration, Repp (1978) found that as the intervocalic period of closure is shortened in VCCV sequences, listeners perceive a single consonant rather than two, even though the onset and offset vowel formant patterns indicate two different places of articulation. Typically, it is the second consonant that is reported (see also Abbs, 1971, Ohala, 1990). Concerning the formant transitions from vowel to consonant, however, Repp (1983) concluded that in VCCV utterances there is no perceptible influence of C2 on the first vowel: listeners could not identify C2 on the basis of the transitional vowel formants preceding the consonant closure. Although he found no perceptual evidence of overlap, Repp did find some statistically significant differences in the formant patterns; in particular, for one speaker F2 was higher preceding /bg/ sequences than preceding /bd/. In addition, Repp used carefully articulated speech in this experiment, in which overlap is predicted to be minimal. Byrd (1991) used speech synthesized by the Haskins computational gestural model, which allows precise control of gestural coordination (Browman, Goldstein, Saltzman, & Smith, 1986), to investigate directly the acoustic and perceptual consequences of changes in gestural overlap. She found that as gestural overlap was increased in the sequence /bæd bæn/, vowel formant transitions into the closure gradually became more labial in character and listeners were increasingly likely to report hearing /bæb bæn/. The experiment to be reported here examines the evidence for such gestural overlap in vowel-to-consonant formant transitions in natural speech. In particular, it investigates whether transitions into a word-final alveolar stop differ as a function of a following word-initial /p/, /t/, or /k/.

The formant transitions that are expected before a single labial, alveolar, or velar stop following a low or mid front vowel (the vowel contexts used in this experiment) are known: F2 and F3 falling for a labial, F2 level and F3 level or slightly rising for an alveolar, and for a velar, F2 rising and F3 falling (Delattre, Liberman, & Cooper, 1955; Fant 1970; Klatt, 1980; Stevens & Blumstein, 1978). Before a consonant sequence such as /dp/ or /dk/ there are two possibilities. If movement toward closure for /p/ or /k/ did not even begin until after closure for the /d/ was reached, there could be no influence of the second consonant on the acoustics of the vowel, and transitions into /dp/ or /dk/ would be identical to transitions into /dt/. If, on the other hand, movement toward /p/ or /k/ began before closure for the /d/ was reached, formant transitions into /dp/ and /dk/ would be expected to differ from those into /dt/. Any influence from a following labial stop would be seen as transitions that are more labial in character: both F2 and F3 falling or falling more sharply. Any influence from a following velar stop would be seen as transitions that are more velar in character; again, F3 would be expected to fall, while F2 would be expected to rise. It was hypothesized that these differing patterns would be found in the formant transitions preceding the different consonant sequences, indicating a substantial amount of overlap.

The second part of the prediction of articulatory phonology concerning casual speech processes is that overlap increases in casual, fast speech. While not all fast speech is casual speech (or all casual speech fast), researchers have found a relationship between an increase in speaking rate and changes in gestural organization in the direction of greater overlap. Engstrand (1988) found that an increase in speaking rate resulted in "active motor reorganization" such that "at the faster speaking rate, vowel- and consonant-related gestures were coproduced to a greater extent than at the slower rate" (pp. 1872, 1863). Similar results were obtained by Gay (1978, 1981). Although the relationship between rate and gestural organization is complex (for example Kuehn & Moll, 1976, Ostry & Munhall, 1985, and Fowler, 1980 found that speakers differed in the way an increase in rate affected gestural organization; see the discussion section below), for this experiment it was hypothesized that the manipulation of rate would result in greater temporal overlap between gestures, consistent with the findings of Engstrand and of Gay.

Acoustically, evidence consistent with an increase in overlap at a fast rate of speech would be found in a greater divergence of the formant patterns before the different consonant sequences. Patterns characteristic of the second consonant are predicted to increasingly dominate the transitions as rate increases. If F2 and F3 are expected to be lower for /dp/ than for /dt/, the formants would be expected to fall more sharply at the fast rate than at the slow. Similarly, for /dk/ the convergence of F2 and F3 would be expected to be more pronounced at the fast rate than at the slow. It is not predicted that the formant transitions in the /dt/ sequence would be affected by a greater or lesser amount of overlap between the two consonantal gestures, because /d/ and /t/ involve the same tongue tip gesture at the same place of articulation.

To test these two predictions, this experiment examines tokens of utterances that differ in containing /dt/, /dp/, and /dk/ sequences, produced at fast and slow rates. Formant transitions are analyzed to determine if there are significant differences between transitions into the three sequences, and whether rate has any effect on the amount of divergence.

2. Method

2.1 Materials. Three series of word pairs juxtaposing a final /d/ and an initial /p/, /t/, or /k/ were constructed. The nine word pairs consisted of a single-syllable modifier followed by a single-syllable noun. The three sets differed in phonological context; that is, in the surrounding vowels and consonants, and in the stress pattern. Two of the sets had stress on the second word, the other set had stressx on the first. The nine word pairs were:

A. bad pick	B. bad pen	C. bed pan
bad tick	bad ten	bed tan
bad kick	bad ken	bed can

Groups A, B, and C will be referred to subsequently as the badCick, badCen, and bedCan contexts. Only front vowels were used, in order to avoid any complications due to rounding: the lowering of the formants associated with rounding might be confounded with any effect of the labial consonant. Different combinations of low and high vowels were chosen to allow for possible effects of vowel-to-vowel coarticulation. Additionally, the phonological contexts were chosen so that for each word pair a plausible meaning could be constructed. Each word pair was placed in a sentence, in which the syntactic structure was kept as constant as possible. Each sentence was then placed in a paragraph, with the test utterances in the final clause. The paragraphs were designed to provide a plausible context for the appearance of each word pair. The full paragraphs, with the test utterances shown in italics, are given in Table 1.

2.2 Procedure. The subjects were four undergraduates, two men and two women, all native speakers of American English, who volunteered their time. The subjects were taped in a sound-treated room. In order to familiarize the subjects with the utterances, each subject was first given a set of nine cards, with one paragraph on each card. The tape recorder was turned on and the experimenter asked the subject to read each paragraph aloud at a conversational rate. The order of the cards was randomized, with the condition that the two stress patterns were kept separate: group C was not mixed in with groups A and B. Half the subjects read group C first and half read it last. After reading the paragraphs the subject was given a second set of cards. On each of these cards one of the target sentences was printed ten times. Again, order was randomized, with the separate stress patterns presented in the same order as in the paragraphs. The experimenter instructed the subject to repeat each sentence ten times, reading at a normal, conversational rate. After having read ten repetitions of all the sentences, the subject was given the same set of cards in the same order and was asked to read each sentence (again repeating it ten times) at a very rapid rate. In total, twenty-one utterances of each word pair were obtained from each of the four subjects: one in a paragraph, ten at a normal rate of speech, and ten at a rapid rate of speech. As the paragraphs were used only to familiarize the subjects with the utterances, tokens from the reading of the paragraphs were not used in the analysis.

Table 1. *Paragraphs used in the experiment.*

Molly liked her job as a nurse at the hospital. Her duties were usually interesting, but *she hated when she had to clean a bed pan for a patient.*

Mary and Jim were planning to go to Hawaii for their vacation, but Mary got sick. She went to the doctor, and *he told her to stay home and get a bed tan for a change.*

A man walked into the furniture store. Bob, the salesman, recognized immediately that he was a foreigner because of his clothes. His suspicions were confirmed when the man asked him for a bed can. After several moments, Bob realized that what he wanted was a waste basket. *The man left the store very happy to have found a bed can for his house.*

Susan is an architect who specializes in drawing blueprints. When the apprentice she was training came and asked why his drawings were always smudging, *she explained that he had chosen a bad pen for the job.*

Mary is in a calligraphy class. The teacher asked the class to practice writing the numerals one through ten across a sheet of paper. When she showed the teacher her paper, *he told her there was a bad ten in the set.*

George works in the Barbie doll factory. He works in the inspection department, inspecting Ken dolls. If any of the workers find a defective doll, they leave it for him. On Wednesday after lunch, *he was annoyed to find a bad Ken on his desk.*

The nominating committee met last Wednesday to discuss candidates for the new position. They considered resumes from Smith, Johnson, and Jones. The committee decided that Smith and Johnson were qualified, but *they all agreed that Jones would be a bad pick for the job.*

David hasn't really recovered from his car accident last spring. He was in the hospital for a long time. Even now he has to take a muscle relaxant, because *the doctors say he still has a bad tic on one side.*

Jim is the second-string field goal kicker for his high school football team. Though he comes to all the practices, he hasn't played since the Thanksgiving game. He isn't allowed to play, because *the coach thinks he made such a bad kick in the game.*

2.3 Analysis. For each talker, eight tokens of each word pair at the slow rate and eight at the fast rate[2] were digitized at a 10-kHz sampling rate, and analyzed by LPC analysis in the ILS program. For each analysis frame, a 20 ms Hamming window was used, with twelve filter coefficients for the female talkers and fourteen for the males. Ten milliseconds separated successive analysis frames.

In order to quantify the formant transitions, F2 and F3 were measured at two points: in the middle of the vowel preceding the consonant sequence, and immediately before closure. The point immediately before closure was defined as the last analysis frame in which three distinct formants were visible in the spectrogram, that coincided with a steeply declining amplitude envelope in the vocalic portion of the waveform, and that had a residual energy 10 to 25% of that found in the steady state portion of the vowel but still greater than that found during closure. The residual energy is an approximate estimate of source amplitude, determined by LPC analysis through inverse filtering. In nearly all cases the three selection criteria pic,.ed out the same point. In the few cases where they did not, two of the three criteria were considered sufficient to determine the endpoint.[3] The first frame after the release of the initial /b/ in "bad" or "bed" was chosen as the beginning of the vowel, and the frame halfway between release and closure was chosen as the midpoint. (For an even number of frames the frame nearer the end of the vowel was chosen as the midpoint.) The formant transition was taken to be the difference between the final value and the midpoint value.[4] Analysis of variance was carried out on F2 and F3 to determine if the transitions differed significantly according to following consonant sequence.

In order to quantify the rate effects, the durations of the vowel preceding the consonant sequence and of the consonant closure were measured in the dig:tized waveform of each token. The vowel duration was measured from the beginning of the release burst to the end of the vocalic portion of the waveform, which was defined by a sharp reduction in the amplitude envelope, and the consonant closure was measured from the end of the vowel to the release of the second consonant.

3. Results

3.1 Differences in transitions preceding the consonant sequences. The predicted contrasts were indeed found in the formant transitions before the consonant sequences in these data. The measured values of the change in F2 and F3 for each subject and phonological context are given in Table 2:

Table 2. *Mean measured values for F2 and F3: vowel midpoint, vowel offset, and change from midpoint to offset.*

	F2			F3		
	midpoint	final	Δ F2	midpoint	final	Δ F3
1 adpi slow	1642	1586	-56	2495	2544	49
fast	1569	1544	-25	2500	2490	-10
adti slow	1649	1623	-26	2504	2628	124
fast	1607	1588	-19	2508	2573	65
adki slow	1675	1801	126	2551	2589	38
fast	1644	1796	152	2476	2470	-6
1 adpe slow	1632	1611	-21	2503	2591	88
fast	1600	1575	-25	2521	2584	63
adte slow	1610	1627	17	2513	2675	162
fast	1595	1555	-40	2532	2583	51
adke slow	1636	1658	22	2486	2561	75
fast	1620	1654	34	2490	2543	53
1 edpa slow	1711	1700	-11	2585	2683	98
fast	1670	1618	-52	2568	2597	29
edta slow	1769	1702	-67	2644	2694	50
fast	1711	1671	-40	2634	2686	52
edka slow	1785	1802	17	2603	2616	13
fast	1734	1738	4	2580	2572	-8

Table 2. *(continued).*

		F2			F3	
	midpoint	final	ΔF2	midpoint	final	Δ F3
2 adpi slow	1915	1773	-142	2828	2878	50
fast	1932	1767	-165	2882	2874	-8
adti slow	1928	1850	-78	2840	2955	115
fast	1902	1792	-110	2890	2939	49
adki slow	2023	2114	91	2836	2977	141
fast	2021	2115	94	2844	2799	-45
2 adpe slow	1924	1792	-132	2868	2923	55
fast	1918	1749	-169	2859	2875	16
adte slow	1933	1855	-78	2910	3046	136
fast	1908	1808	-100	2808	2909	101
adke slow	1981	2051	70	2841	2896	55
fast	1994	2011	17	2920	2905	-15
2 edpa slow	2042	1999	-43	2889	2966	77
fast	2028	1959	-69	2922	2952	30
edta slow	2026	1991	-35	2924	2944	20
fast	2023	1984	-39	2926	2977	51
edka slow	2049	2144	95	2916	2964	48
fast	2059	2144	85	2916	2952	36
3 adpi slow	1730	1760	30	2769	2887	118
fast	1783	1777	-6	2790	2815	25
adti slow	1772	1875	103	2843	2887	44
fast	1794	1852	58	2824	2944	120
adki slow	1821	2241	420	2840	2707	-133
fast	1888	2476	588	2851	2765	-86
3 adpe slow	1764	1790	26	2871	2906	35
fast	1736	1708	-28	2850	2877	27
adte slow	1800	1924	124	2859	2863	4
fast	1772	1886	114	2855	2964	109
adke slow	1776	2128	352	2806	2855	49
fast	1807	2118	311	2854	2744	-110
3 edpa slow	1924	1884	-40	2887	2913	26
fast	1844	1804	-40	2866	2903	37
edta slow	1992	1995	3	2937	2997	60
fast	1944	1939	-5	2911	2958	47
edka slow	1968	2191	223	2914	2873	-41
fast	1965	2260	295	2896	2819	-77
4 adpi slow	1509	1456	-53	2398	2512	114
fast	1552	1468	-84	2406	2426	20
adti slow	1517	1476	-41	2424	2558	134
fast	1544	1486	-58	2409	2486	77
adki slow	1577	1602	25	2402	2321	-81
fast	1698	1865	167	2412	2241	-171
4 adpe slow	1483	1430	-53	2416	2518	102
fast	1528	1454	-74	2411	2432	21
adte slow	1506	1463	-43	2456	2575	119
fast	1565	1485	-80	2424	2464	40
adke slow	1557	1570	13	2442	2327	-115
fast	1618	1716	98	2339	2258	-81
4 edpa slow	1579	1491	-88	2454	2468	14
fast	1598	1529	-69	2379	2441	62
edta slow	1577	1513	-64	2505	2588	83
fast	1616	1555	-61	2421	2501	80
edka slow	1650	1636	-14	2482	2416	-66
fast	1631	1696	65	2410	2287	-123

A negative value indicates a falling transition, a positive value a rising transition. Figure 1 shows the mean formant transitions preceding /dp/, /dt/, and /dk/ for all subjects across both rates. The plot shows the formant frequency value at the vowel midpoint connected by a line to the value at the vowel offset. For the transitions into /dk/ F3 fell and F2 rose sharply. In the transitions into /dp/, while the differences were smaller, both F2 and F3 ended lower than F2 and F3 preceding /dt/.

To test for the significance of these patterns, analyses of variance were carried out on the differences between the final value and midpoint value for each token. An overall ANOVA, including data from the four subjects in three phonological contexts at both rates, reveals a significant main effect of consonant sequence for both F2 and F3 (for F2, $F_{2,6}$ = 10.48, p = .011; for F3, $F_{2,6}$ = 6.05, p = .036). However, for both F2 and F3 there was also a highly significant main effect of subject ($F_{3,503}$ = 251.27 for F2 and 15.578 for F3, p < .001) and a highly significant interaction of subject and consonant ($F_{6,503}$ = 53.94 for F2 and 18.402 for F3, p < .001), as well as several other significant interactions. These statistics indicate that the formant transitions for each subject differed in some respects, and that the relationships between the transitions into /dt/, /dp/, and /dk/ also differed for each subject. When the subjects are considered separately, these differences become apparent. Figure 2 a — d plots the transitions for each of the four subjects, again including the three phonological contexts and both rates. The subjects differed in the extent to which effects of consonant sequence were seen in both F2 and F3, and in the extent to which /dp/ and /dk/ were distinct from /dt/. For subject 1, the F3 values at the vowel offset are very similar for the three consonants, but when the midpoint values are considered, F3 for /dk/ is seen to fall, distinguishing it from the slightly rising patterns of both /dp/ and /dt/. While subject 2 shows little or no separation by consonant sequence in F3, subjects 3 and 4 show a clear divergence in F3 in the predicted directions. For all four subjects, the endpoints for /dp/ and /dk/ are lower than those for /dt/, as expected. In F2, for all subjects the transition into /dk/ ends higher than the transitions into /dp/ and /dt/, but /dt/ and /dp/ are widely divergent only for subject 3. Overall, subject 3 fits the hypothesis of differing formants almost perfectly, showing a clear separation of both formants in the predicted directions for the three consonant sequences. The plots of the data for the other subjects do not show all of these distinctions as clearly, especially the distinction between /dt/ and /dp/ in F2.

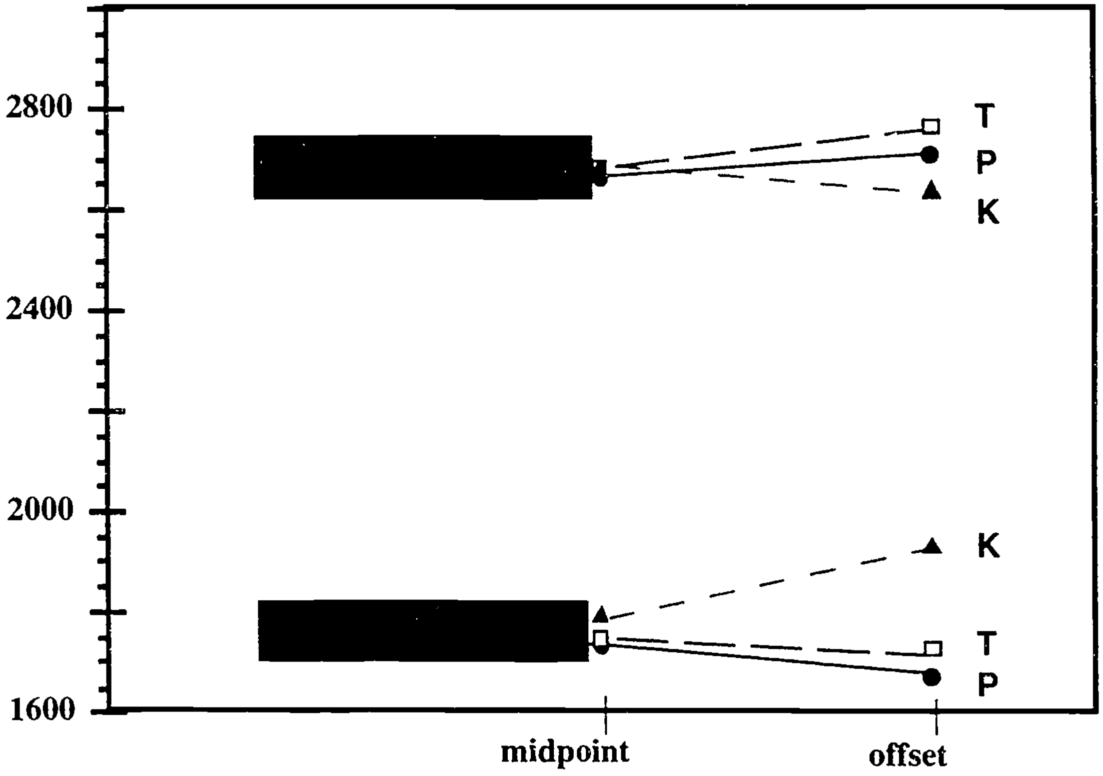


Figure 1. Formant transition: Mean for all subjects.

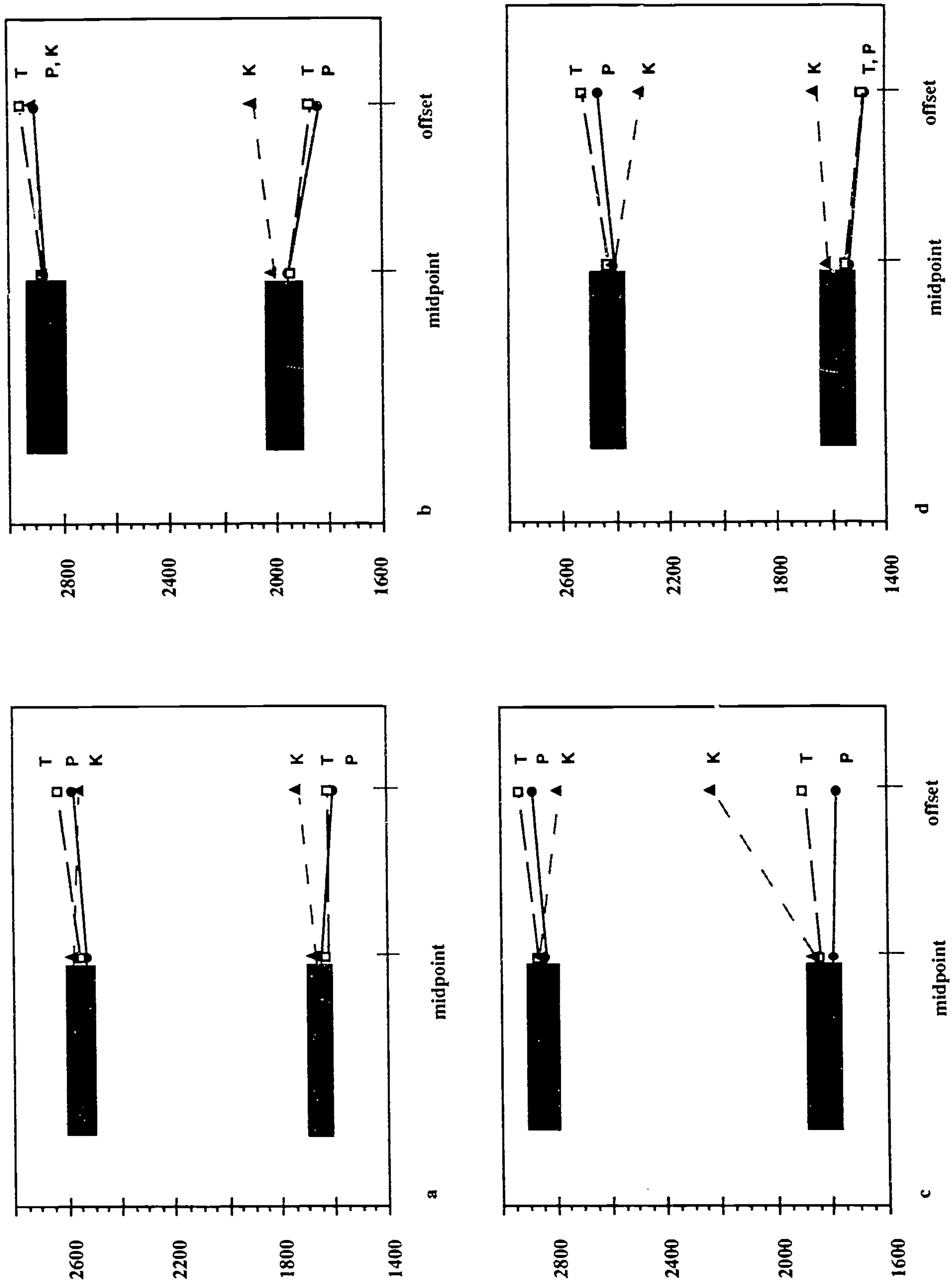


Figure 2. Mean formant transitions. a. Subject 1. b. Subject 2. c. Subject 3. d. Subject 4.

To simplify the statistical analysis, the data from each subject were analyzed separately. For each of the four subjects an analysis of variance (Table 3) revealed highly significant effects on both F2 and F3 due to the word-initial consonant. Although for each subject the consonant effect interacted with phonological context, or phonological context and rate, simple main effects revealed that the consonant effect was robust across these other factors.[5]

These effects were further analyzed by post-hoc comparisons, to determine if both initial /p/ and initial /k/ differed significantly from initial /t/ in the expected direction, or whether the significance came solely from the difference between /t/ and /k/. Newman-Keuls analysis was used to compare the means of the formant transitions both between /t/ and /k/ and between /t/ and /p/. In each case, the post-hoc test was done separately in the smallest cell for which analysis of variance showed no interactions. Thus for each subject each phonological context was considered separately by an analysis of variance. When the ANOVA for a given context indicated a significant interaction of consonant sequence and rate, the two rates were also considered separately.

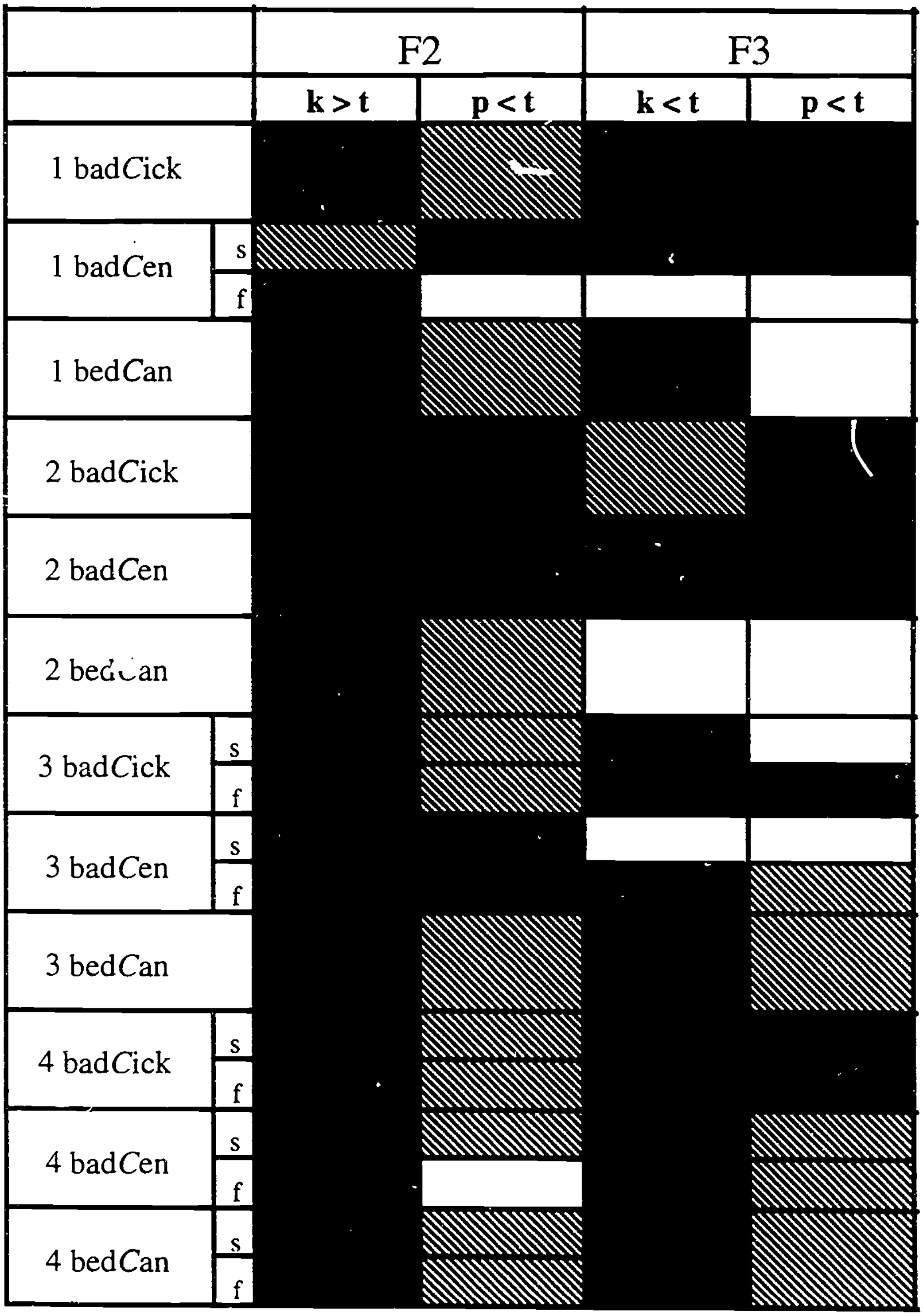
Figure 3 shows the results of these tests. The hypothesis being tested predicts that both between /t/ and /k/ and between /t/ and /p/ there will be significant differences. Both F2 and F3 are predicted to be lower for /p/ than for /t/. For /k/, F3 should be lower than for /t/, F2 higher. In the figure, the diagonally shaded blocks mark a difference in the expected direction; the gray blocks a *significant* difference (p < .05) in the expected direction. There were no significant differences in the direction opposite to that predicted. As was seen in the graphs of these data, the distinction between /t/ and /k/ is the more robust. In F2, the formant transitions were significantly higher (more sharply rising) for /k/ than for /t/ in almost all cases. In F3, /k/ was significantly lower than /t/ in a majority of cases. While both F2 and F3 were lower for /p/ than for /t/ in almost every case, the difference was significant at the .05 level in half or fewer cases.

Overall, however, these analyses of variance indicate that all subjects and phonological contexts show a highly significant variation due to following consonant sequence. The formant transitions in a vowel preceding a consonant sequence do differ depending on the second consonant, in the predicted direction. The influence of a following velar consonant is especially clear.

Table 3. *Subject by subject ANOVA on formant transitions, showing only significant effects.*

Source	df	F-ratio	Probability
Subject 1			
F2			
Consonant	2,126	61.08	0.000
Phol. Context	2,126	14.40	0.000
Ctx*Cns	4,126	13.49	0.000
Ctx*Cns*Rate	4,126	2.34	0.059
F3			
Consonant	2,126	16.28	0.000
Phol. Context	2,126	11.46	0.000
Ctx*Cns	4,126	3.63	0.007
Rate	1,126	31.57	0.000
Ctx*Cns*Rate	4,126	2.84	0.027
Subject 2			
F2			
Consonant	2,125	147.27	0
Phol. Context	2,125	14.35	0.000
Ctx*Cns	4,125	4.48	0.002
F3			
Consonant	2,125	5.67	0.004
Ctx*Cns	4,125	4.33	0.003
Rate	1,125	19.50	0.000
Ctx*Rat	2,125	4.45	0.014
Subject 3			
F2			
Consonant	2,126	241.11	0
Phol. Context	2,126	24.08	0.000
Ctx*Cns	4,126	8.16	0.000
Cns*Rat	2,126	4.61	0.017
F3			
Consonant	2,126	29.47	0.000
Cns*Rat	2,126	4.68	0.011
Ctx*Cns*Rate	4,126	4.16	0.003
Subject 4			
F2			
Consonant	2,126	132.93	0
Phol. Context	2,126	6.36	0.002
Ctx*Cns	4,126	2.82	0.028
Rate	1,126	12.05	0.001
Cns*Rat	2,126	29.34	0.000
F3			
Consonant	2,126	106.67	0.000
Rate	1,126	12.50	0.001
Ctx*Rat	2,126	3.39	0.037
Ctx*Cns*Rate	4,126	2.99	0.022

▨ = difference in predicted direction ■ = significant difference in predicted direction



		F2		F3	
		k > t	p < t	k < t	p < t
1 bad*C*ick		■	▨	■	■
1 bad*C*en	s	▨	■	■	■
	f	■			
1 bed*C*an		■	▨	■	
2 bad*C*ick		■	■	▨	■
2 bad*C*en		■	■	■	■
2 bed*C*an		■	▨		
3 bad*C*ick	s	■	▨	■	
	f	■	▨	■	■
3 bad*C*en	s	■	■		
	f	■	■	■	▨
3 bed*C*an		■	▨	■	▨
4 bad*C*ick	s	■	▨	■	■
	f	■	▨	■	■
4 bad*C*en	s	■	▨	■	▨
	f	■		■	▨
4 bed*C*an	s	■	▨	■	▨
	f	■	▨	■	▨

Figure 3. Results of the post-hoc tests. The distinction between fast and slow is made only when there is a significant interaction of rate and consonant sequence within each phonological context. There were no significant differences in the direction opposite to that predicted.

3.2 Rate. When asked to speak "at a rapid rate," the subjects did indeed speed up. They differed however, in the extent to which their rate of speech changed. Phonological context also had an effect. Table 4 shows, for each subject and phonological context, the mean duration of the vowel in the first word of the pair (either "bed" or "bad") at both the conversational and rapid rates, the difference in mean vowel duration between the two rates, and the ratio of mean duration at the fast rate to mean duration at the slow rate. Consonant closure duration for the two rates is also reported, as well as the ratio of mean vowel duration to mean consonant closure duration. The actual vowel length is of little interest, except to show that there is no overlap between the rates: for each phonological context, the slowest talker at the fast rate shows a shorter vowel duration than the fastest talker at the slow rate. The concern here is not how fast in absolute terms the talker spoke, but rather how the talker's rate of speech changed between the two conditions. It was hypothesized that a greater change in rate would lead to more overlap and greater acoustic influence of the second consonant in the sequence. The ratio of durations at the fast and slow rates is thus more interesting, as it may be taken as a measure of the change in rate from fast to slow for each subject.[6] An analysis of variance on the mean ratios of fast to slow vowel duration for each subject, phonological context, and consonant revealed that with respect to changes in rate subjects again behaved differently.

Table 4. *Change in rate: Duration measurements.*

		Mean vowel duration slow rate	Mean vowel duration fast rate	Difference in mean vowel duration slow - fast rate	Ratio of mean vowel dur. fast rate / vowel dur. slow rate
1	bedCan	111.5	94.9	16.6	.851
	badCen	156.9	106.9	50.0	.681
	badCick	148.8	108.2	40.6	.727
2	bedCan	113.6	87.4	26.2	.769
	badCen	172.7	132.6	40.1	.767
	badCick	174.7	123.7	51.0	.705
3	bedCan	101.2	90.1	11.1	.890
	badCen	170.8	132.0	38.8	.770
	badCick	154.2	126.0	28.2	.817
4	bedCan	111.7	85.7	26.0	.767
	badCen	171.2	98.7	72.5	.577
	badCick	157.1	93.5	63.6	.595
		Mean C closure dur. slow rate	Mean C closure dur. fast rate	Ratio of mean V dur. / C dur. slow rate	Ratio of mean V dur. / C dur. fast rate
1	bedCan	108.3	82.1	1.03	1.16
	badCen	114.7	88.2	1.37	1.21
	badCick	114.8	84.2	1.30	1.28
2	bedCan	134.3	98.0	.846	.891
	badCen	138.2	112.0	1.25	1.18
	badCick	137.8	99.6	1.27	1.24
3	bedCan	87.4	73.3	1.16	1.22
	badCen	84.0	69.3	2.03	1.90
	badCick	94.8	71.1	1.63	1.77
4	bedCan	106.4	66.0	1.05	1.30
	badCen	106.2	64.6	1.61	1.53
	badCick	98.8	60.4	1.59	1.54

The analysis indicated significant main effects for both subject ($F_{3,12}$ = 12.5, p = .005) and phonological context ($F_{2,12}$ = 8.0, p < .02). The subjects speeded up by different amounts: as can be seen from Table 4, subject 4 showed the greatest change from slow to fast, speaking, in one context, almost twice as fast at the rapid rate (indicated by a vowel duration only 58% as long), while subject 3 showed the least change, with a vowel duration at the fast rate in one context nearly 90% as long as the vowel duration at the slow rate. The main effect of phonological context indicates that the different vowels showed different amounts of change. This is presumably due at least in part to stress. Note that for each subject, it is the bed*C*an context, in which the vowel being measured is stressed, that shows the least change from fast to slow. Probably because they are less subject to reduction, the stressed vowels retain a longer duration at the fast rate. The analysis of variance found no significant interactions, and no significant effect of consonant sequence.[7] This indicates that, for each subject, the consonant to be articulated did not affect the way the change in rate was implemented.

3.3 Rate and Formant Transitions. For these data, the hypothesized relationship between rapid speaking rate and increased formant change held for some subjects and contexts, but not for all. Statistically, this effect should be evidenced as an interaction of rate and consonant sequence in the analyses of variance for each subject (Table 3).[8] The interaction of consonant and rate was rarely significant, however: F2 and F3 for subject 3, and F2 for subject 4. For subject 1, however, the three-way interaction of consonant*rate*phonological context is significant for F3 and just misses significance for F2, and for subject 4 this three-way interaction was significant for F3 . Subject 2 shows no interactions between consonant and rate at all. Where the interaction of consonant*rate*context was significant for a given subject, the consonant*rate interaction was tested separately for each phonological context. The results are given in Table 5. A significant interaction of consonant and rate was found in fewer than half of the cases analyzed. For subject 1, the interaction in the bed*C*an context approached significance, but for bad*C*en, the post-hoc analysis showed that the effect of rate was the opposite of that predicted: there was a significantly greater difference between the formant transitions at the slow rate than at the fast (see Figure 3).[9] It is clear that a change in rate produced different results for different subjects and different phonological contexts.

Table 5. *The interaction of rate and consonant sequence compared to the effect of rate on the ratio of vowel duration to consonant duration.*

		Interaction of Consonant and Rate				Main Effect of Rate on V duration / C duration	
		F2		F3			
		F(2.42)	*p*	F(2.42)	*p*	F(1.42)	*p*
Subject 1	bed*C*an	3.08	**.056**	2.51	**.094**	10.3	**.002**
	bad*C*en	4.48	**.017***	5.70	**.001***	29.1	**.001***
	bad*C*ick	.176	.8394	.110	.896	.009	.922
Subject 2	bed*C*an	(1.125)		(1.125)		2.68	.108
	bad*C*en	.411	.663	1.98	.142	.103	.593
	bad*C*ick					.737	.396
Subject 3	bed*C*an	(1.126)		.829	.443	5.63	**.022**
	bad*C*en	4.61	**.016**	5.10	**.010**	1.34	.253+
	bad*C*ick			4.26	**.021**	5.62	**.022**
Subject 4	bed*C*an	(1.126)		1.53	.228	17.2	**.001**
	bad*C*en	29.3	**0.00**	5.72	**.006**	.018	.893
	bad*C*ick			.311	.734	.054	.816

*This indicates a significant change in the direction opposite to that predicted: both the influence of the second consonant and the ratio of vowel duration to consonant duration were larger at the slow rate than at the fast.

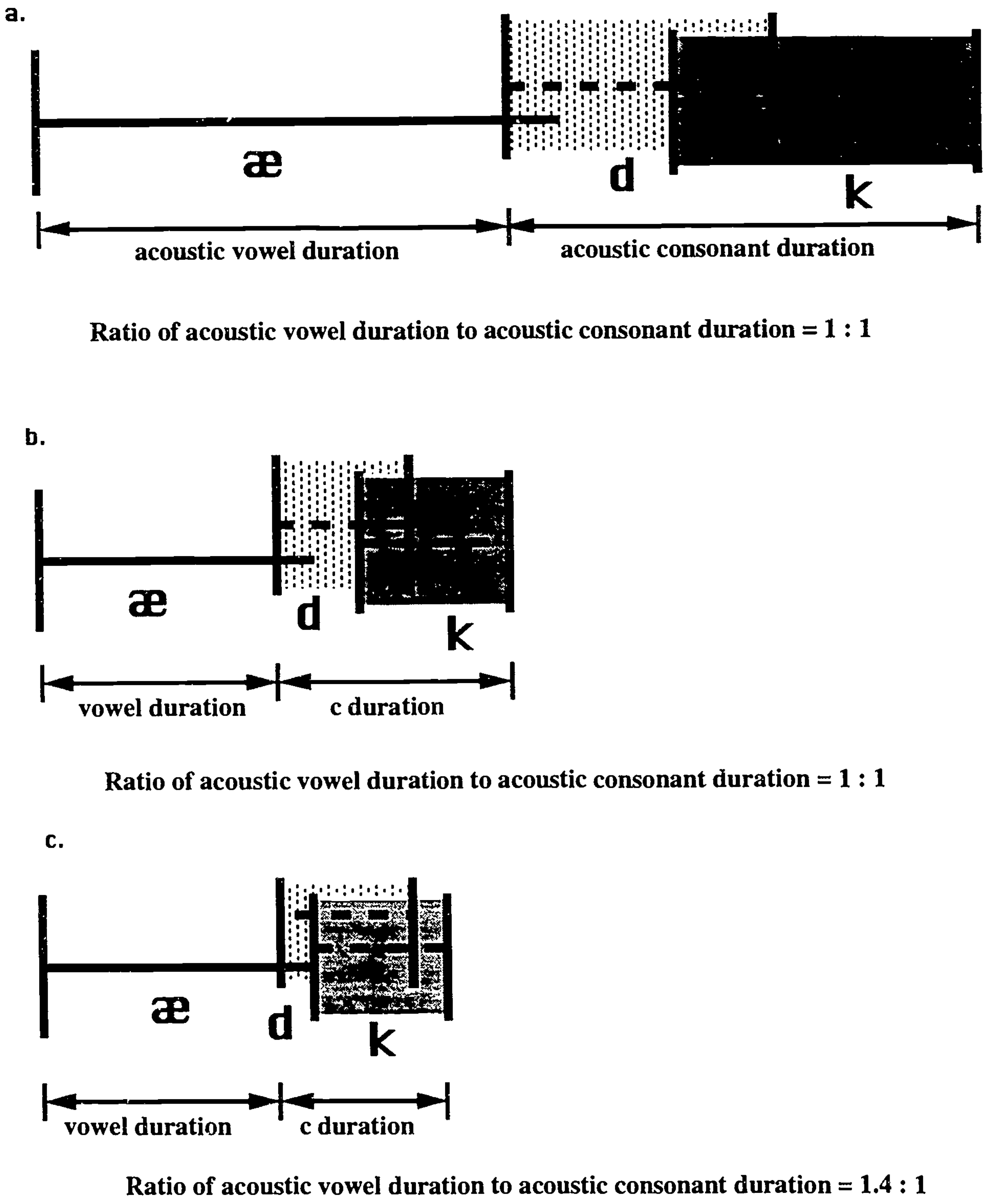
+The interaction of consonant and rate was significant for this cell: F(2.42) = 6.34, p < .004.

A comparison of Tables 4 and 5 shows that whether or not there was a greater effect of consonant sequence at the fast rate is not directly related to how much the talker sped up at the fast rate. A talker could speak very quickly and still not show any interaction of rate and consonant sequence. Subject 2, for example, showed a large difference in rate, but the change in rate did not produce significant changes in the pattern of formant transitions preceding the different consonant sequences.

Another factor, however, *was* correlated with a significant difference in formant transitions between slow and fast rate: the ratio of vowel duration to consonant closure duration. This ratio may be an indication of the style a speaker uses when speaking quickly. Two possibilities are diagrammed in Figure 4. Figure 4a illustrates a hypothetical articulatory relationship between a vowel and two consonants in a cluster at a conversational rate of speech. The shaded blocks indicate the intervals of closure for the two consonants, showing some overlap between the two. (The vowel is represented by a line only. It will overlap considerably with the articulations of the consonants, although the extent of that overlap is not the focus here. The right-hand end of the vowel line in this figure is not assumed to be meaningful.) Below the diagram the acoustic results of this articulatory organization are indicated: the ratio of acoustic vowel duration to acoustic consonant closure duration is 1:1 (within the range of values for these data). Figures 4b and 4c indicate two possible strategies for increasing rate. When speaking at a fast rate, one possibility would be for a speaker to execute each articulation twice as fast, without changing the relative temporal relationship between the articulations. If this were the case, there would be no difference between the ratio of acoustic vowel duration to acoustic consonant closure duration at the slow and fast rates. This strategy is diagrammed in 4b. The time taken for each articulation is half that taken in 4a, but the percentage of overlap remains constant and the vowel to consonant ratio remains 1:1. Using another strategy, a speaker, in addition to speeding up each individual gesture, might change the temporal relationship between them. If this change produced increased overlap between the consonants, as shown in 4c, the actual duration of the consonant closure would be shorter, and the ratio of acoustic vowel duration to acoustic consonant closure duration would be larger.

A relationship was found to hold between a larger ratio of vowel duration to consonant closure duration and a greater formant change due to following consonant sequence. This relationship is suggested by a comparison of the contexts in which formant transitions show a significant interaction of consonant effect and rate with those contexts in which there was a significant effect of rate on the ratio of vowel to consonant duration. (Analysis of variance on the ratio of vowel duration to consonant closure duration for each token shows highly significant main effects of subject ($F_{3,499} = 174.9$, $p < .001$) and phonological context ($F_{2,499} = 14.7$, $p < .005$). Each subject and context was thus analyzed separately.) A significant effect of rate indicates that the ratio of vowel duration to consonant closure duration was different at the two rates. The absence of a significant effect indicates that the ratio remained the same. Table 5 compares those contexts in which the interaction of rate and consonant sequence is significant with those in which the interaction of rate significantly affects the ratio of vowel to consonant duration. In many cases, those subjects who showed an interaction of consonant effect and rate in some contexts also showed an effect of rate on the ratio of vowel to consonant duration in those contexts. Where there was no interaction of consonant and rate, rate also had no effect on the vowel to consonant ratio. In the one condition where an increase in rate led to a significant *reduction* in the effect of following consonant sequence, the ratio of vowel duration to consonant closure duration was found to be significantly smaller at the fast rate. This suggests a relationship between a change in articulatory organization (evidenced by a change in vowel-to-consonant ratio) and a change in the effect a following consonant sequence has on formant transitions.

The relationship between the hypothesized spectral and temporal indices of overlap was tested directly by correlating the mean transition in F2 and F3 for each subject, phonological context, and rate with the mean ratio of vowel to consonant duration in each of these environments. The scatterplots in Figure 5 plot the relationship between formant transition and the ratio of vowel to consonant duration for /dk/. The data points plot the means for a given rate*subject*consonant condition. A significant correlation was found between a larger vowel to consonant ratio and a steeper rise in F2 and a steeper fall in F3.

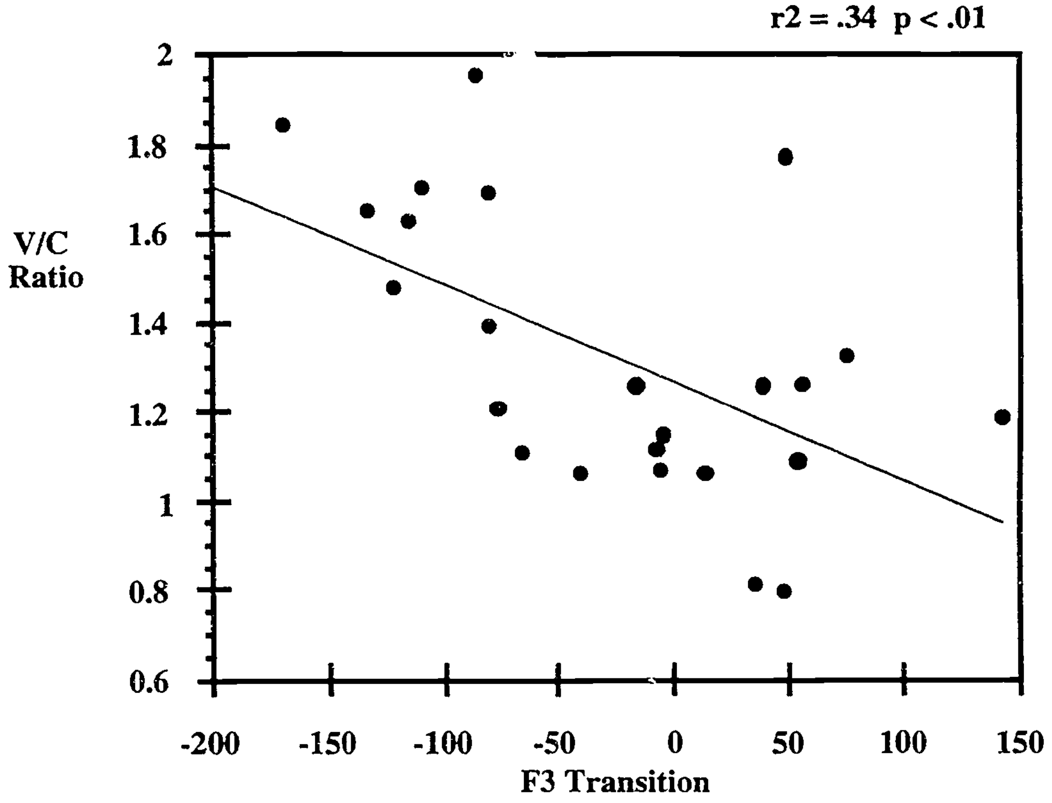


Ratio of acoustic vowel duration to acoustic consonant duration = 1 : 1

Ratio of acoustic vowel duration to acoustic consonant duration = 1 : 1

Ratio of acoustic vowel duration to acoustic consonant duration = 1.4 : 1

Figure 4. Strategies for increasing rate: Patterns of overlap and ratios of vowel duration to consonant closure duration.



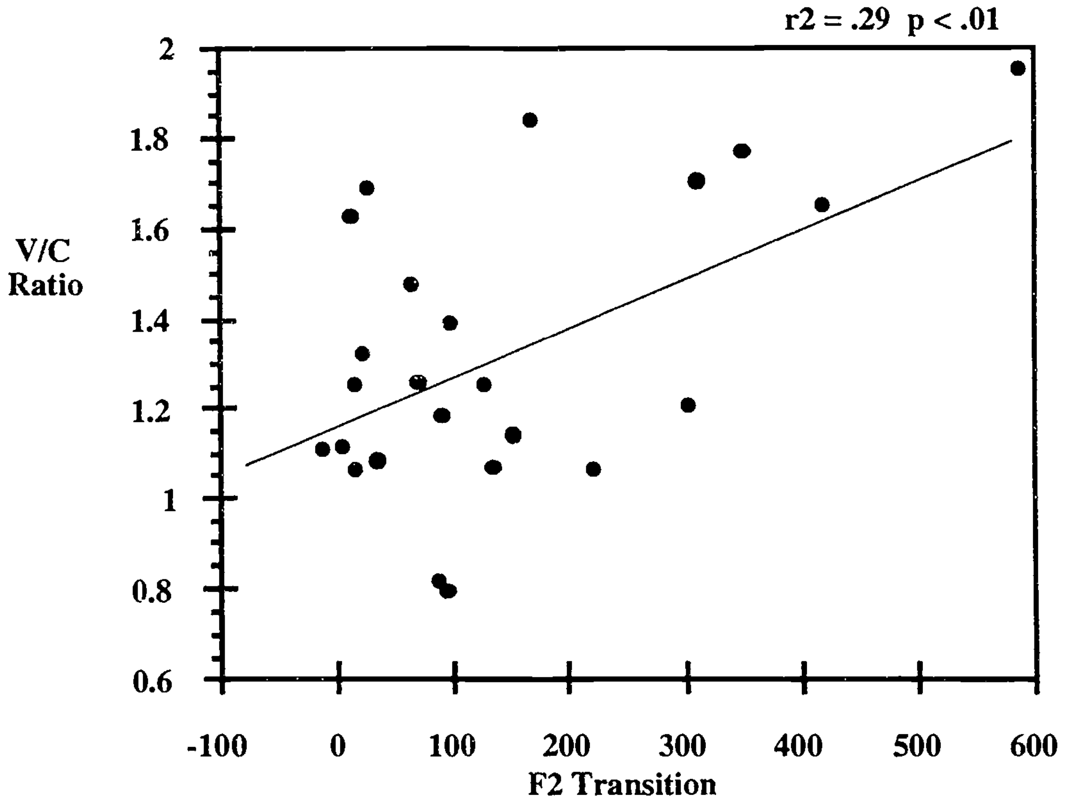


Figure 5. Correlation for /dk/ between increased formant change and ratio of vowel duration to consonant closure duration.

That is, over all subjects and vowels, F2 rose more and F3 fell more when the duration of the /dk/ sequence was shorter with respect to the vowel duration. As the regression lines show, the change in F2 becomes more and more positive, and the change in F3 more and more negative, as vowel to consonant ratio increases.

For /dp/, the correlation between formant change and the ratio of vowel to consonant duration was not significant, as shown in Figure 6. The flat regression line in 9a shows that there was no relationship between a larger vowel to consonant ratio and the amount of change in F3. Although the regression line for F2 in 6b is slightly rising, this trend is not significant.

These analyses show that a simple increase in speaking rate is not correlated with a greater acoustic influence of the second consonant in a sequence. Speakers differ in the effect that change in rate has on formant transitions. The strategy a speaker uses in speeding up, as evidenced in the ratio of vowel duration to consonant duration, seems to determine whether or not the formant transitions into a consonant sequence differ more at a rapid speaking rate than at a slow, at least for a following /k/.

One further test of the influence of a following consonant sequence was conducted. In a small perceptual experiment, the word "bad" was excised from the badCick utterances of subject 4. No final release burst was included in the excised syllables. The syllables were randomized and played back to four phoneticians, who were asked to transcribe the words they heard. The results of this informal test are given in Table 6. The subjects overwhelmingly reported the final consonant to be alveolar. The results of this perceptual experiment must be reconciled with the measured differences found in the formant transitions. Although acoustic measurements showed changes toward formant transitions that had labial or velar characteristics in /dp/ and /dk/ sequences, listeners still perceived alveolar stops.

4. Discussion

As was stated earlier, the prediction of articulatory phonology that apparent consonant deletions in fluent speech may be the result of increased gestural overlap has two parts: first, that there is a substantial degree of overlap between two sequential consonant gestures, and that this overlap will be evidenced in the acoustic output as influence of the second consonant on the vowel-to-consonant formant transitions; second, that overlap will increase in more fluent speech, with a concomitant increase in the acoustic influence of the second consonant. The acoustic evidence examined in this experiment is consistent with the hypothesis of substantial gestural overlap, the first part of the prediction. The formant transitions into a word-final alveolar stop were found to differ significantly as a function of a following word-initial /p/, /t/, or /k/. These differences are consistent with the hypothesis that movement of the lips toward closure for the /p/, and movement of the tongue body toward closure for the /k/, begin, in /dp/ and /dk/ sequences, before closure for the /d/ is achieved. Comparison of consonant effect in slow and fast tokens of the same utterances, however, revealed that there is no evidence of a direct relationship between increased speaking rate and increased gestural overlap. For some subjects in some contexts there was a significant interaction of rate and consonant effect; for other subjects and contexts there was no significant difference in formant transitions at the slow and fast rates; for a few contexts, the relationship was the opposite of that predicted, showing a greater effect of consonant sequence at the slow rate.

These results add to the evidence that the relationship between fast speech and casual speech is not straightforward. One can speak very quickly and yet very precisely. The formal, experimental setting in which this speech was recorded may very well have influenced at least some of the talkers to speak carefully. Researchers have found that asking talkers in an experimental situation to speak rapidly will not necessarily result in fluent speech. Rather, speakers will use different strategies in speeding up articulation (Engstrand, 1988; Gay, 1978, 1981; Kuehn & Moll, 1976; Ostry & Munhall, 1985). For example, Kuehn and Moll (1976, p. 320) report that

> The subjects were found to use different physiological methods of changing speaking rate. With an increase in speaking rate each subject reduced transition time by the same amount but the velocity and displacement variabl :s were changed in different proportions to each other depending on the individual speaker.

In this experiment it was found that an increase in the ratio of vowel duration to closure duration is a better indicator of an increase in gestural overlap than is a simple increase in rate. As was shown in Figure 4, as overlap between consonant closure gestures increases, the duration of the consonant closure will be shorter, and the ratio of vowel duration to closure duration will increase.

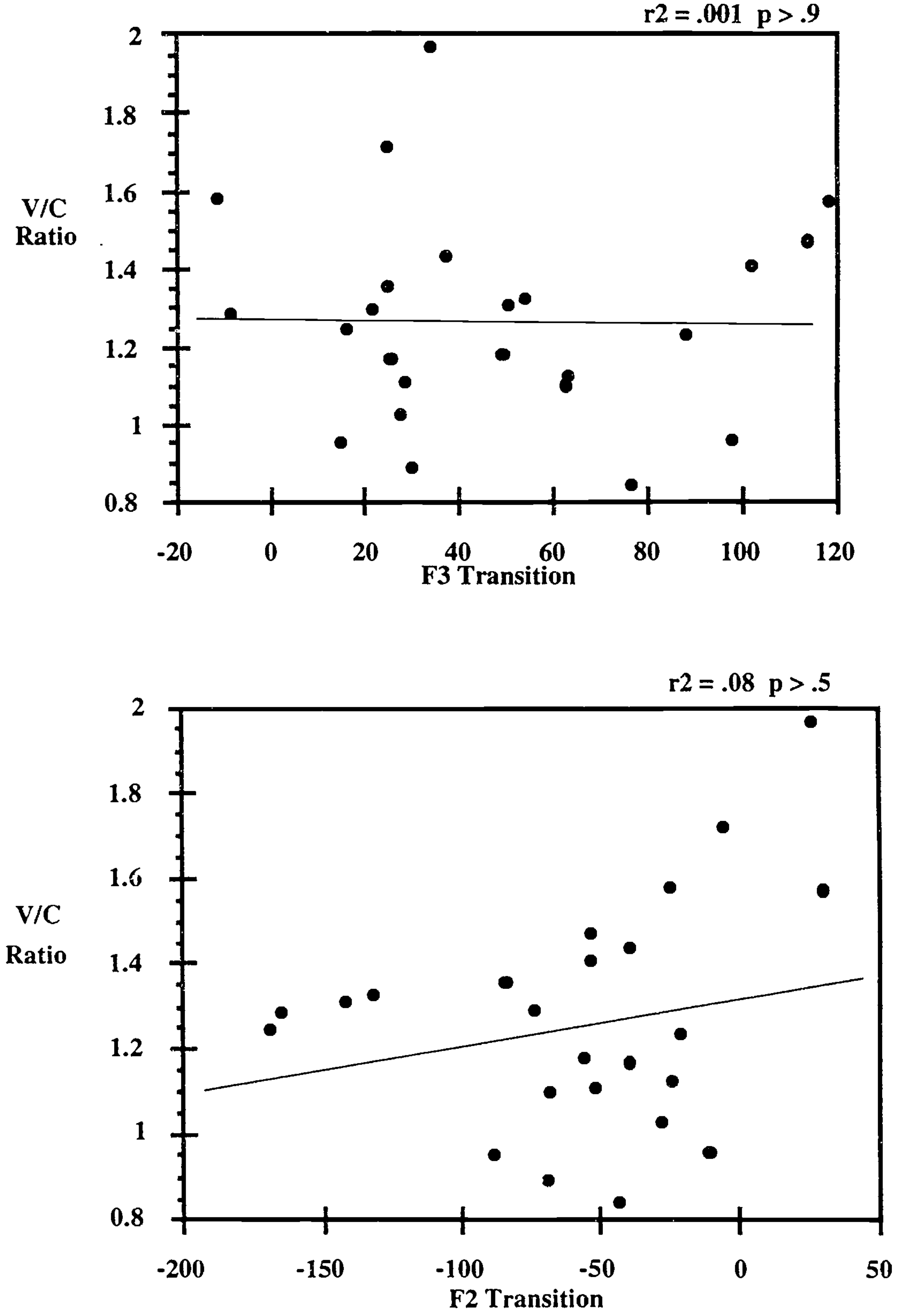


Figure 6. Correlation for /dp/ between increased formant change and ratio of vowel duration to consonant closure duration.

Table 6. *Results of the perceptual experiment.*

	Consonant Reported (out of 48)		
Subject	Labial	Alveolar	Velar
AF	2	42	4
CS	2	44	2
DW		48	
JK	2	44	2

In this experiment it was found, at least for the /dk/ sequences, that an increase in vowel to consonant ratio did correlate significantly with an increased influence of the second consonant on the vocalic formants. If a larger ratio of vowel duration to consonant duration is taken as a measure of increased overlap in fluent speech, as indicated by Figure 4, this correlation provides evidence for a relationship between increasing fluency and formant transitions increasingly characteristic of the overlapping consonant.

For the /dp/ sequences, however, no significant correlation was found. In the analysis of variance on the effects of consonant sequence as well, there were fewer instances of significant differences between /dp/ and /dt/ sequences than between /dt/ and /dk/ sequences. There are two possible explanations for the difference in results between /p/ and /k/. The first is that there is less overlap in alveolar-labial sequences than in alveolar-velar sequences. This would mean that, in the temporal coordination of consonant gestures, movement of the tongue toward closure for a /k/ begins sooner with respect to a preceding consonant than does movement of the lips toward closure for a /p/. While the claim that there is less overlap in /dp/ than in /dk/ sequences accounts for why the difference in formant transitions before /dp/ and /dt/ sequences was less often significant than the difference in transitions before /dk/ and /dt/ sequences (Figure 3), it does not account for the lack of correlation between temporal and spectral measures of overlap in the /dp/ sequences (Figure 6). As Figure 6 shows, the ratio of vowel duration to consonant closure duration did change; no direct relationship could be found, however, between this change and the measured values of the formant transitions.

A second (and more likely) possibility is that the temporal coordination of labial and velar articulations is not different, but that the effects of increased overlap of a labial gesture are less evident in these contexts than the effects of increased overlap of a tongue body gesture. Lack of evidence in the acoustic record of increased labial/alveolar overlap would help account for the lack of a significant correlation between the temporal and spectral measures for /dp/. Possibly, movement of the lips has less of an effect on F2 and F3 in these environments than does tongue body movement; the acoustic effects of two simultaneous closures in the vocal tract has not been extensively investigated. (While Byrd, 1991 measured the effects of varying degrees of overlap between labial and aiveolar closures in synthetic speech, she did not compare the timing or magnitude of these effects to the influence of a velar closure). It is also possible that, because the tongue dorsum and tongue tip are connected, movement of the dorsum has a greater articulatory influence on the tongue tip closure gesture than does movement of the lips, which are relatively independent.

In addition to the effects of the following consonants on transitional vowel formants, vowel-to-vowel coarticulation may have had an influence. In two of the three phonological environments examined (badCick and badCen), the measured vowel was followed by a vowel higher and further forward. Overlap between the vowel articulations in these environments would result in a higher F2 in the first vowel (Choi and Keating 1991, Manuel and Krakow 1984). Such an influence might counteract any lowering of F2 induced by labial closure, but would enhance raising of F2 caused by movement toward velar closure. There is some statistical support in these data for the influence of a following vowel: in F2 all four subjects show a significant effect of phonological context and of the interaction of context and consonant sequence (Table 3). Vowel-to-vowel coarticulation cannot account for all of the discrepancy in results between /dp/ and /dk/, however. As Figure 3 shows, the results for /dp/ were no better in the bedCan context, where no vowel-induced raising of F2 is expected, than they were in the badCick context. A further complication is introduced by the observation that for subjects two and four the lack of a significant difference in the formant transitions preceding /dt/ and /dp/ sequences does not come from the fact that F2 fails to fall in the alveolar-labial sequences, but from the fact that F2 *does* fall in the alveolar-alveolar sequences (see Figure 2). This result was unexpected, and cannot be attributed to the effects of overlap of either the consonants or of the vowels. Given the several factors that might be involved—differences in acoustic influence, differences in articulatory influence, or differences in the influence of the vowel context—an explanation of the divergent results obtained for /dp/ and /dk/ sequences must await further study.

A final problem remains to be addressed. There is an alternative hypothesis that could account for the differences found between transitions into /dp/, /dt/, and /dk/. If the alveolar consonant had been completely deleted in some cases, leaving only the word-initial consonant, transitions into the closure would certainly be those characteristic of that consonant. The overall means would then combine cases where deletion occurred with those where it did not, resulting in intermediate formant transitions. There is evidence, however, that the alveolar consonant was not deleted. First, the /d/ is present perceptually, as was seen in the transcription experiment. An account that posits deletion of the alveolar consonant must explain why subjects still heard a /d/. Second, the influence of the word-initial consonant varies along a continuum, at least for the /dk/ case. The correlation plotted in Figure 5 shows that formant change increases gradually as the ratio of vowel to consonant duration changes.[10] Had the alveolar consonant been deleted in some cases, the points would cluster in two groups: one with a low ratio of vowel duration to consonant duration and little change in F2 or F3, representing cases where the /d/ was present, and the second with a much higher ratio of vowel duration to consonant duration and considerable change in F2 or F3, representing cases where only a /k/ was present. Rather than showing evidence of a sudden change, indicative of complete consonant deletion, the data are more consistent with the hypothesis of overlap. Overlap can vary in its extent, with the second consonant gradually showing more acoustic influence as overlap increases. Byrd (1991) found that, for synthesized speech, as overlap increases the formants gradually become more like those characteristic of the second consonant. Although changes in the formants began as soon as any movement of the articulator for the second consonant preceded the complete closure for the first, listeners continued to perceive the original final consonant until overlap was well advanced.

While the data are not consistent with abrupt deletion of the word-final consonant, reduction in its gestural magnitude may well be involved. The results described here are consistent with an alveolar closing gesture that is shorter in duration or perhaps incomplete before a competing velar closure. X-ray microbeam data for these utterances, which will allow direct measurement of the temporal relations between gestures, as well as of their relative magnitudes, have been collected. Further research, involving both acoustic and physiological measurements, is planned.

5. CONCLUSION

This experiment has shown that there is acoustic evidence for gestural overlap in consonant sequences. There are significant differences in the formant transitions into a word-final /d/ before initial /p/, /t/ or /k/. These differences are consistent with the hypothesis that the gestures for the second consonant begin before closure for the first consonant is reached. Further, comparisons of formant changes across different rates showed that while increased rate does not necessarily result in increased gestural overlap, overlap may increase with rate. A more pronounced gestural overlap, as evidenced, at least for velar consonants, in a larger vowel to consonant duration ratio, does result in more pronounced differences in the vowel to consonant transitions.

REFERENCES

Abbs, M. H. (1971). A *study of cues for the identification of voiced stop consonants in intervocalic contexts*. Doctoral dissertation, University of Wisconsin.

Avery, P., & Rice, K. (1989). Segment structure and coronal underspecification. *Phonology, 6(2)*, 179-200.

Barry, M. (1985). A palatographic study of connected speech process. *Cambridge Papers in Phonetics and Experimental Linguistics, 4*, 1-16.

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3*, 219-252.

Browman, C. P., & Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. In O. Fujimara (Ed.), *Articulatory organization—Phonology tospeech signals*. Basel: S. Karger.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6(2)*, 201-51.

Browman, C. P., & Goldstein, L. (1990a) Gestural structures and phonological patterns. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception* Hillsdale, NJ: Lawrence Erlbaum.

Browman, C. P., & Goldstein, L. (1990b). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and the physics of speech*. Cambridge: Cambridge University Press.

Browman, C. P., Goldstein, L., Saltzman, E., & Smith, C. (1986). GEST: A computational model for speech production using dynamically defined articulatory gestures. *Journal of the Acoustical Society of America, 80*, S97.

Byrd, D. (1991.) Perception of assimilation in consonant clusters: a gestural model. *UCLA Working Papers in Phonetics, 78*, 97-126

Catford, J. C. (1977). *Fundamental problems in phonetics*. Bloomington: Indiana University Press.

Choi, J., & Keating, P. (1991) Vowel-to-vowel coarticulation in Slavic languages. *UCLA Working Papers in Phonetics, 78* 78-86.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27*, 769-73.

Engstrand, O. (1988). Articulatory correlates of stress and speaking rate in Swedish VCV utterances. *Journal of the Acoustical Society of America, 83*, 1863-75.

Fant, C. G. M. (1970). Analysis and synthesis of speech processes. In B. Malmberg (Ed.), *Manual of phonetics*. Amsterdam: North Holland.

Fowler, C. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics, 8*, 113-133.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America, 63*, 223-30.

Gay, T. (1981). Mechanisms in the control of speech rate. *Phonetica, 38*, 148-58.

Gimson, A. C. (1962). *An introduction to the pronunciation of English*. London: Edward Arnold.

Guy, G. R. (1980). Variation in the group and in the individual: The case of final stop deletion. In W. Labov (Ed.), *Locating language in time and space*. New York: Academic Press.

Hardcastle, W. J. (1985). Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. *Speech Communication, 4*, 247-63.

Hardcastle, W. J., & Roach, P. J. (1977). An instrumental investigation of coarticulation in stop consonant seqences. *University of Reading Working Papers*.

Kaisse, E. (1985). *Connected speech: The interaction of syntax and phonology*. NY: Academic Press.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America, 67*, 971-95.

Kuehn , D. P., & Moll, K. (1976). A cineflourographic investiagion of CV and VC articulatory velocities. *Journal of Phonetics, 3*, 303-20.

Manuel, S. Y., & Krakow, R. A. (1984). Universal and language-particular aspects of vowel-to-vowel coarticulation. *Haskins Laboratories Status Report on Speech Research , SR77/78*, 69-78.

Marchal, A. (1988). Coproduction: Evidence from EPG data. *Speech Communication, 7*, 287-295.

Ohala, J. J. (1990). The phonetics and phonology of aspects of assimilation. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and the physics of speech*. Cambridge: Cambridge University Press.

Öhman, S. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America, 39*, 151-168.

Ostry, D. J., & Munhall. K. G. (1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America, 77*, 640-48.

Paradis, C., & Prunet, J.-F. (Eds.). (1991). *The special status of coronals*. New York: Academic Press.

Perkell, J. (1969). *Physiology of speech production: Results and implications of a quantitative cineradiographic study*. Cambridge, MA: MIT Press.

Repp. B. (1978). Perceptual integration and differentiation of spectral cues for intervocalic stop consonants. *Perception and Psychophysics, 24*, 471-85.

Repp, B. (1983). Bidirectional contrast effects in the perception of VC-CV sequences. *Perception and Psychophysics, 33*, 147-55.

Stevens, K. N., & Blumstein, S. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64*, 1358-68.

FOOTNOTES

[*] To appear in *Journal of Phonetics*.

[†] Department of Linguistics, Yale University.

[1] That is, no alveolar stop is heard. In this context, the process might be described as either deletion or assimilation, and no distinction between the two will be made here.

[2] For subject 2, only 7 tokens of "bad kick" at the fast rate were clear enough to be used in the analysis.

[3] For example, formants of a very weak amplitude might be visible some frames into what the waveform indicated was closure. These frames were not counted as vocalic. In the course of the analysis, each token was independently measured twice. There was exact agreement on the frame chosen 90% of the time, and in no case did the points chosen differ by more than one frame.

[4] While defining the formant transitions in this way allowed for consistency across tokens, the shape of the transition could not be taken into account; that is, a smooth decline from midpoint to endpoint could not be distinguished from a steady state followed by a sharp fall. In the judgement of the investigator, however, the transitions were smooth and consistent.

[5] In F2, the consonant effect was significant for all cells except subject 1 adCe slow, where p =.06. In F3, the consonant effect was significant for all cells except subject 1 adCe fast, subject 2 edCa, and subject 3 adCe slow.

[6] The absolute difference between the two rates (column 3) is of course another measure of change in rate. For comparison between the talkers, however, the ratio is a more accurate measure, as it abstracts away from differences between subjects in overall rate of speech. The same absolute difference in msec means a greater relative change for a talker who speaks quickly than for one who speaks slowly. For example, a change of 40 msec translates into a larger percent change for the faster speaking subject 2 than for subject 1. For these data, however, the distinction appears to be small. In fact, statistics done on the difference and statistics done on the ratio show very similar results, with no disparity in the variables found to be significant.

[7] The effect of consonant approached significance ($F_{2,12}$ = 3.6, p = .0944). This small effect may be due to the relative strangeness ("bed tan") or familiarity ("bad pick") of the phrases.

[8] In the overall ANOVA, the interaction of consonant sequence and rate was only marginally significant for F2 ($F_{2,6}$ = 5.31, p = .047) and was not significant for F3 ($F_{2,6}$ = 1.06, p = .402), although the triple interaction of subject*consonant*rate was significant for both F2 and F3 (for F2, $F_{6,503}$ = 2.88 p = .009, for F3, $F_{6,503}$ = 2.47, p = .023).

[9] It can be also be seen from Figure 3 (where each phonological context was examined individually for an interaction with rate, regardless of whether the rate*context interaction was significant overall for that subject) that for subject 4 badCen as well, the difference between /dt/ and /dp/ in F2 was greater at the slow rate than at the fas' although the difference was not significant at either rate.

[10] While each point on this graph represents the mean of eight tokens, a token by token analysis reveals the same pattern.

Acoustic Evidence for the Development of Gestural Coordination in the Speech of 2-Year-Olds: A Longitudinal Study*

Elizabeth Whitney Goodell† and Michael Studdert-Kennedy

Studies of child phonology have often assumed that young children first master a repertoire of phonemes and then build their lexicon by forming combinations of these abstract, contrastive units. However, evidence from children's systematic errors suggests that children first build a repertoire of words as integral sequences of gestures and then gradually differentiate these sequences into their gestural and segmental components. Recently, experimental support for this position has been found in the acoustic records of the speech of 3-, 5- and 7-year-old children suggesting that even in older children some phonemes have not yet fully segregated as units of gestural organization and control. The present longitudinal study extends this work to younger children (22- and 32-month-olds). Results demonstrate clear differences in the duration and coordination of gestures between children and adults, and a clear shift toward the patterns of adult speakers during roughly the third year of life. Details of the child-adult differences and developmental changes vary from one aspect of an utterance to another.

INTRODUCTION

From a study of word-initial consonants in the early words of three children learning English, Ferguson and Farwell (1975) concluded that the initial unit of linguistic contrast in child phonology was not the phoneme, but the word. Implicit in this proposal was the notion that the word (or phrase) is the domain over which a child initially organizes its articulations (cf. Waterson, 1971). A good deal of evidence has now accumulated to support this view (e.g., Macken, 1979: McCune & Vihman, 1987; Menn, 1983, 1986; Menyuk, Menn, & Silber, 1986; Vihman & Velleman, 1989).

Also implicit in the apparent primacy of the word is the notion that smaller units of articulatory organization, whether consonants and vowels (Davis & MacNeilage, 1990) or the gestures that compose them (Studdert-Kennedy, 1987) gradually emerge as independently controllable units through local differentiation of the CV syllable into its onset and nucleus (cf. Lindblom, MacNeilage, & Studdert-Kennedy, 1983). The process of differentiation, or gestural segregation, may begin with variegated babble (Davis & MacNeilage, 1990) and continues as the child's lexicon grows.

By this account, the early course of phonological development is one in which the child gradually narrows its minimal domain of articulatory organization from the syllable, or syllable string, to the segment, (cf. Menn, 1986). If this is so, we might expect spatiotemporal overlap of gestures to diminish as children come to segregate consonantal from vocalic gestures and to

The research reported here was part of a dissertation completed by the first author and submitted to the University of Connecticut in partial fulfillment of the requirements for the doctoral degree in Psychology. Some of the research was presented in April 1990 at the International Conference on Infant Studies, Montreal, Canada and in April 1991 at the Biennial Meeting of the Society of Research in Child Development, Seattle, Washington. We thank the children who gave us their utterances and the parents who lent us their children; Arthur Abramson, Cathi Best, Carol Fowler, Len Katz, and Richard McGowan for suggestions on data analysis and for comments on earlier drafts of the paper; James Flege, Megan Hodge and an anonymous reviewer for instructive remarks. The work was supported in part by NICHD Grants HD-01994 and DC-00403 to Haskins Laboratories.

coordinate them into the precise temporal patterns typical of adult speech (Browman & Goldstein, 1986, 1989; see also Studdert-Kennedy & Goodell, in press). Evidence for such a decline has indeed come from acoustic analyses of fricative-vowel syllables spoken by young children and adults (Goodell, 1991; Goodell & Studdert-Kennedy, 1991; Nittrouer, Studdert-Kennedy, & McGowan, 1989; Siren, 1991), although an attempt to replicate Nittrouer et al. (1989), under slightly different experimental conditions, found no differences between children and adults (Katz, Kripke, & Tallal, 1991).

Before reviewing other relevant studies, we should note a terminological issue of some theoretical importance. The standard term "coarticulation" refers to the interaction and supposed mutual adjustments in articulatory form between nearby phonetic segments (consonants and vowels). Moreover, the commonly used terms "anticipatory coarticulation" and "perseveratory coarticulation" imply (incorrectly, in our view) that the beginnings and ends of these segments are phonetically irrelevant intrusions into a neighboring segment rather than necessary and intrinsic portions of the articulatory act. Such views are inevitable as long as consonants, vowels and features—the entities customarily said to be coarticulated—are physically undefined elements of abstract linguistic description.

In the present paper we take the segment to be a recurrent pattern of gesture that gradually emerges, as a potential unit of phonetic representation, through differentiation and integration of the gestures that form a child's early words (Studdert-Kennedy, 1987). Following Browman and Goldstein (1986, 1989), we take a gesture to be the formation and release of a constriction within the oral (lip, tongue tip, tongue body), velic or laryngeal articulatory subsystems. We assume, further, that acoustic vectors commonly attributed to coarticulation arise not from articulatory adjustments between neighboring segments, but from the coproduction, or temporal overlap, of invariant neighboring gestures. Accordingly, we use the term "gestural overlap" to describe our own data, reserving the term "coarticulation" for studies where that word has been used.

Several studies have used acoustic analysis to compare coarticulation in children's and adult's utterances. Here, we briefly review work directly relevant to the present paper, namely, studies of anticipatory vowel-to-schwa gestures (where the effect of the stressed vowel on a preceding schwa in an iambic əˈCV sequence is measured), of anticipatory vowel-to-stop-consonant gestures (where the effect of a vowel on the preceding stop consonant in a CV sequence is measured), and of anticipatory stop consonant-to-vowel gestures (where the effect of a stop-consonant on the preceding vowel in a VC sequence is measured). We should note that the studies to be reviewed differ considerably in their numbers of subjects, and so in their statistical power, or probability of correctly rejecting the hypothesis of no difference between children and adults. We leave it to the reader to adjudicate among their findings in light of these differences.

Three studies have examined the development of intersyllabic stressed-vowel-to-schwa effects in iambic disyllables. Repp (1986) analyzed coarticulation in one adult and two children (ages 4;8 and 9;5). Second formant (F2) estimates for the schwa in two-word sequences such as [ə#ˈCV] (in which # represents a word boundary) showed that the adult and the older child anticipated the front-back tongue position of the stressed vowel transconsonantally, while the younger child did not. In addition, first formant (F1) estimates for the schwa showed that only the adult anticipated tongue height. In a far more extensive study of 3-, 5- and 9-year olds and adults (n=10 for each group), Hodge (1989) estimated F2 values in bark at the midpoint of the schwa in utterances of [ə#ˈstV] (where V=i, u); only the 9-year-olds and adults gave significant evidence of anticipating the stressed vowel in the schwa.

Such results suggest that overlap of vocalic gestures in neighboring syllables may develop as speakers mature, with tongue front-back position emerging earlier than overlap of tongue height. However, Repp's results for tongue height are contradicted by those of Flege (in preparation) who conducted a study with a larger sample of 8 adults and 8-year-old children (n=8). Flege used glossometry (a technique in which the vertical distance between tongue and hard palate is measured with an artificial palate or "glossometer") to measure tongue height in /əˈhVp/ utterances. The children showed significantly greater assimilation of stressed vowel tongue height to preceding schwa than did adults. Flege's results therefore suggest that cross-syllabic vowel coarticulation decreases with age.

Five studies have examined intrasyllabic stop-vowel coarticulation in children and adults (Hodge, 1989; Repp, 1986; Sereno, Baum, Marean, & Lieberman, 1987; Sereno & Lieberman, 1987; Turnbaugh, Hoffman, Daniloff, & Absher, 1985). Four of these studies report that children coarticulate virtually the same as adults, and one study suggests that children may coarticulate more than adults. Turnbaugh et al. (1985) investigated within-syllable CV coarticulation in three 3- and 5-year-old children and 3 adult males who produced repetitions of CVC nonsense monosyllables. They found no differences due to age in the degree or pattern of coarticulation across consonants, and concluded that within-syllable CV stop coarticulation is virtually the same for children and adults. They suggested that within-syllable CV coarticulation may develop before the age of three.

As described above, Repp (1986) examined lingual coarticulatory effects in [ə#CV] repetitions produced by a 4-year-old, a 9-year-old and an adult. Estimates of F2 close to consonant release revealed significant consonant-vowel coarticulation in the younger child and adult, but only a marginal degree of coarticulation in the older child. If we assume that the older child's marginal effect would have proved significant on more extensive sampling, these results are consistent with those of Turnbaugh et al., suggesting no difference between 4-year-old children and adults.

Sereno et al. (1987) studied anticipatory lip rounding in consonant-rounded-vowel syllables, spoken by 4 adults and 8 children (three to seven years old). The authors excluded "unintelligible tokens" spoken by the children, combined the children's data for statistical analysis, and concluded that coarticulation of lip rounding for the vowel and tongue release for the consonant was virtually the same for the two groups. The lack of child-adult differences may have been due to the selection of adult-like tokens from the children's utterances, and to the averaging of children's data over a four year range. However, Sereno et al.'s finding is consistent with a similar result for overlap of lip rounding and tongue constriction by Nittrouer et al. (1989) in larger age-segregated groups (n=8) of 3-, 4-, 5-, 7-year-olds and adults, uttering fricative rounded-vowel syllables.

Sereno and Lieberman (1987) investigated coarticulation in [ki] and [ka] syllables spoken by 5 adults and 14 children (approximately two and a half to seven years). The adults consistently exhibited coarticulation of consonant and vowel, but the children were more variable, some displaying adult-like patterns, others quite different patterns than adults. The mean difference between groups was not significant, and the authors concluded that coarticulatory patterns are similar in adults and children. Again, the lack of child-adult differences may have been due to data selection and averaging across a four to five year age range.

One study has departed from the consensus of the preceding four. Hodge (1989) reported a study of 3-, 5-, 9-year-olds and adults repeating the monosyllables: [di], [dæ], and [du]. Results indicated that variation in F2 onset values, as a function of following vowel, was greatest for the 3-year-olds and declined with age, but the author did not report the statistical significance of these effects.

Finally, Kent (1983) has reported the apparent effect of a final stop on a preceding vowel in a CVC syllable, by inference from formant values over the final portion of the vowel immediately before closure for [k] in the word *box* [bɑks], spoken by 3 4-year-olds and 3 adults. For the adults he found that F2 rose into the closure, while for the children F2 was relatively flat throughout the vowel. He interpreted this as evidence of greater coarticulation in the adults than in the children. An alternative interpretation, suggested by Nittrouer et al. (1989), is that the children's lack of a final transition was due to a raised F2 throughout the vowel in anticipation of the /k/ closure, and therefore reflected more rather than less gestural overlap.

In short, four of the six studies of stop-vowel or vowel-stop coarticulation report no differences between children and adults. The two discrepancies, depending on interpretation, either are themselves discrepant or agree in suggesting that children coarticulate more than adults.

Whatever conclusions we draw from the studies reviewed here are limited, however, by the fact that, with the exception of one 32-month-old in Sereno and Lieberman (1987), none of the subjects was less than three years old. By this age a child typically has a sizeable receptive and expressive vocabulary (Templin, 1957) and a basic command of syntax (Limber, 1973). Elsewhere it has been argued that consonants and vowels first emerge as integrated patterns of gesture in child speech in response to at least two pressures: one pressure toward economy of storage as the lexicon increases in size, another toward rapid lexical access in the formation of multi-word utterances (Studdert-Kennedy, 1987, 1989; cf. Branigan, 1979; Donahue, 1986). If this is so, and if we want to understand the early development of gestural organization in speech, we need to examine the acoustic records of children most of whose utterances are still single words.

Two further cautions emerge from the studies we have reviewed. First, if we are to trace development in any detail, we must use relatively homogeneous groups well defined by age and by an appropriate linguistic measure, such as vocabulary size or mean length of utterance (MLU). Grouping data from 3- through 7-year-olds (as was done by Sereno et al. (1987) and by Sereno and Lieberman (1987)) can be seriously misleading. Second, even within a developmentally homogeneous group, children are likely to differ from one another more than adults do. High variability, combined with the small samples that labor-intensive measurement procedures often necessitate, add up to low statistical power and a high probability of failing to detect true child-adult differences. One way to increase statistical power is to control for individual differences in rate and style of phonological development by conducting longitudinal studies that permit comparison of each child with herself.

With these considerations in mind, a 10-month longitudinal study was designed to examine gestural coordination in children who were within a few months of their second birthday at the start of the study, close to the earliest age at which we could expect reliable cooperation from the subjects. The following questions were addressed: Does gestural coordination in the speech of children at this age differ from that of adults? How does gestural organization change over a ten-month interval within roughly the third year of life? Do measurements from the acoustic records of these children support the inference from studies of child phonology that their minimal domain of gestural organization is wider than that of adults?

Method

Subjects. Twelve subjects (6 girls 20-27 months old at the start of the study and 6 adult females) participated in the study. All subjects were monolingual English speakers from the southern New England region. One sex (female) was chosen in order to avoid confounding by possible sex differences in development. The time span of the study (10 months, beginning during the one-word stage) was chosen as a convenient period long enough, at the age under study, for a fair amount of developmental change tc be observed. Previous studies have shown that for many children at roughly this stage of development MLU increases substantially over a 10-month period (Brown, 1973; Miller & Chapman, 1981).

Table 1 lists the children's ages and MLUs. The girls were roughly matched for age (mean=22 months) and level of language development at the start of the study, as assessed by estimates of MLU, reported in morphemes, (mean=1.4) taken from a 30-minute recording session; 10 months later mean MLU had increased to 4.4. Thus it was expected that changes in the children's phonological skills would also be apparent.

Table 1. *Children's ages in months and their corresponding mean lengths of utterance (MLU) at Time 1 and Time 2.*

	Time 1		*Time 2*	
SUBJECT	*Age (months)*	*MLU*	*Age (months)*	*MLU*
BO	20	1.4	30	4.2
SK	20	1.0	30	3.8
LH	21	1.5	31	4.5
SR	21	1.9	31	4.6
EA	23	1.1	33	4.7
JM	27	1.5	37	4.6
Mean	22	1.4	32	4.4

Materials. The test utterances for the present experiment were the following minimal-pair nonsense disyllables, with stress on the second syllable: [bəˈbɑ], [bəˈbi], [bəˈdɑ], [bəˈdi], [bəˈgɑ], [bəˈgi]. The intervocalic labial, alveolar, and velar consonants were chosen in order to compare the effects of consonant place of articulation on vocalic formant transitions; all subjects at both 22 and 32 months of age were able to produce velars. An iambic stress pattern was chosen because adult studies (e.g., Alfonso & Baer, 1982) have found that a neutral schwa is highly susceptible to the effects of a following stressed vowel; a listener, naive to the pruposes of the study, spot checked the experimenter's model utterances in several sessions and confirmed that they were indeed iambic. The vowels [i] and [ɑ] were chosen because the former is the highest front vowel and the latter the lowest back vowel in English, so that any intersyllabic context effects on the preceding schwa should be quite evident.

Instruments. A Marantz PMD 430 portable tape recorder was used to collect all speech samples. The adults were recorded using an Audio Technica AT9300 microphone, while the children were recorded wearing a Samson Stage II wireless microphone with an Audio Technica 831 lapel microphone sewn into a vest.

Procedures. Each child's utterances were recorded in two half-hour sessions with the first author in the child's home, in each of the first and tenth months. Only the first session from each month was fully analyzed; data from the second session were drawn on only when a subject did not produce at least two repetitions of a given utterance type in the first session. Specially designed stuffed animals with names corresponding to the minimal pairs were used to elicit as many productions as possible through games with puppets, puzzles, books, and a doll house. The subjects repeated the test utterances after the experimenter, so that all utterances were immediate repetitions, or imitations. Since imitations may be closer to their targets than spontaneous utterances (cf. Leonard, Schwartz, Folger, & Wilcox, 1978), this procedure may lead to an underestimate of the child-adult differences likely to be observed in spontaneous speech. The order of the test utterances differed across children, but the experimenter attempted to elicit the same number of repetitions of each utterance type in each session. The number of acceptable tokens of each utterance type for each child ranged from 1 to 15, with a mean of 7. Data entered into the statistical analyses for individual children were means based on all acceptable utterances. No utterances were rejected on the basis of the perceived quality of the first vowel because this procedure might have biassed the sample toward adult-like forms from which vowel harmony and spondaic stress patterns (precisely the possible effects of interest) had been eliminated. Utterances were rejected only when formants could not be estimated in the Discrete Fourier Transform (DFT) spectra, or when the first author judged a subject's response to differ from the adult target in the second vowel (e.g., [biˈdɑ] instead of [bəˈdi]), or in syllable structure (e.g., [ˈbəˈdəˈgə] instead of [bəˈgɑ]). Seven unacceptable responses out of a total of 515 responses from the children's data were excluded from the analysis. The adult subjects recorded the set of disyllables in random order six times. The adult subjects repeated the test words after the experimenter, and no adult responses were excluded.

Data Analysis

Acoustic measurements. All tokens were digitized at a 20-kHz sampling rate on a VAX 780 computer at Haskins Laboratories. The Haskins Waveform Editing and Display system was used to measure utterance durations and to locate the midpoint of each syllable. Durations were measured on the waveform of each utterance for the first syllable (from the onset of the first full period of voicing to the offset of the last full period of voicing before closure), for the medial stop closure (voice offset to voice onset), and for the second syllable (voice onset to voice offset). Estimates of the center frequencies of the first and second formants were made from DFT spectra, computed with a 25.6 ms. Hamming window and a 3.2 ms slide between windows, at five locations: onset, midpoint and offset of the schwa, onset and midpoint of the stressed vowel. These formant estimates were made by first locating on the DFT spectral display the highest amplitude harmonic in the region of a given formant at a given point in the utterance, together with its two neighboring harmonics, immediately above and below it. A special purpose program then computed the amplitude-weighted mean of the three harmonics by summing over the corresponding bins of the DFT transform of the frequencies. An informal test of the reliability of the procedure (by having a second judge locate the main harmonic for a given formant in a few randomly chosen tokens) yielded 100% agreement.)

Statistical analyses. A set of three separate analyses of variance, in addition to a number of t-tests, was carried out on each of a dozen different aspects of the data. These analyses compared: (i) the children at Times 1 and 2 (repeated measures); (ii) the children at Time 1 with the adults; (iii) the children at Time 2 with the adults. Obviously, these analyses are not independent. However, none of the standard procedures for reducing the risk of a Type I error (false rejection of the null hypothesis) across multiple non-orthogonal comparisons, by adjusting the significance level, is applicable to a set of analyses that combines both repeated and non-repeated measures. Nor indeed are these procedures designed to apply to sets of comparisons across different, yet correlated, aspects of a body of data, such as (for example) formant patterns at schwa midpoint and formant patterns at schwa offset. In the absence of any generally accepted procedures, we note simply that most of the effects reported below are highly significant. Importantly, they also reveal systematic patterns of change with age that encourage belief in their reliability. Nonetheless, we recommend that readers bear in mind the risks of both Type I and Type II error (false acceptance of the null hypothesis) in interpreting the results.

Results

The results are reported in three separate main sections (Durations, Overlap of Vowel Gestures, Overlap of Consonant and Vowel Gestures), each followed by its own Discussion. The paper concludes with a General Discussion.

Results: Durations

Table 2 lists the mean absolute durations and the mean proportions of the utterances accounted for by each syllable and by the closure durations. The mean proportions of the total durations assumed by the first syllable, closure and second syllable are illustrated in Figure 1. For the 22-month-olds the first syllable accounts for 31% of the total utterance, closure for 19% and the second syllable for 50%. For the 32-month-olds: first syllable 20%, closure 20%, second syllable 60%. For the adults: first syllable 18%, closure 15%, second syllable 67%. Comparing the children and adults, we find that the proportion of the utterance taken up by the first syllable and closure is greatest for the children at Time 1 (50%) and diminishes with age to 33% in the adults.

Table 2. *Group mean durations for the whole utterance, for each syllable and for closure in milliseconds, listed by medial consonant and stressed vowel. Proportions of the total duration for syllables and closure are given in parentheses, while standard deviations are given in italics.*

	Duration										
	Whole Utterance		*Syllable 1*			*Closure*			*Syllable 2*		
CHILDREN											
22 months											
Consonant											
b	560	*134*	183	(.32)	*61*	96	(.18)	*46*	280	(.49)	*86*
d	631	*114*	195	(.30)	*55*	110	(.18)	*42*	326	(.52)	*95*
g	640	*95*	192	(.30)	*50*	140	(.21)	*55*	307	(.48,	*72*
Stressed Vowel											
i	633	*113*	199	(.30)	*60*	126	(.20)	*54*	310	(.48)	*82*
ɑ	587	*120*	181	(.31)	*47*	105	(.17)	*44*	305	(.51)	*89*
Mean	**610**		**190**			**116**			**308**		
32 months											
Consonant											
b	522	*87*	125	(.22)	*21*	104	(.21)	*20*	293	(.56)	*79*
d	539	*110*	100	(.19)	*23*	104	(.19)	*31*	334	(.62)	*88*
g	573	*73*	109	(.19)	*27*	109	(.20)	*23*	359	(.61)	*59*
Stressed Vowel											
i	554	*103*	119	(.20)	*26*	116	(.22)	*26*	319	(.57)	*90*
ɑ	538	*80*	102	(.19)	*23*	96	(.18)	*19*	339	(.62)	*67*
Mean	**546**		**111**			**106**			**329**		
ADULTS											
Consonant											
b	481	*52*	90	(.19)	*14*	81	(.17)	*14*	310	(.64)	*50*
d	484	*52*	90	(.19)	*21*	72	(.15)	*14*	321	(.66)	*51*
g	481	*53*	89	(.18)	*18*	65	(.14)	*12*	328	(.68)	*51*
Stressed Vowel											
i	484	*53*	94	(.19)	*14*	75	(.16)	*14*	314	(.65)	*54*
ɑ	480	*51*	85	(.17)	*20*	70	(.15)	*14*	325	(.67)	*47*
Mean	**482**		**90**			**73**			**320**		

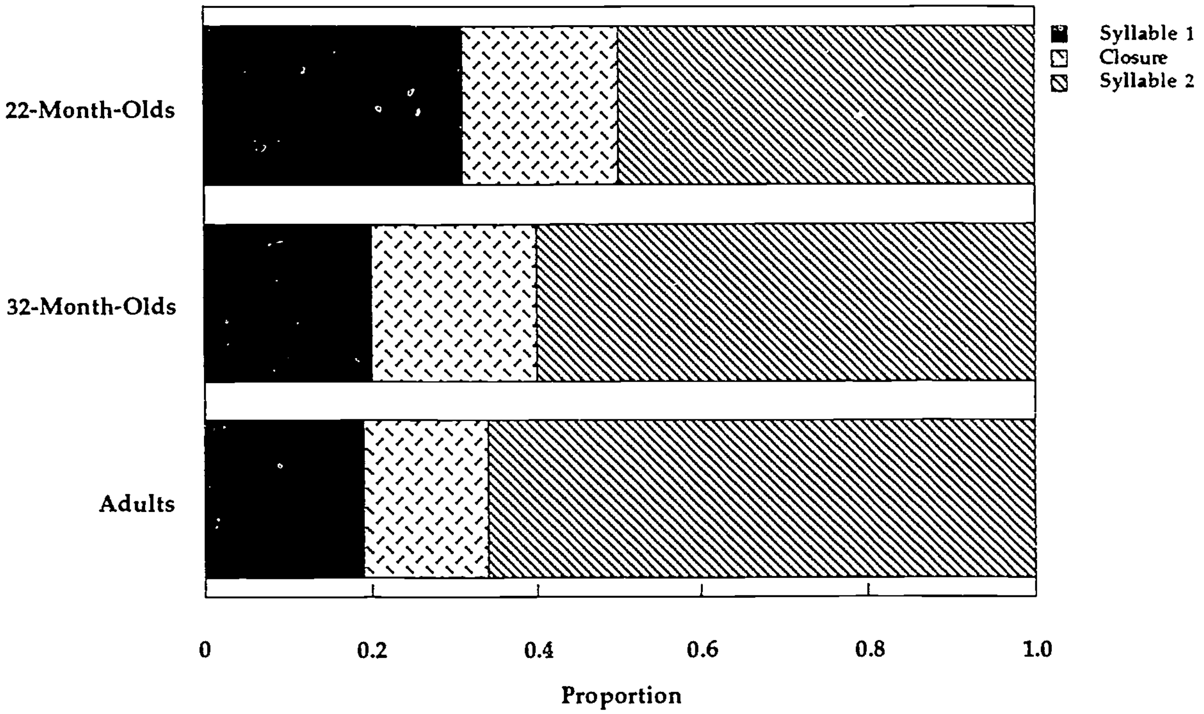


Figure 1. Mean syllable and closure durations as a percentage of the mean total duration for 22-, and 32-month-olds, and adults.

As we would expect, overall duration decreases with age: 610 to 546 to 482 ms. For the adults all total duration means were within 3 ms of each other (481 ms, 484 ms, 481 ms for utterances carrying the medial consonants [b], [d], and [g], respectively). By contrast, the 22-month-olds' durational means for tokens with [d] and [g] were roughly the same (631, 640, respectively), but the mean for utterances with labials was 70-80 ms shorter (560 ms).

Analyses of variance on the total durations with three factors (Consonant × Stressed Vowel × Age) revealed significant effects of consonant ($F_{2,20}$ = 6.96, *p*<.005) and age ($F_{1,10}$ = 8.31, *p*<.02) for the 22-month-old and adults, and of age ($F_{1,10}$ = 4.79, p=.05) for the 32-month-olds children and adults. A consonant-by-age interaction for the 22-month-olds and adults ($F_{2,20}$ = 6.37, *p*<.01) reflected the different pattern of durations for these age groups noted above. Two-tailed t-tests on the first syllable ratios revealed significant differences for between children at Time 1 and adults (*t*(10)=4.422, *p*<.001), between the children at Time 1 and Time 2 (*t*(5)=3.932, *p*<.01), but not between the children at Time 2 and adults. For closure duration ratios two-tailed t-tests revealed a significant difference between the 32-month-olds and adults (*t*(10)=2.856, *p*<.02), but not between the 22-months olds and adults, or between the children at Time 1 and Time 2. Finally, two tailed t-tests on the second syllable ratios revealed significant differences between the age groups: 22-month-olds and adults (*t*(10)=6.252, *p*<.0001), 32-month-olds and adults (*t*(10)=3.028, *p*<.01), and 22- and 32-month-olds (*t*(5)=3.603, *p*<.02).

Discussion: Durations

The results agree with previous studies reporting that children's utterances tend to be longer than adults' (DiSimoni, 1974; Eguchi & Hirsh, 1969; Kubaska & Keating, 1981; Smith, Sugarman, & Long, 1983). The shorter total durations for [bə'bV] than for [bə'dV] and [bə'gV]. spoken by the younger children, perhaps stem from repetition of the initial consonant, in the manner of canonical babble, facilitating relatively rapid execution.

Interestingly, the children's greater duration is not distributed evenly across the two syllables. The younger children have difficulty producing weak or reduced syllables (as reported also by Allen & Hawkins, 1980), so that they tend to equate the duration of the first syllable and the following closure with the duration of the second syllable. The decline in the proportion of the total duration assigned to the first syllable, and the roughly corresponding increase in the proportion assigned to the second, from 22 months to 32 months, combined with the lack of a significant difference in the first syllable proportion for 32-month-olds and adults, indicate that the children's ability to make a stress contrast has become virtually adult-like over this 10-month interval.

Results: Overlap of vowel gestures

Formant paths and vowel plots. Tables 3 and 4 list mean formant values for F2 and F1 at the five measurement points for all subjects. Peterson and Barney (1952) report mean first and second formant frequencies for their 28 female speakers of the high front vowel [i] as 310 Hz and 2790 Hz respectively, and for the low back vowel [ɑ] as 850 Hz and 1200 Hz, respectively. The mean adult formant values for the midpoint of the stressed vowel in the present study correspond roughly to the above values: 317 and 2513 Hz for [i], 772 and 1399 Hz for [ɑ].

Table 3. *Group mean F2 values at the five measurement points, listed by medial consonant and stressed vowel. Standard deviations are given in parentheses.*

	Schwa		*Schwa*		*Schwa*		*Stressed vowel*		*Stressed vowel*	
	Onset		*Midpoint*		*Offset*		*Onset*		*Midpoint*	
CHILDREN										
Mean Age 22 Months										
Consonant										
b	2063	(556)	2224	(614)	1997	(582)	2097	085)	2887	(1036)
d	1761	(230)	2194	471)	2449	(394)	2944	(747)	3068	(948)
g	1857	(327)	2325	(466)	2500	(629)	3028	(970)	3111	(1145)
Stressed Vowel										
i	2108	(473)	2471	(615)	2611	(584)	3336	(686)	3995	(288)
ɑ	1679	(127)	2023	(225)	2019	(458)	2043	(376)	2048	(235)
Mean Age 32 Months										
Consonant										
b	1783	(281)	1875	(252)	1600	(244)	2194	(609)	2897	(1042)
d	2051	(213)	2287	(184)	2490	(285)	2889	(562)	3043	(980)
g	1856	(232)	2180	(371)	2402	(429)	3037	(900)	2971	(1052)
Stressed Vowel										
i	1994	(300)	2277	(332)	2337	(559)	3306	(582)	3930	(236)
ɑ	1799	(179)	1950	(223)	1990	(416)	2107	(392)	2010	(229)
ADULTS										
Consonant										
b	1438	(165)	1424	(157)	1361	(203)	1767	(484)	1904	(637)
d	1606	(93)	1705	(179)	1805	(117)	1996	(258)	1973	(584)
g	1561	(144)	1658	(240)	1907	(244)	2175	(443)	1991	(593)
Stressed Vowel										
i	1573	(163)	1655	(235)	1783	(318)	2331	(248)	2513	(252)
ɑ	1497	(136)	1536	(208)	1599	(267)	1631	(274)	1399	(203)

Table 4. *Group mean F1 values at five measurement points, listed by medial consonant and stressed vowel. Standard deviations are given in parentheses.*

	Schwa Onset		Schwa Midpoint		Schwa Offset		Stressed Vowel Onset		Stressed Vowel Midpoint	
CHILDREN										
Mean Age 22 Months										
Consonant										
b	725	(170)	999	(299)	622	(173)	537	(154)	856	(358)
d	755	(147)	970	(208)	584	(137)	612	(133)	832	(401)
g	710	(157)	994	(225)	596	(174)	611	(181)	849	(407)
Stressed Vowel										
i	694	(175)	894	(251)	556	(145)	486	(90)	504	(93)
ɑ	765	(127)	1081	(193)	645	(163)	687	(145)	1187	(195)
Mean Age 32 Months										
Consonant										
b	736	(163)	833	(106)	588	(127)	638	(166)	823	(383)
d	679	(121)	739	(138)	520	(82)	547	(144)	884	(449)
g	711	(124)	760	(173)	509	(128)	605	(260)	841	(429)
Stressed Vowel										
i	673	(138)	726	(154)	493	(118)	450	(78)	457	(70)
ɑ	744	(127)	829	(115)	585	(98)	743	(162)	1241	(123)
ADULTS										
Consonant										
b	495	(144)	543	(84)	395	(123)	471	(156)	549	(241)
d	454	(117)	510	(83)	330	(90)	370	(93)	545	(252)
g	448	(121)	479	(84)	301	(84)	373	(119)	540	(251)
Stressed Vowel										
i	463	(127)	504	(101)	335	(107)	303	(46)	317	(43)
ɑ	467	(132)	517	(71)	349	(107)	506	(113)	772	(127)

With regard to adult vowel-vowel overlap (Figure 2), early onset of the stressed vowel gesture is evident for the front-back dimension in the second formant paths, estimated from formant values at onset, midpoint and offset of the first syllable schwa. These paths lie approximately 200 Hz higher in [bəˈbi], [bəˈdi], and [bəˈgi] than they do in the corresponding members of the pairs: [bəˈbɑ], [bəˈdɑ], and [bəˈgɑ]. No anticipatory effects appear in the high-low tongue dimension: measurements of F1 in the schwa before stressed [ɑ] and [i] are roughly the same, a finding which accords with a previous adult study (Alfonso & Baer, 1982).

In contrast, mean formant estimates for the six children show clear overlap of the stressed vowel with schwa in both dimensions at both ages (Figures 3 and 4). For the 22-month-olds (Figure 3) in [bəˈbV], F2 before stressed [i] is approximately 700 Hz higher than F2 before stressed [ɑ] at schwa onset, midpoint and offset; in [bəˈdV] the corresponding [i]-[ɑ] difference is about 120 Hz at onset, 500 Hz at midpoint, and 350 Hz. at offset; in [bəˈgV] the difference is 500-600 Hz throughout the schwa. For schwa F1, the largest difference between the [i]-[ɑ] formant paths is found in [bəˈbV] with a difference of about 300 Hz at vowel midpoint; for [bəˈdV] there is a difference of about 100 Hz throughout the schwa, and in [bəˈgV] a difference of about 120 Hz at schwa offset. The F1 formant paths indicate that these children anticipate the stressed vowel in tongue height, as well as in the front-back tongue dimension (as indicated by F2).

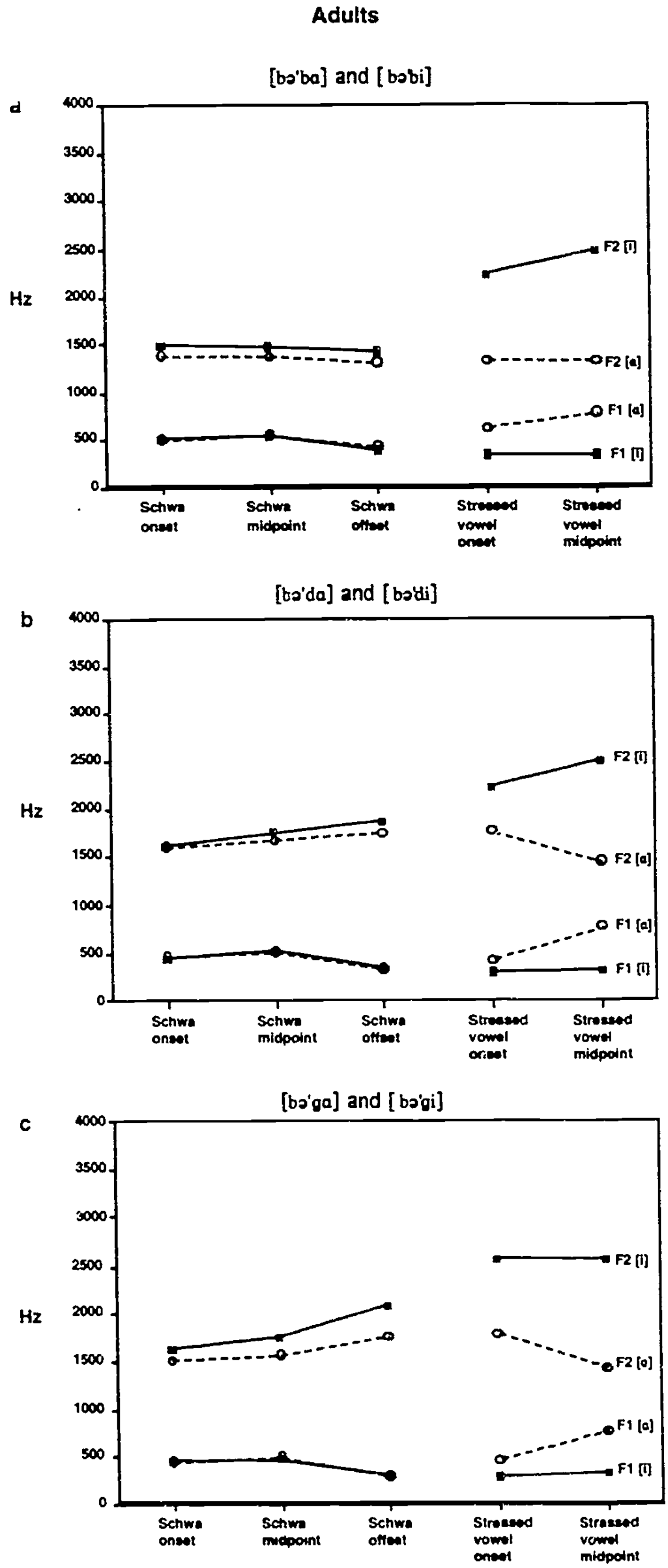
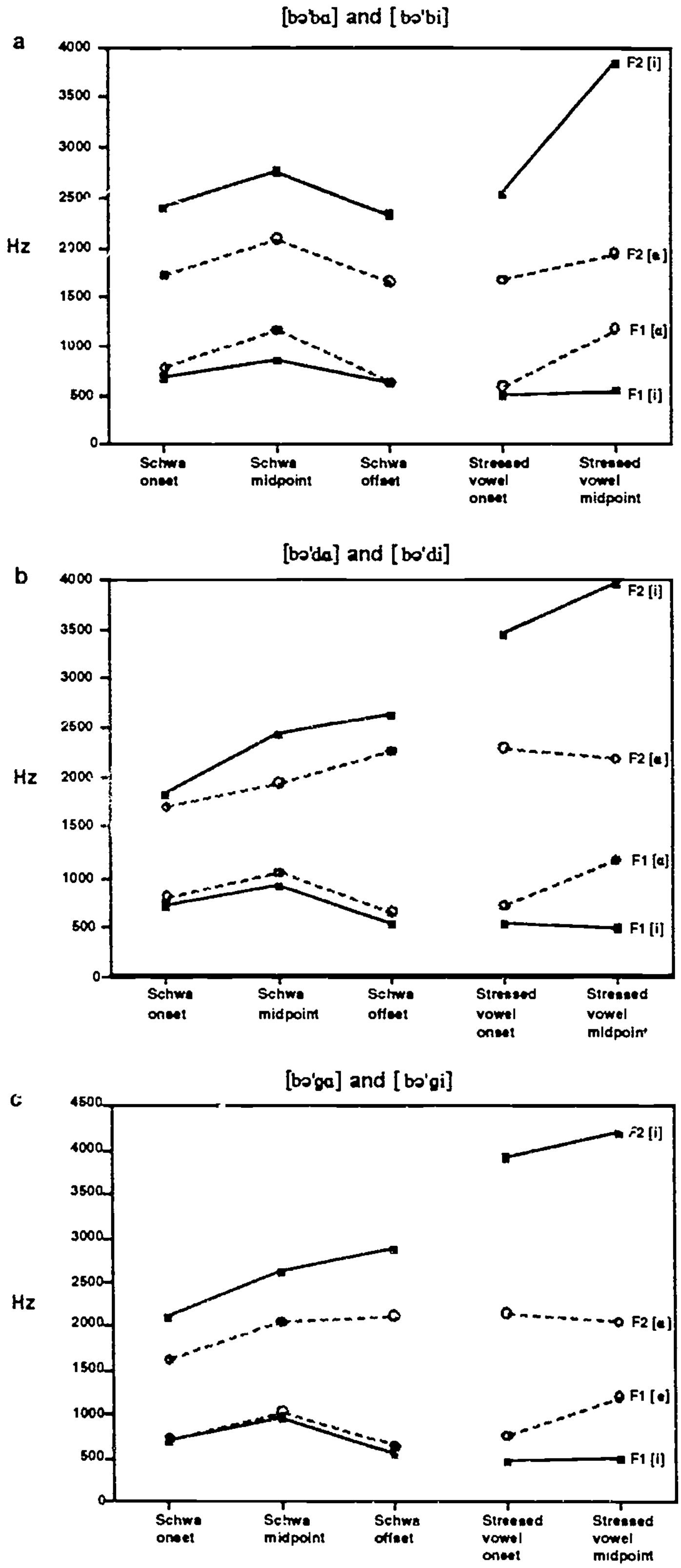


Figure 2. Mean formant paths for adu.'ts for[bə'bV],[bə'dV],[bə'gV].

22-Month-Olds


[bəˈbɑ] and [bəˈbi]

[bəˈdɑ] and [bəˈdi]

[bəˈgɑ] and [bəˈgi]

Figure 3 Mean formant paths for 22-month-olds for [bəˈbV], [bəˈdV], [bəˈgV].

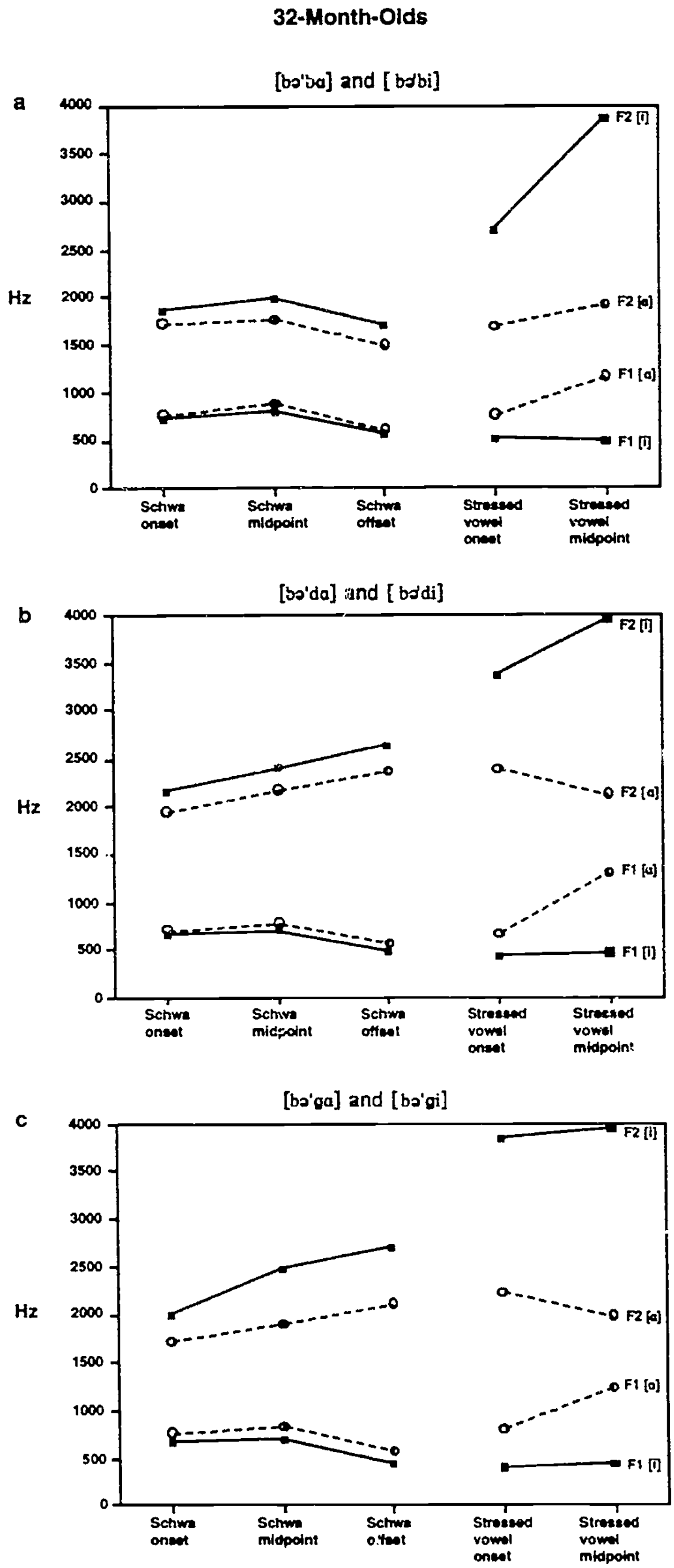


Figure 4. Mean formant paths for 32-month-olds for [bə'bV], [bə'dV], [bə'gV].

The mean formant paths for the children at Time 2 are shown in Figure 4. For the stressed vowels mean formant values remain virtually unchanged over the 10-month period, indicating that any differences in spectral structure at other points in the utterance are not a consequence of vocal tract growth. (Two tailed t-tests at stressed vowel midpoints revealed a significant difference only for F1 [i] ($t(5)=4.017, p<.01$), none for F1 [ɑ], for F2 [i] or for F2 [ɑ]).

The most salient differences between the two ages are in the schwa. Consider, for example, the schwa midpoints. For [bəˈbV] mean F2 at the schwa midpoint of the 32-month-olds has dropped from its value 10 months earlier by about 800 Hz before [i], and by about 300 Hz before [ɑ] ; for [bəˈdV] the drop before [i] is 200 Hz, while F2 before [ɑ] has risen by over 200 Hz; for [bəˈgV] F2 values at schwa midpoint have dropped by about 150 Hz before both vowels. F1 values at schwa midpoint have dropped roughly 50 Hz before [i], 300 Hz before [ɑ] in [bəˈbV], and by about 200 Hz before both [i] and [ɑ] in [bəˈdV]. In [bəˈgV] F1 values at schwa midpoints are lower by about 150 Hz before both vowels. The difference between F1 values at schwa midpoints before the two vowels has decreased by about 200 Hz in [bəˈgV], but is roughly the same at both ages in [bəˈdV]. The absolute, non-normalized formant values suggest then that the 22-month-olds overlap gestures for the stressed vowel and the schwa more strongly in both dimensions of tongue placement than the older children, and that children at both ages do so more strongly than the adults.

We see the effects of age, and the different effects of the two stressed vowels, quite clearly if we plot the mean formant values at schwa midpoint before [i] and [ɑ] separately for each age group: in Figure 5 F2-F1 is plotted on the abscissae, estimating relative front-back tongue position (Ladefoged, 1984), and F1 on the ordinates, estimating relative tongue height. Due to the reversed axes, the plots of the vowels correspond roughly to vowel position in the oral cavity. For the 22-month-olds the schwa values before [ɑ] lie close to those of the target stressed vowel itself, while before [i] they are appropriately more central. For the 32-month-olds, schwa values before [ɑ] have moved forward and upward in the oral cavity and are beginning to cluster with schwa values before [i], much in the fashion of the adults. Notice that, even in the adults, for each place of consonantal closure, schwa before [i] lies closer to [i] than to [ɑ], and schwa before [ɑ] lies closer to [ɑ] than to [i]. Notice too that for both older children and adults schwa before both [i] and [ɑ] is slightly lower and clearly more backed before labial than before alveolar or velar closure.

To determine the statistical significance of the differences between children and adults described above, we must remove the effects of differences in vocal tract size on spectral range. We therefore turn now to our normalization procedures.

Normalization Procedures and Statistical Tests

Tongue front-back position (F2-F1). Ladefoged (1984) has proposed that the difference between F2 and F1 provides a more direct estimate of front-back tongue position across the cardinal vowels within a speaker than simple F2. Accordingly, mean F1 was subtracted from mean F2 for each utterance type for each subject at schwa onset (1), schwa midpoint (2), and stressed-vowel midpoint (5) in order to estimate tongue position in the front-back dimension. The estimates at the midpoint of the stressed vowel represent the subject's "target" formant values for [i] and [ɑ]. In order to get a normalized measure of the stressed vowel effects, F2-F1 values at the other two points were expressed as proportions of this "target" value: $F2_{i1}\text{-}F1_{i1}/F2_{i5}\text{-}F1_{i5}$, $F2_{\alpha5}\text{-}F1_{\alpha5}/F2_{\alpha1}\text{-}F1_{\alpha1}$, etc. The letters and numbers in the subscripts refer to stressed vowels and measurement points. Thus, $F2_{i1}\text{-}F1_{i1}$ refers to the tongue position at the onset of the schwa before stressed [i], while $F2_{i5}\text{ -}F1_{i5}$ refers to tongue position in stressed [i], etc. Note that, for [ɑ] utterances mean F2-F1 is expected to be lower in the stressed vowel than at the other measurement points; accordingly, the inverse ratio was formed. If the value of F2-F1 in the schwa is identical to the value of F2-F1 in the stressed vowel (i.e., if the gestures for schwa and stressed vowel maximally overlap, yielding front-back tongue harmony), the ratio will be 1. Conversely, any reduction in the degree of gestural overlap will yield a corresponding reduction in the value of the ratio. This procedure for self-normalization of vowels is analogous to one developed by Gerstman (1968) to normalize Peterson and Barney (1952) vowels within and across speakers.

Table 5 lists the indices for the first two measurement points in the schwa. At the onset and midpoint of the schwa before [ɑ]—the points at which gestural overlap should be apparent—the children at 22 months have values of 1.00 and .99 respectively, indicating complete vowel harmony, but by 32 months the values have dropped to .79 and .75, while the adult values of .63 and .64 are even lower. Thus, for [ɑ] the younger children display the largest effect of the stressed vowel in tongue front-back position, while the adults display the least. For [i], by contrast, the values for the adults are somewh·it closer to 1 at both these points than are those of the 22- and 32-month-olds.

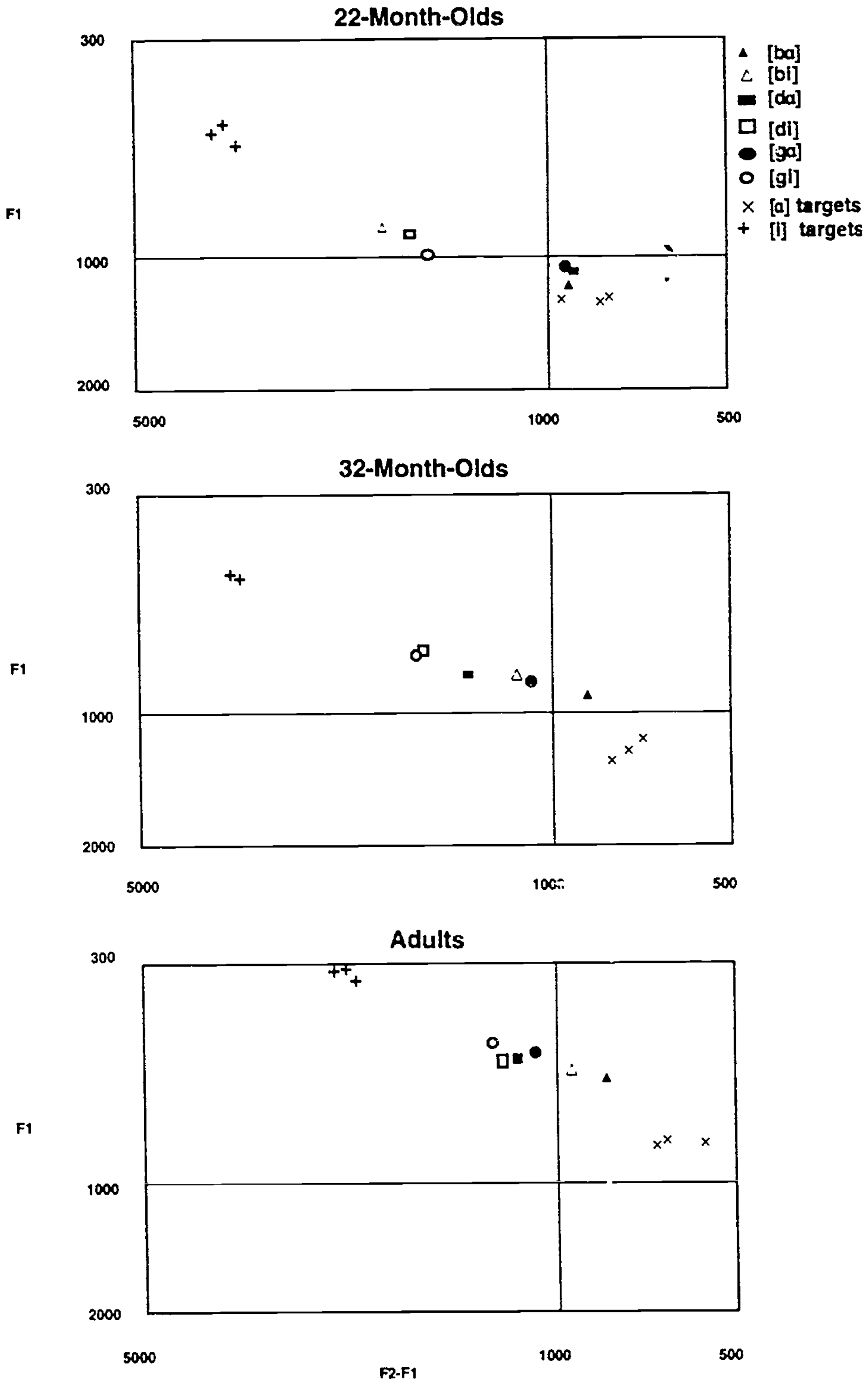


Figure 5. Mean F2-F1 plotted against mean F1 for schwa midpoints and stressed vowel midpoints for 22- and 32-month-olds and adults.

Table 5. *Mean ratios of individual mean F2-F1 values at indicated points to individual mean F2-F1 values at stressed vowel midpoint, listed by stressed syllable and vowel.*

	Schwa Onset	*Schwa Midpoint*
CHILDREN		
22 months		
Stressed Syllable		
bi	0.53	0.60
di	0.32	0.45
gi	0.39	0.46
bɑ	0.82	0.86
dɑ	1.19	1.21
gɑ	1.00	0.90
Stressed Vowel		
i	0.42	0.51
ɑ	1.00	0.99
32 months		
Stressed Syllable		
bi	0.35	0.35
di	0.44	0.49
gi	0.38	0.51
bɑ	0.85	0.88
dɑ	0.72	0.63
gɑ	0.80	0.74
Stressed Vowel		
i	0.39	0.45
ɑ	0.79	0.75
ADULTS		
Stressed Syllable		
bi	0.51	0.46
di	0.57	0.58
gi	0.52	0.57
bɑ	0.64	0.69
dɑ	0.60	0.61
gɑ	0.63	0.62
Stressed Vowel		
i	0.54	0.54
ɑ	0.63	0.64

Separate age-pair analyses of variance (Stressed Vowel × Age) were carried out on the F2-F1 ratios at the first two measurement points. See Table 6 for all significant effects and interactions for the three groups. For the 22-month-olds and adults there were significant main effects of both age and stressed vowel as well as a significant stressed vowel by age interaction at both onset and midpoint of the schwa. A simple effects analysis on the ratios for [i] revealed no significant differences between the groups at schwa onset or midpoint. The interactions therefore reflect the fact that the children's ratios are closer to 1 than the adults' before [ɑ], while before [i] they do not differ significantly.

For the 32-month-olds and the adults there were significant stressed vowel effects and stressed vowel by age interactions at schwa onset and midpoint, but no main effects of age. The interactions at both points reflect the fact that before [ɑ], the children's ratios are closer to 1, while before [i] the adult values are closer to 1. A simple effects analysis on values before [i] revealed a significant difference at schwa onset, indicating earlier anticipation of front-back tongue position of the stressed vowels by the adults (cf. Alfonso and Baer, 1982).

Finally, repeated measures analyses of variance on the 22- and 32-month-olds' ratios at schwa onset and midpoint revealed significant effects of age and stressed vowel at both points, but no significant interactions. Thus, averaged across vowels, the children's ratios in the first half of the schwa were closer to 1 at Time 1 than at Time 2, indicating a decrease in gestural overlap with age. At the same time, children at both ages tended to anticipate front-back tongue position more extensively in schwa before [ɑ] than in schwa before [i].

Tongue height (F1). Ratios were also formed to normalize estimates of tongue height. For [ɑ], mean F1 values in the schwa were placed over mean F1 values at the midpoint of the stressed vowel; for [i], since its F1 at midpoint is typically the lowest F1 in the utterance, an inverse ratio was formed: $\mathrm{F1}_{ɑ1}/\mathrm{F1}_{ɑ5}$ and $\mathrm{F1}_{i5}/\mathrm{F1}_{i1}$, etc. Once again, the closer the ratio is to 1, the more the gestural overlap between schwa and stressed vowel.

Table 7 lists the F1 indices at schwa onset and midpoint. At schwa onset, the indices reveal little difference among groups in the anticipation of tongue height for either vowel, and these indices were not further analyzed. But for [ɑ] at schwa midpoint the younger children display almost complete harmony in tongue height between schwa and stressed vowel with a ratio of .99.

Table 6. *Summary of significant effects in analyses of variance for F2-F1 ratios at schwa onset and schwa midpoint.*

Measurement Point	*Independent variable*	*Degrees of Freedom*	*F*	*p*
22-month-olds and Adults				
Schwa Onset	Stressed Vowel	1,10	38.93	<.001
	Age	1,10	7.63	<.02
	Stressed Vowel × Age	1,10	20.65	<.001
Schwa Midpoint	Stressed Vowel	1,10	27.44	<.001
	Age	1,10	7.95	<.02
	Stressed Vowel × Age	1,10	11.07	<.007
32-month-olds and Adults				
Schwa Onset	Stressed Vowel	1,10	34.45	<.001
	Stressed Vowel × Age	1,10	13.64	<.004
Schwa Midpoint	Stressed Vowel	1,10	22.24	<.001
	Stressed Vowel × Age	1,10	5.26	<.04
22- and 32-month-olds				
Schwa Onset	Stressed Vowel	1,5	117.47	<.001
	Age	1,5	8.05	<.04
Schwa Midpoint	Stressed Vowel	1,5	41.55	<.001
	Age	1,5	12.73	<.02

Table 7. *Mean ratios of individual mean F1 values at indicated points to individual mean F1 values at stressed vowel, listed by stressed syllable and vowel.*

	Schwa Onset	*Schwa Midpoint*
CHILDREN		
22 months		
Stressed Syllable		
bi	0.87	0.74
di	0.78	0.60
gi	0.81	0.58
bɑ	0.69	1.02
dɑ	0.75	1.03
gɑ	0.66	0.90
Stressed Vowel		
i	0.82	0.64
ɑ	0.70	0.99
32 months		
Stressed Syllable		
bi	0.76	0.64
di	0.80	0.72
gi	0.71	0.70
bɑ	0.74	0.84
dɑ	0.62	0.68
gɑ	0.63	0.68
Stressed Vowel		
i	0.76	0.68
ɑ	0.66	0.73
ADULTS		
Stressed Syllable		
bi	0.73	0.63
di	0.75	0.65
gi	0.73	0.69
bɑ	0.68	0.75
dɑ	0.61	0.67
gɑ	0.61	0.67
Stressed Vowel		
i	0.73	0.66
ɑ	0.63	0.70

Analyses of variance (Stressed Vowel × Age) were carried out for each age-pair on the F1 ratios at schwa onset and midpoint. There were no significant main effects or interactions at schwa onset. At schwa midpoint for the 22-month-olds and the adults, a main effect of age ($F_{1,10} = 7.01$, $p<.02$), was almost entirely due to the ratios in the [ɑ] context, with means of .99 and .70 for the children and adults, respectively; a significant effect of stressed vowel ($F_{1,10} = 6.27$, $p<.03$) indicated less extensive anticipation of [i] than of [ɑ] by both groups; a stressed vowel by age interaction was only marginally significant ($F_{1,10} = 3.76$, $p < .08$), but the value of the ratio in the [ɑ] context for the 22-month-olds (.99) indicates that the vowel harmony observed for F2 at schwa midpoint was also present for F1. The corresponding analysis for the 32-month-olds and adults yielded no significant effects or interactions. Finally, repeated measures analyses of variance on the 22- and 32-month-olds' ratios at midpoint revealed significant effects of stressed vowel ($F_{1,5} = 6.68$, $p<.05$) and age $F_{1,5} = 7.67$, $p<.04$), but no significant interactions. The means for the 22-month-olds and 32-month-olds (averaged across vowels) are .82 and .71, respectively, indicating that schwa and stressed vowel tongue heights are better differentiated at the older age.

Individual differences. Up to this point we have been discussing general trends across the six children. Let us now consider some of the individual differences. Figures 6 and 7 show how two children's vowel plots at schwa midpoint before the two stressed vowels changed over the ten-month interval. In Figure 6a for subject EA overlap of schwa and stressed vowel gestures has resulted in very similar formant values for the midpoints of the schwas and the midpoints of the stressed vowels, a case of near vowel harmony. By 32 months (Figure 6b) the schwa values have assumed a more central position, approaching the pattern of the adults (cf. Figure 5c). Notice however that schwa before labial closure is appreciably lower and more backed than before alveolar or velar closure. The backing pattern is similar to, but more extreme than, that noted above for the adults.

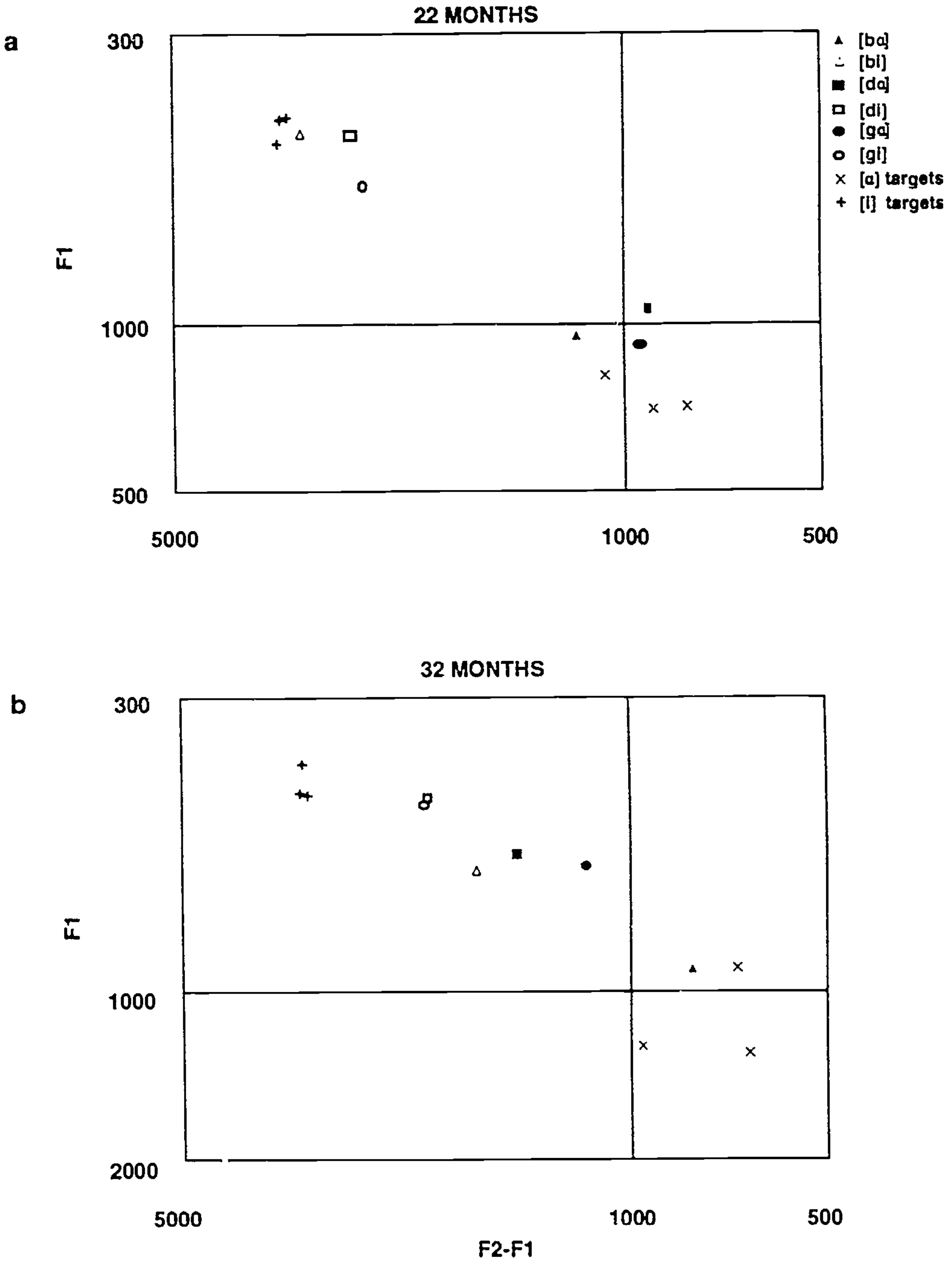


Figure 6. Mean F2-F1 plotted against mean F1 for schwa midpoints and stressed vowel midpoints for subject EA at 22 and 32 months of age.

In Figure 7a (subject SR), a very different pattern is evident. At 22 months schwa measures for all tokens are closer to [ɑ] than to [i]: this subject completely eschews vowel harmony before [i], executing an [ɑ]-like schwa before both [i] and [ɑ]. The data at 32 months (Figure 7b) reveal a more centralized schwa, higher and more forward in the oral cavity. Once again, as in the adults and in EA at 32 months, schwa is more backed before labial than before alveolar or velar closures (cf. Figures 5c and 6b).

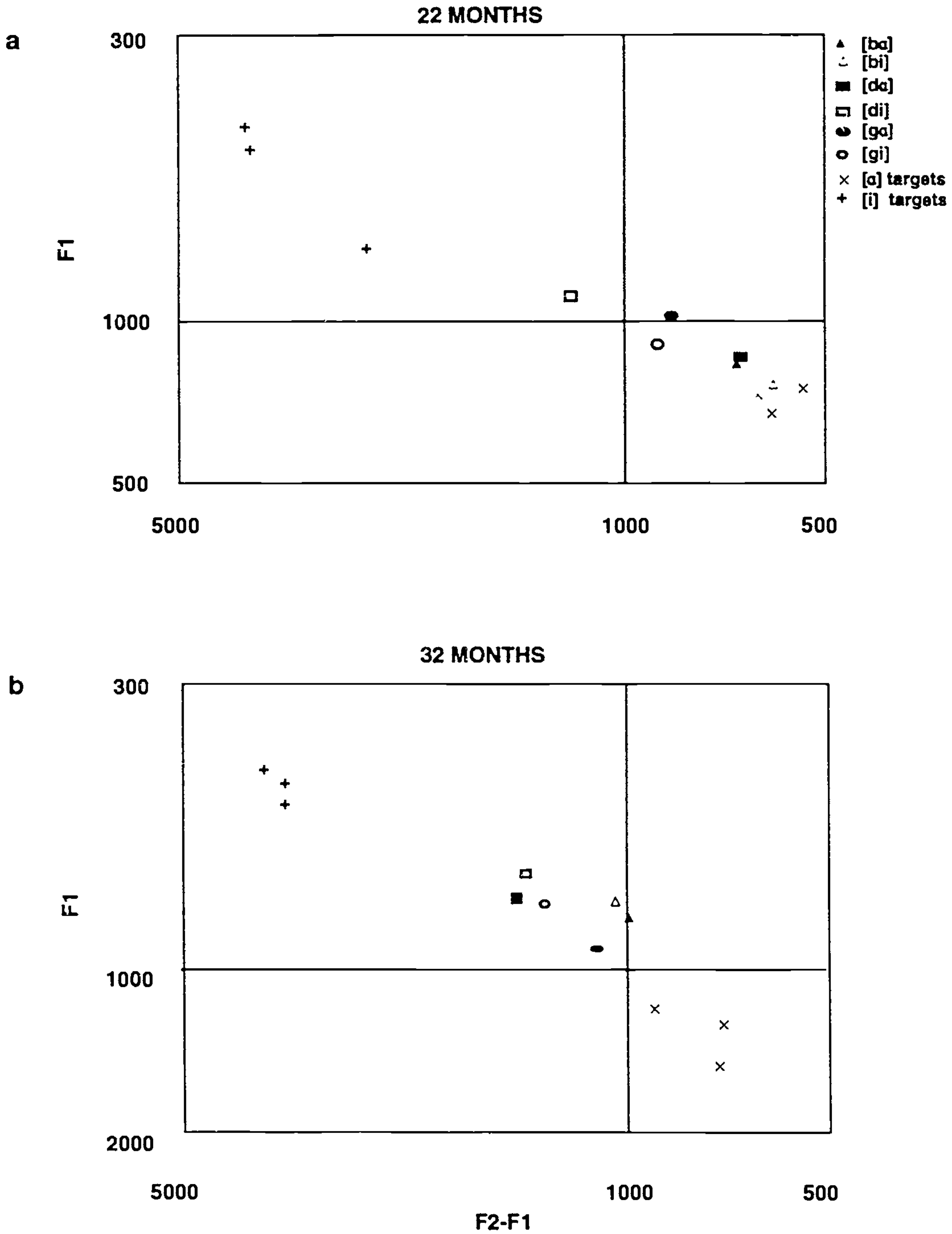


Figure 7. Mean F2-F1 plotted against mean F1 for schwa midpoints and stressed vowel midpoints for subject SR at 22 and 32 months of age.

Discussion: Overlap of vowel gestures. The formant estimates for the subjects as illustrated in the formant path figures prompted two questions: one concerning child-adult differences in gestural overlap, the other concerning differences in the effects of the target stressed vowels, [i] and [ɑ]. The results show that the younger children anticipated the stressed vowel [ɑ] in the schwa in both tongue front-back position and tongue height significantly more than the older children and adults. The lack of significant effects for the older children and adults indicates that over this 10-month period the children's control of tongue position and height in the schwa had become virtually adult-like. For schwa before [i], all three groups displayed roughly the same amount of gestural overlap in tongue position and tongue height.

We can perhaps gain insight into the differential effects of [ɑ] and [i] on the preceding schwa in the children's data, if we consider in more detail the differences between subjects EA and SR, described above. The almost complete harmony between target schwa and its following stressed vowel, displayed by subject EA at 22 months before both [i] and [ɑ], demonstrates that she could not segregate schwa from its following vowel in either tongue front-back position or tongue height. As may be judged by comparing the more or less central position of schwa values before [i] in the 22-month-old group data (Figure 5a) with their harmonized pattern for EA (Figure 6a), and their [ɑ]-like values for SR (Figure 7a), EA was alone in this tendency to harmonize schwa with the color of its following vowel. Ten months later she still could not control tongue height before a labial consonant followed by [ɑ], but was well on the way to controlling both tongue front-back position and tongue height in all other contexts. By contrast, subject SR at 22 months showed no tendency to anticipate either tongue front-back position or tongue height before [i]; instead, she lowered and backed her tongue in target schwa almost as far before [i] as before [ɑ]—in fact, even further if the following closure was labial (see Figure 7a). The greater lowering and backing of the tongue before [b] than before [d] or [g] (evident in both children and adults, as remarked above) may reflect the lower jaw position at closure for labials than for linguals (cf. Sussman, MacNeilage & Hanson, 1973), implicating jaw rather tongue action in the acoustic effect. If this is so, SR's difficulty with schwa at 22 months was perhaps not so much in positioning the tongue to achieve a neutral vowel, as in a tendency to anticipate the jaw lowering of the following stressed vowel, whatever its color. SR's difficulties with schwa would then have arisen from a general tendency to harmonize the jaw height of the two syllables, while EA's difficulty, at least with schwa before [i], if not also with schwa before [ɑ], would have arisen from a tendency to harmonize their tongue positions.

In their difficulties with schwa before [i] these two children are unlike the 22-month-olds as a group, each in her different way. But in their difficulties with schwa before [ɑ] they are typical (see Figure 5a). We may therefore suspect that the older children's difficulty with schwa before [ɑ] also arose, like that of SR, from a tendency to harmonize jaw height, that is, to anticipate in schwa the jaw lowering of the following stressed vowel. Their ability to block jaw lowering before [i] would then be due to the relatively little jaw movement that stressed [i] entails, as compared with the extensive movement for stressed [ɑ] (cf. Farnetani & Faber, in press). (The effect evidently does not arise from a tendency to anticipate phonetic stress more strongly before [ɑ] than before [i], since the durations of schwa before [i] and [ɑ] are roughly equal (see Table 2).) Imprecise control of the jaw, once extensive jaw-lowering has been initiated, may indeed be characteristic of young children's speech. Several authors have reported that children's tongue height for [i] is remarkably accurate, while tongue height for [ɑ] and schwa is more variable (Hare, 1983; Otomo & Stoel-Gammon, 1992; Paschall, 1983).

Finally, we may note that subject EA, who alone tended to harmonize schwa with the color of its following vowel, was less phonologically advanced at the start of the study than most of the other children, as indicated by MLU and vocabulary estimates, while subject SR who displayed no general tendency to harmonize vowel color, was the most phonologically advanced. This result is consistent with the hypothesis that less phonologically advanced subjects display greater vowel-to-vowel gestural overlap.

Results. Overlap of Consonant and Vowel Gestures

For normalized estimates of the effect of the stressed vowel gesture on the preceding schwa we had a reference point within the same utterance type, namely, mean formant values at the midpoint of the stressed vowel itself. For normalized estimates of the interaction between vowels and medial consonant gestures, we have no appropriate reference points within an utterance. (As noted above, the extent of formant transitions into or out of a consonant closure, proposed by Kent

(1983) as a measure of consonant-vowel coarticulation, is ambiguous precisely because variations in vowel formant values across subjects cannot be normalized against a reference point within the same utterance.) Accordingly, we had recourse to two across utterance measures by which a change in F2 at a given point in the utterance, as a function of a change in either the medial consonant or the following stressed vowel, served as a normalized index of the influence of the changed gesture on tongue position at that point.

Following Nittrouer et al. (1989), we computed F2 ratios for corresponding points in different utterances. To assess the effect of the medial consonant at schwa midpoint, schwa offset (consonant closure) and stressed vowel onset (consonant release), ratios were formed across consonant contexts with following stressed vowel fixed (bV/dV,bV/gV,dV/gV). (For example, at schwa midpoint for one 32-month-old subject, the mean F2 values for [bi] (1758) and [di] (2300), gave a ratio of .76 and a consonant index of (1-0.76)=.24; for [bɑ] (1701) and [dɑ] (2060) the ratio was 0.83, and the consonant index, (1-0.83)=0.17. The corresponding group mean indices in Table 8 are 0.27 and 0.18.). The classification of utterances according to stressed vowel was necessary due to the demonstrated effect of the stressed vowel on schwa formant values. To assess the effect of the stressed vowel at consonant closure, consonant release and stressed vowel midpoint, ratios were formed across vowels with consonant fixed (Cɑ/Ci). If there were no effects of varied context on such ratios (indicating complete absence of gestural overlap), their values would be equal to 1. To bring them into conformity with the within-utterance measures described above (and so to make for easier reading), indices of gestural overlap were formed by subtracting each ratio from 1; complete gestural overlap would then give an index of 1, complete lack of overlap an index of 0 (cf. Turnbaugh et al., 1985).

The effect of consonant on vowel gestures

Table 8 lists the consonant indices by age, measurement point and vowel. We note first that the d/g indices are uniformly low, oscillating around zero for all groups at all measurement points. Evidently, differences in F2 alone do not suffice to make this lingual contrast; in fact, the importance of F3 in distinguishing between alveolar and velar stops is well known. Accordingly, we confine our analysis to the labial-lingual contrasts (b/d, b/g).

The expected effect of a labial relative to a lingual constriction is to lower F2, giving rise to a positive index between 0 and 1. At consonant closure and release all indices are indeed positive for all groups and for both vowels; with the exception of the low index (.01) for bi/di at stressed vowel onset in the adults, all indices at these points are significantly different from zero by standard t-tests, indicating a substantial effect of the consonant gestures at all ages. At schwa midpoint, by contrast, the indices are positive and significant only for the 32-month olds and adults; for the 22-month olds at this point indices are negative (significantly so before /i/).

Table 8. *Mean F2 consonant indices ($1-[C_1V/C_2V]$) estimating relative F2 shift for three consonant contrasts at schwa midpoint, schwa offset (consonant closure), and stressed vowel onset (consonant release) before each stressed vowel.*

	Age								
	22-months			32-months			Adults		
Consonant ratio	Midpoint	Offset	Onset	Midpoint	Offset	Onset	Midpoint	Offset	Onset
bi / di	-0.18	0.11	0.27	0.27	0.34	0.20	0.15	0.24	0.01
bɑ / dɑ	-0.05	0.26	0.26	0.18	0.36	0.30	0.17	0.25	0.26
Mean b / d	-0.11	0.19	0.26	0.23	0.35	0.25	0.16	0.24	0.13
bi / gi	-0.13	0.29	0.36	0.19	0.37	0.29	0.15	0.35	0.14
bɑ /gɑ	-0.03	0.21	0.20	0.06	0.29	0.33	0.13	0.25	0.26
Mean b / g	-0.08	0.25	0.28	0.13	0.33	0.31	0.14	0.30	0.20
di / gi	0	0.08	0.12	0.03	0.01	0.12	-0.01	0.09	0.13
dɑ /gɑ	0.04	0.06	-0.10	-0.15	-0.12	-0.10	-0.06	0.01	0
Mean d / g	0.02	0.07	0.01	-0.06	-0.06	0.01	-0.03	0.05	0.07

Table 9 summarizes the significant results of each age-pair analysis of variance (Consonant Ratio × Vowel × Age) on the labial-lingual indices at the three measurement points. There are no systematic main effects of consonant ratio or vowel, although there are several significant interactions (Consonant by Vowel, Consonant by Age, Vowel by Age). None of the interactions lends itself to ready interpretation, because the pattern of interaction tends to vary from one analysis to another (see Table 8). Similarly, the significant effect of age in the 22- vs. 32-month-old analysis at schwa offset is anomalous, because neither 22- nor 32-month-olds differ from the adults at this point.

In fact, the most striking results are the effects of age at schwa midpoint, where the children's indices at Time 1 are significantly less than those at Time 2 or of the adults. Evidently, consonant closure began relatively earlier in the schwa for the adults and older children than for the 22-month-olds. Yet by the time closure was more-or-less complete (at schwa offset) the difference between the younger children and the adults had disappeared, and at consonant release before [i] it was even reversed (see the Vowel × Age interaction), indicating a stronger effect of the consonant contrast in the children (cf. Figures 2 and 3).

The effect of vowel on consonant gestures

Table 10 lists the vowel indices by age, measurement point and consonant. Table 11 summarizes the significant results of each age-pair analysis of variance (Vowel Ratio × Consonant × Age) on these indices at schwa offset and stressed vowel onset; there were no significant effects or interactions at stressed vowel midpoint.

The expected effect of the low back vowel [ɑ] with respect to the high front vowel [i] is to lower F2, giving rise to a positive index. All indices are indeed positive and, with the exception of those for the adults in the [d] context at schwa offset, significantly different from zero by standard t-tests. Notice further that, for the most part and as expected, the indices increase (that is, the vowels are increasingly differentiated) for all groups in all consonant contexts, as we move from schwa offset to stressed vowel midpoint.

Figure 8 displays the pattern of results across ages at schwa offset and stressed vowel onset. At schwa offset (Figure 8a) there is a clear tendency for the indices to decrease with age, and the differences between the 22-month-olds and adults are significant (Table 11). All three analyses also yielded a significant effect of consonant. In each age group the indices are lowest for [d], presumably due to the tight constraints exerted by alveolar constrictions on consonant-vowel coarticulation (Stevens, House, & Paul, 1966; Recasens, 1991). For the 22-month-olds, labial and velar indices are equal; for the adults and 32-month-olds velar indices are higher than labials. However these differences were not reliable enough to induce significant consonant by age interactions.

Table 9. Summary of significant effects in analyses of variance for consonant indices at schwa midpoint, schwa offset and stressed vowel onset.

Measurement Point	*Independent variable*	*Degrees of Freedom*	*F*	*p*
22-month-olds and adults				
Schwa midpoint	Age	1,5	24.42	.0043
Schwa offset	Consonant × Vowel	1,5	6.95	.0462
Stressed vowel onset	Age	1,5	13.65	.0141
	Consonant × Vowel	1,5	35.10	.0020
	Vowel × Age	1,5	82.10	.0003
32-month-olds and adults				
Schwa midpoint	Consonant × Vowel	1,5	18.69	.0016
Schwa offset	Consonant × Vowel	1,5	15.10	.0116
	Consonant × Age	1,5	7.32	.0425
Stressed vowel onset	Consonant × Vowel	1,5	28.09	.0032
	Vowel × Age	1,5	8.25	.0349
22-and 32-month olds				
Schwa midpoint	Age	1,10	18.49	.0016
Schwa offset	Age	1,10	4.77	.0538
	Consonant × Vowel	1,10	6.41	.0298
Stressed vowel onset	Consonant × Vowel	1,10	12.77	.0051

Table 10. *Mean F2 vowel indices (1-[Ca/Ci]) estimating F2 shift due to the vowel contrast at schwa offset, stressed vowel onset and stressed vowel midpoint, in three consonant contexts.*

	Vowel index								
	Schwa offset			Stressed vowel onset			Stressed vowel midpoint		
Age	ba/bi	da/di	gagi	ba/bi	da/di	gagi	ba/bi	da/di	ga/gi
22 months	.25	.12	.25	.33	.33	.45	.50	.45	.51
32 months	.11	.09	.20	.37	.29	.41	.50	.46	.49
Adults	.08	.06	.15	.40	.19	.30	.42	.42	.44

Table 11. *Summary of significant effects in analyses of variance for vowel indices at schwa offset and stressed vowel onset.*

Measurement Point	*Independent variable*	*Degrees of Freedom*	*F*	*p*
22-month-olds and Adults				
Schwa offset	Consonant	2.20	3.33	<.056
	Age	1.10	7.01	<.024
Stressed vowel onset	Consonant	2.20	9.90	<.001
	Age	1.10	7.42	<.021
	Consonant × Age	2.20	8.76	<.002
32-month-olds and Adults				
Schwa offset	Consonant	2.20	6.03	<.009
Stressed vowel onset	Consonant	2.20	12.53	<.001
	Age (marginal)	1.10	3.33	<.098
	Consonant × Age (marginal)	2.20	3.29	<.058
22- and 32-month-olds				
Schwa offset	No significant effects			
Stressed vowel onset	Consonant	2.10	6.15	<.02

At stressed vowel onset (Figure 8b), the moment of consonant release into the vowel, the indices are appropriately higher for all groups in all contexts, than at the moment of consonant constriction, and indices for lingual consonants again decrease with age, as indicated by a significant effect of age for the 22-month-olds and adults and a marginally significant effect of age for the 32-month-olds and adults. As at schwa offset, all analyses give a significant effect of consonant, with a similar pattern for the linguals ([d] lower than [g]) at all ages. For the labial, by contrast, the pattern reverses: gestural overlap is least at Time 1, larger at Time 2 and largest for the adults, giving rise to a significant consonant by age interaction for the younger children and adults, and a marginally significant interaction for the older children and adults. Separate analyses revealed no effects of age for the labials, but significantly greater overlap in the children at Time 1 than in the adults for [d] and [g] (alveolars F(1,10)=9.18, *p*<.01, velars F(1,10)=9.41, *p*<.01), and for [g] the children at Time 2 (F(1,10)=5.72, *p*<.04). Comparison of the indices for [b] and [g] by *t*-tests within groups revealed significant differences only for the adults (*t*(5)=2.63, *p*<.05), indicating greater overlap for labials.

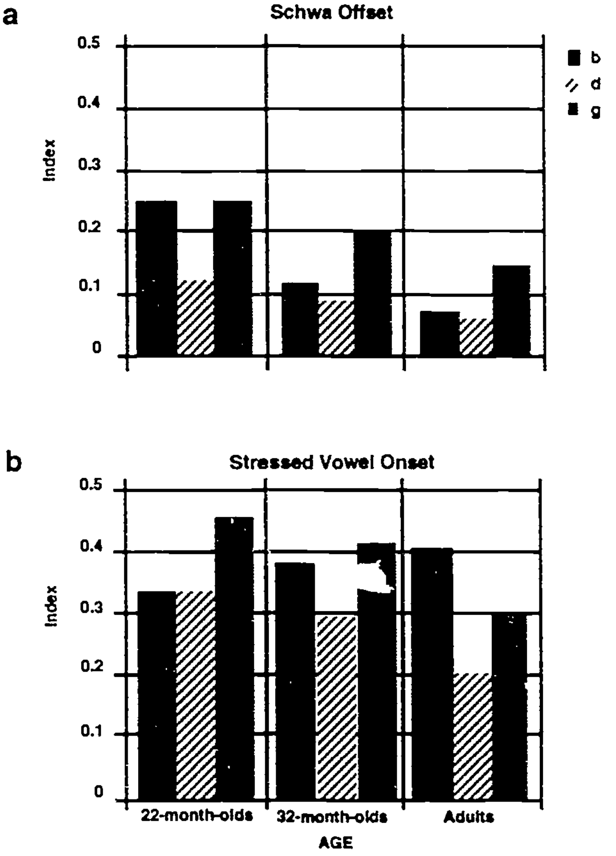


Figure 8. Mean vowel indices ([1-Ca/Ci)], estimating relative F2 lowering at schwa offset (top, a) and stressed vowel onset (bottom, b) for 22- and 32-months olds and adults.

Discussion: Overlap of Consonant and Vowel Gestures

The consonant indices showed that, at the midpoint of the schwa, the children at Time 1, unlike Time 2 and the adults, had not yet begun to move toward closure for the medial consonant. Yet, at the time of consonant closure, the effect of the consonant contrast was not significantly different across ages. This result (paradoxical, if we assume that articulatory movements are slower in younger children) makes sense when we recall that schwa midpoint was (on average) 95 msec before consonant closure in the 22-month-olds, but only 55 msec and 43 msec, respectively, before consonant closure for the 32-month-olds and adults (see Table 2). Thus, all three age groups may have begun to close for the consonant roughly the same number of milliseconds before complete closure. It may even be that, in absolute time measured from consonant closure, the children at Time 1 began the movement *before* they did at Time 2. But we do not have the data to test this. All we can conclude from the data we do have is that the absolute extent of gestural overlap in an intersyllabic vowel-stop sequence may be no different in 2-year-old children than in adults. Certainly, we have no evidence for a decline in overlap with age.

Turning to the vowel indices, we recall first that Turnbaugh et al. (1985) found no overall age-related differences in intrasyllabic CV coarticulation for CVC syllables spoken by 3- and 5-year-old children and adults. They suggested that adult-like coarticulation may develop at an earlier age. The results of the present study support this hypothesis for the lingual consonants, [d] and [g]: overlap of consonant release with the following vowel gesture decreases from 22-months-olds to 32-months-olds to adults. For the labials, by contrast, there is no significant decrease; on the contrary, there is a non-significant trend toward an increase in CV overlap with age, culminating in significantly greater overlap for [b] than for [g] in the adults, a finding in accord with the results of Turnbaugh et al. (1985).

In this difference between lingual and labial consonants, we have another clear counter-example to the general hypothesis that gestural overlap decreases after two years of age. Evidently, an age-related decrease may occur when consonant and vowel gestures engage the same articulatory sub-system (the tongue), but an increase may occur when they engage different sub-systems (tongue and lips). The former observation is consistent with the results of another experiment with the same subjects as the present one, in which the adults began the vowel gesture later in the fricative of an alveolar/palatal fricative-vowel syllable than did the children at Time 1 and at Time 2 (Goodell, 1991; Goodell & Studdert-Kennedy, 1991). The same result with fricatives has been reported for 3-year-olds by Nittrouer et al. (1989) and by Siren, (1991). This interpretation is also consistent with the finding of Stevens et al. (1966), confirmed by Recasens (1991) electropalatographically, by Turnbaugh et al. (1985) and the present study acoustically, that consonant-vowel overlap is lowest in a stop-vowel syllable when demands for precision of tongue placement are greatest, namely, in the execution of closure by the tongue-tip at the alveolar ridge.

GENERAL DISCUSSION

Three questions were posed at the end of the introduction: Does gestural coordination in the speech of 2- to 3-year-old children differ from that of adults? How does gestural organization change over a 10-month interval within roughly the third year of life? Do measurements from the acoustic records of young children support the inference from studies of child phonology that children around this age organize their gestures over a wider domain than adults?

The results demonstrate clear differences in speech gestural coordination between 2- to 3-year-old children and adults. They also demonstrate a clear shift in gestural coordination toward that of adult speakers during roughly the third year of life. Generally, we may say that 2- to 3-year-old children do not organize their utterances over a wider domain than adults, but do tend to produce longer utterances with different degrees of overlap between neighboring gestures than adults. Details of the child-adult differences and developmental changes vary from one aspect of an utterance to another, as reviewed below.

Gestural Organization Across Ages

The effect of stressed vowel on preceding schwa. In the vowel-to-vowel comparison, the 22-month-olds displayed stronger effects of the stressed vowel on the preceding schwa than did the 32-month-olds or adults for both tongue front-back position and tongue height when this vowel was low back [ɑ]; for schwa before [i] the pattern of results was the same in all three age groups. There were also no differences between the 32-month-olds and the adults in front-back tongue position or tongue height before either [ɑ] or [i], indicating that older children were executing the schwa in much the same way as the adults. We attributed the differential effects of stressed [ɑ] or [i] on preceding schwa in the younger children to anticipation of the more extensive jaw-lowering characteristic of [ɑ], and to poor control of the jaw once extensive jaw-lowering had been initiated.

At the same time, we found extensive gestural overlap in the schwa before both vowels, [ɑ] and [i], for one of the less phonologically advanced subjects at 22 months, but not at 32 months, suggesting that gestural overlap might be found in [i] tokens for other less phonologically advanced subjects. Certainly, at the onset of speech, young children learning English display a bias toward vowel harmony that is absent from the surrounding language: Kent and Bauer (1985) reported that 44% of vowel pairs in the disyllabic babbling of five 13-month-olds were reduplications. Perhaps development of vowel-to-vowel effects in iambic disyllables proceeds from schwa being assimilated with the stressed vowel (vowel harmony) in younger children to schwa assuming a more central vowel position in older children. Harmony might then break earlier before [i] than before [ɑ] because the former's greater acoustic-articulatory distance from schwa, facilitates its perceptual and aeticulatory segregation.

Yet there are discrepancies with previous studies of vowel-to-vowel overlap in older children: studies of two girls, four and nine years of age by Repp (1986) and eight 8-year-old children by Flege (in preparation). Repp, analyzing [ə#CV] utterances where V was [i,æ,u], found anticipatory effects of front-back tongue position only for the older child, and no effects of tongue height in either child. Flege found that 8-year-olds showed greater anticipation of tongue height in a schwa-stressed vowel context, [əˈhVp], (where V was [i,ɪ,o,u,ɑ]) than did adults. If anticipatory effects for tongue position are found in 22- and 32-month-olds (as reported in the present study), why are such effects absent in a 4-year-old, but present in a 9-year-old (as reported by Repp)? If vowel-to-vowel gestural overlap in tongue height has diminished to adult levels by 32 months of age (as reported in the present study), why is it more salient in 8-year-olds than in adults (as reported by Flege)? Here, utterance type is a likely factor: medial stops probably block gestural overlap between flanking vowels more effectively than medial /h/. Of course, discrepancies may also arise from differences in measurement technique, cross-sectional sampling bias, statistical test power, and so on. Systematic longitudinal studies hold the best promise of resolving these issues.

The effect of medial stop consonant on preceding schwa. Indices of the effects of the medial consonant on preceding schwa at midpoint and offset revealed that the adults and older children had already initiated consonant closure at schwa midpoint, where the younger children had not, but that by schwa offset the age difference had disappeared. Since the duration of the schwa was nearly twice as long in the younger children as in the older children and adults, this result tells us nothing about age differences in the absolute temporal extent of overlap. The uncertainty arises because we made our schwa measurements at the same relative point in each utterance (schwa midpoint) rather than at a fixed point, such as a certain number of milliseconds before consonant closure.

However, since the younger children's utterances were longer than those of their elders, their gestures were probably slower, so that movement toward consonant closure must have begun earlier in the vowel in order for it to be completed at the same time. The observed result is therefore consistent with (although it provides no direct support for) the hypothesis of an earlier absolute onset time of the consonant gesture in the children at Time 1, and perhaps in the children at Time 2, than in the adults. Such a result would also be consistent with the interpretation offered by Nittrouer et al. (1989), as noted above in the introduction, of the data from Kent (1983).

The effect of stressed vowel on preceding medial consonant. For the lingual consonants at closure and release, the pattern of overlap with the stressed vowel was the same at all ages (greater for [g] than for [d]), but its extent decreased with age. For the labial, by contrast, overlap decreased significantly with age at closure, but tended to increase with age at release. The similar patterns for the lingual consonants reflect the fact that initiation of the vowel gesture is more tightly constrained by alveolar than by velar gestures at every age. The age-related changes, though opposite in their effects on labial and lingual consonants, evidently reflect the same developmental process, namely, a growing capacity to segregate, or differentiate, successive gestures. Adults have learned to take advantage of the independence of tongue and lips to execute a more extensive portion of the vowel gesture while their lips are closed than have the children. By the same token, they have learned to delay the onset of the vowel gesture, when closure is executed by the tongue.

As already noted, several cross-sectional studies of so-called "anticipatory coarticulation" have been carried out on older children and adults (Hodge, 1989; Sereno et al., 1987; Sereno & Lieberman, 1987; Turnbaugh et al., 1985). The last three of these studies found no age differences. Hodge (1989), comparing F2 trajectories at vowel onset in /di/, /dæ/, /du/, found more coarticulation in 3-year-olds than in older children and adults, a result consistent with our own for children close to three years of age. However, the lack of age effects in cross-sectional studies of older children, suggests once again that, if we are to understand the development of gestural organization, we must study children longitudinally over their first two to three years of life.

The Domain of Gestural Organization

The results of this study do not support the general hypothesis that the temporal domain of gestural organization is wider in 2- to 3-year-olds than in adults. The durations of the children's utterances were certainly greater, and the younger children, particularly, tended to assimilate the duration of the unstressed to the duration of the following stressed syllable. Yet at whatever point in an utterance we found evidence of gestural overlap in the children, we also found corresponding evidence in the adults. Nor do the results support the general hypothesis that the degree of overlap at any given point in an utterance is greater in 2- to 3-year-olds than in adults. The extent may be equal (as in the overlap of the stressed vowel [i] with the schwa), greater (as in the overlap of the stressed vowel [ɑ] with the schwa, and of the stressed vowel with lingual stop gestures at consonant closure and release), or less (as in the overlap of the stressed vowel with the labial stop gesture). Whether gestural overlap is greater or less, the differences seem largely to arise from the children's difficulties in timing both the precise duration of a gesture and its onset or offset with respect to other gestures. Learning to talk evidently entails (among other things) learning to differentiate, and to bring under independent control, the several gestures that compose the sequence of syllables in an utterance.

REFERENCES

Alfonso, P. J., & Baer, T. (1982). Dynamics of vowel articulation. *Language and Speech, 25*, 151-173.

Allen, G. D., & Hawkins, S. (1980). Phonological rhythm: Definition and development. In G. Yeni-Komshian, J. F. Kavanagh & C. A. Ferguson (Eds.), *Child phonology, Volume 1: Production* (pp. 227-256). New York: Academic Press.

Branigan, G. (1979). Some reasons why successive single word utterances are not. *Journal of Child Language, 6*, 411-421.

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3*, 219-252.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*, 151-206.

Brown, R. (1973). *A first language.* Cambridge, MA: Harvard University Press.

Donahue, M. (1986). Phonological constraints on the emergence of two-word utterances. *Journal of Child Language, 13*, 209-218.

Davis, B. D., & MacNeilage, P. F. (1990). Acquisition of correct vowel production: A quantitative case study. *Journal of Speech and Hearing Research, 33*, 16-27.

DiSimoni, F. G. (1974). Influence of vowel environment on the duration of vowels in the speech of three-, six-, and nine-year-old children. *Journal of the Acoustical Society of America, 55*, 362-363.

Eguchi, S., & Hirsh, I. J. (1969). Development of speech sounds in children. *Acta Otolaryngologica*, Supplement 257.

Farnetani, E., & Faber, A. (in press). Tongue-jaw coordination in vowel production: Isolated words vs. connected speech. Speech Communication.

Ferguson, C. A., & Farwell, C. B. (1975). Words and sounds in early language acquisition. Language, 51, 419-439.

Flege, J. E. (in preparation). Children show greater anticipatory lingual coarticulation in VCVs than adults.

Gerstman, L. H. (1968). Classification of self-normalized vowels. IEEE Transactions on Audio- and Electroacoustics, AU-16, 16-19.

Goodell, E. W. (1991). Gestural organization in the speech of 22- to 32-month-old children. Unpublished doctoral dissertation, University of Connecticut.

Goodell, E. W., & Studdert-Kennedy, M. (1991). Articulatory organization in early words: From syllable to phoneme. Proceedings of the XIIth International Congress of Phonetic Sciences, Vol. 4 (pp. 166-169) Aix-en-Provence, France: Universite de Provence.

Hare, G. (1983). Development at 2 years. In J. V. Irwin & S. P. Wong (Eds.), *Phonological development in children: 18-72 months* (pp. 55-85). Carbondale, IL: Southern Illinois University Press.

Hodge, M. M. (1989). *A comparison of spectral-temporal measures across speaker age: Implications for an acoustic characterization of speech maturation.* Unpublished doctoral dissertation, University of Wisconsin-Madison.

Katz, W. F., Kripke, C., & Tallal, P. (1991). Anticipatory coarticulation in the speech of adults and young children: Acoustic, perceptual, and video data. *Journal of Speech and Hearing Research, 34,* 1222-1232.

Kent, R. D. (1983). Segmental organization of speech. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 57-89). New York: Springer-Verlag.

Kent, R. D., & Bauer, H. R. (1985). Vocalizations of one-year-olds. *Journal of Child Language, 12,* 491-526.

Kubaska, C., & Keating, P. (1981). Word duration in early speech. *Journal of Speech and Hearing Research, 24,* 615-621.

Ladefoged, P. (1984). *A course in phonetics.* San Diego: Harcourt, Brace, Jovanovich.

Leonard, L. B., Schwartz, R. G., Folger, M. K., & Wilcox, M. J. (1978). Some aspects of child phonology in imitative and spontaneous speech. *Journal of Child Language, 5,* 403-415.

Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (1983). Self-organizing processes and the explanation of language universals. In B. Butterworth, B. Comrie, & O. Dahl (Eds.), *Explanations for language universals* (pp. 181-203). The Hague: Mouton.

Limber, J. (1973). The genesis of complex sentences. In T. E. Moore (Ed.) *Cognitive development and the acquisition of language.* New York: Academic Press.

Macken, M. A. (1979). Developmental reorganization of phonology: A hierarchy of basic units of acquisition. *Lingua, 49,* 11-49.

McCune, L., & Vihman, M. (1987). Vocal motor schemes. *Papers and Reports in Child Language Development.* (26).

Menn, L. (1983). Development of articulatory, phonetic and phonological capabilities In B. Butterworth (Ed.), *Language production* (pp. 3-50). London: Academic Press.

Menn, L. (1986). Language acquisition, aphasia and phonotactic universals. In F. R. Eckman, E. A. Moravcsik, and J. R. Wirth, *Markedness.* New York: Plenum Press. pp. 241-255.

Menyuk, P., Menn, L., & Silber, R. (1986). Early strategies for the perception and production of words and sounds. In P. Fletcher & M. Garman (Eds.) *Language acquisition* (pp. 223-239) (2nd ed.). Cambridge: Cambridge University Press.

Miller, J. F., & Chapman, R. S. (1981). The relations between age and mean length of utterance. *Journal of Speech and Hearing Research, 24,* 154-161.

Nittrouer, S., Studdert-Kennedy, M. & McGowan, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research, 32,* 120-132.

Otomo, K., & Stoel-Gammon, C. (1992). The acquisition of unrounded vowels in English. *Journal of Speech and Hearing Research, 35, 604-616.*

Paschall, L. (1983). Development at 18 months. In J. V. Irwin & S. P. Wong (Eds.), *Phonological development in children: 18-72 months* (pp. 27-54). Carbondale, IL: Southern Illinois University Press.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175-184.

Recasens, D. (1991). An electropalatographic and acoustic study of consonant-to vowel coarticulation. *Journal of Phonetics, 19,* 177-192.

Repp, B. (1986). Some observations on the development of anticipatory coarticulation. *Journal of the Acoustical Society of America, 79,* 1616-1619.

Sereno, J. A., Baum, S. R., Marean, G. C., & Lieberman P. (1987). Acoustic analyses and perceptual data on anticipatory labial coarticulation in adults and children. *Journal of the Acoustical Society of America, 81,* 512-519.

Sereno, J. A., & Lieberman, P. (1987). Developmental aspects of lingual coarticulation. *Journal of Phonetics, 15,* 247-257.

Siren, K. A. (1991). *Coarticulation in the speech of children and adults: Developmental trends and associated linguistic factors.* Unpublished doctoral dissertation, University of Kansas.

Smith, B. L., Sugarman, M. P., & Long, S. H. (1983). Experimental manipulation of speaking rate for studying temporal variability in children's speech. *Journal of the Acoustical Society of America, 74,* 744-749.

Stevens, K., House, A., & Paul, A. (1966). Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation. *Journal of the Acoustical Society of America, 40,* 123-132.

Studdert-Kennedy, M. (1987). The phoneme as a perceptuomotor structure. In Allport, A., MacKay, D., Prinz, W., & Scheerer, E. (Eds.). *Language perception and production* (pp. 67-84). London: Academic Press.

Studdert-Kennedy, M. (1989). The early development of phonological form. In C. von Euler, H. Forssberg & H. Lagercrantz (Eds.), *Neurobiology of early infant behavior* (pp. 287-301). MacMillan: Basingstoke, England.

Studdert-Kennedy, M., & Goodell, E. (in press). Gestures, features and segments in early child speech. In B. deGelder & J. Morais (Eds.), *Language and literacy: Comparative approaches.* Cambridge: MIT Press.

Sussman, H. M., MacNeilage, P. F., & Hanson, R. J. (1973). Labial and mandibular dynamics during the production of bilabial consonants. *Journal of Speech and Hearing Research, 16,* 397-420.

Templin, M. (1957). Certain language skills in children. Minneapolis: University of Minnesota Press.

Turnbaugh, K., Hoffman, P., Daniloff, R. D. & Absher, R. (1985). Stop-vowel coarticulation in 3-year-olds, 5-year-olds, and adults. *Journal of the Acoustical Society of America, 77,* 1256-1258.

Vihman, M. M., & Velleman, S. (1989). Phonological reorganization: A case study. *Language & Speech, 32,* 149-170.

Waterson, N. (1971). Child phonology: A prosodic view. *Journal of Linguistics, 7,* 179-211.

FOOTNOTES

**Journal of Speech and Hearing Research,* in press.

† Also University of Connecticut, Storrs. Now at the University of Massachusetts, Amherst.

Gestures, Features and Segments in Early Child Speech*

Michael Studdert-Kennedy and Elizabeth Whitney Goodell

Alphabetic orthographies represent speech at the level of the phoneme. Yet the definition and functional status of the phonemic segment are widely viewed as problematic. Studies of child phonology have shown that the initial domain of a child's articulatory organization is the word rather than the segment, and have attributed a child's errors to failure to organize the features of a target word. The paper rejects the feature, as a unit of articulatory organization, on both rational and empirical grounds, proposing in its stead the gesture. Evidence from a child (1;9-2;1) learning American English shows how deviant forms, puzzling for a featural account, are readily explained as errors in the execution and timing of gestures. Phoneme-sized phonetic segments are then viewed as emergent properties of the gestural routines from which the child constructs its lexicon.

INTRODUCTION

Preliminary

That an alphabetic orthography represents speech at the level of the phoneme seems to be generally agreed. But the definition of the phoneme, and even its functional status, are still matters of contention among linguists. We do not propose to join the linguistic argument here. We take the facts of reading and writing to be sufficient evidence for the functional reality of the phoneme as a perceptuomotor control structure representing a class of phonetic segments (cf. Studdert-Kennedy, 1987). We assume, further, that phonetic segments (consonants and vowels) are not the irreducible elements of which speech is composed. Rather, segments are complex structures, implicit in the gestural patterns of speech, that gradually emerge and take on their perceptuomotor functions over the first few years of life. The present chapter attempts to justify this claim by applying a developing theory of articulatory phonology (Browman & Goldstein, 1986, 1989) to a small set of data drawn from the utterances of a two year old child.

Background

A child, learning to talk, often says the same word in several different ways. Indeed, variability of phonetic form has been a commonplace of child language studies since their inception (e.g., Albright & Albright, 1956, 1958; Cohen, 1952; Leopold, 1953). On the other hand, a child, learning to talk, often says several different words in the same way, and this homonymy is also a commonplace of child language studies (see Vihman, 1981 for review). While variability and homonymy may reflect many factors, including the communicative situation, whether the utterance is spontaneous or imitated, its meaning, phonetic structure, phonetic context, and so on (Schwartz, 1988), none of these factors would matter, if it were not that "...a child's phonemic system is in the process of development, and the sound patterning is probably less regular than that of adult speech" (Albright & Albright, 1956:382). But what, in fact, is developing, and what is the nature of the irregularity? What varies in the execution of a target word from one occasion to another? What does the child find that different words have in common?

Some of the data reported here were first presented at the Fifth International Congress for Child Language, Budapest, July 1990. We thank "Emma" and her family for their cheerful cooperation; Cathe Browman, Charles Ferguson, Louis Goldstein, Marlys Macken, Lise Menn, Elliot Saltzman, Marilyn Vihman and Natalie Waterson for the insights we have gained from their papers; Catherine Best, Cathe Browman, Ignatius Mattingly and Natalie Waterson for their instructive comments; Zefang Wang for assistance with the figures. The research and the preparation of this paper were supported, in part, by NIH grants to Haskins Laboratories, HD-01994 and NS-24655.

Let us begin with the observation, now supported by a variety of evidence, that the unit of phonological contrast, and therefore the unit of articulatory organization, in early child speech is the whole word or phrase rather than the segment (Ferguson, 1963, 1986; Ferguson & Farwell, 1975; Goodell & Studdert-Kennedy, 1991, in press; Macken, 1979; Menn, 1983; Menyuk, Menn, & Silber 1986; Nittrouer, Studdert-Kennedy, & McGowan, 1989; Waterson, 1971). To say that a child utters a word as a "whole," or Gestalt, cannot mean, however, that the child has not broken the word into at least some of its parts, because even a partly correct utterance requires coordination of independent, or partially independent, actions of lips, tongue, jaw, velum and larynx. Accordingly, while the word may be the domain over which a child organizes its articulations, it cannot be the basic unit of production (or, *a fortiori*, of perception). Nor, since the segment is no less compounded of independent articulatory actions than the word, can the basic unit be the segment itself.

The standard unit adopted in studies of child phonology is, of course, the feature. By this we cannot mean the abstract feature of generative phonology, a relational property fulfilling the linguistic function of contrast across a phonological system, because we are dealing with a child for whom such a system does not yet exist. We must therefore mean the concrete feature, an absolute property located "...within the speech sounds, be it on their motor, acoustical or auditory level" (Jakobson & Halle, 1956, p.8). However, at least two facts make this proposal unacceptable. First, the feature has no independent existence: it is a property of a larger unit and is carried into existence on that unit, as an adjective, not a noun (cf. Fowler, Rubin, Remez, & Turvey, 1980). This fact is implicit in the adjectival terminology of all feature theories: grave, acute, compact, coronal, nasal, strident, and so on.

A second, closely related objection is that the properties to which featural terminology customarily refers are purely static, devoid of temporal extension, and therefore intrinsically unfit to define the dynamic properties of speech either as a motor act or as an acoustic signal. Not surprisingly, none of several sets of acoustic and articulatory definitions of phonological features (e.g., Chomsky & Halle, 1968; Jakobson, Fant, & Halle, 1951/1963; Stevens, 1972, 1975, 1989) has proved precise or full enough to support a procedure for speech synthesis or speech recognition by machine, let alone a theory of speech production or speech perception.[1] The "autonomous features" of autosegmental and other forms of non-linear phonology (e.g., Clements, 1985; Goldsmith, 1990; McCarthy, 1988; cf. Menn, 1978) might seem to promise a solution. However, these features are abstract units, the temporal analogs of points in Euclidean space, admitting of sequence, but not of extension. In short, as Ladefoged (1980:485) has remarked, "...phonological features are certainly not sufficient for specifying the actual sounds of a language." They can hardly therefore serve the turn of a child striving to learn how to perceive and produce those sounds.

Very much the same holds for the informal, and perhaps intuitively more appealing, phonetic features adopted by, for example, Waterson (1971, p.183; cf. Ferguson & Farwell, 1975; Macken, 1979). She analyzes a child's word forms into: "Various features of articulation, such as nasality, sibilance, glottality, stop (complete closure), continuance, frontness, backness, voicing, voicelessness, labiality, rounding, non-rounding." Waterson goes on to group the child's forms into "structures" or "schemata", corresponding to the adult "prosodic" patterns in which features are distributed over a word. She shows how in a child's utterance the "prosody" may be disrupted, so that features lose their temporal order and recombine into patterns quite unlike the adult model. In this respect, Waterson's work has stimulated the approach taken in the present paper. Nonetheless, her schemata are purely descriptive, indications of, but certainly not specifications for, the spatio-temporal pattern of movements by which a speaker, child or adult, executes a word. They offer, at most, a sketch of the high points of a word, rendered in the familiar language of traditional articulatory phonetics. We conclude that, despite the utility of the feature as a descriptive and classificatory element in phonetic theory, it cannot guide a child into speech.

In fact, what a child quite evidently needs, to imitate an adult word, is a grasp on which articulators to move where and how, and on when to move them. And what we, for our part, need, to understand the child's attempts, is a description of the target word in terms of the units of articulatory action, and their relative timing, necessary to utter it. No generally agreed upon set of articulatory units exists, although several have been proposed. Ladefoged (1980), for example, offered a tentative list of 17 "articulatory parameters" that he judged necessary and sufficient to specify the sounds of a wide range of languages, but he did not develop them into a functional model. Here we adopt the framework of the most explicit model of speech production currently available, the gestural phonology being developed by Browman, Goldstein, Saltzman and their colleagues at Haskins Laboratories (Browman & Goldstein, 1986, 1987, 1989, 1990; Saltzman, 1986; Saltzman & Munhall, 1989), in which the basic phonetic and phonological unit is the gesture. We illustrate the approach with a small set of data, drawn from the utterances of a 21-25 month old girl, learning American English. But before we come to the data we must give a brief sketch of the gestural framework.

Gestures as basic units of articulatory action

If we watch, or listen to, someone speaking, we see, or hear, the speaker's mouth repeatedly closing and opening, forming and releasing constrictions. In the framework of gestural phonology, each such event, each formation and release of a constriction, is an instance of a gesture. Constrictions can be formed within the oral, velic or laryngeal articulatory subsystems; within the oral subsystem, they can be formed by the lips, the tongue tip (blade) or the tongue body. The function of each gesture, or act of constriction, is to set a value on one or more vocal tract variables that contribute to the shaping of a vocal tract configuration, by which (in conjunction with pulmonic action) the flow of air through the tract is controlled, so as to produce a characteristic pattern of sound. Presumably, this pattern of sound specifies for a child (or an adult) the gesture that went into its making.

Figure 1 displays the tract variables and the effective articulators of a computational model for the production of speech, at its current stage of development (Browman & Goldstein, 1990).

	tract variable	articulators involved
LP	lip protrusion	upper & lower lips, jaw
LA	lip aperture	upper & lower lips, jaw
TTCL	tongue tip constrict location	tongue tip, tongue body, jaw
TTCD	tongue tip constrict degree	tongue tip, tongue body, jaw
TBCL	tongue body constrict location	tongue body, jaw
TBCD	tongue body constrict degree	tongue body, jaw
VEL	velic aperture	velum

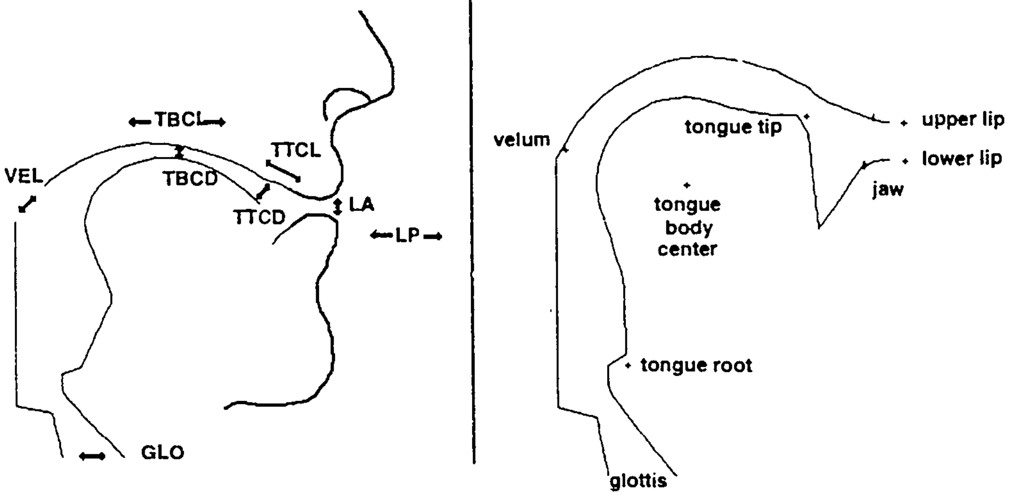


Figure 1. Tract variables and associated articulators used in the computational model of phonology and speech production discussed in the text. (Adapted from Browman & Goldstein, 1990).

The inputs to the model are the parameters of sets of equations of motion for gestures. A gesture is an abstract description of an articulator movement,[2] or of a coordinated set of articulator movements, that unfolds over time to form and release a certain degree of constriction at a certain location in the tract. Settings of the parameters permit constriction degree to vary across five discrete values (closed, critical, narrow, mid, wide), and constriction location for oral gestures to vary across nine values (protruded (lips), labial, dental, alveolar, postalveolar, palatal, velar, uvular, pharyngeal).[3] The reader may observe that the degree and location of an oral constriction roughly correspond to the manner and place of articulation of a segment in standard terminology.

The gestures for a given utterance are organized into a larger coordinated structure, represented by a gestural score. The score specifies the values of the dynamic parameters for each gesture, and the period over which the gesture is active. Figure 2 (center) schematizes the score for the word *nut* ([nʌt]), as a sequence of partially overlapping gestural activation intervals; possible free variation in the duration of the velic gesture, and the resulting nasalization of the vowel, is indicated by extending the velic activation interval with a dashed line.

We cannot here go into detail on the workings of the model (for which the reader is referred to the papers cited above). We note only the following further points that, taken with the preceding sketch, may suffice for an intuitive grasp on how a gestural framework can contribute to an understanding of the nature and origin of irregularities in a child's early words.

1. An instance of a gesture is an objective, observable event. We can observe a gesture by ear, and this is the usual basis of both imitation and phonetic transcription. We can observe a gesture by eye, either unaided, as in lipreading, or with X-ray cinematography. We can observe a gesture by touch, as in the Tadoma method of speech perception. Finally, we can observe a gesture by sensing our own movements. However, if gestures drawing on the same, or closely neighboring, neuromuscular sets overlap in time (as, for example, in certain lingual gestures for the consonantal onset and vocalic nucleus of a syllable), the individual gestures may merge, so that we can observe only the resultant of their vectors.

2. The articulator sets and their dimensions given above are not exhaustive. For example, the tongue root must ultimately be included in the model to handle variations in pharynx width. Also, constriction shape will have to be included, to handle the tongue bunching, narrowing or hollowing, necessary in the formation of certain complex gestures (cf. Ladefoged, 1980). Even the definition of the gesture itself may have to be revised to permit independent control of the formation and release of a constriction.

3. A gesture is larger than the properties of constriction location, degree and shape that describe it, but smaller than the segment. Several independent gestures are required to form a segment, syllable, or syllable string.

4. Each gesture has an intrinsic duration that varies with rate and stress. Correct execution of an utterance requires accurate timing of the gesture itself, and accurate phasing of gestures with respect to one another.

5. By adopting the gesture as a primitive of articulatory action we can predict what types of errors children are likely, or not likely, to make in their early attempts at adult words. To this topic we now turn.

The gestural origins of errors in a child's early words

Phonetically, speech emerges over the first year of life from the lip smacks, tongue clicks and pops associated with the vegetative processes of eating and breathing, combined with the stereotyped vocalizations of cries and comfort sounds (Stark, 1986), through the reduplicated syllables of canonical babble, into the brief strings of phonetically contrastive elements that make up early words. Articulatorily, the progression is a cyclical process of differentiation and integration by which the child moves toward finer modulation of individual gestures and more precise phasing of their sequence. (For a fuller account of the hypothesized developmental course, see Studdert-Kennedy, 1991 a, b.)

Here, two steps are of interest. First is the shift in gestural timing associated with the integration of prebabbling oral and laryngeal gestures into the canonical syllable, usually around the seventh month (Holmgren, Lindblom, Aurelius, Jalling, & Zetterström, 1986; Koopmans van Beinum & van der Stelt, 1986). Earlier vocalizations, termed "marginal babble" by Oller (1980), are commonly longer than adult syllables, but display adult-like properties of resonance, intensity and fundamental frequency contour.

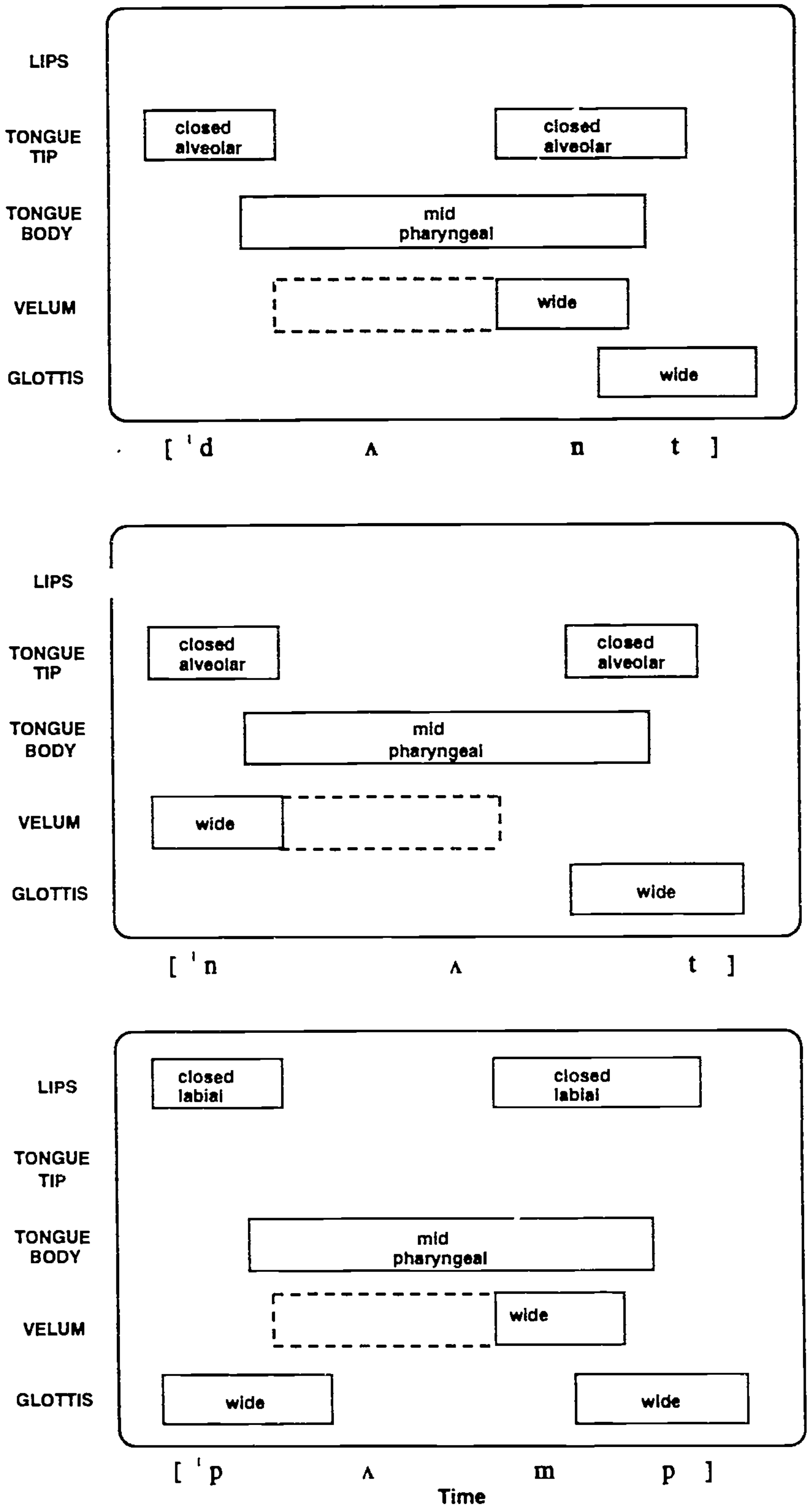


Figure 2. Schematic gestural score for the word *nut* ['nʌt], (center), and for *nut* as spoken by Emma in doughnut, ['dʌnt], (top), and *peanut*, ['pʌmp], (bottom). The extensions of the velic activation intervals by dashed lines indicate possible free variation in the duration of velic gestures.

Canonical babble is marked by integration of a resonant nucleus with rapid (25-120 ms) closing gestures at its margins to form a syllable with adult-like duration (100-500 ms) (Oller, 1986). The canonical syllable is the first step in the emergence of two major classes of oral gesture: the narrow or complete constriction of consonants and the wider constriction of vowels (cf. MacNeilage & Davis, in press).

The early canonical syllable is often, though not always, one of a rhythmic, reduplicated string of identical syllables. Reduplication indicates first that the child may lack independent control of the closing constriction at the margin and the wider constriction at the nucleus of a syllable; second, that the child cannot easily switch gestures in successive syllables. The tendency to reduplicate may continue for many months, or even years, as evidenced by the commonly reported harmony in early words between consonants (Vihman, 1978) and vowels, and even within consonant-vowel sequences. The last is revealed, for example, by certain children's preference for high front vowels after alveolar closures, for low back vowels after the relatively extensive jaw-lowering release of labial closures (Davis & MacNeilage, 1990; cf. Jakobson, 1941/1968:29, 50).

Integration of prebabbling gestures into the canonical syllable is a necessary condition of a second step: differentiation of the syllable into independent gestural components. Differentiation gives rise to what Oller (1980) terms "variegated babble" in which the consonant-like syllable onset and the vowel-like nucleus, or both, differ in successive syllables. The process may begin soon after, or even at the same time as, the onset of canonical babble, but typically comes to predominate in the fourth quarter of the first year, and continues over many months in both babble and early words (Davis & MacNeilage, 1990; Vihman, Macken, Miller, Simmons, & Miller, 1985).

Before we consider the types of error that a child may make we should note that, although we shall appeal to similarity among gestures as the basis of a child's confusions, we will not spell out the dimensions of similarity. In fact, we shall deliberately avoid the question of whether those confusions reflect an incomplete percept (under which we may include incomplete storage of the percept in memory) or inadequate articulatory control (under which we may include failure to recover stored motor commands from memory). Perhaps, indeed, there is no general answer to this question: similar errors may reflect different processes in different words and in different children. Here we adopt a neutral stance. We suppose that learning the phonology of a language is a matter both of learning to perceive the acoustic pattern that specifies a talker's gestures, and of learning to plan and produce that pattern oneself. Both these processes can be a source of error.

Differentiation itself has two aspects, each open to characteristic forms of error. First is paradigmatic differentiation among individual gestures. Possible errors here follow from failure to identify or execute the location, shape or degree of a gesture; in the limit, an error of degree (or amplitude) may yield complete omission. The second aspect is syntagmatic differentiation among gestures in a particular utterance. Possible errors here include gestural reduplication (harmony), errors of timing (duration and relative phasing, including metathesis), and errors of amplitude or degree. The consonant-vowel (or vowel-consonant) harmony noted above may be viewed as a syntagmatic error arising from incomplete differentiation of the syllable into its component consonantal and vocalic gestures. Our purpose in what follows is to illustrate how the erroneous forms of a child's early words can be perspicuously described as arising from gestural errors such as these.

Method

The subject "Emma" is a second child, born in Connecticut to parents who had moved there from Vancouver, British Columbia. Emma's mother was the full time caregiver for the child. The second author lived with the family before and during the study and spent several hours a day observing Emma at meals, watching her play with her older brother and interact with her parents, and occasionally looking at picture books with her. One of these books (*Richard Scarry's Best Word Book Ever*) was a rich source of new words.

Audio recordings began when Emma had a vocabulary of about 100 words, mostly monosyllables understood primarily by her mother and brother. The size of her vocabulary was assessed with the MacArthur Communicative Development Inventory for Toddlers, and by maternal report. In the ninety-first week her mean length of utterance (MLU) was 1.00 and at the end of the study (week 106), 1.15. For the weekly audio taping sessions, lasting from 30 minutes to an hour, she wore a wireless 831 Audio

Technical lapel microphone concealed in a vest. E.W.G. was present at all sessions and kept a diary of the subject's phonological development to supplement the recordings.

To facilitate the transcription and analysis of Emma's utterances, recordings of the sessions were digitized on a VAX 780 computer, at a 20 kHz sampling rate, to yield a total of some 950 utterances of which the experimenter and a colleague independently transcribed roughly 250. Transcription followed the principles of the International Phonetic Association (1989), with some elaborations according to the Stanford system for transcribing consonants in child language (Bush, Edwards, Edwards, Luckau, Macken, & Peterson, 1973). Each utterance was coded as either spontaneous or imitated; an utterance was assumed to be spontaneous unless it immediately followed the adult target; all the examples reported in this paper were spontaneous, unless otherwise indicated. We report only utterances on which the two transcribers independently agreed. The transcriptions will be given in square brackets, following the convention for adult phonetic segments. We emphasize that our use of phonetic symbols does not imply that segments were already established in the child as discrete units of perception and production. A phonetic symbol is simply a convenient shorthand for combinations of laryngeal, oral or velic gestures.

Results and Discussion

Many researchers have described how a child, making the transition from reduplicated babble to variegated speech, discovers a pattern that roughly fits a fair number of adult words, and so can serve as a bridge into the lexicon. These patterns, variously termed prosodic schemata (Waterson, 1971), canonical forms (Ingram, 1974), articulatory routines, programs, templates (Menn, 1978, 1983), word patterns (Macken, 1979) or vocal motor schemes (McCune & Vihman, 1987) will be treated here as routinized gestural scores. (Much of our analysis will indeed follow the lead of Waterson (1971), Macken (1979), and, particularly, Menn (1978, 1983) whose attention to the articulatory organization of a child's utterances anticipates our own.) Gestural routines support both stereotypy (including homonymy) and variability in a child's early attempts at words; they are of interest because the gestural properties common to a particular score and to the different target words for which it is used reveal the scope of a child's gestural conflations.

Stereotypy

Gestural routines in babble and word play. During the sessions themselves Emma did not babble much, but diary entries from the first month of the study (weeks 91-94) often record Emma's babbles while quietly playing. For example:

(1) [ɑˈbiːnˈɑˈbiːnˈɑˈbiːn]
(2) [ˈbeːdəˈbeːdəˈbeːdəˈbeːdə]

These utterances happen to consist of repetitions of one of Emma's forms for *elephant* ([ˈɑˈbin]) and *playdough* ([ˈbeːdə]), but she chanted the sequences in a sing-song, with no apparent communicative intent. Both utterances contain the alternating sequence of constrictions at the lips and at the alveolar arch that proved to be Emma's most productive gestural routine. (For an example of another child, learning Mexican Spanish, with a similar routine, see Macken, 1979).

Strings of similar alternations occasionally occurred in taping sessions over the same period:

(3) [ˈmʌtʃːˈmʌtʃːˈmʌtʃː]
(4) [ˈɑˈbuːdiːˈɑˈbuːdiːˈɑˈbuːkuːkiː]
(5) [ˈweːdɑˈwiːdɑˈmeːnɑˈmiːnəˈmuːniːˈmiːniːˈmiːniː]

Emma repeatedly produced (3) in weeks 92 and 93, elongating the final frication, as though savoring the flow of air, and with no apparent referent. (4) contains two of Emma's words (see Table 1 for [ˈɑˈbuːdiː]), while (5) is a mixture of apparent nonsense syllables and word forms (see Table 2), but none of the objects to which the words refer was present. In (4) she abruptly broke off her labial-alveolar chant when a cookie came into view. In (5) she seemed to be playing with the location and degree of labial constriction ([w]-[m]), the degree of accompanying velic constriction ([w]-[m], [d]-[n]), and (in an apparent instance of vowel-consonant harmony) the front-back location of narrow constrictions at the syllable nucleus before an alveolar constriction ([iː]-[eː]-[uː]).

Table 1. *Active use of the labial-alveolar routine as a bridge into the lexicon.*

New Words	*Adult Target*	*Emma's Attempts*
Cranberry	[ˈkrænbɛri]	[ˈbeːbi] [ˈboːbeːbi] [ˈɑˈbuːdiː]
Red Lights	[ˈrɛdˈlaɪts]	[ˈweːjaɪ] [ˈbɛˈtθaɪts]
Hippopotamus	[ˈhɪpəˈpatəmʌs]	[ˈɑˈpɪnz] [ˈhɪpas]

Table 2. *Words attempted by means of the labial-alveolar routine in weeks 91-94.*

	Emma's attempts	*Adult targets*
*	['buːˈdiː]	berry, bird, booster
*	['beːˈdə]	pillow, playdough
	['beːˈdiː]	umbrella
	['peːˈdə]	peanut
	['pəˈtə]	puppet
	['meːˈnə]	tomato
	['meːˈniː]	medicine
	['muːˈniː]	money
	['weːˈdə]	playdough
	['weːˈdiː]	raisin
*	['ɑˈmiːn]	elephant, airplane
	['ɑˈbiːn]	elephant
	['ɑˈpiːn]	airplane
*	['ɑˈbuːˈdiː]	Happy Birthday, cranberry, raspberry
*	Homonyms	

These examples illustrate the emergence of gestural stereotypy in babble and word play. They also illustrate the familiar, but important fact that a listener often cannot distinguish, by phonetic form alone, between syllables that are babbled and syllables that should count as a word.[4] Despite the discontinuity of function that Jakobson (1941/1968) noted many years ago, babble and speech are formally continuous.[5] An adequate account of the shift in function must therefore posit units of action that can be comfortably engaged by both babble and speech. Phonemes, phonetic segments and features are unsuited to this task: they cannot properly be adopted for prelinguistic babble because they are defined in terms of language and speech. Moreover, as already noted, segments are complex units, customarily analyzed into their featural predicates, while features have no existence independently of the syllables and segments they describe. By contrast, the posited gesture is an integral unit of action that can serve equally as a primitive unit of both babble and speech.

Gestural routines as bridges into the lexicon. A child who has discovered a gestural routine, such as the labial-alveolar sequence described above, will often extend it to a surprisingly diverse collection of new words in which she recognizes the appropriate pattern (Macken, 1979)—in Emma's case words as diverse as *cranberry*, *red lights*, and *hippopotamus* (Table 1). Thus in the recording session of week 92, Emma's mother showed her a cranberry for the first time, repeated the word and asked her to say it. First Emma attempted the word with gestural harmony, repeating the labial closure of the second syllable, ['beːˈbiː], then she perfected the number of syllables, ['boːˈbeːˈbiː], finally she reverted to a three-syllable labial-alveolar routine, ['ɑˈbuːˈdiː], transposing the postalveolar retroflex constriction of [r] into the alveolar closure of [d]. She used this form for *cranberry* for approximately the next two months. (We may note, incidentally, that Emma here adopted a tactic that recurred in her attempts at several other words, usually words of three or more syllables: she lowered her jaw and substituted the wide vocalic gesture of [ɑ] as a sort of place-holder for the initial syllable or syllables.)

In week 95, hearing her mother point out the red lights on the tape recorder, Emma spontaneously attempted *red lights* as ['weːjɑɪ], and seconds later, without correction, as ['bet'θɑɪts]. Here, for [r], she first picked up the narrow constriction at the protruded lips, but omitted the accompanying postalveolar retroflexion, giving [w], then fell back on full labial closure, giving [b]. The alveolar closure for [d] she omitted on the first attempt, while successfully executing the nearby palatal glide of [j] in place of [l]—a common shift in the exact location and shape of the gesture for [l] in early speech (e.g., Vihman & Velleman, 1989). On the second attempt she achieved full alveolar closure, but anticipated the glottal opening and critical fricative constriction of final [ts], giving the sequence [tθ] instead of [dl].

A final, more complicated example occurred in week 101. Seeing a familiar picture of a hippopotamus, Emma spontaneously pronounced ['ɑpɪnz]. Here she substituted her favored wide vocalic constriction for the first one or two syllables. The remaining three or four syllables she collapsed into one, built around her labial-alveolar routine. For this she correctly executed the labial closure and glottal opening of [p], as well as the alveolar constrictions of medial [t] and final [s]. But she omitted the labial closure of [m] and the glottal openings of [t] and [s]; she roughly harmonized the syllable nucleus to the following alveolar closure; and she erroneously synchronized alveolar closure for [t] with velic opening for [m]. The outcome of these maneuvers was [pɪnz], a syllable composed of four apparent segments, three of which do not occur in the target word—a result difficult to understand if we assume segmental primitives, but readily intelligible in gestural terms.

To Emma's spontaneous ['ɑpɪnz] E.W.G. replied: "Oh, hippopotamus," eliciting a form that Emma had used on previous occasions: ['hɪpɑs], repeated four times. Here, with the model freshly in mind, Emma recaptured the first syllable, but omitted the second, as well as the medial alveolar and velic gestures of the final three syllables which she collapsed into the bare routine of initial labial and final alveolar gestures.

These three examples illustrate Emma's active use of a routinized gestural score as an armature, or skeleton, around which to construct her articulation of words presumably otherwise too difficult, whether perceptually or motorically, to attempt. We have characterized the routine in terms of rough gestural location, disregarding differences in precise location and in degree or shape. Thus, we have treated [b/p/m] and [w] as equivalently labial in Emma's utterances, [d/t/n], [j], [r], and [s] as equivalently alveolar. These equivalences are justified by Emma's gestural alternations both in the examples above and in her other uses of the labial-alveolar routine to which we now turn.

A gestural routine as a source of homonyms. Table 2 lists the entire set of recorded words to which Emma applied the labial-alveolar routine. Some of these were in Emma's repertoire before the study began (according to maternal report) and, with the exception of ['ɑmin], *elephant*, all were recorded during the first month of the study. We have grouped them according to the similarity of their phonetic patterns, making clear that in addition to the actual sets of homonyms, marked with asterisks, there are several sets of near-homonyms (Emma's forms for *pillow* and *umbrella, tomato* and *medicine, playdough* and *raisin*, where each member of a pair differs from the other only in its final vocalic gesture). These homonymous groups, clearly not semantically based, validate the proposed routine as a functional process in Emma's attack on the lexicon, by drawing attention to gestural similarities among target words that, at first glance, are quite dissimilar (cf. Vihman, 1981). Thus, we find alveolar [d] for [r] in *berry*, for [st] in *booster*, for [l] in *pillow* and *umbrella*, for [s] in *raisin*. At the same time, the labial grouping is justified by Emma's own use of [b] and [w] for [pl] in *playdough*, of [p] and [m] for [pl] in *airplane*, of [w] for initial [r] in *raisin*, of [m] and [b] for [f] in *elephant*.

Several of these substitutions can, of course, be interpreted in featural terms. However, substitution of the narrow labial constriction of [w] for [r] in initial position, but of full alveolar closure for [r] in medial position, would not be expected on a featural account, since a given segment carries the same featural predicates regardless of context, and so should be subject to the same perceptual or motoric confusions.[6] A gestural account, on the other hand, predicts such syntagmatic errors precisely because it views the task of learning to talk as quite largely one of learning to coordinate gestures that may differ in their articulatory compatibility (cf. Menn, 1983). We shall see further examples of contextual effects in our discussion of variability.

Finally, we must remark another process, difficult for a featural account, and important to our later discussion: the tendency for gestures to "slide" along the time line (Browman & Goldstein, 1987) into misalignment with other gestures, often giving rise to apparent segments not present in the target word. We have already noted this process in Emma's ['ɑpmz] for hippopotamus. Here (Table 2) we find it in ['meː'nə], *tomato*, where velic lowering for [m] extends into the alveolar closure for [t], yielding [n]; and in ['ɑmiːn], *elephant, airplane*, where velic lowering for [n] slides into alignment with labial closures for [f] or [pl], yielding [m]. In these examples the effect is of gestural harmony, and so may be due not only to an error of gestural phasing, but also to "...the difficulty in planning and production of rapid changes of articulation in a short space of time" (Waterson, 1972:13, cited by Menn, 1983:30). Of course, this too is a form of timing error.

Variability

Spontaneous variations. While a gestural routine may afford a child initial access to difficult words of similar gestural pattern, it cannot solve all the problems of gestural selection and phasing with which the child must contend in moving toward an acceptable pronunciation. Variability within the constraints of the routine is an important part of this process.

For example, Emma's attempts at *elephant* in a single session in week 91 included: ['ɑbin], ['ɑm'bin], ['ɑmin], ['ɑfin], and ['ɑpin], all of which are formed by combining her initial vocalic placeholder and her labial-alveolar constriction routine with her favored medial vowel-consonant harmony. Yet within these limits she seemed to be trying to hit upon the exact location of the gesture for labial [f], and the relative phasing of glottal opening for [f] and velic lowering for [n]. She experimented with the timing of velic action again in her forms for *raisin*: ['weː'ni], ['weːn'di], ['weːdi]. And in ['beːdə], ['weːdə], *playdough*, she seemed to be trying to simulate the labial alveolar sequence in the cluster [pl] by playing with the exact location and degree of labial constriction.

Further examples of variability within the constraints of a stereotyped routine come from Emma's attempts to execute the syllable [ˈnʌt] in the words *doughnut* and *peanut*. We might have expected these words to be relatively easy, the first because it calls for three harmonious alveolar gestures, the second because it fits Emma's labial-alveolar routine, already well established when she met the word. But in fact they proved to be quite difficult, both overall and in their identical final syllable in particular. This syllable elicited very different patterns in the two contexts—a type of result, as we have already remarked, readily intelligible on a gestural, but not on a featural account of her errors.

Doughnut was introduced in week 92. Emma's first attempts were [ˈduːˈdʌtʃ] and [ˈdoːˈdʌts]. The final critical alve..ar constrictions added apparent segments not present in the model. They were not attempts at the plural, because she was given only part of a doughnut to eat and only heard the word in the singular. Rather, they seem to have resulted from a relatively slow release of [t], making the fricative portion of the release (Fant, 1973:112) more salient. Steriade (1989) offers a similar analysis for derived affricates, proposing that they "...differ from stops in the quality of their release." Over the next 10 weeks Emma's attempts at this word varied over forms as diverse as [ˈduːdə] and [ˈduːnˈdʌnt]. The latter seems to result from prolongation of the alveolar closure for medial [n] after velic release, giving an unwanted [d], combined with prolongation of the alveolar closure for final [t] and a shift in (or harmonious repetition of) the medial velic gesture, giving the unwanted final cluster. Figure 2 (top) displays a schematic gestural score illustrating the timing errors required to make the shift from [ˈnʌt] to [ˈdʌnt], and Table 3 lists in chronological order some of the variations on *nut* in *doughnut* for comparison with those elicited by *peanut*.

Emma encountered a peanut in a picture book in week 94. Drawing appropriately on her labial-alveolar routine, she first tried [ˈpeːdə], omitting velic action, and later that week, [ˈpeːnˈtə], where prolongation of the medial alveolar closure, combined with a shift in the timing of the final glottai opening, relative to velic closure and the tongue body gesture, gives rise to an apparent shift in the ordering of the target consonant-vowel-consonant sequence—a result difficult to explain in either segmental or featural terms. In week 96 she offered [ˈpiːˈpʌp], omitting the velic gesture and succumbing to labial harmony, and [ˈpeːmˈpʌmp]. The latter, formally analogous to [ˈduːnˈdʌnt], *doughnut*, with its velic harmony, mistimed velic action and resulting unwanted segments, is further complicated by the substitution of harmonized labial closures for the alveolar closures called for by the target, and proper to her routine. Figure 2 (bottom) illustrates the errors of gestural location and timing required to make the shift from [ˈnʌt] to [ˈpʌmp].

Table 3. *Spontaneous variability within and between words: The same target syllable executed d'':rently in different phonetic contexts. The utterances are listed chronologically, but the columns for* ***doughnut*** *and* ***peanut*** *are not synchronized.*

Nut as in *doughnut* and *peanut*

doughnut [ˈdoːnʌt] ----→	*nut*	*nut*	←--- *peanut* [ˈpiːnʌt]
[ˈduːdə]	də	də	[ˈpeːdə]
[ˈduːnˈdʌnt]	dʌnt	tə	[ˈpeːnˈtə]
[ˈdoːdiːdʌt]	dʌt	de	[ˈpeːdeː]
[ˈduːdʌtʃ]	dʌtʃ	pʌmp	[ˈpeːmˈpʌmp]
[ˈduːdʌts]	dʌts	pʌp	[ˈpiːpʌp]
[ˈdoːnʌt]	nʌt	nʌt	[ˈpiːnʌt]

Other examples of Emma's errors, evidently due to a variety of gestural processes, including harmony and the slow release of alveolar closures, include: [ˈduːˈdətʃiz], [ˈdoːnʌˈtʃfiz], *doughnut please*, and [ˈsɛlzɛˈtʃiz], [ˈsɛpəˈpiz], *seltzer please*. Since isolated forms for both *seltzer* and *please* occurred in Emma's repertoire, the last example nicely illustrates a child organizing her articulations over a phrase of several syllables (cf. the "coalesced word patterns" of Macken, 1979).

Variability in imitations. As a final example, let us consider six of Emma's repeated attempts to imitate a word that did not lend itself either to gestural harmony or to the labial-alveolar routine: *apricot*, [ˈæprɪˈkɑt]. (All these examples were recorded in week 95, except for the third which was recorded in week 105.) The word is challenging because it calls not only for three different locations of gestural closure, irregularly ordered (labial, velar, alveolar), but also for an alternating pattern of glottal closure and release.

Table 4 lists Emma's attempts. With several exceptions she captures certain properties of the word quite accurately: the number of syllables (2-6), the stress pattern (1,4), an initial vocalic gesture (1-4), the constriction degree of the final vowel (1,3-6), the rough location of at least two out of three consonantal constrictions (2-6) and, omitting the initial velar intrusions of 5 and 6, the labial-lingual sequence of these constrictions.

Table 4. *Imitation: Within-word variability in Emma's attempts at* ***apricot.***

Adult target	*Order of closed constrictions in target*	*Emma's imitations*	*Order of closed constrictions in imitations*
['æprɪkɑt]	Labial---Velar---Alveolar	1. ['aɪbə^{w}ɑ^{h}ɑː]	L
		2. ['ɑ̤pə'gʌ]	L---V
		3. ['ə'fu'kɑː]	L---V
		4. [^{h}ʌfə'tsɑː]	L---A
		5. ['gɛɾgʌ'pɑː]	V---V---L
		6. ['ŋə'··ɑp^{w}ɚɾ^{h}ɑː]	V---L---A

Apparent consonantal segments in Emma's responses not found in adult target: [b], [g], [f], [ts], [ŋ]

Yet every attempt contains at least one apparent segment not present in the model: [b], [g], [f], [s] or [ŋ]. With the exception of [f] (an error in the exact location and degree of the word's labial constriction), all these errors arise from a failure of gestural timing or coordination: for [b], [g] and [ŋ], a failure to open the glottis during oral closure; for the affricate [ts] (in 4), a relatively slow release of [t], as in the examples above. Other indications that Emma had difficulty in managing the alternating pattern of glottal action in the word come from the brief periods of aspiration (superscript [h]) inserted in 1, 4 and 6. Finally, the whispered initial vowel of 2 presumably reflects a delay in glottal closure, while the initial velar nasal of 6 reflects a delay in velic closure, as the child moves from silent breathing to speech. Thus, the principal source of error (apart from errors in the location and degree of vocalic constrictions) was gestural phasing. No doubt we could construct a set of "rules" relating the observed segments to the supposed underlying forms of the target utterance. But the task would be laborious, and completely *ad hoc*. A gestural account, by contrast, is simple and readily intelligible.

GENERAL DISCUSSION

The relation of gestures to features and segments

We have tried to show how a child's errors in early words can arise from paradigmatic confusions among similar gestures in a child's repertoire and from syntagmatic difficulties in coordinating the gestures that form a particular word. Yet a reader accustomed to think in terms of features and segments may see little difference between our approach and those of previous researchers. For example, Waterson (1971:181) proposes that "...a child perceives only certain of the features of the adult utterances and reproduces only those that he is able to cope with"; Macken (1979:29) writes of a child's "...tendency to combine features from different segments of the adult word"; Ferguson and Farwell (1975:426), commenting on a child's diverse forms for a single word, suggest that she "...seems to be trying to sort out the features of nasality, bilabial closure, alveolar closure and voicelessness."

What is missing in all these formulations is an explicit statement of how a percept is linked to its articulation. Their implicit conception of the link seems to be close to that of K. N. Stevens who answered a conference questic [illegible] is matter as follows: "I would say th [illegible] e lexicon is represented in abstract units that are neither directly articulatory nor directly acoustic. A relation projects these abstract units both to the acoustics and to the articulation. As you can see, I am taking the view of Jakobson, originally postulating something like features which have both acoustic correlates and articulatory correlates and must have both" (Mattingly & Studdert-Kennedy, 1991, p.194).

We have already stated the key objection to this position: a feature is a property, not an entity. Phonetic features are not like facial features—eyes, nose, mouth—each of which can, at least in principle, be removed from one face and transferred to another. Rather, phonetic features are like the size and shape of a nose: we cannot remove either without removing the nose in which they are embodied. In short, features are attributes, not substantive components.

Of what substantive object or event, then, is the feature an attribute? The customary answer, the phonetic segment or phoneme, will not do, because segments are defined by their features: the answer is circular as long as we have no independent (and no substantive) definition of a segment We propose, instead, that a feature is an attribute of a gesture. We assume that gestures, like Jakobson's features, "...have both acoustic correlates and articulatory correlates and must have both."

Because these two sets of correlates are necessarily isomorphic, the gesture is the link between a speech percept and its articulation. In this respect, speech gestures resemble every other imitable act: their perceptual representation specifies their motor form.

Adopting the gesture as a vehicle for the feature also permits an independent and substantive definition of the segment. We noted earlier that the canonical syllable was the first step toward differentiation of two major classes of oral gesture: vocalic and consonantal. Let us now note, further, that although consonant and vowel gestures may interact (as in Emma's preferences for particular consonant-vowel combinations) they are not interchangeable: we do not find a child (or an adult) making the mistake of replacing a narrow/mid/wide vocalic gesture with a closed/critical consonantal gesture, or vice versa. No doubt such errors are blocked by the biophysical structure of the syllable, that is, by its alternating pattern of opening and closing the mouth. In any event, we view differentiation of the syllable into its closed and open phases as a move toward the formation of gestural routines with a narrower domain than the word, namely, the encapsulated patterns of precisely phased laryngeal, velic and oral gesture that we term segments (cf. Menn, 1986; Studdert-Kennedy, 1987).

The emergence of segments

The emergence of segments as elements of word formation in a child's lexicon seems, then, to have two aspects. First, is the grouping of all instances of a particular sound-gesture pattern into a single class, presumably on the basis of their perceptuomotor, or phonetic, similarity (e.g., grouping the initial or final patterns of *dad, dog, bed,* etc. into the class /d/). Second is the distributional analysis and grouping of these gesture-sound patterns into higher order classes (consonants, vowels) on the basis of their syllabic functions (onset, nucleus, coda).

Two possible selection pressures may precipitate formation of these categories. One pressure is toward economy of storage. As the lexicon increases, words may organize themselves according to their shared gestural and sound properties. Recurrent patterns of laryngeal and supralaryngeal gesture would thus form themselves into classes of potential utility for recognition or activation of lexical items (Lindblom, 1989; Lindblom, MacNeilage, & Studdert-Kennedy, 1983).

A second pressure may be toward rapid lexical access in the formation of multiword utterances. Several authors (e.g. Branigan, 1979; Donahue, 1986) have reported evidence that the form of early multiword combinations may be limited by the child's ability to organize the required articulatory sequences. Donahue, for example, reports her son's "adamant refusal" to attempt two successive words with different initial places of articulation. Such findings imply that the integration of gestures into independent phonemic control structures, or articulatory routines (Menn, 1983), may serve to insulate them from articulatory competition with incompatible gestures, and so facilitate their rapid, successive activation in multiword utterances.

SUMMARY

We have presented three lines of evidence for a gestural model of phonological development that can deal coherently with (i) the continuous transition from babbling to speech, and (ii) the word as the contrastive unit of early phonology. For the transition from babbling to speech, details of the developmental course may vary from child to child: not every child displays gestural harmony, not every child who does so escapes from harmony into the lexicon by a non-harmonious gestural routine. Nonetheless, every child does have to find a path from babbling to speech. We have argued from one child's path that the gesture, with its roots in the child's prelinguistic mouthings and vocalizations, is a more valid unit of linguistic function than the feature with its roots in the formalisms of adult phonology.

With regard to the word as the contrastive unit of early phonology, we have reviewed two lines of evidence that the gesture, rather than the feature, is the basic unit of a word's articulatory organization. First, the same syllable may take different forms as a function of the target word, or phonetic context, in which it appears. A featural account would not predict this outcome, because a given segment carries the same featural predicates regardless of context; a gestural account, on the other hand, with its emphasis on the syntagmatic processes of articulatory action, finds the outcome natural. Second, in our subject's attempts to articulate a word with a pattern of alternating glottal gestures and a varied sequence of oral constrictions, the attempts were so diverse, so variable from occasion to occasion, that a featural account of her utterances would be little more than a list of arbitrary deletions, additions, and substitutions. By contrast, the present approach attributing the child's errors to imprecise execution and timing of the gestures that form the target word, offers a simple and perspicuous account.

Finally, we have argued that the feature can be ruled out as a basic unit of either speech perception or speech production on rational grounds, because it is, by definition, an attribute that has no existence independently of the object or event that it describes. We reject the segment as the primary vehicle of the feature because a segment is (circularly) defined by its features. We propose instead that a feature be viewed as an attribute of a gesture, and that segments be defined, superordinate to the gesture, as emergent structures, comprising recurrent, spatiotemporally coordinated, gestural routines.

REFERENCES

Albright, R. W., & Albright, J. B. (1956). The phonology of a two-year-old child. *Word 12*, 282-390.

Albright, R. W., & Albright, J. B. (1958). Application of descriptive linguistics to child language. *Journal of Speech and Hearing Research, 1*, 257-61.

Branigan, G. (1979). Some reasons why successive single word utterances are not. *Journal of Child Language, 6*, 411-421.

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook , 3*, 219-252.

Browman, C. P., & Goldstein, L. (1987). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology, I* (pp. 341-376). New York: Cambridge University Press.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*, 151-201.

Browman, C. P., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics, 18*, 299-320.

Bush, C. N., Edwards, M. L., Edwards, J. M., Luckau, C. M., Macken, M. A., & Peterson, J. D. (1973). On specifying a system for transcribing consonants in child language. Stanford Child Language Project, Department of Linguistics. Stanford University, Stanford, CA.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook, 2*, 225-252.

Cohen, M. (1952). Sur l'etude du language enfantin. *Enfance, 5*, 181-249.

Davis, B. L., & MacNeilage, P. F. (1990). Acquisition of correct vowel production: A quantitative study. *Journal of Speech and Hearing Research, 33*, 16-27.

Donahue, M. (1986). Phonological constraints on the emergence of two word utterances. *Journal of Child Language, 13*, 209-218.

Fant, G. (1962). Descriptive analysis of the acoustic aspects of speech. *Logos, 5*, 3-17.

Fant, G. (1973). *Speech sounds and features*. Cambridge: M.I.T. Press.

Ferguson, C. A. (1963). Contrastive analysis and language development. *Georgetown University Monograph Series, 21*, 101-112.

Ferguson, C. A. (1986). Discovering sound units and constructing sound systems: It's child's play. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 36-51). Hillsdale, NJ: Erlbaum.

Ferguson, C. A , & Farwell, C. B. (1975). Words and sounds in early language acquisition. *Language, 51*, 419-439.

Fowler, C. A., Rubin, P. E., Remez, R. & Turvey, M. (1980). Implications for speech production of a general theory of action. In B. Butterworth (Ed.), *Language production* (pp. 373-420). New York: Academic Press.

Goldsmith, J. A. (1990). Autosegmental and metrical phonology. Oxford: Basil Blackwell.

Goodell, E. W., & Studdert-Kennedy, M. (1991). Articulatory organization of early words: From syllable to phoneme. In *Proceedings of the XIIth International Congress of Phonetic Sciences* (pp. 166-169). Aix-en-Provence, France: Université de Provence.

Goodell, E. W., & Studdert-Kennedy, M. (in press). Acoustic evidence for the development of gestural coordination in the speech of 2-year-olds: A longitudinal study. *Journal of Speech and Hearing Research.*

Holmgren, K., Lindblom, B., Aurelius, G., Jalling, B., & Zetterström, R. (1986). On the phonetics of infant vocalization. In B. Lindblom & R. Zetterström (Eds.), *Precursors of early speech* (pp. 51-63). New York: Stockton.

Ingram, D. (1974). Fronting in child phonology. *Journal of Child Language, 1*, 233-241.

International Phonetic Association (1989). Report on the 1989 Kiel Convention. *Journal of the International Phonetic Association, 19*, 67-80.

Jakobson, R. (1941/1968). Child language, aphasia and phonological universals. [Trans. of Kindersprache, Aphasie und allgemeine Lautgesetze. Uppsala: Almqvist & Wiksell, 1941]. The Hague: Mouton.

Jakobson, R., Fant, G., & Halle, M. (1951/1963). *Preliminaries to speech analysis:* The distinctive features and their correlates. Cambridge, MA: MIT.

Jakobson, R., & Halle, M. (1956). *Fundamentals of language.* The Hague: Mouton.

Kent, R. D., & Bauer, H. R. (1985). Vocalizations of one year olds. *Journal of Child Language, 12*, 491-526.

Koopmans, van B., Florien. J., & van der Stelt, J. M. (1986). Early stages in the development of speech movements. In B. Lindblom & R. Zetterström (Eds.), *Precursors of early speech* (pp. 37-50). New York: Stockton.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language, 56*, 485-502.

Leopold, W. F. (1953). Patterning in children's language learning. *Language Learning, 5*, 1-14.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. J. Ohala & J. J. Jaeger (Eds.), *Experimental Phonology* (pp. 13-44). New York: Academic Press.

Lindblom, B. (1989). Some remarks on the origin of the "phonetic code." In C. von Euler, I. Lundberg, & G. Lennerstrand (Eds.), *Brain and reading* (pp. 27-44). Basingstoke, England: MacMillan.

Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (1983). Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie, & Ö. Dahl (Eds.), *Explanations of linguistic universals* (pp. 181-203). The Hague: Mouton.

MacArthur Communicative Development Inventory: Toddlers. 1989. Center for Research in Language, UCSD C-008, San Diego, CA 92093.

Macken, M. A. (1979). Developmental reorganization of phonology: A hierarchy of basic units of acquisition. *Lingua, 49*, 11-49.

MacNeilage, P. F., & Davis, B. L. (in press). Acquisition of speech production: Frames, then content. In M. Jeannerod (Ed.), *Attention and performance XIII: Motor representation and control.* Hillsdale, NJ: Erlbaum.

MacNeilage, P. F., Hutchinson, J, & Lasater, S. (1981). The production of speech: Development and dissolution of motoric and premotoric processes. In J. Long & A. Baddeley (Eds.), *Attention and Performance IX* (pp. 503-520). Hillsdale, NJ: Erlbaum.

Mattingly, I. G., & Studdert-Kennedy, M. (1991) (Eds.). *Modularity and the Motor Theory of Speech Perception*. Hillsdale, NJ: Erlbaum.

McCarthy, J. J. (1988). Feature geometry and dependency: A review. *Phonetica, 43,* 84-108.

McCune, L., & Vihman, M. M. (1987). Vocal motor schemes. *Papers and Reports in Child Language Development, 26.*

Menn, L. (1978). Phonological units in beginning speech. In A. Bell & J. Hooper, (Eds.) *Syllables and segments* (pp. 157-171). Amsterdam: North Holland.

Menn, L. (1983). Development of articulatory, phonetic and phonological capabilities. In B. Butterworth, (Ed.), *Language production* (pp. 3-50). London: Academic Press.

Menn, L. (1986). Language acquisition, aphasia and phonotactic universals. In F. R. Eckman, E. A. Moravcsik, & J. R. Wirth (Eds.), *Markedness* (pp. 241-255). New York: Plenum Press.

Menyuk, P., Menn, L., & Silber, R. (1986). Early strategies for the perception and production of words and sounds. In P. Fletcher & M. Garman (Eds.), *Language acquisition* (pp. 198-222) (2nd ed). New York: Cambridge University Press.

Nittrouer, S., Studdert-Kennedy, M., & McGowan., R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research, 32,* 120-132.

Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology, Vol. 1: Production* (pp. 93-112). New York: Academic Press.

Oller, D. K. (1986). Metaphonology and infant vocalizations. In B. Lindblom & R. Zetterström (Eds.), *Precursors of early speech* (pp. 21-35). New York: Stockton.

Oller, D. K., Wieman, L. A., Doyle, W., & Ross, C. (1975). Infant babbling and speech. *Journal of Child Language, 3,* 1-11.

Saltzman, E. (1986). Task dynamic coordination of the speech articulators: A preliminary model. Generation and modulation of action patterns. In H. Heuer & C. Fromm (Eds.), *Experimental Brain Research, Series 15* (pp. 129-144). New York: Springer-Verlag.

Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1,* 333-382.

Scarry, R, (1980). *Richard Scarry's Best Word Book Ever.* New York: Western Publishing Company.

Schwartz, R. G. (1988). Phonological factors in early lexical acquisition. In M. D. Smith & J. L. Locke (Eds.), *The emergent lexicon,* (pp. 185-222). New York: Academic Press.

Stark, R. E. (1986). Prespeech segmental feature development. In P. Fletcher & M. Garman (Eds.), *Language acquisition* (pp. 149-173) (2nd ed). New York: Cambridge University Press.

Steriade, D. (1989) Affricates. Paper read at Conference on Feature and Underspecification Theories, Massachusetts Institute of Technology, October 7th-9th.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David & P. B. Denes (Eds.), *Human communication: A unified view,* (pp. 51-66). New York: McGraw-Hill.

Stevens, K. N. (1975). The potential role of property detectors in the perception of consonants. In G. Fant & M. A. A. Tatum (Eds.), *Auditory analysis and perception of Speech* (pp. 303-330). New York: Academic Press.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17,* 3-45.

Studdert-Kennedy, M. (1987). The phoneme as a perceptuomotor structure. In A. Allport, D. MacKay, W. Prinz, & E. Scheerer, (Eds.), *Language perception and production* (pp. 67-83). London: Academic Press.

Studdert-Kennedy, M. (1991a.) A note on linguistic nativism. In R. R. Hoffman & D. Palermo, (Eds.) *Cognition and the symbolic processes* (pp. 39-58). Hillsdale, NJ: Erlbaum.

Studdert-Kennedy, M. (1991b). Language development from an evolutionary perspective. In N. A. Krasnegor, D. M. Rumbaugh, R. Schiefelbusch, & M. Studdert-Kennedy, (Eds.), *Biological and behavioral determinants of language development* (pp. 5-28). Hillsdale, NJ: Erlbaum.

Vihman, M. M. (1978). Consonant Harmony: Its Scope and Function in Child Language. In J. Greenberg (Ed.), *Universals of human language, Volume 2: Phonology* (pp. 282-334). Stanford: Stanford University Press.

Vihman, M. M. (1981). Phonology and the development of the lexicon: Evidence from children's errors. *Journal of Child Language, 8,* 239-264.

Vihman, M. M, Macken, M., Miller, R., Simmons, H., & Miller, J. (1985). From babbling to speech: A re-assessment of the continuity issue. *Language, 61,* 397-445.

Vihman, M. M., & McCune, L. (in press). When is a word a word? *Journal of Child Language.*

Vihman, M. M., & Velleman, S. (1989). Phonological reorganization: A case study. *Language and Speech, 32,* 149-170.

Waterson, N. (1971). Child Phonology: A prosodic view. *Journal of Linguistics, 7,* 179-211.

FOOTNOTES

*In B. de Gelder & J. Morais (Eds.), *Language and literacy: Comparative approaches.* Cambridge, MA: MIT Press (in press).

[1] Fant (1962, p. 4) remarked many years ago concerning the theory of Jakobson, Fant and Halle (1952/1963): "...its formulations are made for the benefit of linguistic theory rather than for engineering or phonetic applications. Statements of the acoustic correlates to distinctive features have been condensed to an extent where they retain merely a generalized abstraction insufficient as a basis for practical applications." The same is true of subsequent attempts to formulate acoustic and articulatory correlates of the features.

[2] In what follows we use the term "gesture" to refer either to an underlying abstract control structure, or to a concrete instance of a gesture activated by this structure, relying on context to make clear which is intended.

[3] These categorical values, axiomatic within gestural phonology, may have emerged evolutionarily, and may still emerge ontogenetically, through auditory and articulatory constraints on individual gestures (Stevens, 1989), and on the entire set of gestures within the child's developing lexicon (Lindblom, 1986; Lindblom et al., 1983).

[4] The problem and criteria for its solution are thoroughly discussed by Vihman and McCune (in press).

[5] Much of the controversy over the issue of continuity between babble and speech has arisen, in our view, from a misreading of Jakobson's claims, and from a failure to distinguish between phonetic form and phonetic function. Jakobson himself drew this distinction quite clearly. Although he believed that "...a short period may sometimes intervene...in which children are completely mute," he also recognized that: "For the most part...one stage merges unobtrusively into the other, so that the acquisition of vocabulary and the disappearance of the prelanguage inventory occur concurrently" (Jakobson, 1941/1968, p.29). In fact, he assumed what later studies have conclusively demonstrated (e.g., MacNeilage, Hutchinson, & Lasater, 1981; Oller, Wieman, Doyle, & Ross, 1976; Vihman, Macken, Miller, Simmons, & Miller, 1985) that listeners often cannot distinguish, by phonetic form alone, a "child's embryo-words from the pre-language residue" (Jakobson, 1941/1968, p. 29). The discontinuity that Jakobson (correctly) posited was a discontinuity of function, not of form.

[6] We thank Susan Brady for pointing this out to us.

Haskins Laboratories Status Report on Speech Research
1992, SR-111/112, 103-110

An Aerodynamic Evaluation of Parkinsonian Dysarthria: Laryngeal and Supralaryngeal Manifestations*

L. Carol Gracco,† Vincent L. Gracco, Anders Löfqvist, and Kenneth Marek‡

The speech of individuals with Parkinson's Disease (PD) is characterized by reduced stress, increased rate, monotonic pitch and loudness and imprecise consonant production (see Darley, Aronson, & Brown, 1975 for review). Acoustically, speech has decreased duration of voiced segments, reduced fundamental frequency variations and limited formant trajectories at consonant-vowel transitions as compared to normal age-matched controls (Canter, 1967; Darley et al., 1975; Forrest, Weismer, & Turner, 1989; Logeman, Fisher, & Boshes, 1978; Ludlow, Bassich, Connor, Coulter, & Lee, 1987; Ludlow & Schulz, 1989; Ramig, Scherer, Titze, & Ringel, 1988; Weismer, 1983). Although specific and detailed analysis of Parkinsonian deficits have been limited, existing studies of larygneal and supralaryngeal structures suggest that the reduction in intelligibility characteristic of Parkinson's disease may be a result of manifestations of this disorder throughout the entire vocal tract musculature. For example, studies of lip and jaw movements have reduced amplitude and velocity (Caligiuri, 1987; Connor, Abbs, Cole, & Gracco, 1989) while cinegraphic studies of laryngeal kinematics in PD (Hanson, Gerratt, & Ward, 1984) reveal a correlation between abnormalities in the phonatory posture of laryngeal structures and voicing deficits. These manifestations may involve the control and coordination of laryngeal and supralaryngeal events. Taken together, these factors result in the overall reduction in speech intelligibility in individuals with PD. Hence, the simultaneous evaluation of upper articulator and laryngeal dynamics may give a more complete analysis of deficit behaviors that ultimately influence intelligibility.

The perceptual significance of aerodynamic events and the utility of aerodynamic measures as a basis for understanding speech and voice articulation has been realized for some time.

With few exceptions, however, attempts to associate complex articulatory and phonatory changes with time varying changes in supraglottal air pressure and air flow has seen little progress. Much of the basic and applied literature investigating pressure and flow parameters has focused on differences in peak amplitude as a function of some variable such as vocal intensity, place, and manner of production, or has used peak amplitude to provide measures of glottal resistance. In isolation, peak measures are little more than a description of system output providing limited information regarding the underlying articulatory dynamics.

In a clinical setting, procedures which sample both temporal aspects of laryngeal and supralaryngeal dynamics as well as peak measures associated with various speech motor disorders are regarded as time consuming. This study attempts to provide the basis for a relatively easy and efficient evaluation of laryngeal and supralaryngeal articulation in individuals with Parkinsonian dysarthria. The methodology is based on previous work by Müller and Brown (1980) which illustrated the significance of assessing pressure and flow characteristics correlated with their time varying changes. Laryngeal factors are included in this analysis, reflecting the integration of vibratory characteristics with upper articulator dynamics.

Methods

Subjects. The salient speech characteristics and demographic data for five adult subjects are summarized in Tables 1 and 2 respectively.

This work was supported in part by NIH Grant No. DC-00044, DC-00594, DC-00865, and DC-00121 from the National Institute of Deafness and Other Communication Disorders.

Table 1. *Summary of Speech Characteristics for 5 subjects with Parkinsonian's Disease.*

Subjects	Speech Characteristics
JH	• minimal reduction in intelligibility • breathy vocal quality • accurate consonant production with infrequent alterations of speech rate
AD	• minimal reduction in intelligibility • breathy vocal quality • accurate consonant production with infrequent alterations of speech rate
HM	• minimal reduction in intelligibility • hoarse and breathy vocal quality • accurate consonant production with occasional alterations of speech rate
AB	• intelligibility moderately to severely impaired • reduced intensity • imprecise consonants • accelerated speech rate
JC	• intelligibility moderately to severely impaired • reduced vocal intensity • imprecise consonant production • inappropriately slowed speech rate

Table 2. *Demographic data.*

Subjects	Sex	Age	Duration PD	H/Y[a]	UPDRS[b]
AD	M	74	2	1	19
JM	M	60	5	2.5	38
AB	F	50	10	4	82
HM	M	50	13	4	79
JC	F	49	22	4	93

[a] Hoehn and Yahr (H/Y)

[b] United Parkinson's Disease Rating Scale (UPDRS)

Three subjects (JC, HM, and AB) had symptoms involving the speech production mechanism and all extremities. These subjects, each on high and frequent dosages of Sinemet, were characterized as having severe bradykinesia, masked faces bilaterally, and pronounced cogwheel rigidity with tremor. Speech intelligibility was moderate to severely impaired in both female subjects, (JC and AB), characterized by reduced intensity, imprecise consonant production and inappropriately slowed or accelerated speech rate. Subject HM showed minimal reduction in speech intelligibility despite the severity of symptoms in the limbs. Vocal quality was characterized as hoarse and breathy, with accurate consonant production and occasional alteration of speech rate. The two remaining subjects, (JH and AD) were mildly impaired, with mild symptoms specific to one upper extremity that were well controlled on low dosages of Sinemet. Vocal quality for these two subjects was characterized as mildly hoarse and breathy.

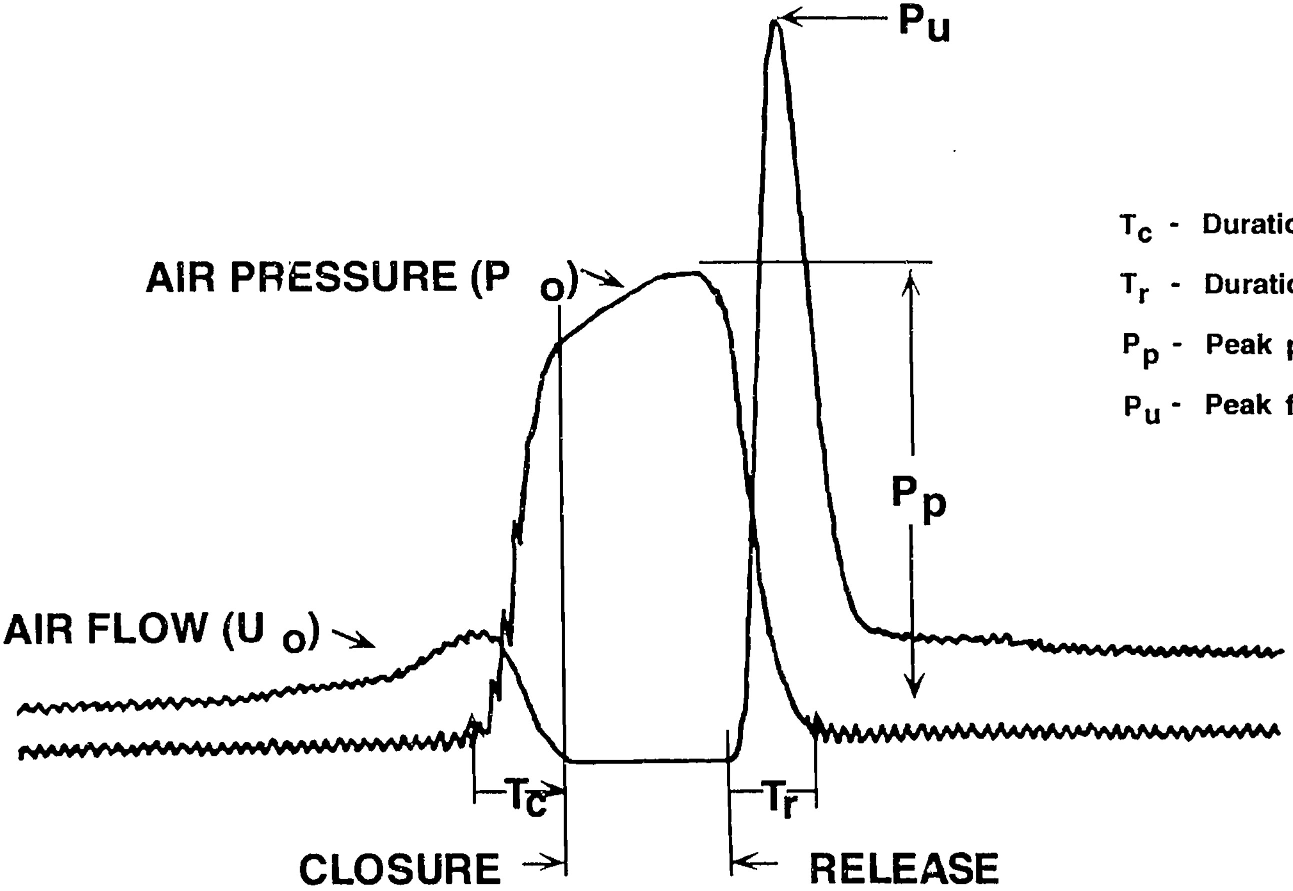
Tasks and measures. Subjects were instructed to produce a VCV disyllable with V being /ae/ or /i/ and C being /p/ or /b/. Testing was consistently accomplished within 1 hour of a medication cycle. Ten repetitions of each disyllable were obtained at comfortable speaking rate. One utterance per breath was sampled. Subjects were instructed to breath as they would in any normal speaking situation. Four repetitions of sustained vowels /a/ and /i/ of four second length were acquired at conversational pitch and intensity levels. In addition, seven consecutive repetitions of the syllable /pae/ at the same requested pitch and intensity levels were obtained in single trials.

Laryngeal / supralaryngeal timing. As illustrated in Figure 1, temporal and magnitude measures were made that indicated the following: 1) T_c - duration of the closing phase defined as the time difference between the initial registration of pressure (P_o) and the associated timing of minimum air flow 2) Tr - duration of the release phase defined as the time between the onset of the pressure drop (at release) and the return of pressure to baseline. In addition, peak intraoral pressure (P_p) and peak air flow (P_u) at the instant of release were obtained.

Mean flow rate / laryngeal resistance. Aerodynamic measures were used to evaluate changes in the physiological state of the vocal folds during sustained phonation. One of the consequences of glottic insufficiency or inadequate closure of the glottis is greater than normal mean air flow rate (MFR) (Isshiki & von Leden, 1964; Hirano, 1981; Iwata, von Leden, & Williams, 1972; Shigemori, 1977; Yoshioka, Sawashima, Hirose, Ushijima, & Honda, 1977).

MFR was based on an averaged 50 ms sample from the mid portion of the vowel. In addition, during sustained phonation, laryngeal resistance (defined as the ratio of transglottal pressure to average glottal air flow) during phonation was estimated from simultaneous measures of air flow and intraoral air pressure during repetitions of the syllable /pae/ based on the method outlined by Smitheran & Hixon, 1981).

GRAPHIC SUMMARY OF MEASUREMENT SCHEME



T_c - Duration of the clo

T_r - Duration of the ope

P_p - Peak pressure

P_u - Peak flow at releas

Figure 1. A summary of the measurement scheme employed for the analysis of the time-varying pressure/flow variations.

Equipment

Aerodynamic events were obtained using a Rothenberg mask equipped with two pressure transducers to sense air flow at the mouth (Microswitch model 163) and air pressure in the oral cavity (Microswitch model 162). A short (approximately 10 cm) polyethylene tube, placed in the oral cavity behind the lips was used to sense the pressure associated with bilabial closure. The acoustic signal was transduced with a microphone at a distance of approximately 15 cm from the subjects' lips.

All signals were digitized at 5000 Hz. with 12 bit resolution. Once acquired, filtered versions of the pressure and flow wave forms were generated and stored as separate files for analysis. For the calculation of laryngeal resistance, MFR, and laryngeal/supralaryngeal timing, the pressure and air flow signals were software filtered at 80 Hz to remove the fundamental frequency variations.

Results

A number of laryngeal and supralaryngeal sequelae were observed in varying degrees in the five subjects. While the intersubject variability was high, intrasubject variability was generally low. Each subject presented a rather consistent set of behaviors across repetitions and across tasks. Subjects HM and AB represent the extremes for this limited sample and their results will be focused on below. In addition, as a result of insufficiency in the two female subjects, duration of the occluded phase was often not possible to measure and will not be presented.

Amplitude measures/pressure and flow. Figures 2 and 3 summarize the peak intraoral air pressure and peak air flow during bilabial consonant production for the five subjects. All subjects demonstrate a voiced/voiceless difference with voiceless pressures and air flow higher than their voiced counterparts. Peak intraoral air pressure for the voiced and voiceless bilabials ranged from 6 to 8 cm H_2O for /p/ and 2 to 5 cm H_2O for /b/. With the exception of peak pressure for the voiced bilabial for subject AB (2 cm H_2O) these values are within the range found for normal speakers (Subtelney, Worth, & Sakuda, 1966). Peak flow rates ranged from 100 to 1180 ml/sec for /p/ and 40 to 550 ml/sec for /b/. Again, with the exception of the flow rates for subject AB, these values are within the range found for neurologically normal speakers (Gilbert, 1973; Isshiki & Ringel, 1964).

Mean flow rates obtained during the mid-portion of the vowels /ae/ and /i/ varied for the different subjects. Flow rates for Subject HM were essentially normal, ranging from 150 to 250 ml/sec during the steady-state portion of the vowels. Flow rates for subjects JC, AB, and AD were quite variable, ranging from 40 to 100 ml/sec. Given the presence of adequate peak intraoral pressures it can be assumed that the respiratory driving force was not the major contributor to the reduced flow. Rather, mean flow rates suggest elevated resistance to air flow.

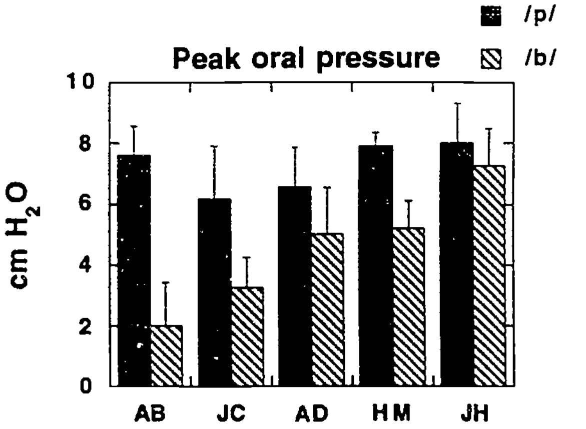


Figure 2. Peak intraoral air pressure (cm H_2O) for consonant production of /p/ and /b/ for five subjects.

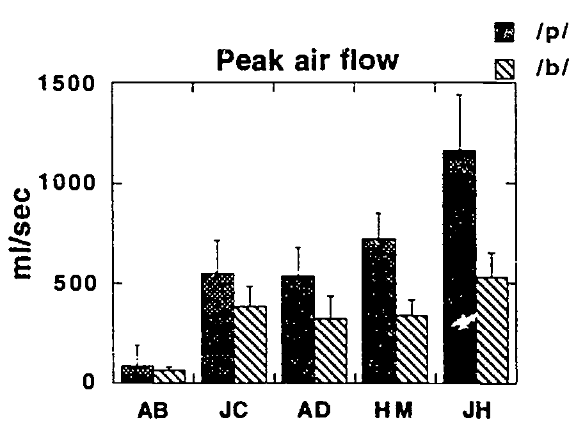


Figure 3. Peak air flow (ml/sec) for the five subjects for the stop consonant productions /p/ and /b/.

In support of this interpretation were the estimates of laryngeal resistance. Figure 4 represents the mean flow rates, peak intraoral pressure, and derived laryngeal resistance measures obtained during /pae/ repetitions. In general, the laryngeal resistance values obtained were higher than those reported for normal subjects (Hillman, Holmberg, Perkell, Walsh, & Vaughan, 1989; Smitherman & Hixon, 1981) and ranged from 20 to 58 cm H_2O/L/sec for the five subjects.

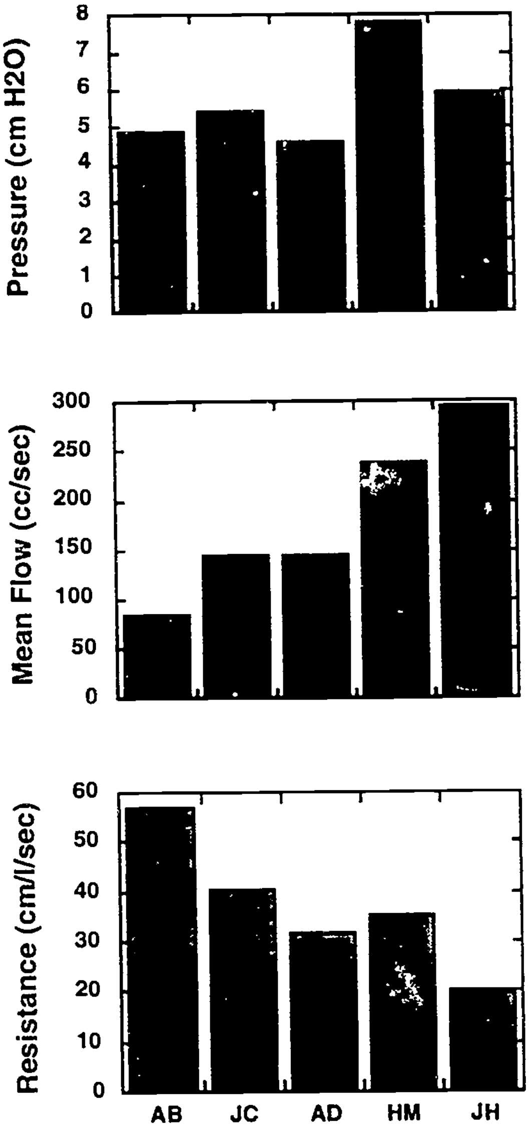


Figure 4. Mean air flow rates, peak intraoral pressure and derived laryngeal resistance measures obtained during /pae/ repetitions.

Shown in Figure 5 are examples from seven serial repetitions of /pae/ for subjects AB and HM illustrating the extremes. Laryngeal resistance for this sample for subject HM was calculated at 30 cm H_2O/L/sec. As can be seen, flow rates averaged approximately 250 ml/sec and pressures ranged from 7.4 tn 9.2 cm H_2O. A distinct voiceless interval can be seen during the closure based on the air flow sig; al. In contrast, laryngeal resistance values for subject AB were much higher, averaging almost 60 cm H_2O/L/sec. Peak pressures were much more variable and decline rapidly over the course of the series of repetitions.

It can also be seen that flow rates are extremely low, averaging approximately 60 ml/sec. Interestingly, it appears that as articulation continues, voicing becomes continuous, and voiceless /p/ becomes the voiced cognate /b/. This is possibly due to increased laryngeal resistance or glottal spasm. It can also be seen that repetition rate for subject AB increases over the five second interval consistent with her tendency toward acceleration.

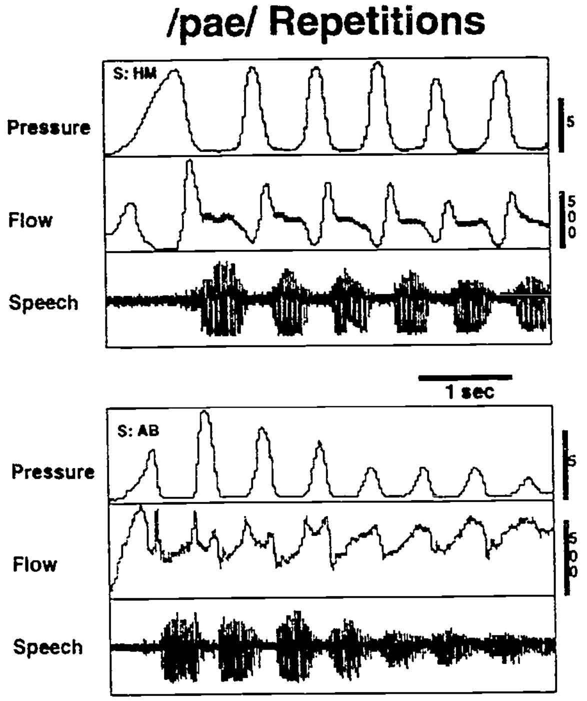


Figure 5. Pressure, flow and acoustic speech signals for /pae/ repetitions for two subjects. For HM /p/ closures are all associated with laryngeal devoicing while AB tends to continue voicing into the voiceless consonant after two repetitions.

Temporal Measures/ Pressure and Flow Variations

Figure 6 is a summary of the duration of the closing phase (T_c) for the lips during voiced and voiceless consonants /p/ and /b/. This value reflects a portion of the change in cross-sectional area at the lips during the oral closing for the stop. The horizontal line in the graph is the average value for this variable reported by Müller and Brown (1980). There was a tendency for all subjects with the exception of JH to display values that are significantly longer than those obtained from normal subjects. These values reflect, in part, overall rate of articulation such that slow oral closing movements will be reflected as higher T_c values. Of the five subjects in the present study, subjects HM, AD and JH display normal speaking rate while JC and AB display slowed speaking rate and associated slowed lip and jaw movements. However, though the rate of speech of subject AB was decreased, this rate reduction was not as dramatic as that of subject JC. In this case one would not expect to see the greater T_c values for AB as compared to JC. Inspection of the pressure/flow waveforms from subject AB reveal the reason for the longer T_c values. For most of her VCV productions, complete cessation of air flow was not achieved, apparently due to velar insufficiency. As such, T_c values reflect a combination of the closing interval and a portion of the occluded phase of stop consonant production. As noted previously, subject JC also demonstrated apparent velar insufficiency, but to a lesser degree than AB. In this case, occlusion was apparent from the rapid decrease in oral air flow.

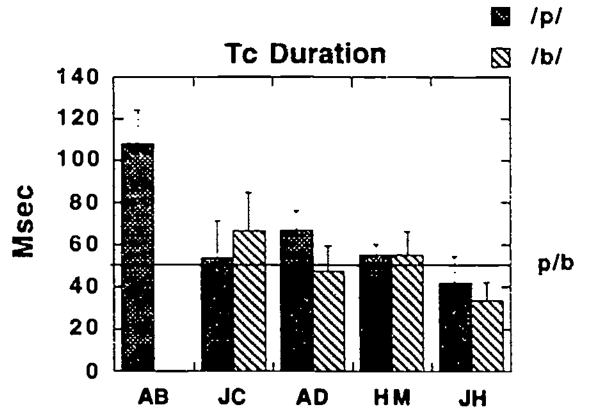


Figure 6. Mean duration of the closing phase (T$_c$) for the five subjects collapsed across the two vowels. No significant vowel affects were noted ($p > .05$). Horizontal line indicates the average T_c values reported by Müller and Brown (1980).

As suggested above, in the absence of kinematic data, simultaneous examination of pressure/flow interactions can be used to infer laryngeal/supralaryngeal coordination and timing. Presented in Figure 7 are examples from two of the five subjects. During the occlusion phase, flow was noted to decrease due to oral closing and was appropriately timed with laryngeal devoicing (see Figure 1 for example). For the two subjects presented in Figure 7, two different patterns are seen. For AB, laryngeal devoicing begins much earlier than oral closing as evidence by the increase in flow noted by the arrows. In contrast, HM displays a decrease in flow and a rather abrupt cessation in voicing at the moment of pressure rise as the lip area decreases to the minimum value to reflect pressure change. In the case of AB, oral/laryngeal actions appeared discoordinated or slowed. For HM, however the transition was more normal although accelerated.

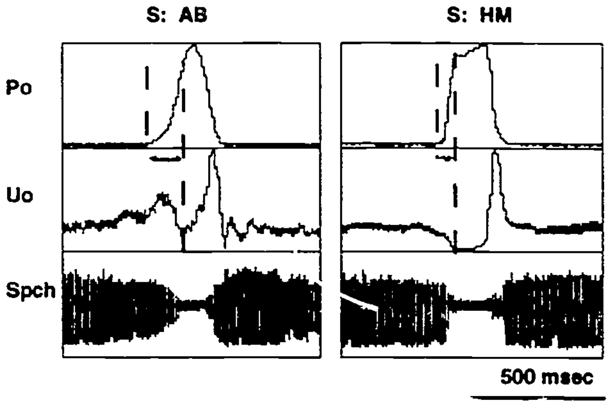


Figure 7. Pressure, flow, and acoustic signals for /aepae/ illustrating examples of the T_c measure for subjects AB and HM.

Figure 8 presents a summary of the results for the duration of the release phase (T_r); the horizontal lines reflect the average T_r values for /p/ and /b/ from Müller and Brown (1980).

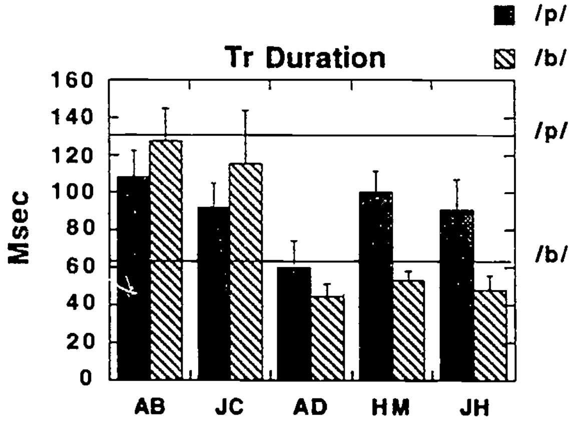


Figure 8. Mean duration of the release phase (T$_r$) for the five subjects collapsed across the two vowels. No significant vowel affects were noted ($p > .05$). Horizontal lines indicates the average Tr values reported by Müller and Brown (1980) for /p/ and /b/.

For the T_r measure, subjects AB and JC displayed longer than normal durations for the time required for the pressure to return to baseline following oral release of the consonant. Time for release is influenced by numerous factors and reflects not only the release gesture at the point of articulation but the possibility of a superimposed breath pulse as well as any variable that can influence the time constant of the decay rate of the Po such as glottal resistance. In contrast to the results from Müller and Brown (1980) for normal subjects and Gracco and Müller (1981) for a group of spastic dysarthrics, two of the three PD subjects displayed longer T_r values for the voiced stop. The longer T_r values reflect a combination of slowed oral opening movements and higher glottal resistance. Interestingly, the higher resistance is most notable when voicing is maintained during the voiced consonant.

Discussion

The present study constitutes an initial attempt to investigate the laryngeal and supralaryngeal deficits in a group of Parkinsonian individuals based in part on previous work by Müller and Brown (1980). A variety of laryngeal and surpalaryngeal impairments were noted. Elevated laryngeal resistance measures may reflect excessive muscle tension either at the level of the glottis or supraglottis, but at the least represent vocal tract constriction. These examples were consistent with limb symptoms of muscular rigidity. That is, for one subject, instances where measured vocal tract resistance was high, limb symptoms of muscular rigidity were also present. For three of the five subjects, vocal tract resistance fell within a range consistent with non-neurologically involved subjects. For these subjects limb rigidity was only mildly present. However, an exception to the limb/bulbar consistency was subject HM. This subject had moderate-to-severe limb involvement; the lower limbs more severely impaired than the upper limbs, and the upper limbs more impaired than the bulbar musculature. However, vocal tract resistance and glottal/supraglottal timing measures were essentially normal.

Supralaryngeal differences were noted in two of the five subjects. Inferences from the time-varying pressure/flow waveforms suggested that oral closing and opening movements were slowed, a finding consistent with previous speech movement studies. In addition, there was some indication of laryngeal/supralaryngeal discoordination.

From this limited sample it is possible to suggest but not to confirm that laryngeal or vocal tract resistance measures may be useful in documenting a variety of the perceptual voice characteristics previously reported for individuals with PD. However, the speech symptoms may not always correlate or correspond to those in the limbs. The time-varying characteristics of the supraglottal pressure and flow waveforms are the consequence of the concomitant articulatory events associated with stop consonant production. Simultaneous recordings and analysis of the pressure and flow events may serve as easily obtained indicators of global system performance aiding in the diagnosis of certain speech related disorders and provide insight into the abnormal articulatory process. For example, the Tr measure or duration of the release phase is defined as the time between the onset of air flow and return to baseline. All of the subjects with one exception showed short Tr values for the voiceless /p/. This measure reflects not only the release gesture at the point of articulation but any variable that can influence the time constant of the decay rate of Po. In the present context, the short Tr values for /p/ may reflect a rapid devoicing gesture perhaps coupled with discoordination of the lips and larynx. For the two most severely involved subjects (AB and JC), the Tr durations for the voiceless /b/ were longer than normal suggesting elevated laryngeal/vocal tract resistance, a slowed release gesture, or the presence on a expiratory breath pulse. Given the low peak pressure values for these subjects, excessive vocal tract resistance seems the most plausible conclusion.

It is especially important to consider that these measures did differentiate subjects within a group of Parkinsonian dysarthrics, though not entirely. Just as there exist subgroups of patients with Parkinson's Disease and various subgroups of Parkinson's syndrome, it appears that acoustic/perceptual and aerodynamic data may be useful in further differentiating these populations. Additionally, pressure and flow information can aid in identifying laryngeal manifestations of pathophysiology affecting phonatory characteristics and glottal efficiency. The preceding dynamic analysis scheme can be used to provide specific information on the general functioning of the speech production mechanism as well as interarticulatory timing from an objective set of data. An analysis scheme of this type in conjunction acoustic and perceptual performance indices, may generate more informed hypotheses concerning the nature of the underlying motor deficit(s) as they affect the speech mechanism.

REFERENCES

Caligiuri, M. P. (1987). Speech labial kinematics in Parkinsonian rigidity. *Brain, 110,* 1033-1044.

Canter, G. J. (1967). Neuromotor pathologies of speech. *American Journal of Physical Medicine, 46,* 659-666.

Connor, N. P., Abbs, J. H., Cole, K. J., & Gracco, V. L. (1989). Parkinsonian deficits in serial multiarticulate movements for speech. *Brain, 112,* 997-1009.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1975). *Motor speech disorders.* Philadelphia: W. B. Saunders.

Forrest, K., Weismer, G., & Turner, G. S. (1989). Kinematic, acoustic, and perceptual analyses of connected speech produced by Parkinsonian and normal geriatric adults. *Journal of the Acoustical Society of America, 85(6),* 2608-2622.

Gilbert, H. R. (1973). Oral airflow during stop consonant production. *Folia Phoniatrica,* 25, 288-301.

Gracco, V. L., & Müller, E. M. (1981). Analysis of supraglottal air pressure variations in spastic dysarthria. Paper presented at the 1981 Convention of the American Speech-Language-Hearing Association, Los Angeles, CA.

Hanson, D. G., Gerratt, B. R., & Ward, P. H. (1984). Cinegraphic observations of laryngeal dysfunction in Parkinson's disease. *Laryngoscope, 94,* 348-353.

Hillman, R. E., Holmberg, E. B., Perkell, J. S., Walsh, M., & Vaughan, C. (1989). Objective assessment of vocal hyperfunction: An experimental framework and initial results. *Journal of Speech and Hearing Research, 33,* 373-392.

Hirano, M. (1981). *Clinical examination of voice.* New York: Springer-Verlag.

Isshiki, N., & Ringel, R. (1964). Air flow during the production of selected consonants. *Journal of Speech and Hearing Research, 7,* 233-244.

Isshiki, N., & von Leden, H. (1964). Hoarseness aerodynamic studies. *Arch. Otolaryngol., 80.* 206-213.

Iwata, S., von Leden, H., & Williams, D. (1972). Air flow measurement during phonation. *Journal of Communication Disorders, 5,* 67-79.

Leanderson, R., Meyerson, B. A., & Persson, A. (1972). Lip muscle function in Parkinsonian Dysarthria. *Acta Otolaryngal, 74,* 354-357.

Logeman, J. A., Fisher, H. B., & Boshes, B. (1978). Frequency and co-occurrence of vocal tract dysfunction in the speech of a large sample of Parkinson's patients. *Journal of Speech and Hearing Disorders, 43(1),* 47-57.

Ludlow, C. L., Bassich, C. J., Connor, N. P., Coulter, D. C., & Lee, Y. J. (1987). The validity of using phonatory jitter and shimmer to detect laryngeal pathology. In T. Baer, C. Sasaki, & K. Harris (Eds.), *Laryngeal function in phonation and respiration* (pp. 463-474). Boston: Little, Brown.

Ludlow, C. L., & Schulz, G. M. (1989). Stop consonant production in isolated and repeated syllables in Parkinson's disease. *Neuropsychologia, 27(6),* 829-838.

Müller, E. M., & Brown, W. S. (1980). Variations in the supraglottal air pressure waveform and their articulatory interpretation. In N. Lass (Ed.), *Speech and language: Advances in basic research and practice, Vol. 4.* New York: Academic Press.

Shigemori, Y. (1977). Some tests related to the air usage during phonation. Clinical investigations. *Otologia (Fukuoka), 23,* 138-166.

Smitheran, J. R., & Hixon, T. J. (1981). A clinical method for estimating laryngeal airway resistance during vowel production. *Journal of Speech and Hearing Disorders, 46,* 138-146.

Subtelney, J. D., Worth, J. H., & Sakuda, M. (1966). Intraoral pressure and rate of flow during speech. *Journal of Speech and Hearing Research,* (9) 498-518.

Ramig, L. A., Scherer, R. C., Titze, I. R., & Ringel, S. P. (1988). Acoustic analysis of voice patients with neurological disease: Rationale and preliminary data. *Annals of Otology, Rhinology, and Laryngology, 97,* 164-171.

Weismer, G. (1983). *Acoustic descriptions of dysarthric speech: Perceptual correlates and physiological inferences.* SMCL. Madison, WI: Wasiman Center.

Yoshioka, H., Sawashima, M., Hirose, H., Ushijima, T., & Honda, K. (1977). Clinical evaluation of air usage during phonation. *Japanese Journal Logopedics and Phoniatrics, 18,* 87-93.

FOOTNOTES

*To appear in J. A. Till, K. R. Yorkston, & D. R. Beukelman (Eds.), *Motor speech disorders: Advances in assessment and treatment.* Baltimore, MD: Brooks Publishing Co.

† Also Yale University School of Medicine, Department of Surgery, Otolaryngology.

‡ Yale University School of Medicine, Department of Neurology.

Effects of Alterations in Auditory Feedback and Speech Rate on Stuttering Frequency*

Joseph Kalinowski,† Joy Armson,† Andrew Stuart,† Marek Roland-Mieszkowski,† and Vincent L. Gracco

This study investigated the effects of altered auditory feedback on stuttering frequency during speech production at two different speech rates. Nine stutterers, who exhibited at least 5% dysfluency during a reading task, served as subjects. They read eight different passages (each 300 syllables in length) while receiving four conditions of auditory feedback: nonaltered, masking, delayed, and frequency altered. For each auditory feedback condition, subjects read at both a normal and a fast rate. Results indicated that stuttering frequency was significantly decreased during conditions of delayed and frequency altered auditory feedback at both speech rates ($p < 0.05$). These findings refute the notion that a slowed speech rate is necessary for fluency enhancement under conditions of altered auditory feedback. Considering previous research and the results of this study, it is proposed that there may be two interdependent factors that are responsible for fluency enhancement: alteration of auditory feedback and modification of speech production.

INTRODUCTION

The finding that stuttering is ameliorated when stutterers speak under conditions of altered auditory feedback has been well documented (see reviews by Starkweather, 1987a; Van Riper, 1982). Two conditions which have been investigated extensively are delayed auditory feedback (DAF), in which there is a delay imposed on the delivery of the feedback speech signal to a speaker's ears, and presentation of a masking noise, which serves to compete with a speaker's auditory feedback.

Recently, the fluency enhancement effect of a different altered auditory feedback condition, frequency alteration, was described by Howell, El-Yaniv, and Powell (1987). Specifically, the frequency components of the speaker's voice were shifted down an octave. The results suggest that frequency altered feedback may be as effective as DAF and masking in the reduction of stuttering. To date, however, there has been no attempt to further investigate the fluency enhancing properties of frequency altered feedback.

The fact that frequency altered feedback has received scant attention in the research literature may be attributed to a general lack of interest in the relationship between altered auditory feedback and stuttering. This attitude stands in direct contrast to the research climate of previous years. Immediately following the initial studies of altered auditory feedback, researchers were intrigued by the possibility that stuttering might be attributed to disordered auditory function, and a number of theoretical models were developed to explain the nature of a possible cause/effect relationship (e.g., Cherry & Sayers, 1956; Mysak, 1966; Webster & Lubker, 1968). A decline in the initial enthusiasm seems to have occurred for two reasons. One reason is that theorists have minimized the role of audition in the control of normal speech production: It has been suggested that auditory information is unlikely to contribute to on-line regulation of speech sound production because information via this modality is not available to motor control centers rapidly enough to be useful (see a review by Borden, 1979). A

Preparation of this paper was supported in part by NIH Grant DC-00121 awarded to Haskins Laboratories. Joy Armson. Andrew Stuart, and Joseph Kalinowski should be considered equal authors of this paper.

second reason relates to widespread acceptance of an idea first advanced by Wingate (1970, 1976): Wingate proposed that artificial disturbance in auditory feedback leads to fluency enhancement only because it alters the speech production characteristics of stutterers. Wingate (1976) argued that the specific change in speech pattern that is induced by altered auditory feedback is "emphasis on phonation...[which] is expressed primarily in slowing down as developed through extended syllable duration" (p. 237).[1] He also argued that increased emphasis on phonation as expressed through slowed speech rate is the primary agent of change which is common to all fluency enhancement conditions (i.e., conditions under which fluency is induced temporarily).

Wingate's suggestion that slowed speech rate is important to fluency enhancement has received powerful empirical support. For example, Perkins, Bell, Johnson, and Stocks (1979) found that stuttering was essentially eliminated when speakers reduced speech rate by approximately 75%. They also found that slow rate achieved through extended syllable duration (as opposed to increased pause time) was optimally effective in promoting fluency. In addition, reductions in speech rate and/or extended segment durations have been found to characterize most conditions of fluency enhancement: for example, singing, and rhythmic speech (Andrews, Craig, Feyer, Hoddinott, Howie, & Neilson, 1983; Andrews, Howie, Dosza, & Guitar, 1982). The importance of slow rate to fluency enhancement is further underscored by the fact that almost all stuttering therapies from the 1800's to the present day have used slow speech rate in some form as a therapeutic strategy (see Van Riper, 1973, and Peters & Guitar, 1991 for reviews).

Consistent with findings for other fluency enhancement conditions, changes in speech rate have been found to occur under conditions of altered auditory feedback. For example, the speech rate of stutterers speaking under DAF has been reported as slow relative to normal rates (see Wingate, 1976, and Andrews et al., 1983, for reviews). According to Wingate (1976), "a general slowing down of speech" (p. 236) is the most commonly reported effect of DAF. With respect to speech produced by stutterers under masking noise, Brayton and Conture (1978) found an increase in vowel durations compared to a control condition, though the differences were not statistically significant. They did report, however, that increases in vowel duration were correlated with decreases in stuttering frequency The temporal characteristics of speech produced under frequency altered auditory feedback have not been reported.

While it is undeniable that slow rate can induce a reduction in stuttering, there is evidence that it may not be the only variable which is involved in fluency enhancement. Support for this notion is provided by Andrews et al. (1982). These authors studied the temporal patterns underlying 15 conditions known to be fluency enhancing and found that slowed speech rate, characterized either by extended syllable duration or increased pause time, occurred in only seven conditions. The question arises, then, as to how necessary a slow rate is to the artificial inducement of fluency which occurs under conditions of altered audition. Correlation of two variables, here parallel reductions in speech rate and stuttering frequency, does not necessarily constitute causality. As such, the argument that slow speech rate may be responsible for inducing fluency under conditions of altered auditory feedback should be viewed as only one possible interpretation. Another interpretation is that speech rate slowing may be merely a naturally occurring by-product of conditions of altered audition which is not necessary for fluency enhancement under these conditions. In such a case, the primary effect of fluency enhancement would be related to some other variable(s). In order to determine how necessary slow speech rate is to the fluency inducing effects of altered auditory feedback, it seems reasonable to investigate whether stuttering decreases when speakers do not exhibit a slow speech rate. One way of counteracting any natural tendency that speakers may have to reduce speech rate under conditions of altered auditory feedback would be to instruct them to speak as quickly as possible.

To the best of our knowledge, an investigation in which stutterers are required to speak as quickly as possible under conditions of altered auditory feedback has not been conducted. There is, however, the practical problem of eliciting a fast rate of speech from a speaker while he/she is experiencing altered auditory feedback. For example, the typical response by speakers to DAF is to slow speech rate as a means of compensating for the imposed feedback delay (see Van Riper, 1982, for a review). However, there is evidence to suggest that normal speakers can override the typical speed-governing effects of DAF. In a study by Siegel, Fehst, Garber, and Pick (1980), adult normal speakers produced speech under delays of

250, 375, 500, and 675 ms at two different speech rates. When instructed to speak as rapidly as possible, these nonstutterers increased their speech rate by 34% as compared to conditions in which they were instructed to speak normally. In contrast to DAF, masking and frequency altered auditory feedback do not involve temporal alterations in the structure of the auditory signal, and there is no reason to believe that a speaker's rate of speech would necessarily be limited by this type of feedback. Thus, these latter two altered auditory feedback conditions seem appropriate to employ in an investigation requiring a fast speech rate.

The purpose of this study, therefore, was to investigate the relationship between speech rate and stuttering frequency under conditions of altered auditory feedback. Specifically, we wanted to determine the effects of masking, frequency altered auditory feedback, and DAF on stuttering frequency when stutterers were instructed to speak at a normal rate and as fast as possible. In keeping with the results of previous investigations, it was predicted that, when instructed to speak at a normal rate, stutterers would exhibit less stuttering under conditions of altered auditory feedback compared to a control condition of nonaltered auditory feedback. Further, if slow rate is a necessary prerequisite to the fluency enhancement process, then altered auditory feedback should not promote fluency when subjects attempt to speak as rapidly as possible. If, on the other hand, slow speech rate is not essential to the induction of fluency under conditions of altered audition, stutterers should exhibit reduced stuttering frequency while speaking rapidly. The latter finding would suggest that the fluency enhancing effect of altered auditory feedback is not solely contingent on decreasing speech rate.

Methods

Subjects. Seven male and two female stutterers, ages 16 to 52 years, participated. For inclusion in the study, subjects were required to demonstrate at least 5% stuttering frequency during a reading task under conditions of nonaltered auditory feedback. All subjects had a history of therapy although none had been enrolled in a program in the last two years.

All but one subject presented with normal bilateral hearing sensitivity, defined as thresholds of 20 dB HL (American National Standards Institute, 1969) or better at octave frequencies of 250 to 8000 Hz. The remaining subject, a 52 year old female, displayed a flat mild sensorineural loss on one side and normal hearing to 2000 Hz with a mild high frequency loss at 4000 Hz on the other. All subjects presented with normal bilateral middle ear function (American Speech-Language-Hearing Association, 1990).

Apparatus. All testing was conducted in a double-walled audiometric test suite (Industrial Acoustics Corporation). The equipment array provided the subjects with four auditory feedback conditions: nonaltered auditory feedback (NAF), masking auditory feedback (MAF), delayed auditory feedback (DAF), and frequency altered auditory feedback (FAF). In all conditions subjects spoke into a microphone (AKG Model C460B) held with a boom on a stand. The distance from the subjects' mouth was 15 cm with an orientation of 330° azimuth and -30° altitude. The microphone output was fed to an audio mixer (JVC Model MI-5000) and routed to a processor and amplifier (Yamaha Model AX-630) before being returned to the subjects' ears through insert earphones (EAR Tone Model 3A). Subjects' speech samples were video recorded with a camera (JVC Model S-62U) and video stereo cassette recorder (Sony Model SL-HF860D).

For the NAF and MAF conditions the speech input was routed through the processor unaltered. For the MAF condition a white noise masker, generated by an audiometer (Beltone Model 10D), was fed to the mixer and routed to the subjects via the same pathway as for their speech input. The masking output from the insert earphone was calibrated to a level of 85 dB SPL in a 2 cm^3 coupler (Brüel and Kjær Model DB-1038) employing a precision sound level meter (Brüel and Kjær Model 2209) and pressure microphone (Brüel and Kjær Model 4144). A level of 85 dB was used for the masking noise in an attempt to produce a noise-to-speech ratio of zero (see below).

In the DAF condition, the processor introduced a 50 ms delay to the speech input. Delays of 50 to 200 ms have previously been reported as optimally effective in enhancing fluency (see Starkweather, 1987a, for review). The smallest of these delays was selected for the present experiment because it was reasoned that a short delay would be less likely than a long delay to limit or restrict speech rate. This was further supported by pilot data which revealed that speakers were able to read rapidly while experiencing a 50 ms delay in the return of auditory feedback.

In the FAF condition, speech input was shifted up in frequency one half an octave by the processor. In contrast, in the Howell et al. (1987) experiments, frequency was shifted down an octave. The latter type of frequency alteration, in our opinion, results in less intelligible speech feedback as opposed to smaller shifts in frequency (e.g., one half an octave). It was considered preferable to use intelligible rather than unintelligible feedback in a FAF condition. Further, our preliminary testing showed that stutterers experienced a reduction in stuttering under the auditory feedback condition of a one half octave frequency shift. The decision to shift frequency up, rather than down, was arbitrary.

During all auditory feedback conditions the amplifier gain for speech input was preset. The output to the insert earphones was calibrated such that a speech signal input of 75 dB SPL to the microphone had an output in a 2 cm^3 coupler of approximately 85 dB SPL. The calibration procedure attempted to approximate real ear average conversation SPLs of speech outputs from normal hearing talkers. That is, an attempt was made to provide a speech level output to the speakers' ears that is consistent with auditory self -monitoring during their normal conversation.[2]

Procedures. While seated in the audiometric test suite, subjects read eight different passages taken from two junior high school level texts (Sims, G. [1987]. *Explorers*, Creative Teaching Press Inc. and Taylor, C. [1985]. *Inventions,* Creative Teaching Press Inc.). Each passage was slightly in excess of 300 syllables. Subjects were instructed to read a given passage at one of two speech rates: normal and fast. At each speech rate, subjects received four conditions of auditory feedback: NAF, MAF, DAF, and FAF. Passages were randomized with respect to auditory feedback condition. Between passage readings, subjects produced approximately one to two minutes of self-formulated monologue speech under NAF in order to minimize any possible carry-over of fluency enhancement from one auditory condition to the next. The auditory feedback conditions were randomized across subjects for each speech rate. Speech rate conditions were counterbalanced for all subjects.

During the fast speech rate condition, subjects were asked to read as fast as they possibly could while maintaining intelligible speech. In the normal speech rate condition, subjects were asked to read at their "usual" or "normal" reading rate. In order to minimize use of fluency-facilitating motor strategies learned in therapy (e.g., slowed speech, gentle voice onset), subjects were instructed to speak as naturally as possible and not to "control" or attempt to minimize their stuttering.

Stuttering was defined as part-word repetitions, part-word prolongations, and inaudible postural fixations. The frequency of stuttering was determined for the first 300 syllables of each video-taped sample by the second author, a certified speech-language pathologist. For one third of the samples, stuttering events were counted a second time by the same judge. Intrajudge reliability for total dysfluencies was .98. A second judge independently determined stuttering event frequency for all samples. Interjudge reliability for total dysfluencies was .95.

Analogue audio signals, from the video recordings of each subject, were digitized at a sampling rate of 10 kHz and analyzed at Haskins Laboratories using an in-house software waveform editing application (WENDY). To determine speaking rate, sections of fluent speech were identified within passages such that the fluently produced syllables were contiguous and the entire fluent speech sample was separated from stuttering episodes by at least one syllable. Separation between fluent speech samples and stuttering episodes was undertaken because it has been shown that the duration of a fluently produced syllable is greater when it is adjacent to a stuttering episode than when it is adjacent to fluent speech (Viswanath, 1986). In most cases, the fluent speech samples consisted of 50 contiguous fluently produced syllables. Identification of samples on the basis of multiple, contiguous fluent syllables was considered important in order to allow speakers to "get up to speed" following a stuttering episode. Fifty syllables was an upper limit for such a sample because of the large number of stutterings which occurred in many of the conditions. As it was, when stuttering frequency was very high, it was not always possible to find 50 fluent syllables which were contiguous. In a few cases, when the fluent syllable count was close to, though less than 50, a slightly smaller syllable count was accepted. In no cases were fewer than 43 syllables used. For some subjects there were conditions for which no samples of fluent syllables could be identified. Durations calculated for the fluent speech samples obtained represented the time between acoustic onset of the first syllable and the acoustic offset of the last fluent syllable, minus pauses that exceeded 100 ms. Most pauses were between 300 and 800 ms and were typically used by the speakers for an inspiratory gesture. Because most of these pauses had an audible inspiratory record, it is unlikely that they were silent stuttering moments. Fluent speech rate in syllables per second was calculated by dividing the duration of each fluent speech sample by the number of syllables in the sample.

Results

Stuttering frequency and syllable rate for each subject by condition are presented in Table 1.

Table 1. *Individual syllable rates and stuttering frequencies as a function of altered auditory condition and speech rate condition(n=9).*

		Syllable Rate				Stuttering Frequency	
Subject Number	Condition	Normal		Fast		Normal	Fast
1	NAF	5.17	(50)	*		2	102
	MAF	4.73	(50)	5.73	(50)	5	4
	DAF	4.57	(50)	4.75	(50)	2	7
	FAF	4.15	(50)	6.22	(50)	2	2
2	NAF	6.30	(45)	*		27	27
	MAF	5.50	(50)	7.68	(50)	10	8
	DAF	5.66	(50)	6.44	(50)	4	1
	FAF	5.85	(50)	7.06	(50)	2	0
3	NAF	*		*		28	36
	MAF	4.87	(50)	5.72	(50)	21	10
	DAF	4.54	(50)	6.00	(50	16	15
	FAF	5.07	(50)	5.48	(50)	6	5
4	NAF	*		*		73	57
	MAF	*		*		67	51
	DAF	4.81	(50)	6.75	(50)	3	0
	FAF	5.05	(50)	7.45	(50)	1	1
5	NAF	4.81	(50)	*		12	33
	MAF	4.50	(50)	*		13	20
	DAF	4.12	(50)	6.61	(50)	0	0
	FAF	4.91	(50)	6.02	(50)	0	0
6	NAF	5.40	(43)	*		11	24
	MAF	5.11	(50)	5.57	(50)	11	17
	DAF	5.38	(50)	5.42	(50)	12	10
	FAF	5.75	(50)	6.22	(50)	9	9
7	NAF	6.46	(50)	*		31	69
	MAF	7.54	(50)	*		18	86
	DAF	4.68	(50)	7.63	(50)	13	9
	FAF	5.34	(50)	*		14	50
8	NAF	3.89	(50)	*		7	37
	MAF	4.45	(50)	4.93	(50)	1	18
	DAF	4.11	(50)	4.84	(50)	1	2
	FAF	3.77	(50)	4.51	(50)	3	2
9	NAF	5.04	(50)	*		13	24
	MAF	5.55	(50)	6.99	(50)	5	13
	DAF	5.12	(50)	5.59	(50)	8	9
	FAF	5.17	(50)	5.94	(50)	5	10

Note: Numbers in parentheses represent contiguous syllable sample size. * represents samples where contiguous sample size criteria were not met.

The means and standard deviations for stuttering frequency and syllable rate, as a function of altered auditory feedback and speech rate conditions, are shown in Figures 1 and 2 respectively. Due to the fact that a number of subjects could not produce the required number of contiguous fluent syllables under certain conditions, means were calculated from eight values for the MAF-normal and FAF-fast speech rate condition, seven values for the NAF-normal speech rate condition, and six values for the MAF-fast speech rate condition. No subject met criterion for a fluent speech sample under the NAF-fast speech rate condition.

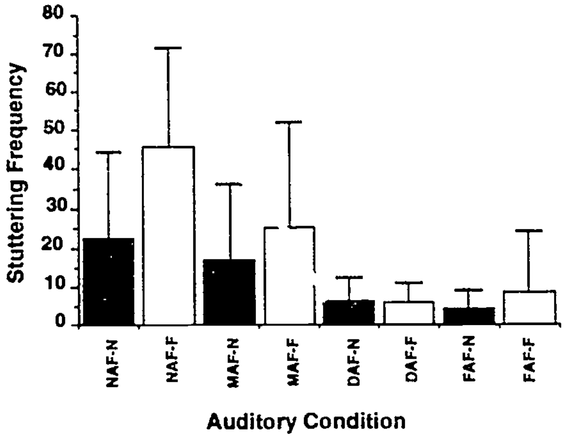


Figure 1. Mean values for stuttering frequency as a function of speech rate and auditory condition (n=9). Error bars represent plus one standard deviation.

Differences in stuttering frequencies and syllable rates, as a function of altered auditory feedback and speech rate, were tested using separate two-factor repeated measures analyses of variance (ANOVA). The NAF condition was not included in the ANOVA for syllable rate because, as noted above, there were no values for the NAF-fast speech rate condition.

The main effect of auditory condition on stuttering frequency was found to be significant [F (3,24) = 11.32, p<0.0001]. Paired t-tests were performed to ascertain which auditory conditions differed significantly. The mean stuttering frequency for NAF was found to be significantly higher than the mean stuttering frequencies for both DAF and FAF (8) = -4.46, p < .01 (p < 0.01).[3] All other pair-wise comparisons were nonsignificant (p >0.01).

A probability value of 0.096 was found for the main effect of speech rate on stuttering frequency [F(1,8) = 3.55] and a probability value of 0.080 was found for the interaction of speech rate and auditory condition [F(3,24) = 2.54]. Thus, both the main effect and interaction approached significance. Inspection of mean values of stuttering frequency at normal and fast rates revealed marked differences for at least two auditory conditions (see Figure 1). Because of this observation and the near-significance of the main effect and interaction, paired t-tests were performed to test the differences in stuttering frequency as a function of speech rate for each auditory condition. A probability value of 0.072 was found for the difference between NAF-normal speech rate and NAF-fast speech rate. No other t- tests yielded results which approached significance.

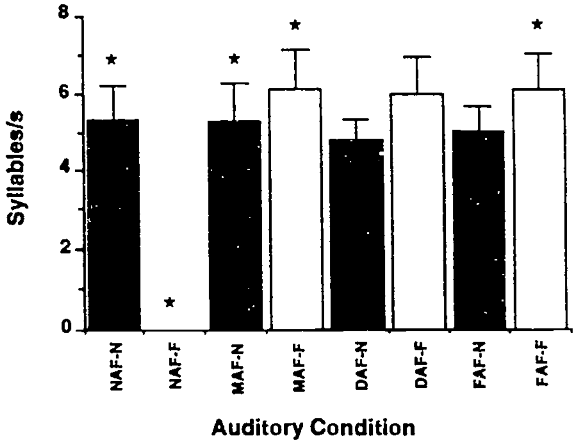


Figure 2. Mean values for fluent syllable rate as a function of speech rate and auditory condition (n=9). Error bars represent plus one standard deviation. (Note: * as some subjects could not produce the required number of contiguous fluent syllables, means were calculated from eight values for the MAF-normal and FAF-fast speech rate condition, seven values for the NAF-normal speech rate condition, and six values for the MAF-fast speech rate condition. No subject met criterion for a fluent speech sample under the NAF-fast speech rate condition).

The main effect of speech rate (i.e., normal vs. fast) was significant for syllable rate [F(1,5)=34.5, p <0.05]. There was no significant difference in syllable rate as a function of auditory condition [F(2,10)=2.54,p>0.05] as well as no significant interaction of speech rate and auditory condition [F(2,10)=1.00, p > 0.05].

DISCUSSION

There are three principal findings in this study: (a) Subjects were able to achieve normal and fast speech rates under conditions of altered auditory feedback, including DAF; (b) there was a tendency for subjects to exhibit more stuttering under instructions to speak rapidly than under instructions to speak normally; and (c) stuttering decreased under conditions of DAF and FAF without respect to speaking rate.

Confirmation that subjects were successful in attaining the targeted speech rates was provided in two ways: First, for the three conditions of altered auditory feedback, all subjects increased speech rate in the fast rate condition compared to the normal rate condition. Second, the absolute values for both speech rates achieved by subjects in this study were consistent with norms from other sources. Specifically, the mean normal speech rates of the present subjects were in the range of 4.8 - 5.3 syllables/s (s/s), which are comparable to values of 4 - 5 s/s found to be characteristic of normal conversational speech (Netsell, 1981; Pickett, 1980; Walker & Black, 1950). Subjects in this study also exhibited mean fast speech rates in the range of 6.0 - 6.1 s/s. These rates clearly exceed the values typically cited as representing a normal conversational speech rate. In addition, these fast rates may also be compared to values provided by Pickett (1980) who judged his own rate of 5.6 s/s as "fast conversational" and 6.7 s/s as the "fastest clear articulation possible" (p.166). It is also interesting to note that the mean fast speech rate achieved under DAF was minimally different from the mean fast rates achieved under MAF and FAF (see Figure 2). Thus, for subjects in this study, a 50 ms delay in return of auditory feedback did not impose a limit on speech rate.

With respect to the second finding, the tendency for subjects to exhibit more stuttering under instructions to speak rapidly than when speaking at a normal rate, it may be noted that differences as a function of rate approached significance for speech produced under NAF only, although this trend was also observed under MAF. It was not possible to verify that, under NAF, subjects were able to comply with instructions to increase speech rate under because for many subjects, sections of fluent speech were too short to permit calculation of this measure, according to our criterion. However, it should be noted that subjects were able to increase speech rate in all other conditions and reported a similar intent under NAF-fast speech rate.

Previous findings regarding stuttering frequency as a function of instructions to increase speech rate are equivocal. Johnson and Rosen (1937) reported that stuttering frequency increased above a baseline level when stutterers were instructed to read as fast as possible. These authors described one subject in particular for whom stuttering increased over 100% in the fast rate condition compared to the baseline conditions. More recently, Armson (1991) reported fluent speech rates and stuttering frequencies for two stuttering speakers who had been asked to read utterances at slow, normal, and fast rate of speech. One of these subjects was able to increase his speech rate to a level which was commensurate with the fast rate of normal speakers. This subject experienced a dramatic increase in stuttering frequency at the fast speech rate compared to the normal rate. On the other hand, Young (1974) instructed stutterers to speak at a normal rate, a faster rate, and finally as fast as possible, and found that the mean ratings of stuttering severity did not differ significantly across these conditions. Ingham, Martin, and Kuhl (1974) reported that two out of three subjects increased speech rate relative to a base rate using words per minute (WPM) as a measure. They stated that "under no circumstances did stuttering frequency exceed that of the initial base rate sessions" (p. 495). Thus, the accumulated data suggest that there may be considerable intersubject variability with respect to the effect of increasing speech rate on stuttering frequency. It may be noted that in the present study, under conditions of nonaltered auditory feedback, seven of the nine subjects exhibited an increase in stuttering frequency as a function of increased speech rate, one subject exhibited fewer stutterings and the remaining subject exhibited no change. Further investigation of this effect is clearly warranted.

The most important finding of this investigation is that stutterers decreased stuttering frequency under certain conditions of auditory feedback alteration at both fast and normal speech rates. As such, the present results refute the notion, first advanced by Wingate (1970, 1976), that under conditions of auditory alteration, a slowed speech rate is a necessary antecedent for fluency improvement.

It is important to note that conditions of altered auditory feedback were similarly efficacious in reducing stuttering at both normal and fast speech rates. That is, the mean percent reductions in stuttering which were associated with a particular auditory feedback condition were not found to differ significantly across speech rates (although there was tendency for greater reductions in stuttering frequency to occur at the fast rate than at the normal rate; e.g., relative to the control condition, stuttering reduced by 87% under DAF at a fast rate and 72% under DAF at a normal rate). Further, the amounts of stuttering reduction which occurred in the present experiment are similar to levels which have been previously reported. For example, Andrews et al. (1983), in a review of studies of fluency enhancing conditions, concluded that reductions in the range of 50-80% are characteristic of speech produced under auditory feedback delays of 50-150 ms and masking. Similar values were obtained in this study under DAF and FAF. Percent reductions in stuttering under MAF were less robust but are consistent with the values reported by Martin and Haroldson (1979). Hence, use of normal and fast rates does not appear to reduce the fluency enhancement effect of these conditions.[4]

If speech rate reduction is not a necessary accompaniment for stuttering amelioration under conditions of altered auditory feedback, then another explanation must be offered. It seems reasonable to speculate that the relevant variables for fluency enhancement under conditions of auditory feedback pertain to auditory function. As such, it is important to search for clues within the present data which might lead to their identification. Toward this end, it may be recalled that the auditory conditions of the present experiment were not equally effective in reducing stuttering frequency. Reduction in stuttering frequency was greater and more consistent across subjects for DAF and FAF than for MAF. Only mean stuttering frequencies for DAF and FAF were found to be significantly different from the control NAF conditions. Given that FAF and DAF are similarly effective in inducing fluency, it is important to consider what properties these conditions may share and how they may differ from MAF.

During MAF, speakers receive two auditory signals: a masking noise and their own speech signal. It is assumed that the masking noise interferes with or impedes reception of the speaker's speech. During DAF and FAF, on the other hand, speakers receive only one auditory signal: their own speech, which is altered slightly in terms of either temporal or frequency characteristics. Therefore, common to DAF and FAF may be the fact that stutterers use their own speech signal (albeit in slightly altered form) to enhance fluency. Further, the fact that alterations in two different parameters of the acoustic signal (i.e., frequency and temporal structure) yield similar fluency enhancing effects seems worthy of note. It may suggest that the effects are not due to the correction of some specific deficit in the stutterer's auditory perceptual processes but to something more global. In this context, it is interesting to note that several subjects reported that speaking under FAF was similar to unison speech. It may be speculated that both DAF and FAF are essentially electronic forms of the "double speaker" phenomenon. The double speaker phenomenon refers to the finding that stuttering is markedly reduced when stutterers speak in combination with another speaker. These phenomena have typically been referred to as "shadow speech" (when the stutterer is slightly delayed relative to the other speaker) and "choral (or unison) speech" (when the two speakers are nearly synchronous) (see Andrews et al., 1983). It is suggested that DAF is a form of "inverse shadow speech" in that the "other speaker" is slightly delayed (e.g., 50 ms) and FAF is a form of choral speech in that the other speaker is speaking simultaneously using either a higher voice (the present study) or a lower voice (Howell et al., 1987).

These speculations about the role of altered auditory feedback in fluency enhancement, in combination with the finding that rate reduction is not necessary for fluency improvement, have important implications with respect to the prevailing view of the nature of stuttering, specifically that stuttering is a disorder of speech timing. In the past, explanations of fluency enhancement have been used to support this view. For example, Kent (1983) argued that "fluency enhancing conditions generally reduce temporal uncertainty ...or allow more time for the preparation of temporal programs" (p. 253), and concluded that stutterers may have "reduced capacity to generate temporal programs" (p. 253). In other words, the key variables responsible for fluency enhancement are either slow speech, which provides a stutterer with more time for motor planning, or an external source of support, which assists the stutterer in generating the temporal patterns of speech. Examples of conditions involving an external source of temporal support are choral and shadow speech. In these conditions, the external source of support is another speaker. That is, it is speculated that the stutterer receives temporal support by relying on the unimpaired timing system of the other speaker (Starkweather, 1987b). According to Kent's argument, the fact that stutterers benefit from an increase in motor planning time or a reliance on externally generated timing patterns indicates that stuttering must somehow result from a speaker's deficiencies in generating these patterns. The present study shows, however, that fluency enhancement may be achieved under certain auditory conditions both when stutterers speak rapidly and when there is no external source of support for generating the temporal patterns of speech (i.e., when subjects use altered/processed versions of their own rapidly produced speech to enhance fluency). If fluency enhancement does not depend on either reduction in temporal uncertainty or more time for preparation of motor programs, then one must question the claim that the underlying deficit in stuttering is reduced capacity to generate temporal programs.

It is evident that studying the fluency enhancement process may be an important route to understanding the nature of stuttering. As noted above, researchers have attempted to identify motor factors which underlie conditions of fluency enhancement and have used this information to develop a theoretical framework for explaining the cause of the disorder. Of the motor patterns which have been studied, slow speech rate is probably the most frequently reported. However, it is likely that other motor patterns are important for fluency enhancement as well. For example, changes in diaphragmatic breath control (Story, 1990), reduction in peak velocity and displacement of the upper lip, lower lip and jaw (Kalinowski, Alfonso, & Gracco, 1991; Story, 1990), and low levels of laryngeal muscle activity (Armson, 1991) have been found to be associated with fluency improvements. One or all of these variables may ultimately prove to be necessary to such improvements. Undoubtedly there are other changes in speech motor patterns which may be also associated with reductions in stuttering.

It is proposed that in addition to modification of speech production characteristics, alterations of auditory feedback may also be responsible for fluency enhancement. These changes in auditory feedback may involve external alterations in the stutterer's own voice (e.g., DAF, FAF) or careful monitoring of another speaker (e.g., choral speech and shadow speech). There are probably other changes to auditory input, as well, which are fluency enhancing. It is further suggested that if auditory and motor factors are inseparable aspects of the speech motor control process, as seems likely, modifications in one will automatically produce changes in the other. In some manner, these sensory-motor modifications may stabilize an intermittently unstable system.

The idea that stuttering is a disorder involving both sensory and motor factors is not novel. Neilson and Neilson (1987) proposed a theoretical account which attributes the problem to "inadequate resources for sensory-motor information processing" (p. 325). According to their model, "feedback...participates in establishing, verifying, and, if necessary, modifying the relationship between motor commands and their... sensory consequences" (p. 327). These sensory-motor relationships are represented in the nervous system by internal models. Neilson and Neilson proposed that stuttering occurs when demands for modeling new sensory-motor relationships exceed the stutterer's limited resources. They suggested that fluency is assured only if some central processing resources which are usually used to form sensory-motor models are freed, or else if sensory-motor processing can be extended in time. However, the existence of fluency enhancement conditions which involve alterations of a speaker's own voice, independent of speech rate reduction, as found in the present experiment, seem to pose a problem for Neilson and Neilson's theory. Specifically, it is difficult to understand how central resources used for modeling sensory and motor relationships would be freed while the speaker is receiving altered auditory feedback and at the same time is speaking rapidly. On the contrary, it would seem that additional, rather than fewer, demands would be made on neuronal resources for sensory-motor processing. Thus, while the present data may be interpreted as supporting the general notion that stuttering is a disorder involving both sensory and motor factors, they do not support the details of Neilson and Neilson's theory. Further research is needed to explore the relationship between speech input/output changes in order to better delineate how this relationship may affect stuttering, and to suggest the nature of a specific sensory-motor deficit which may underlie the disorder.

REFERENCES

American National Standards Institute (1969). *Specifications for audiometers* (ANSI S3.6 - 1969). New York, NY

American-Speech-Language-Hearing Association (1990). Guidelines for screening for hearing impairments and middle ear disorders. *ASHA, 32* (Suppl. 2), 17-24

Andrews, G., Craig, A., Feyer, A., Hoddinott, P., & Neilson, M. (1983). Stuttering: A review of research findings and theories circa 1982. *Journal of Speech and Hearing Disorders, 45* , 287-307.

Andrews, G., Howie, P.M., Dozsa, M., and Guitar, B. E. (1982). Stuttering: Speech pattern characteristics under fluency-inducing conditions. *Journal of Speech and Hearing Research, 25,* 208-215.

Armson, J. (1991). *A study of laryngeal muscle activity during stuttering episodes. Searching for an invariant physiological correlate.* Doctoral dissertation, Temple University, Philadelphia, PA.

Bentler, R. A., & Pavlovic, C. (1989). Transfer functions and correction factors used in hearing aid evaluation and research. *Ear and Hearing, 10,* 58-63.

Borden, G. J. (1979). An interpretation of research on feedback interruption in speech. *Brain and Language, 7,* 307-319.

Brayton, E. R., & Conture, E. G., (1978). Effects of noise and rhythmic stimulation on the speech of stutterers. *Journal of Speech and Hearing Research,* 21, 285-294.

Cherry, E., & Sayers, B. (1956). Experiments upon total inhibition of stammering by external control and some clinical results. *Journal of Psychomotor Research, 1,* 233-246.

Cornelisse, L. E., Gagné, J. P., & Seewald, R. C. (1991). Ear level recordings of the long-term average spectrum of speech. *Ear and Hearing, 12,* 47-54.

Howell, P., El-Yaniv, N., & Powell. D J. (1987). Factors affecting fluency in stutterers. In H. F. M. Peters & W. Hulstijin (Eds.), *Speech motor dynamics in stuttering* (pp. 361-369). New York: Springer-Verlag.

Ingham, R. J., (1984). *Stuttering and behavior therapy: Current status and experimental foundations.* San Diego, CA: College Hill Press.

Ingham, R. J., Martin, R. R., & Kuhl, P. (1974). Modification and control of rate of speaking by stutterers. *Journal of Speech and Hearing Research, 17,* 489-496.

Johnson, W., & Rosen, P. (1937). Studies in the psychology of stuttering: VII. Effect of certain changes in speech pattern upon stuttering frequency. *Journal of Speech and Hearing Disorders, 2,* 105-109.

Kalinowski, J. S., Alfonso, P. J., & Gracco, V. L. (1991). *Modifications to stutterers lip and jaw kinematics following therapy.* Paper presented at the Annual Convention of the American Speech-Language-Hearing Association, Atlanta, Georgia.

Kent, R. D. (1983). Facts about stuttering: Neurolinguistic perspectives. *Journal of Speech and Hearing Disorders , 48,* 249-255.

Lotzman, V. G. (1961). Zur Anwendung variierter Verzögerungszeiten bei balbuties. *Folia Phoniatrica, 13,* 276-312.

Martin, R., & Haroldson, S. K. (1979). Effects of five experimental treatments of stuttering. *Journal of Speech and Hearing Research,* 22,.132-66.

Mysak, E. D. (1966). *Speech pathology: Feedback theory.* Springfield, IL: Charles C. Thomas.

Neilson, M. D., & Neilson, P. D. (1987). Speech motor control and stuttering: A computational model of adaptive sensory-motor processing. *Speech Communications, 6,* 325-333.

Netsell, R. (1981). The acquisition of speech motor control: A perspective with directions for research. In R. Stark (Ed.), *Language behavior in infancy and early childhood* (pp. 127-153). Amsterdam: Elsevier-North Holland.

Perkins, W. H., Bell, J., Johnson. L., & Stocks. J. (1979). Phone rate and the effective planning time hypothesis of stuttering. *Journal of Speech and Hearing Research, 29,* 747-755.

Peters, T. J., & Guitar, B. (1991). *Stuttering: An integrated approach to its nature and treatment.* Baltimore, MD: Williams and Wilkins.

Pickett, J. M. (1980). *The sounds of speech communication: A primer of acoustic phonetics and speech perception.* Baltimore: University Park Press.

Seigel, G. M., Fehst, C. A., Garber, S. R., & Pick, H. L. (1980). Delayed auditory feedback with children. *Journal of Speech and Hearing Research.* 23. 802-813.

Starkweather, C. W. (1987a). *Fluency and stuttering.* Englewood Cliffs, NJ: Prentice Hall.

Starkweather, C. W. (1987b). Laryngeal and articulatory behavior in stutterers. In H. F. M. Peters & W. Hulstijin (Eds.), *Speech motor dynamics in stuttering* (pp. 3-18). New York: Springer-Verlag.

Story, R. (1990). *A pre- and post-therapy comparison of respiratory, laryngeal, and supralaryngeal kinematics of stutterers' fluent speech.* Doctoral dissertation, University of Connecticut, Storrs.

Van Riper, C, (1973). *The treatment of stuttering.* Englewood Cliffs, NJ: Prentice Hall.

Van Riper, C, (1982). *The nature of stuttering.* Englewood Cliffs, NJ Prentice Hall.

Viswanath, N. S. (1986). *An investigation of pre- and post-cursive effects of a stuttering event in the context of a planning unit and temporal reorganization of adapting utterances.* Doctoral dissertation, City University of New York.

Walker, C., & Black, J. (1950). *The intrinsic intensity of oral phrases* (Joint Project Report No. 2). Pensacola, FL: Naval Air Station, United States Naval School of Aviation Medicine.

Webster, R. L., & Lubker, B. B. (1968). Interrelationships among fluency producing variables in stuttered speech. *Journal of Speech and Hearing Research, 11,* 754-66.

Wingate, M. E. (1970). Effect on stuttering of changes in audition. *Journal of Speech and Hearing Research, 13,* 861-863.

Wingate, M. E. (1976). *Stuttering: Theory and treatment.* New York: Irvington.

Young, M. A. (1974). Stuttering severity and instructions to increase speaking rate. *The Illinois Speech and Hearing Journal, 8,* 3-6.

FOOTNOTES

[*] *Language and Speech, 36(1),* 1-16 (1993).

[†] Dalhousie University.

[1] It has been shown that extended syllable duration and an increase in pause time are the primary strategies for reducing speech rate (Ingham, 1984).

[2] A speech signal input of 75 dB SPL to the microphone was selected to be consistent with the average conversation levels from adults and children to a microphone located 15 cm from the talker's mouth (Cornelisse, Gagné, and Seewald, 1991). The 2 cm^3 output level of 85 dB SPL was derived with a 2 cm^3 behind-the-ear hearing aid microphone to ear canal correction from Bentler and Pavlovic (1989) applied to the same ear level recordings from the average conversational speech found by Cornelisse et al. The input signal to the microphone was a speech weighted composite noise generated by a Fonix 6500 Hearing Aid Test System. The stimulus was composed of frequencies from 100 to 8000 Hz (in 100 Hz intervals) with a flat amplitude for the low frequency components and a roll-off slope of 6 dB per octave starting at 1000 Hz. The output of the insert earphones was measured in the 2 cm^3 coupler with the precision sound level meter and pressure microphone.

[3] In order to account for multiple t-tests, a significant alpha level of 0.01 was adopted.

[4] Andrews et al. (1983) reported reductions in the range of 90-100% as characteristic of prolonged speech under DAF (i.e., speech produced under delays which exceed 150 ms). This finding suggests that the fluency enhancement effects under conditions of altered auditory feedback may be enhanced by a subject's use of exaggerated prolongation of speech segments.

Haskins Laboratories Status Report on Speech Research
1992, SR-111/112, 121-134

Phonetic Recoding of Phonologically Ambiguous Printed Words*

Ram Frost[†] and Michal Kampf[†]

Speech detection and matching simultaneously presented printed and spoken words were used to examine phonologic and phonetic processing of Hebrew heterophonic homographs. In Experiments 1 subjects were presented with spoken words-plus-noise and with noise-only trials, and were required to detect the speech in the noise. The spoken words were masked by amplitude modulated noises. The auditory stimuli were presented simultaneously with printed letter strings that represented two meaningful phonological structures, one dominant and the other subordinate. Subjects detected a correspondence between the ambiguous letter string and between the amplitude envelopes of both dominant and subordinate phonological alternatives. In Experiment 2, when the homographs were phonologically disambiguated by adding vowel marks, similar effects were obtained. Experiment 3 revealed that matching the unpointed printed forms of heterophonic homographs to the dominant and subordinate spoken alternatives presented auditorily, was as fast as matching the pointed unambiguous forms to the respective spoken words. This outcome was not obtained in Experiment 4 where print and speech were not presented simultaneously. These results suggest that printed heterophonic homographs activate the two spoken alternatives they represent, and provide further confirmation for fast phonetic recoding in reading.

Most studies of lexical ambiguity have examined the processing of printed homophonic homographs embedded in text or presented in isolation (e.g., Onifer & Swinney, 1981; Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982; Simpson & Burgess, 1985; Swinney, 1979). Homophonic homographs (e.g., "bug"), are characterized by an orthographic structure which has one pronunciation but two different meanings in semantic memory. Research with homophonic homographs has focused on whether the two meanings related to the orthographic structure are activated in parallel, or whether one meaning acquires dominance at some stage after the presentation of the ambiguous letter string. Several studies have suggested that, even in a biasing context, all the meanings of a homograph may be automatically activated and retrieved (e.g., Onifer & Swinney, 1981; Seidenberg et al., 1982; Swinney, 1979; Tanenhaus, Leiman, & Seidenberg, 1979). In contrast, it has been shown that biasing contextual information affects lexical processing of homographs at an early stage, selecting only contextually appropriate meanings (e.g., Glucksberg, Kreuz, & Rho, 1986; Schvaneveldt, Meyer, & Becker, 1976). A third approach posits exhaustive access which does not occur in parallel, but is determined by the relative frequency of the two meanings related to the ambiguous word (e.g., Hogaboam & Perfetti, 1975; Forster & Bendall, 1976; Simpson, 1981; Neil, Hilliard, & Cooper, 1988; Duffy, Morris, & Rayner, 1988; and see Simpson, 1984, for a review).

Homophonic homographs are not the only forms of word ambiguity. Ambiguity can also exist in the

This work was supported in part by a grant awarded to the author by the Hebrew national Academy of Science, and in part by the National Institute of Child Health and Human Development Grant HD-01994 to Haskins Laboratories. I am indebted to Liora Yancovitch for her help in running the experiments, and to Shlomo Bentin, Asher Cohen, Bruno Repp, Derek Besner, Charles Perfetti, and Alexander Pollatsek for their criticism of earlier drafts of this paper.

relationship between the orthographic and the phonologic forms of a word, forming a heterophonic homograph. In contrast to homophonic homographs, heterophonic homographs (e.g., "wind," "bow") are characterized by an orthographic structure that is related to two or more phonological structures. Each of these phonological realizations addresses a different meaning in semantic memory. Since heterophonic homographs form a small and non-representative group of words in English orthography, few studies have examined their processing. Kroll and Schweickert (1978) showed that heterophonic homographs take longer to pronounce than homophonic homographs. Similar results were reported in Serbo-Croatian by Frost, Feldman, and Katz (1990), who demonstrated that subjects are slower to match phonologically ambiguous printed words with their spoken forms.

In contrast to English or Serbo-Croatian, the unpointed Hebrew orthography presents an opportunity to explore the processing of heterophonic homographs. In Hebrew, letters represent mostly consonants while most of the vowels can optionally be superimposed on the consonants as diacritical marks. In most printed material (except for poetry and children's literature), the diacritical vowel-marks are usually omitted. Since different vowels may be added to the same string of consonants to form different words or nonwords, the Hebrew unpointed print cannot specify a unique phonological unit. Therefore, a printed letter string is always phonologically ambiguous and often represents more than one word, each with a different meaning (though frequently related). An example of Hebrew homography is presented in Figure 1.

Unpointed print	חלב (CH-L-V)	
pointed phonological alternatives	Dominant	Subordinate
	חָלָב	חֵלֶב
	(/chalav/)	(/chelev/)
Semantic meaning	"milk"	"grease"

Figure 1. Example of phonological ambiguity in Hebrew.

In a recent study Frost and Bentin (1992) examined the processing of Hebrew heterophonic homographs by using a semantic priming paradigm. Subjects were presented with heterophonic homographs as primes, whereas the targets were related to only one of the primes' possible meanings. The targets followed the primes at different SOAs. It was assumed that if a specific meaning of the prime was accessed, lexical decisions for targets related to that meaning would be facilitated. Frost and Bentin reported that, in the absence of biasing context, both meanings of heterophonic homographs were active at SOAs ranging from 250 to 750 ms from stimulus onset, whereas at a short SOA of 100 ms only the dominant meaning was active.

One characteristic of many studies concerned with lexical ambiguity is the use of semantic priming for examining the processing of an ambiguous prime (for a review see Simpson, 1984). The experimental strategy employed in most cases consisted of monitoring lexical decisions or naming latencies for targets that are related to ambiguous primes embedded in text or presented in isolation (Onifer & Swinney, 1981; Seidenberg et al., 1982; Simpson & Burgess, 1985; Swinney, 1979). Several studies used other behavioral measures such as event-related potentials (Van Petten & Kutas, 1987) or eye movements (Duffy et al., 1988). However, in all of these studies the processing of lexically ambiguous letter strings was investigated by examining semantic facilitation. Although semantic priming effects may well indicate how the two meanings of homographs are entertained during lexical access, their interpretation is not always unequivocal. One major problem relates to the possibility of backward semantic priming. Backward semantic priming refers to a situation in which the target word reactivates the meaning of the prime. The lexical decision, in this case, is facilitated by processing the target in the presence of a related reactivated meaning rather than the result of a direct pre-activation of the target (Koriat, 1981; see Neely, 1990, for a review). Applied to the disambiguation of homographs presented in isolation, the backward priming hypothesis suggests, for example, that the activation of the subordinate meaning of the prime might be initiated by the presentation of the related target, rather than being the result of a context-independent automatic lexical process. Hence, semantic facilitation cannot unequivocally provide evidence in support of an exhaustive or an ordered access model of lexical disambiguation.

Evidence from semantic priming effects is even more problematic when the *phonologic* processing of heterophonic homographs is investigated. It has been suggested that the orthographic structure of heterophonic homographs is linked with two or more lexical entries in the phonologic lexicon, each of which is unequivocally related to one meaning in semantic memory (see Frost & Bentin, 1992, for a discussion of this point). However, it is often assumed that, with the possible exception of very infrequent words, printed words activate orthographic units that are *directly* related to meanings in semantic memory (e.g., Seidenberg, 1985; Seidenberg, Waters, Barnes, & Tanenhaus, 1984). Recently, Jared and Seidenberg (1991) suggested that the activation of meanings precedes phonologic activation. Jared and Seidenberg (1991) employed the semantic decision task developed by Van Orden (Van Orden, 1987) and reported that phonologically based activation of meanings is limited to low-frequency words and nonwords. Thus, according to this direct access model, the visual presentation of a heterophonic homograph might directly activate its two semantic meanings, or at least its dominant meaning, without a prior activation of the word (phonologic structure) related to it. Similarly, if backward priming should occur, the meanings of the primes would be initiated by the disambiguating targets, and not necessarily the unequivocal phonological structures that are related to the ambiguous primes. Therefore, the measurement of semantic facilitation does not indicate whether the presentation of the ambiguous letter string has caused the activation of the two phonologic structures related to it, or merely the activation of the two semantic meanings which were accessed directly from the print.

The aim of the present study was to investigate whether the two alternative phonemic realizations of heterophonic homographs are activated following the visual presentation of the ambiguous orthographical pattern, yet avoiding the problems inherent to semantic priming methods. For this purpose two tasks that directly tap phonologic and phonetic activation were employed. Experiments 1 and 2 employed the speech detection task (Frost, 1991; Frost, Repp, & Katz, 1988). Experiments 3 and 4 employed the matching task (Frost, Feldman, & Katz, 1990; Frost & Katz, 1989).

The detection task

Frost et al. (1988) reported an auditory illusion occurring when printed words and masked spoken words appear simultaneously. In a set of experiments, subjects were presented with speech-plus-noise and with noise-only trials, and were required to detect the masked speech in a signal detection paradigm. The auditory stimuli were accompanied by print which either matched or did not match the masked speech.

The noise used in this experiment was amplitude-modulated (i.e., the spoken word was masked by noise with the same amplitude envelope). A word's amplitude envelope is mainly the variation of its amplitude over time. It represents the dynamic property of the acoustic signal. Amplitude-modulated noise is a stretch of white noise with an amplitude that is correlated with the word's amplitude fluctuations. Thus, amplitude-modulated noise (representing the word's amplitude envelope) does not provide any spectral information and therefore cannot convey the explicit phonemic or syllabic structure of the word. Rather, it retains some speechlike features and conveys mostly prosodic and stress information. Since in the Frost et al. (1988) study the words were masked by their own amplitude envelopes, when a printed word matched the spoken word it also matched the amplitude envelope of the noise generated from it.

The results suggested that subjects automatically detected a correspondence between noise amplitude envelopes and printed stimuli when they matched. The detection of this correspondence made the amplitude-modulated noise sound more speechlike, causing a strong response bias: Whether speech was indeed present in the noise or not, subjects had the illusion of hearing it when the printed stimuli matched the auditory input. In order to match the visual to the auditory information, subjects had to generate from the print the relevant amplitude envelope. The process of matching the auditorily presented envelopes to the lexically addressed envelope representations derived from the print, was probably performed in one single match of the overall acoustic shapes. This is because in the speech detection task the visual presentation occurs at the onset of speech presentation, which unfolds over time. Thus, the subject often retrieved the complete phonetic representation of the printed word before the complete presentation of the auditory stimulus. Regardless of how the matching process occurs, it is the positive matching of the envelope representation derived from the print with the envelope provided auditorily that causes the illusion.[1]

Frost (1991) replicated these findings in Hebrew. As in the original study, subjects listened to speech-plus-noise and noise-only trials, accompanied by pointed or unpointed Hebrew print that either matched or did not match the masked speech. The stimuli in the study consisted of high-frequency words, low-frequency words, and nonwords. The results showed that matching print caused a strong bias to report speech in noise for words regardless of frequency, but not for nonwords. The bias effect was not affected by spelling-to-sound regularity—that is, similar effects were obtained in the pointed and the unpointed conditions. Note that in this and previous studies using the speech detection technique subjects were not required to respond to the printed information. Moreover, they were informed of the equal distribution of signal and noise trials in the different visual conditions. Nevertheless, the correspondence between the visual stimulus and the speech envelope affected their response criterion. This outcome was interpreted to suggest that the generation of a phonetic representation following the presentation of a printed word occurred automatically. The bias effect did not appear when the printed words and the spoken words from which the amplitude envelopes were generated were merely similar in their syllabic stress pattern or phonologic structure. Frost and his colleagues therefore concluded that the processing of a printed word results not only in a phonologic code but also in a detailed phonetic speech code which includes the word's amplitude envelope.

Amplitude envelopes representations can assist the listener in the process of spoken word recognition. The envelopes cannot identify a specific lexical candidate, However, they do convey prosodic and segmental information (e.g., speech timing, number of syllables, relative stress, and several major classes of consonant manner), that might help in selecting a lexical candidate among a highly constrained set of response alternatives (Van Tasell, Soli, Kirby, & Widin, 1987). Thus, the amplitude envelope might serve as additional information used by the listener in order to identify spoken words which have several acoustic realizations, or which their phonemic structure was not clearly conveyed (cf. Gordon, 1988). In these cases, a match between the perceived amplitude envelope and the stored template might confirm the identity of a lexical candidate.

The bias to report hearing speech in amplitude-modulated noise when matching print accompanies the auditory presentation was used as the dependent variable in the present investigation. This bias occurs only when subjects detect a correspondence between the printed and the spoken information. Therefore, the present experiments examined whether subjects detected a correspondence between a printed heterophonic homograph and the masked spoken forms of the *two* phonologic alternatives it represents. This could easily be monitored by measuring the amount of bias obtained when the amplitude envelopes derived from each of the two phonologic alternatives were presented in the auditory modality.

EXPERIMENT 1

In Experiment 1 subjects were presented with printed Hebrew heterophonic homographs in the visual modality. Each homograph could be read as two different words, one with a higher frequency of occurrence (dominant phonological alternative) and the other with a lower frequency of occurrence (subordinate phonological alternative). Simultaneously with the visual presentation subjects heard over earphones the possible spoken words related to the homographs, each one masked by noise having the same amplitude envelope. In addition, they heard just the amplitude-modulated noises that were derived from the spoken dominant or subordinate alternatives. Thus, in half of the trials the noise was presented by itself, and in the other half it served as a masker for the spoken forms. The subjects' task was to distinguish between those trials.

If following the visual presentation subjects retrieve the amplitude envelopes (as well as other phonetic information) of *both* the dominant and the subordinate phonologic alternatives that are related to the homograph, then the simultaneous auditory presentation of either of these envelopes should produce a bias to detect speech in the noise. If, on the other hand, the visual presentation of the printed homograph does not result in a phonetic activation of the two spoken words related to the letter string, then no effect of bias should be obtained. Alternatively, if only the dominant word is activated following the presentation of the printed letter string, a bias should be revealed only for the dominant alternative.

Method

Subjects. Twenty-four undergraduate students, all native speakers of Hebrew, participated in the experiment for course credit or for payment.

Stimulus preparation. The stimuli were generated from 24 ambiguous consonant strings each representing two words: one high- and one low-frequency. The two words were not semantically related. In the absence of a reliable frequency count in Hebrew, the subjective frequency of each word was estimated using the following procedure: From a pool of 100 ambiguous consonant strings, two lists of 100 pointed words each were generated. Each list of disambiguated words contained only one form of the possible realizations of each homograph. Dominant and subordinate meanings were equally distributed between the lists. Both lists were presented to 50 undergraduate students, who rated the frequency of each word on a 7-point scale from very infrequent (1) to very frequent (7). The rated frequencies were averaged across all 50 judges. Each of the 24 homographs that were selected for this study represented two words that differed in their rated frequency by at least 1 point on that scale. The validity of this selection was then tested by naming: Twenty-four subjects were presented with the unpointed homographs, and their vocal responses were recorded. The relative dominance of each phonological alternative was assessed by the number of times it was actually pronounced by the subjects. Only those homographs whose frequency judgments coincided with the results obtained in the naming task (i.e. at least 66% of the subjects chose to name the phonological alternative that had a higher frequency rate) were used in the experiment.

The auditory stimuli were originally spoken by a male native speaker in an acoustically shielded booth and recorded on an Otari MX5050 tape-recorder. The speech was digitized at a 20 kHz sampling rate. From each digitized word, a noise stimulus with the same amplitude envelope was created by randomly reversing the polarity of individual samples with a probability of 0.5 (Schroeder, 1968). This signal-correlated noise retains a certain speechlike quality, even though its spectrum is flat and it cannot be identified as a particular utterance unless the choices are very limited (see Van Tasell et al., 1987). The speech-plus-noise stimuli were created by adding the waveform of each digitized word to that of the matched noise, adjusting their relative intensity to yield a signal-to-noise ratio of -10.7 dB.

Each digitized stimulus was edited using a waveform editor. The stimulus onset was determined visually on an oscilloscope and verified auditorily through headphones. A mark tone was inserted at the onset of each stimulus on a second channel, inaudible to the subject. The edited stimuli were recorded at three-second intervals on a two-track audiotape, one track containing the spoken words while the other track contained the mark tones. The purpose of the mark tone was to trigger the presentation of the printed stimuli on a computer screen.

Design. Each of the dominant and subordinate spoken alternatives was presented in two auditory forms: (1) speech-plus-noise trials, in which the spoken stimulus was presented masked by noise; (2) noise-only trials, in which the noise was presented by itself without the speech. Each of these auditory presentations was accompanied by two possible visual presentations: (1) matching print (i.e. the same word that was presented auditorily, and/or that was used to generate the amplitude-modulated noise, was presented in print); (2) nonmatching print (i.e, the printed stimulus was a different word, having the same number of phonemes and a similar phonologic structure, but without sharing any phoneme in the same location with the word that was presented auditorily, or that was used to generate the noise). Thus, there were four combinations of visual/auditory presentations for each of the 48 words, making a total of 192 trials in the experiment.

Procedure and apparatus. Subjects were seated in front of a Macintosh SE computer screen and listened binaurally over Sennheiser headphones. They sat approximately 70 cm from the screen, so that the stimuli subtended a horizontal visual angle of 4 degrees on the average. A bold Hebrew font, size 24, was used. The task consisted of pressing a "yes" key if speech was detected in the noise, and a "no" key if it was not. The dominant hand was always used for the "yes" responses. Although the task was introduced as purely auditory, the subjects were requested to attend carefully to the screen as well. They were told in the instructions that, when a word was presented on the screen, it was sometimes similar to the speech or noise presented auditorily, and sometimes not. However, they were informed about the equal proportions of "yes" and "no" trials in each of the different visual conditions.

The tape containing the auditory stimuli was reproduced by a two-channel Otari MX5050 tape-recorder. The verbal stimuli were transmitted to the subject's headphones through one channel, and the trigger tones were transmitted through the other channel to an interface that directly connected to the Macintosh, where they triggered the visual presentation.

The experimental session began with 24 practice trials. The practice trials were generated from ambiguous words that were not used in the experiment. After the practice all the 192 experimental trials were presented in one block. The duration of a whole experimental session was approximately 20 minutes.

Results and Discussion

The indices of bias in the different experimental conditions were computed following the procedure suggested by Luce (1963). Results computed according to Luce's procedure tend to be very similar to results produced by the standard signal detection computations (e.g., Wood, 1976). However, Luce's indices do not require any assumptions about the shapes of the underlying signal and noise distributions, and are easier to compute relative to the standard measures of signal detection theory. The Luce indices of bias and sensitivity, originally named *lnb* and *lnη*, but renamed here for convenience *b* and *d*, are:

$$b = 1/2 \ln [p(\text{yes/s+n})\, p(\text{yes/n}) / p(\text{no/s+n})\, p(\text{no/n})],$$

and

$$d = 1/2 \ln [p(\text{yes/s+n})\, p(\text{no/n}) / p(\text{yes/n})\, p(\text{no/s+n})],$$

where *s+n* and *n* stand for speech-plus-noise and noise only, respectively. The index *b* assumes positive values for a tendency to say "yes" and negative values for a tendency to say "no." For example, according to the above formula, in order to obtain an average *b* of +0.5, the subject must generate on the average 60% more positive than negative responses. The index *d* assumes values in the same general range as the d' of signal detection theory, with zero representing chance performance.

The average values for the bias indices in each experimental condition are shown in Table 1.

Table 1. *Bias indices (b) obtained in Experiment 1 with unpointed heterophonic homographs in the matching and nonmatching conditions. for dominant and subordinate spoken alternatives.*

	Dominant	Subordinate
Match	0.74	0.59
No Match	0.09	-0.16

There was a bias to say "yes" in the matching condition for both the dominant and subordinate alternatives. There was no bias in the nonmatching condition. Similar differences in bias between the matching and the nonmatching conditions were found for the dominant and the subordinate alternatives. The average *d* in the experiment was 0.15.[2]

The bias indices were subjected to a two-way analysis of variance with the factors of visual condition (matching print, nonmatching print) and dominance (dominant, subordinate).[3] The main effect of visual condition was significant ($F(1,23)=21.4$, $MSe=0.5$, $p<0.001$), as was the main effect of dominance $F(1,23)=4.6$, $MSe=0.2$, $p<0.04$). The main effect of dominance, however, is not of great importance because the bias effects for the dominant and subordinate alternatives are measured in reference to the overall differences between the matching and the nonmatching conditions. These differences were similar for the dominant and subordinate alternatives as reflected by the nonsignificant two-way interaction ($F(1,23)<1.0$). The results of Experiment 1 suggest, then, that both the dominant and the subordinate phonological alternatives were activated to the same extent following the presentation of the ambiguous homograph.

EXPERIMENT 2

In Experiment 2 subjects were presented with the same set of consonantal strings as in Experiment 1 but in a disambiguated form. Hebrew heterophonic homographs can be disambiguated by adding diacritical dots to the ambiguous letter strings. The vowel marks unequivocally determine the phonemic structure of the consonantal cluster, thereby denoting either its dominant or its subordinate reading. The aim of the experiment was to compare the bias effect obtained with unpointed ambiguous letter strings to the bias effect obtained when the print conveyed explicitly the exact phonological alternatives, dominant or subordinate. If indeed both phonological alternatives were activated in Experiment 1 with unpointed print, similar bias effects should emerge in Experiment 2 where the vowels unequivocally denote the dominant and the subordinate readings.

The pointed presentation also allowed the inclusion of an additional experimental control condition. Before speculating about mechanisms of processing heterophonic homographs, it was important to make sure that the dominant and subordinate spoken alternatives in their masked forms were clearly distinguishable from each other. Note that the dominant and the subordinate alternatives of heterophonic homographs have a very similar phonologic structure (identical consonantal cluster) and often differ by only one vowel. It is possible that the amplitude envelopes of the two alternatives were similar to the extent that subjects could not distinguish between them. The presentation of the printed homograph might have resulted in the activation of only one phonological alternative, probably the dominant one, and consequently in the generation of its amplitude envelope alone. If the amplitude envelopes of the dominant and the subordinate alternatives were very similar, even if the subjects had generated the envelope of the dominant alternative only, they would have detected a correspondence between this envelope representation and the envelope of the subordinate alternative that was presented auditorily. By this account, the bias obtained for both dominant and subordinate forms would have resulted merely from the inability of subjects to differentiate between the similar amplitude envelopes, and their falsely detecting a match between the dominant envelope they generated from the ambiguous printed word, and the subordinate envelope they heard. Thus, the second aim of Experiment 2 was to test this possibility directly. In addition to the four experimental conditions of Experiment 1, subjects were exposed to a fifth condition in which the phonological alternatives that were presented explicitly in pointed print were accompanied by the amplitude envelopes of the other phonological alternatives that were related to the same homograph, and vice versa. The purpose of these trials was to examine whether such a countermatch can result in a bias to detect speech in the noise.

The explicit (pointed) presentation of the dominant or the subordinate phonological alternative presumably would result in the activation of a detailed phonetic representation of that word, including the word's amplitude envelope (Frost, 1991; Frost et al., 1988). If the amplitude envelopes of the dominant and the subordinate alternatives cannot be distinguished from each other, then the countermatch condition should produce a bias effect that is similar to the effect obtained in the matching conditions; subjects would detect a correspondence between the envelope generated from one printed phonological alternative and between the envelope of the other alternative, provided auditorily. If, on the other hand, the amplitude envelopes of the two phonological alternatives are perceptibly different, the pattern of bias obtained in the countermatch condition should be more similar to the pattern obtained in the nonmatching conditions.

Method

Subjects. Twenty four undergraduate students, all native Hebrew speakers, participated for course credit or for payment. None of the subjects had participated in Experiment 1.

Stimuli, Design, and Procedure. The stimuli and procedure were identical to those used in Experiment 1, except that all the words were presented in conjunction with vowel marks. Thus, each word was presented in an unequivocal phonological form and had only one meaning.

The design of the experiment included an additional control condition. In this condition for each homograph, subjects were presented with one of the pointed alternatives in print, in conjunction with the masked spoken forms of the other alternatives, and vice versa. Hence, this additional condition (the "countermatch condition") served as an experimental control to measure the bias effect caused by the possible phonetic similarity between the dominant and subordinate phonological alternatives related to the same letter string.

Results and Discussion

The average values for the bias indices in each experimental condition are shown in Table 2. There was a bias to say "yes" in the matching condition for both the dominant and the subordinate phonological alternative, whereas there was no bias in the nonmatching condition. The bias effects found for the dominant and the subordinate alternatives were very similar.

Table 2. *Bias indices (b) obtained in Experiment 2 with pointed heterophonic homographs for dominant and subordinate spoken alternatives in the matching, nonmatching, and countermatch conditions.*

	Dominant	Subordinate
Match	0.55	0.51
No Match	0.07	-0.14
Countermatch	0.16	

The bias indices were subjected to a two-way analysis of variance with the factors of visual condition (matching print, nonmatching print) and dominance (dominant, subordinate). The main effect of visual condition was significant (F(1,23)= 15.0, MSe=0.5, p<0.001). The main effect of dominance and the two-way interaction were not significant (F(1,23)= 2.6, MSe=0.14, p<0.12; F(1,23)<1.0; respectively).

We compared the effects of bias obtained for the dominant and the subordinate alternatives in the unpointed ambiguous print relative to the pointed unequivocal print. For this analysis the relevant data from the two experiments were combined in mixed ANOVA designs in which the print form (pointed or unpointed) was introduced as an additional between-subjects factor. The effect of print form was not significant (F(1,46) <1.0, MSe =1.6). Print form did not interact with visual condition (F(1,46) < 1.0, MSe = 0.22), or with dominance (F(1,46) < 1.0, MSe =0.17), nor was the three-way interaction significant (F(1,46) <1.0, MSe=0.17). This outcome suggests that the unequivocal printed presentation of the dominant or subordinate alternatives in Experiment 2 resulted in bias effects that were very similar to the effects produced by the ambiguous print in Experiment 1.

Another significant outcome of Experiment 2 was that there was almost no bias effect in the countermatch condition (0.16). In order to compare directly the effects of bias obtained in the matching and the countermatch conditions, planned comparisons were conducted. These comparisons revealed a significantly greater effect of bias in the matching condition than in the countermatch condition for both the dominant and the subordinate alternatives (t(23)=2.88, p<0.008; t(23)=3.0, p<0.006, respectively). The difference in bias between the countermatch and the nonmatching condition was significant for the subordinate alternatives (t(23)= 2.1, p< 0.04), but not for the dominant alternatives (t(23) = 0.6, p<0.5). This merely suggests that the countermatch condition was more similar to the nonmatching condition for the dominant than for the subordinate alternatives. Thus, the results of Experiment 2 support the conclusions of Experiment 1.

EXPERIMENT 3

The next two experiments employed the matching task, which directly taps the process of mapping printed words into lexical phonological structures (Frost & Katz, 1989; Frost et al., 1990). In the matching task subjects are simultaneously presented with a printed word on a computer screen, and with a spoken word via headphones. The subjects are asked to decide as fast as possible whether or not the stimuli presented in the visual and the auditory modalities are the same or different. In order to match the spoken and the printed forms of words, they both have to converge at an identical lexical entry. Because the transformation of speech into an orthographic representation is, by far, less practiced than the transformation of spelling into phonology, the common end result of both print and speech processing in the matching task is presumably a phonological representation in the lexicon (see Frost et al., 1990, for a detailed discussion of the matching task).

In Experiment 3 subjects were presented simultaneously with printed heterophonic homographs and with the spoken forms of the dominant and subordinate alternatives. They were instructed to determine whether the printed words and the spoken words were equivalent. In some of the trials the printed homographs were presented in their pointed form and were therefore disambiguated; that is, the vowel marks depicted unequivocally either the dominant or the subordinate alternatives. In these trials the matching of the visual printed words to the spoken words did not involve any ambiguity resolution. In other trials the homographs appeared unpointed, and thus could be read in two ways. In these trials the outcome of matching the visual words to the spoken words was dependent on the specific phonological alternative generated from the ambiguous consonant string. The aim of the experiment was to compare the decision time for pointed and unpointed print.

If only the dominant phonologic alternative is generated from the ambiguous printed homograph, then the vowel marks will not affect the decision time when the dominant spoken word is presented auditorily. The dominant alternative will be generated as a rule from the letter string (pointed or unpointed) and compared to the dominant spoken word presented auditorily, to yield a "yes" response. In contrast, different decision times will be revealed for the pointed and unpointed forms of the subordinate alternatives. This is because with the unpointed ambiguous print, the subject would first generate the dominant alternative, and would consequently find a mismatch with the spoken subordinate alternative presented auditorily. This mismatch will slow his/her decision time relatively to the pointed presentation that depicts the subordinate alternative unequivocally. If, on the other hand, both phonologic interpretations are generated from the consonant strings, then no advantage in decision time for pointed presentations will be found for both the dominant and the subordinate alternatives. As both phonological alternatives are computed from the letter string, their matching with the auditory information should not be affected by the explicit presentation of vowel marks.

Method

Subjects. Sixty undergraduate students, native Hebrew speakers, participated in the experiment for course credit or for payment. None of the subjects had participated in the previous experiments.

Stimuli and Design. The target stimuli were 40 ambiguous consonant strings which represented both a high- and a low-frequency word. Their selection criteria were identical to those used in Experiments 1 and 2. With different vowel marks, these 40 consonant strings represented 80 words, 40 frequent and 40 nonfrequent. Each trial in the experiment consisted of a visual and an auditory presentation of a word. There were 4 experimental conditions: An unpointed printed letter string could appear in conjunction with either the dominant or the subordinate alternative. In addition, the spoken dominant and subordinate alternatives appeared in conjunction with their printed pointed forms. All of these trials were "same" trials. In addition, 40 "different" pairs were introduced as fillers in order to achieve a probability of 0.5 for a "same" response. The "different" trials consisted of ambiguous printed consonant strings that were not used in the "same" trials. These letter strings were half pointed and half unpointed, and were paired with spoken words with the same length and vowel-consonant structure, but differed with respect to one or two phonemes. In order to introduce a higher level of complexity in the experimental task, one third of the "different" trials consisted of a pointed presentation of one phonological alternative of the ambiguous letter string whereas the spoken words were the other phonological alternative related to that homograph. These trials ensured that the matching process could not be performed by a superficial similarity judgment.

Four lists of words were formed: Each list contained 10 pairs in each of the four experimental conditions and 40 mismatch fillers. Each subject was tested in only one list. The pairs were rotated across lists by a Latin Square design, so that each subject was exposed to all experimental conditions, yet avoiding any stimulus repetition effect.

The visual stimuli were presented on a Macintosh II computer screen. They subtended a visual angle of approximately 2.5 degrees on average. The auditory stimuli were originally spoken by a female native speaker in an acoustically shielded booth and digitized at a 20 kHz sampling rate using Macrecorder. Each digitized stimulus was edited. Its onset was determined visually on an oscilloscope, and was verified auditorily through headphones.

Procedure and Apparatus. Subjects wore Sennheiser headphones and sat in a semi-darkened room in front of the Macintosh II computer screen. The auditory stimuli were transmitted binaurally to the subject's headphones by the computer. The visual presentation occurred simultaneously and triggered the computer's clock for reaction time (RT) measurements. The experimental task consisted of pressing a "same" key if the visual and the auditory stimuli were the same word, and pressing a "different" key if they were different. The dominant hand was always used for the "same" responses. The experimental session began with 16 practice pairs. After the practice, all 80 test trials were presented in one block.

Results and discussion

Means and standard deviations of RTs for correct responses were calculated for each subject in each of the four experimental conditions. Within each subject/condition combination, RTs that were outside a range of 2 SDs from the respective mean were excluded, and the mean was recalculated. Outliers accounted for less then 5% of all responses. RTs in the different experimental conditions are shown in Table 3. Decision latencies were identical for pointed and unpointed print for both dominant and subordinate alternatives.

Table 3. *RTs and (errors) for pointed and unpointed dominant and subordinate alternatives in the matching task. Print and speech are presented simultaneously.*

Visual presentation	Auditory presentation	
	Dominant	Subordinate
Unpointed	710 (4%)	736 (4%)
Pointed	709 (4%)	736 (6%)
No Match fillers	685 (11%)	

Hence, no statistical analysis for this effect was needed. Overall, RTs for dominant alternatives were faster than RTs for subordinate alternatives (F1(1,59)= 6.5, MSe = 7131, *p*< 0.01; F2(1,38) = 5.5, MSe = 6199, *p*<0.02; for the subject and stimuli analyses respectively). This is a well documented frequency effect in the matching task (Frost & Katz, 1989; Frost et al. 1990), and merely reflects the slower matching of printed and spoken words that have lower frequency of occurrence.

The results of Experiment 3 strongly support the results of Experiment 1 and 2. The presentation of the spoken subordinate alternative in conjunction with the unpointed heterophonic homograph did not slow the matching process relative to the unequivocal printed presentation. This outcome suggests that *both* phonologic alternatives were generated from the ambiguous letter string. Apparently, when the dominant or subordinate spoken words were presented auditorily (in the matching task the auditory information effectively lags behind the visual information), they matched one of the phonological representations that had already been generated from the print.

EXPERIMENT 4

The aim of Experiment 4 was to examine whether the two phonologic alternatives remain active even at a longer SOA so as to test the limits of the dual activation process. For this purpose the design of Experiment 3 was repeated, but the auditory presentation was delayed by 500 ms relative to the onset of the visual word. Since the auditory presentation in itself was distributed over 400 to 500 ms of time on the average, Experiment 4 examined the activation of the two phonologic alternatives at approximately 800 ms from visual stimulus onset.

Method

Subjects. Sixty undergraduate students, native Hebrew speakers, participated in the experiment for course credit or for payment. None of the subjects had participated in the previous experiments. The stimuli design and apparatus were identical to those of Experiment 3.

Results and discussion

RTs in the different experimental conditions are shown in Table 4. The results with the lagged presentation clearly differ from the results with the simultaneous presentation: Decision latencies with unpointed print were overall slower than decision time with pointed print. This effect was strongest for the subordinate alternatives.[4]

Table 4. *RTs and (errors) for pointed and unpointed dominant and subordinate alternatives in the matching task. Auditory presentation is delayed by 500 ms relative to the onset of the visual presentation.*

Visual presentation	Auditory presentation	
	Dominant	Subordinate
Unpointed	612 (5%)	687 (4%)
Pointed	600 (4%)	633 (7%)
No Match fillers	638 (11%)	

The statistical significance of these results was assessed by an analysis of variance (ANOVA) across subjects (F1) and across stimuli (F2), with the main factors of visual presentation (pointed, unpointed), and dominance of spoken alternatives (dominant, subordinate). Both main effects were significant [F1(1,59)=53.6, MSe=1412, *p*<0.001; F2(1,39)=70.0, MSe=641, *p*<0.00, for visual presentation; and F1(1,59)= 57.2, MSe= 3033, *p*<0.001; F2(1,39)= 27.5, MSe=4350, *p*<0.001, for dominance]. The interaction of visual presentation and dominance was significant as well [F1(1,59)= 19.6, MSe= 1247, *p*< 0.001; F2(1,39)= 25.3, MSe= 664, *p*<0.001].

The relatively slower RTs for unpointed as compared to pointed print for subordinate alternatives suggest that they were less activated at 800 ms from stimulus onset, relatively to the dominant alternatives. Consequently, the following presentation of the spoken subordinate alternative presented auditorily resulted often in a mismatch that caused slower RTs in the unpointed presentation relative to the explicit pointed presentation. This suggestion is further supported by the somewhat slower RTs for unpointed relative to pointed print when the dominant alternatives were presented auditorily. Since the subjective dominance ratings were not always unanimously accepted, these slower RTs were probably due to several trials in which, for some subjects, the subordinate alternatives were in fact considered as the dominant ones and vice-versa. In these few cases, the following auditory presentation of the dominant spoken alternatives resulted in mismatches that caused the overall slower latencies in the unpointed condition.

GENERAL DISCUSSION

The present study examined the processing of Hebrew heterophonic homographs by employing two experimental tasks that directly tap phonetic and phonologic processing. In Experiments 1 and 2, the detection of speech in amplitude modulated noise served as a tool for directly measuring the amount of phonetic activation that developed following the visual presentation of the printed homograph. The extent of phonetic activation was reflected in a response bias to report speech in the noise when the masked speech was accompanied by matching print. In Experiments 3 and 4, RTs for matching the pointed and unpointed printed forms of heterophonic homographs to the dominant and subordinate spoken alternatives reflected the relative activation of the two phonologic structures represented by the ambiguous letter string.

The results of Experiment 1 demonstrated that the pairing of a printed heterophonic homograph with an auditory presentation of both dominant and subordinate masked spoken alternatives led to a similar bias towards a "yes" response. This outcome suggests that subjects derived from the ambiguous letter string two phonetic representations that were subsequently matched to the dominant and subordinate amplitude envelopes presented via the auditory channel. In Experiment 2 the explicit presentation of the dominant and subordinate phonological alternatives using the vowel marks produced bias effects that were very similar to the effects obtained in the unpointed presentation. Moreover, the inclusion of the countermatch condition clarified that the two phonetic representations that were derived from the letter string were clearly distinct, and that the bias obtained for both dominant and subordinate alternatives did not result from a possible acoustic similarity between the amplitude envelopes of the two spoken alternatives.

The results of Experiments 3 and 4 employing the matching task, supported the conclusions of Experiments 1 and 2. If in Experiments 1 and 2 the activation of both phonologic alternatives was inferred from responses to amplitude modulated noises that could not be consciously recognized as words by the subjects, in Experiments 3 and 4 the two spoken forms of the ambiguous letter strings were explicitly and clearly presented. The identical matching latencies of pointed and unpointed print to both dominant and subordinate spoken alternatives in Experiment 3 suggest that both alternatives were indeed generated from the ambiguous consonant cluster. This conclusion is further supported by the contrasting results of Experiment 4, that revealed a much larger activation of the dominant alternatives relatively to the subordinate alternatives, some 800 ms from visual stimulus onset.

The results of the present study thus suggest that both phonological alternatives were activated following the presentation of the printed homographs. This outcome converges with previous studies that examined the activation of dominant and subordinate meanings of homophonic homographs (e.g., Simpson & Burgess, 1985) and of heterophonic homographs (Frost & Bentin, 1992). Using a semantic priming paradigm, Simpson and Burgess (1985) demonstrated that the dominant and subordinate meanings of heterophonic homographs are both active at SOAs ranging from 100 to 250 ms from stimulus onset. However, more relevant to the present study are the recent results reported by Frost and Bentin (1992), who employed a semantic priming paradigm similar to that used by Simpson and Burgess (1985), but with Hebrew heterophonic homographs. Isolated ambiguous consonant strings were presented as primes, and the visual targets which were related to only one of their possible meanings, followed the ambiguous primes at 100 ms, 250 ms, and 750 ms SOA. The results demonstrated that, at SOAs of 250 ms or longer, lexical decision for targets that were related either to the dominant or to the subordinate phonological alternative were facilitated.

The strong bias effect obtained for both dominant and subordinate phonological alternatives in Experiment 1 suggests that the visual presentation of heterophonic homographs resulted not only in the activation of the two meanings related to it, as reported by Frost and Bentin (1992), but also in the activation of the two phonetic alternatives that the letter string represents. Whether this activation reflects a mandatory and automatic process of phonologic recoding of the ambiguous printed word cannot be unequivocally determined by the present results. It is possible that the presentation of auditory stimuli served to *elicit* phonological processing of the printed stimuli. Although we cannot rule out this possible interpretation, we find it less probable. In Experiments 1 and 2 subjects were explicitly instructed not to base their auditory judgments on the visual information, and were informed of the equal distribution of "yes" and "no" trials in the different visual conditions. Nevertheless, the large numbers of false alarms in the matching condition suggests that they could not avoid being influenced by the visual information.

The findings that both phonetic alternatives were activated following the printed presentation provides further confirmation for the notion of fast phonetic recoding in reading offered by a large literature in visual word perception (e.g., Perfetti, Bell, & Delaney, 1988; Van Orden, Johnston, & Halle, 1988), but often challenged by those who find evidence suggesting that printed words activate orthographic units that are directly related to meanings in semantic memory (see Van Orden, Pennington, & Stone, 1990, for a review). What the present findings of Experiments 1 and 2 teach us is that subjects generate the two phonetic representations related to a phonologically ambiguous letter string, even when the experimental task (speech detection in noise) does not require any processing of the printed stimulus. This conclusion is in accordance with the claim put forward by Frost and Bentin (1992), who suggested that the activation of the two phonological entries of heterophonic homographs precedes the activation of their meanings. Frost and Bentin based their conclusions on findings showing a different onset of meaning activation for heterophonic than for homophonic homographs, and on more robust priming effects observed when the primes were heterophonic homographs than when they were homophonic homographs. Since both homophonic and heterophonic homographs address two meanings in semantic memory, models of direct access from print to meaning could not account for the differences in meaning activation obtained for these two types of letter strings. These differences can be accounted for only by assuming that the processing of heterophonic homographs first involves phonological disambiguation. The results of the present study confirm that the phonological processing of heterophonic homographs is characterized by the generation of both of the phonetic representations that the letter string represents.

One important question related to the activation of the two phonological alternatives refers to their relative level of activation with simultaneous and delayed presentation. Note that, in contrast to semantic priming experiments, in which the SOA between visual primes and targets can be explicitly manipulated and monitored, the matching task cannot provide evidence for the exact onset of activation of the two phonologic alternatives. Moreover, unlike the study reported by Frost and Bentin (1992), in which SOA was systematically manipulated, Experiments 3 and 4 examined the activation of the two phonologic alternatives only with simultaneous and delayed presentations. However, the overall range of onset activation can be estimated. The stimuli employed in the present experiments consisted mainly of two-syllable words, and the duration of the auditory presentation ranged from 400 to 500 msec on the average. Assuming that some of the responses were given as soon as the auditory presentation reached the word's recognition point (see Marslen-Wilson, 1987), an overall estimate of the time course of activation of the two phonological alternatives in Experiments 1 and 3 suggests that they were both active at 200 to 500 ms SOA. This estimation, however, cannot rule out the possibility that the two alternatives were available before 250 ms, and remained active later than 500 ms from stimulus onset. The results of Experiment 4 suggest that, at approximately 800 ms from stimulus onset, the activation of subordinate alternatives decayed relatively to the activation of the dominant alternatives.

Another finding of the present study was the very small bias effect obtained in the countermatch condition of Experiment 2. This result suggests that the envelope representations generated from the print were very detailed. The ability of subjects to automatically generate detailed amplitude envelopes from verbal information presented in the visual modality was shown previously in studies that employed printed words (Frost et al., 1988; Frost, 1991) and in a study that employed speechreading (Repp, Frost, & Zsiga, 1992). In these studies the words in the matching and nonmatching condition had the same phonologic structure and stress pattern, but did not share phonemes in the same location. The subjects were shown to be able to discriminate between a representation they generated from a word presented visually and between the envelope of a nonmatching word presented auditorily. This discrimination was reflected by a lower effect of bias in the nonmatching condition relatively to the matching condition. The words in the countermatch condition in the present study were the two different realizations of heterophonic homographs. Thus, although in some pairs there were differences in stress patterns, they shared the same consonants and in many cases they differed only by a single vowel. Nevertheless, it appears that the envelope representation generated from the printed dominant alternative was not confused with the envelope of the subordinate alternative presented auditorily, and vice versa. This outcome suggests a surprising ability of subjects to generate a detailed and specific phonetic representation from printed words .

In conclusion, the present study suggests a novel tool for examining phonetic processing of print. The task of detecting speech embedded in amplitude modulated noise, when matching or nonmatching print is presented simultaneously in the visual modality, has been shown to be able to monitor the generation of phonetic representations from the printed stimuli (Frost et al., 1988; Frost, 1991; but see also Repp et al., 1992). The advantage of the speech detection task is that phonetic processing is not inferred from responses to printed words following orthographic or semantic experimental manipulations, but is measured indirectly in a task that does not explicitly require any response to the printed word. The results from this task converge with the results of the matching task that taps the conversion of printed words into phonologic structures.

REFERENCES

Duffy, S. A., Morris, R. K ., & Rayner, K. (1988). Lexical ambiguity and fixation time in reading. *Journal of Memory and Language, 27*, 429-446.

Forster, K. I., & Bendall, E. S. (1976). Terminating and exhaustive search in lexical access. *Memory & Cognition, 4*, 53-61.

Frost, R. (1991). Phonetic recoding of print and its effect on the detection of concurrent speech in amplitude modulated noise. *Cognition. 39*, 195-214.

Frost, R., & Bentin, S. (1992). Processing phonological and semantic ambiguity: Evidence from semantic priming at different SOAs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*. 58-68

Frost, R., Feldman, L. B., & Katz, L. (1990). Phonological ambiguity and lexical ambiguity: Effects on visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*. 569-580.

Frost, R., & Katz, L. (1989). Orthographic depth and the interaction of visual and auditory processing in word recognition. *Memory & Cognition, 10*, 302-311.

Frost, R., Repp. B. H., & Katz, L. (1988). Can speech perception be influenced by a simultaneous presentation of print? *Journal of Memory and Language, 27*, 741-755.

Glucksberg, S., Kreuz, R. J., & Rho, S. H. (1986). Context can constrain lexical access: Implications for models of language comprehension *Journal of Experimental Psychology: Lerrning, Memory, & Cognition, 12*, 323-335.

Gordon, P. C. (1988). Induction of rate-dependent processing by coarse-grained aspects of speech. *Perception & Psychophysics, 43*, 137-146.

Hogaboam, T. W., & Perfetti, C. A. (1975). Lexical ambiguity and sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 14*, 265-274 .

Jared, D., & Seidenberg, M. S. (1991). Does word identification proceed from spelling to sound to meaning? *Journal of Experimental Psychology: General. 120*, 358-394.

Koriat, A. (1981). Semantic facilitation in lexical decision as a function of prime-target association. *Memory & Cognition, 9*, 587- 598.

Kroll, J. F., & Schweickert, J. M. (1978). Syntactic disambiguation of homographs. Paper presented at the Nineteenth Annual Meeting of the Psychonomic Society, San Antonio, Texas.

Luce, R. D. (1963). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*, 71-102.

Neely, J. H. (1990). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. Humphreys (Eds.), *Basic Processes in Reading: Visual Word Recognition*. Hillsdale, NJ: Erlbaum.

Neill, W. T., Hilliard, D. V., & Cooper, E. (1988). The detection of lexical ambiguity: Evidence for context-sensitive parallel access. *Journal of Memory and Language, 27*, 279-287.

Onifer, W., & Swinney, D. A. (1981). Accessing lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. *Memory & Cognition, 9*, 225-236.

Perfetti, C. A., Bell, L. C., & Delaney, S. M. (1988). Automatic (prelexical) phonetic activation in silent word reading: Evidence from backward masking. *Journal of Memory and Language, 27*, 59-70.

Repp, B. H., Frost, R., & Zsiga, E. (1992). Lexical mediation between sight and sound in speech reading. *Quarterly Journal of Experimental Psychology*.

Schroeder, M. R. (1968). Reference signal for signal quality studies. *Journal of the Acoustical Society of America, 43*, 1735-1736.

Schvaneveldt, R. W., Meyer, D. E., & Becker, C.A. (1976). Lexical ambiguity, semantic context, and visual word recognition. *Journal of Experimental Psychology: Human Perception & Performance, 2*, 243-256.

Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition, 19*, 1-30.

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. M., & Bienkowski, M. (1982). Automatic access of the meaning of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology, 14*, 489-537.

Seidenberg, M. S., Waters, G. S., Barnes, M., & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior, 23*, 383- 404.

Simpson, G. B. (1981). Meaning dominance and semantic context in the processing of lexical ambiguity *Journal of Verbal Learning and Verbal Behavior, 20*, 120-136.

Simpson, G B. (1984) Lexical ambiguity and its role in models of word recognition. *Psychological Bulletin, 96*, 316-340

Simpson, G. B., & Burgess, C. (1985). Activation and selection processes in the recognition of ambiguous words. *Journal of Experimental Psychology: Human Perception & Performance, 11*, 28-39.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*, 645-660

Tanenhaus, M K., Leiman, J. M., & Seidenberg, M S. (1979) Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior, 18*, 427-440.

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound, and reading. *Memory & Cognition, 10,* 434-442.

Van Orden, G. C., Johnston, J. C., & Hale, B., L. (1988). Word identification in reading proceeds from spelling to sound to meaning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 371-386.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review, 97,* 488-522.

Van Petten, C., & Kutas, M. (1987). Ambiguous words in context: An event-related Potential analysis of the time course of meaning activation. *Journal of Memory and Language, 26,* 188-208.

Van Tassel, D. J., Soli, S. D., Kirby, V. M., & Widin, G. P. (1987). Speech waveform envelope cues for consonant recognition. *Journal of the Acoustical Society of America, 82,* 1152-1161.

Wood, C. C. (1976). Discriminability, response bias, and phoneme categories in discrimination of voice onset time. *Journal of the Acoustical Society of America, 60,* 1381-1389.

FOOTNOTES

[*] *Journal of Experimental Psychology: Learning, Memory, & Cognition* (1993).

[†] Department of Psychology, The Hebrew University, Jerusalem, Israel.

[1] We cannot account for the specific mechanism responsible for the bias effect. We can only suggest that it is the convergence of visual and auditory information that is responsible for it. The illusion to detect speech in the noise is not restricted to the presentation of print and speech. Repp, Frost, & Zsiga (in press) have shown similar effects with lipreading, when a visual presentation of a speaker's was employed instead of printed words.

[2] This almost chance level of detection is very similar to the detectability scores reported by Frost (1991) with identical signal-to-noise ratio. But note that the bias for a "yes" response is not affected by the poor level of detection (in fact, effects of bias can emerge only when detection is imperfect). Frost et al. (1988) showed significant bias effects in the speech detection task over a wide range of signal-to-noise ratios.

[3] As in all the previous studies that used the detection task, a stimulus analysis was not carried out. This is because in this task the auditory stimuli cannot be identified by the subjects, and because each stimulus is rotated and repeated in the various auditory forms several times for each subject across all experimental conditions.

[4] Since RTs were measured from the onset of the auditory presentation, the faster RTs in Experiment 4 relative to Experiment 3 reflect the trivial fact that subjects have already processed the visual stimulus for 500 ms before the onset of RT measurement.

Haskins Laboratories Status Report on Speech Research
1992, SR-111/112, 135-146

Reading Consonants and Guessing Vowels: Visual Word Recognition in Hebrew Orthography*

Ram Frost[†] and Shlomo Bentin[†]

For many years studies in the English language have dominated experimental research in visual word recognition. This state of affairs cannot be accounted for by considering merely geographic reasons. Rather, it was partly due to an underlying belief that English was sufficient because reading processes (as well as other cognitive processes) are universal. In recent years, however, studies in orthographies other than English have become more and more prevalent. These studies have the common view that reading processes cannot be explained without considering the reader's linguistic environment. Moreover, it is assumed that reading strategies in one orthography can be understood better when other orthographies provide additional points of reference. It is in this context that recent research in reading Hebrew should be evaluated. In the present chapter we describe the specific characteristics of Hebrew orthography and discuss their origin with regard to the complex morphology of the Hebrew language. We further examine their possible effects on the reading strategies adopted by beginning and skilled readers. Finally, we discuss the processing of morphologic information conveyed by Hebrew print, a particularly interesting contrast to other writing systems that have been studied.

CHARACTERISTICS OF THE HEBREW ORTHOGRAPHY

The orthography of the Hebrew language should be described in reference to its very complex productive morphology (see Katz & Frost, this volume). In Hebrew, as in other Semitic languages, all verbs and the vast majority of nouns and adjectives are comprised of roots which are usually formed of three (sometimes four) consonants.

The three-consonant roots are embedded in pre-existing morphophonological word patterns to form specific words. Phonological patterns can be either a sequence of vowels or a sequence consisting of both vowels and consonants. Thus, in general, Hebrew words can be decomposed into two abstract morphemes, the root, and the phonological pattern. Roots and phonological patterns are abstract structures and only their joint combination (after the application of phonological and phonetic rules) forms specific words. Although these morphemes carry some semantic and morpho-syntactic information, their meaning is often obscure and changes for each root-pattern combination (see Berman, 1980). This is because there are no unequivocal rules for combining roots and phonological patterns to produce specific word meanings. For example, the word KATAVA ("a newspaper article") is composed of the root KTV, and the phonological pattern -A-A-A (the lines indicate the position of the root consonants). The root KTV alludes to anything related to the concept of writing, whereas the phonological pattern A-A-A is often (but not always) used to form nouns that are usually the product of the action specified by the root. It is the combination of both root and word pattern that forms the word meaning "article". Other phonological word patterns may combine with the same root to form different words with different meanings that can be closely or very remotely related to writing. For example, the word KATAV ("press correspondent") is formed by combining the root KTV with the phonologic pattern -A-A-. The phonological pattern -A-A- carries the morpho-syntactic information that the word is a noun which signifies a profession. But

This work was supported in part by National Institute of Child Health and Human Development Grant HD-01994 to Haskins Laboratories. We thank Avital Deutch for her insightful comments on earlier drafts of this paper.

this same phonological pattern is also common in adjectives that signify attributes. Unlike KATAV, the word KTOVET ("address") is formed by combining the same root with a phonological pattern that includes consonants as well as vowels. This pattern carries the morpho-syntactic information of that the word is a feminine noun. Note that the same phonologic pattern can be applied to other roots resulting in various different verbs or nouns, each of which is related to its respective root action. Therefore only the combination of both root and phonological pattern specifies the exact meaning of a word.

Although words in Hebrew are composed of two morphemes, the root and the phonologic pattern, the semantic information conveyed by each morpheme is not equally constraining; the semantic information specified by the root is by far more restricted and more specific than that specified by the phonologic word pattern, and it conveys the core meaning of the word. The word pattern, on the other hand, in many cases carries nothing more than word class information. Therefore, one might assume that the understanding of spoken language is based primarily on the identification of the root. Although speculative, it can be reasonably suggested that this morphologic decomposition characteristic of the Semitic languages had directly influenced the development of the Semitic writing systems.

Because of the productive characteristic of Hebrew morphology, Hebrew orthography was designed to convey to the reader primarily the root information (see Katz & Frost, this volume). Hence, the letters in Hebrew represent mainly consonants. The vowels are depicted by diacritical marks (points and dashes) presented beneath (sometimes above) the letters. Although the diacritical marks carry mainly vowel information, they also differentiate in some instances between fricative and stop variants of consonants. In modern Hebrew, we have lost most of the phonetic differentiation between fricative and stop pronunciations, but it is still kept for 3 consonants, in which the letter indicates two different phonetic realizations of these phonemes: /b/→[b] or [v], /p/→[p] or [f], and /k/→[k] or [x]. In these cases a point is inserted inside the letter to indicate the stop pronunciation. Thus the presentation of vowels reduces considerably several aspects of phonemic ambiguity. The diacritical marks, however, are omitted from most reading material, and can be found only in poetry, children's literature, and religious scripts. Although some of the vowels can also be conveyed by letters, these letters are not regularly used, and are considered optional. Thus the most salient characteristic of the Hebrew orthography is that it presents the reader with only partial phonological information. However, incomplete phonologic information is only one specificity of the Hebrew orthography. Because the same root may be combined with different word patterns, frequently the vowel-sequence is the only difference between several words. Therefore, when the vowel marks are omitted, the same string of letters sometimes denotes up to seven or eight different words. Consequently the Hebrew reader is normally exposed to phonological as well as semantic ambiguity. An illustration of Hebrew unpointed and pointed print is presented in Figure 1.

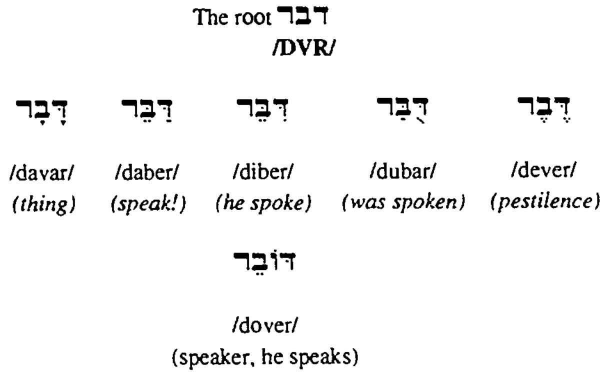


Figure 1. Phonologic ambiguity in unpointed Hebrew print.

The Figure describes the possible reading of one consonant cluster. The unpointed letter string "דבר" has five meaningful possible readings. The letter "ב" can be read either as [v] or [b] which are distinguished by a dot that appears within the letter, but only in pointed print. The triconsonantal root "דבר" can, thus, be read as /dvr/ or /dbr/ and forms 3 clusters of words: Three words inflected from the root /dbr/ which signifies the action of speaking, and two words /davar/ and /dever/ which share the same consonants, but originated historically from different languages, and therefore do not share any semantic features (the former meaning "a thing" while the latter means "pestilence"). An example of a phonologic pattern which is conveyed by letters in addition to diacritical marks underneath the consonants can be seen in the word /dover/. Note that in the present tense of the root /dbr/ the pronunciation of the middle phoneme /b/ changes into a /v/. These interchanges between fricative and stop pronunciations of consonants are very common in Hebrew. For /dover/ the phoneme /o/ is conveyed by the letter ו. In its unpointed form, this letter can represent the phoneme /v/ as well.

The introduction of vowels marks in printed Hebrew

For the above reasons, Hebrew orthography was designed to provide the reader with the abstract root information, regardless of the possible words that the letter string might represent. Thus, the unpointed orthography served the purpose of denoting in print the optimal amount of phonologic information. The gain in omitting the vowels from the print was multiple: First, the set of letters in the alphabet was smaller—Hebrew has only 22 letters, and the written words were shorter. Second, the presentation of consonants alone made the abstract root more salient. Indeed, the original Hebrew writing system was unpointed. It remained unpointed as long as Hebrew was a living language, that is until the second century.

It was only between the second and the tenth century that the vowel marks were introduced into Hebrew orthography (see Morag, 1972). Since after the second century most of the Jewish nation was dispersed in Europe, Asia, and Africa, they no longer spoke Hebrew as their native language. For the fear that the correct pronunciation of the Hebrew words in the holy scriptures might be forgotten, the vowel marks were introduced. The point of interest in this historical analysis is that the vowel marks were not necessary when Hebrew was a living spoken language. Their function of denoting the specific pronunciation of words became a necessity only when Hebrew ceased to be a naturally spoken language. It is worth noting that the vowel marks were used only for writing holy scriptures or poetry. This is because it is only for poetry and religious scripts that the exact phonemic notation is indeed crucial. Nevertheless, as will become evident in the next section, the importance of vowel marks for both beginning and skilled readers is incontestable.

The use of vowel marks by the beginning reader

Vowel marks aid phonologic recoding. Aside from poetry and religious texts, most children's literature in Hebrew is pointed. Traditionally, most schools in Israel have adopted methods of teaching reading which involve the use of vowel marks at the initial stages of reading acquisition. The purpose of this method is two-fold. First, the vowels convey the unequivocal phonemic structure of the printed word to the beginning reader. It is well established today that beginning readers recognize and name printed words through a process of phonological mediation (e.g., Calfee, Chapman, & Venezky, 1972; Conrad, 1972; Shankweiler & Liberman, 1976). Moreover, decoding skills were shown to be a developmental prerequisite for efficient reading for meaning (e.g., Perfetti & Hogaboam, 1975). Phonemic recoding based on grapheme-to-phoneme conversion rules is very simple in pointed Hebrew. In fact, in its pointed form, Hebrew orthography is almost as shallow as the Serbo-Croatian orthography (Katz & Frost, 1992) and allows a simple use of prelexical phonologic processing. Without the vowel marks the beginning reader in Hebrew would have to rely on the holistic identification of consonant clusters and their correspondence to spoken words, which as mentioned above is extremely ambiguous.

Vowel marks affect phonologic awareness. A second gain in teaching children to read with vowel marks is their beneficial effect on the development of phonological awareness. Phonological awareness is the ability to consciously recognize the internal phonemic structure of spoken words (Bentin, this volume). Several authors reported that the ability to manipulate phonemic segments consciously, develops only around the first grade in elementary school (e.g., Liberman, Shankweiler, Fisher, & Carter, 1974), and has been positively correlated with reading ability (e.g., Bertelson, 1986; Bradley & Briant, 1983; 1985; Liberman & Shankweiler, 1985). This correlation was used to develop methods for predicting in kindergarten, how efficiently would the children acquire the reading skills in school (Lundberg, Olofsson, & Wall, 1980; Mann, 1984). Recently, the importance of phonological awareness for reading was demonstrated also in Hebrew (Bentin & Leshem, in press). In that study, the authors found that children who scored low on a phonemic awareness battery administered in kindergarten scored also low on a reading test in school. However, if those children were trained in kindergarten and improved their segmentation skills, they reached the school standards and read as well as children who had scored highly on the initial tests of phonological awareness.

The relationship between phonological awareness and reading is not, however, unidirectional. Several studies have suggested that, in the absence of reading instruction, the ability to isolate and manipulate single phonemes in coarticulated syllables is obstructed (e.g., Bertelson & de Gelder, 1990). Apparently, by being exposed to the alphabetic principle, children become aware that letters are usually mapped into single phonemes rather than into coarticulated phonological units. The emergence of this revelation should be facilitated when the relationship between letters and phonemes is simple and isomorphic (as in a shallow orthography) than when it is complex or partial (as in a deep orthography). The addition of the vowel marks to the consonants changes the Hebrew orthography from being deep to being almost as shallow as Serbo-Croatian or Italian. Therefore, by using the pointed print, teachers help triggering phonemic awareness that is essential for efficient reading acquisition.

The processing of consonants and vowel marks by the skilled reader

A question of great interest in the study of word recognition in Hebrew is how vowel marks are processed by the skilled reader. From the beginning of the third grade children are gradually exposed to unpointed print and by the sixth grade they encounter unpointed print almost exclusively. What is, then, the possible purpose of vowel marks for the skilled reader? How are they processed in print? This question is of special interest because it is often assumed that mature readers rely on fast visual-orthographic cues rather than on phonologic recoding in word recognition (see McCusker, Hillinger, & Bias, 1981, for a review).

Skilled readers cannot disregard vowel information in print

Navon and Shimron (1981) were the first to examine the use of vowel marks by skilled readers in Hebrew print. Interested to see whether readers can disregard the vowel marks while making lexical decisions, they presented undergraduate subjects with pointed letter strings, and instructed them to ignore the vowel marks while making word/nonword discriminations. Their results showed that positive decisions were slowed when the consonants formed a legal word while the marks underneath the letters suggested an incorrect vowel configuration. Consequently, Navon and Shimron (1981) concluded that although the Hebrew skilled reader does not need the vowel marks for fast lexical decisions he or she cannot ignore them even when instructed to do so (see also Navon & Shimron, 1984).

One point of interest in Navon & Shimron's study relates to the recognition of the correct vowel marks in Hebrew. Although modern Hebrew differentiates only between five vowels (/a/, /e/, /i/, /o/, /u/) it has more than five vowel marks. When the vowel marks were introduced into Hebrew between the second and the tenth century, the vocalization system that became the most influential originated from the Tiberias region. This system had two notations for /a/ (т,-) and two notations for /e/ (∵ , ..). These notations probably reflected a Hebrew dialect that was spoken in the northern part of the country and had seven rather than five vowels. Although this dialect had become extinct, the printed notations for these vowels are still used in modern Hebrew and used consistently according to orthographic rules (Morag, 1972). Navon & Shimron's results demonstrated that the Hebrew reader is not sensitive to interchanges in the printed forms of the two vowel marks representing /a/ or the two vowel marks representing /e/, as long as the correct phonemic structure of the word is maintained. This is of special interest because similar ambiguity exists with current Hebrew consonants. Hebrew has two letters representing each of the phonemes /t/, /k/, and /kh/. Similar to the vowels, these letters also reflect a historical distinction between phonemes, a distinction without phonetic reality in modern Hebrew. Nevertheless, in contrast to the insensitivity of the reader to the alternative forms of the vowel marks, the skilled reader makes very few errors in lexical decision when the letters representing these consonants are interchanged. This probably reflects the relative importance given by the skilled reader to the consonants as opposed to the vowel marks.

The inability of Hebrew readers to disregard vowel information was further examined in a study that employed the repetition priming paradigm (Bentin, 1989). In this study subjects were required to make lexical decisions to words and nonwords that were either pointed or unpointed. Orthographic, phonemic, and identity repetitions were examined at lags 0 and 15. Orthographic repetition consisted of a second presentation of the consonants but with different vowel marks. Phonemic repetition consisted of repeating the phonemes but with different letters (Hebrew has several pairs of letters that denote the same phoneme). The results showed differential effects of phonemic and orthographic repetition for pointed and unpointed print. For unpointed print, all three forms of repetition affected lexical decisions at lag 0, whereas at lag 15 only identity repetition was effective. With pointed print, on the other hand, phonemic repetition had a significant effect at lag 15, but orthographic repetition did not. These results suggest that the vowel marks indeed attracted subjects' attention and induced phonologic coding of the printed words. Because the same phonemic cluster appeared at the second presentation, it was recognized faster even though the orthographic spelling referred to a different meaning. When the vowels were not presented to the reader, he or she was encouraged to access the lexicon through a visual-orthographic code, and the effects of phonemic repetition disappeared.

Naming unpointed print involves postlexical phonology

Although the vowels convey to the reader unequivocally the phonemic structure of a printed word, for many words the vowel marks are not essential for locating a specific lexical entry. For these words the consonant structure is sufficient for specifying a unique word. This is because in such cases, only one phonologic pattern can be assigned to the letter string to create a meaningful word. But even considering the prevalence of phonologic ambiguity in Hebrew, the skilled reader does not need the vowel marks for fast reading. A comparison of lexical decision time in the deep unpointed Hebrew orthography and in the very shallow Serbo-Croatian orthography revealed similar, almost identical, performance (Frost, Katz, & Bentin, 1987). Being exposed to unpointed print almost exclusively, the skilled reader in Hebrew has developed reading strategies that allow him to generate the missing vowel information in the print using the lexical route following visual lexical access. This hypothesis was confirmed by a cross-language study that compared naming strategies in deep and shallow orthographies (Frost et al., 1987). In this study lexical decisions and naming performance were examined in unpointed Hebrew, in English, and in Serbo-Croatian. The results showed that, in Hebrew, the lexical status of the word (being a high-frequency word, a low-frequency word, or a nonword) had similar effects on naming and on lexical decision, suggesting that pronunciation was achieved by an addressed routine in which the whole word phonology is retrieved from lexical memory. The lexical status of the word had smaller effects on naming in English and even smaller effect on naming in Serbo-Croatian. Similar results were obtained in a second experiment that showed stronger semantic priming effects on naming in Hebrew relative to English and Serbo-Croatian, again suggesting stronger involvement of the lexicon in naming unpointed words.

Lexical decisions in unpointed print are based on fast orthographic recognition

The use of the lexical route in processing Hebrew print was also demonstrated by Koriat (1984), who examined lexical decision latencies for pointed and unpointed letter strings. In his study, Koriat used Hebrew words that had only one meaningful pronunciation in their pointed form, and found almost identical lexical decision latencies for pointed and unpointed words. Moreover, the presentation of vowel marks had similar effects on words of different length. Koriat has therefore concluded that lexical access in Hebrew is probably visual and direct, not involving phonologic mediation. In a subsequent study, however, Koriat (1985) found that the presentation of vowel marks had some beneficial effect on lexical decisions. The advantage of pointed print was larger for low-frequency words than for high-frequency words, suggesting that the use of prelexical phonology is more prevalent for infrequent words. To summarize Koriat's work, it appears that despite his initial conclusions, his data indicate that the presence of vowel marks affects visual word recognition. This evidence, however, was inconclusive.

Additional and more convincing evidence suggesting that lexical decisions in Hebrew do not involve deep phonologic processing of the printed word, emerges from studies that employed words with two meaningful pronunciations (Bentin, Bargai, & Katz, 1984; Bentin & Frost, 1987). Bentin et al. (1984) examined naming and lexical decision for unpointed consonantal strings. Some of these strings could be read as two words whereas some could only be read as one word only. The results demonstrated that phonologic ambiguity affected naming but not lexical decision performance: Naming phonologically ambiguous strings was slower than naming unambiguous ones. In contrast, phonologically ambiguous letter strings were recognized as fast as letter strings with only one meaningful pronunciation. These results suggested that, although the reader of Hebrew is indeed sensitive to the phonologic structure of the orthographic string when naming is required, lexical decisions are based on a fast familiarity judgment of the consonantal cluster and do not require a detailed phonological analysis of the printed word.

These conclusions were further supported by Bentin and Frost (1987). In this study subjects were presented with phonemically and semantically ambiguous consonantal strings. Each of the ambiguous strings could have been read either as a high-frequency word or as a low-frequency word, depending on the vowel configuration which was assigned to it. Lexical decision time for the unpointed ambiguous consonantal string was compared to lexical decision time for the unequivocal pointed printed forms of the high- or the low-frequency phonological alternatives. The results showed that lexical decisions for the unpointed ambiguous strings were faster than lexical decisions for either of their pointed (and therefore disambiguated) alternatives; explicit presentation of vowel marks did not necessarily accelerate lexical decision time. This result suggests that lexical decisions for Hebrew unpointed words may occur *prior* to the process of phonological disambiguation at least when the letter string represents two different words. In this case, the decisions are probably based on the printed word's orthographic familiarity (cf. Balota & Chumbley, 1984; Chumbley & Balota, 1984). On the basis of those studies we suggest that lexical decisions in Hebrew involve neither a prelexical nor a postlexical phonologic code. They are probably based upon the *abstract* linguistic representation that is common to several phonologic and semantic alternatives. Thus, in addition to a phonologic lexicon the Hebrew reader probably develops an "interface" lexical system that is based on consonantal strings common to several words. Whether the entries in this interface lexicon are orthographic (letters) or phonologic (phonemes that represent the consonants) in nature is hard to determine. Nevertheless, lexical processing occurs, at a first phase, at this morphophonological level. The reader accesses the abstract string and recognizes it as a valid morphologic structure. Lexical decisions are usually reached at this early stage and do not necessarily involve further phonological processing. This possibility is depicted in Figure 2. Although lexical decisions in Hebrew might be based on abstract orthographic representations, there is no doubt that the process of word identification continues until one of several phonological and semantic alternatives is finally determined. This process of lexical disambiguation was more clearly revealed by using the naming task. Bentin and Frost (1987) investigated the process of selecting specific lexical candidates by examining the naming latencies of unpointed and pointed words. The complete phonological structure of the unpointed word that is necessary for naming can only be retrieved postlexically, after one word candidate has been accessed. The selection of a word candidate is usually constrained by context, but we found that in the absence of context it is based on word-frequency. In contrast to lexical decisions, we found that naming ambiguous unpointed strings was just as fast as naming the most frequent pointed alternative, and that the pointed low-frequency alternative was the slowest. In the absence of constraining context, the selection of one lexical candidate for naming seems to be affected by a frequency factor: the high-frequency alternative is selected first.

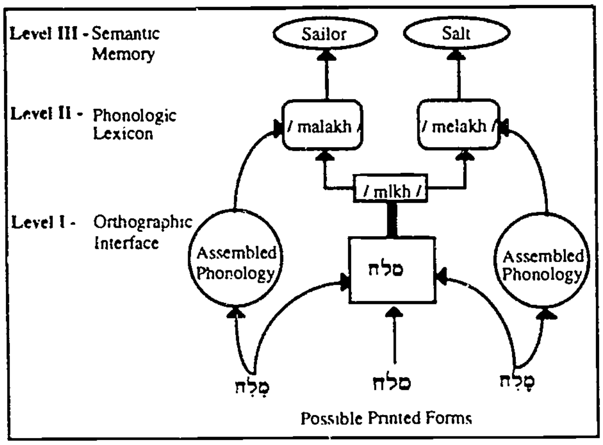


Figure 2. A model of the lexical structure of the Hebrew reader and the possible processing of pointed and unpointed printed words. Naming of phonologically ambiguous words is affected by frequency factors

Naming in pointed Hebrew also involves prelexical phonologic recoding

Another set of experiments recently completed in our laboratory (Frost, in press) provides important insight regarding the use of vowel marks by the skilled reader. In this study subjects were presented with consonantal strings which were followed by vowel marks appearing at different stimulus onset asynchronies (SOA). The vowel marks were superimposed on the consonants at SOAs ranging from 0 ms (simultaneous presentation) to 300 ms from the onset of consonant presentation. In one condition the letter strings represented only one meaningful word, and in another condition the letter strings could represent two meaningful words. Subjects were required either to make lexical decisions or to name the words and nonwords on the computer screen as fast as possible. The aim of this manipulation was to examine whether subjects would be inclined to delay their decisions until the presentation of the vowel marks. The results showed similar decision times for simultaneous presentation of vowel marks and for their very late presentation (300 ms SOA). Thus, lexical decisions were only slightly affected by the delayed presentation of vowels. The effect was especially conspicuous with ambiguous letter strings. These results support the conclusions put forward by Bentin and Frost (1987), suggesting that lexical decisions in Hebrew are based on the recognition of the abstract root or orthographic cluster and do not involve access to a specific word in the phonologic lexicon.

In contrast to lexical decision, a very different strategy was revealed with lagged presentation of vowels in the naming task: the delayed presentation of the vowels delayed naming latencies, and the effects of SOA on RTs were twice as large as the effects found for lexical decisions. Thus, although the phonologic structure of the unambiguous words could be unequivocally retrieved from the lexicon following visual access (postlexical phonology), subjects were more inclined to wait for the vowels to appear in the naming task. Obviously, the longest delays occurred when the words were phonologically ambiguous. Because the correct pronunciation of these words was unequivocally determined only after the presentation of the vowel marks, subjects had to wait for the vowels to appear in order to name those words correctly. Thus, these stimuli provide a baseline for assessing the effect of lagging the vowel marks on naming latencies. When the words were phonologically ambiguous, the effects of lagging the vowel marks on RTs were twice as large as the effects found for unambiguous words, where only one pronunciation was meaningful. These results suggest that subjects adopted two parallel strategies for generating the phonology of the unambiguous printed words: on the one hand they used explicit vowel information using prelexical transformation rules (hence the greater effect of SOA on naming relative to lexical decisions latencies), on the other hand they generated the phonologic structure of the unambiguous words postlexically as well (hence the smaller effects of SOA on naming unambiguous words relative to ambiguous words). These conclusions converge with the results reported by Koriat (1984). Koriat examined the joint effects of semantic priming and vowel mark presentation, and found that semantic priming facilitated naming performance for both pointed and unpointed words, but to the same extent. The presentation of vowel marks speeded naming latencies, but so did a previous presentation of semantic context. Koriat therefore concluded that the pronunciations of unambiguous words are derived both lexically and nonlexically in parallel, and that both processes must be completed and their outcomes compared before the onset of articulation.

Processing lexical ambiguity in Hebrew

Obviously, in the absence of vowel marks, the complete phonemic structure of the letter string in Hebrew cannot be recovered by applying grapheme-to-phoneme conversion rules. Prelexical phonology, therefore, does not appears to be a viable option for the Hebrew reader when presented with unpointed print. He or she is forced to recover the missing phonological information from the lexicon. When the letter string can have only one meaningful pronunciation, the relevant phonologic representation is easy to recover lexically. However, when the letter string has two or more meaningful pronunciations, how does the reader chose among the possible alternatives?

Semantic activation of heterophonic homographs is ordered-accessed

Bentin and Frost (1987) found similar naming latencies for unpointed ambiguous letter strings and for the pointed dominant alternatives. Therefore, they suggested that readers retrieve first the dominant phonological structure of a phonologically ambiguous letter string. The significant delay in naming the subordinate pointed alternatives, relative to the unpointed and the dominant forms of the same letter string, was interpreted as supporting an ordered-access model for the retrieval of phonological information. The naming task, however, cannot disclose covert phonological selection processes. In particular, naming does not reveal whether phonological alternatives, other than the reader's final choice, had been accessed during the process of disambiguation. Although subjects overtly express only one phonological structure, (usually the high-frequency alternative), it is possible that other alternative words were generated but discarded during the output process. Therefore, a more direct measure was necessary to examine whether more than one phonologic alternative of a heterophonic homograph is automatically activated in reading single words.

The possible activation of the two phonologic alternatives related to Hebrew heterophonic homographs was examined by Frost and Bentin (1992) using a semantic priming paradigm. In this study, subjects were presented with isolated heterophonic homographs as primes, whereas the targets were related to only one of the primes' possible meanings. The targets followed the primes at different SOAs ranging from 100 to 750 ms. It was assumed that if a specific meaning of the prime was accessed, lexical decisions for targets related to that meaning would be facilitated. This experimental paradigm is similar to that used by Simpson and Burgess (1985), who examined the processing of English homophonic homographs (letter strings with two meanings but only one pronunciation). Frost and Bentin (1992) reported that, in the absence of biasing context, both meanings of heterophonic homog. ıphs were active at SOAs ranging from 250 to 750 ms from stimulus onset, whereas at a short SOA of 100 ms only the dominant meaning was active.

Phonologic disambiguation of heterophonic homographs precedes semantic activation

In another experiment reported in the same study, the processing of heterophonic homographs was compared to the processing of homophonic homographs using an identical technique. It was found that the decay of activation of subordinate meanings of homophonic and heterophonic homographs followed a similar pattern; all meanings remained active as late as 750 ms from stimulus onset. However, when the onset of activation was examined, a different pattern of results was found for heterophonic and homophonic homographs: in contrast to heterophonic homographs, both subordinate and dominant meanings of homophonic homographs were active as early as 100 ms from stimulus onset. Another finding of interest in that study was that across all SOAs, the effects of semantic priming for heterophonic homographs were larger than the effects found for homophonic homographs. Thus, it appears that both the time-course of activating the different meanings, and the amount of activation were influenced by phonological factors.

These results were interpreted to suggest that heterophonic homographs are phonologically disambiguated *before* the semantic network is accessed. Thus, phonologically ambiguous letter strings refer to different lexical entries, one for each phonological realization (see Figure 2). The alternative lexical entries are automatically activated by the unique orthographical pattern, though at different onset times: in the absence of biasing context the order of activation is determined by the relative word frequency; higher-frequency words are accessed before lower frequency words. As a consequence of the multiple-entry structure and the ordered-access process, heterophonic homographs are phonologically disambiguated prior to any access to semantic information. The overall greater priming effects found for heterophonic than for homophonic homographs suggests that when one lexical unit activates two or more semantic nodes, each of these nodes is activated less than nodes which are unequivocally related to phonological units in the lexicon. Thus, in contrast to lexical decisions, the retrieval of meaning requires the activation of the phonological structure to which the unpointed printed word refers. Note that if meaning were retrieved directly from the orthographic input, no difference should be found between processing homophonic and heterophonic homographs.

One intriguing outcome of the study with Hebrew homographs was that subordinate meanings of both heterophonic and homophonic homographs were still available and used 750 ms from stimulus onset. This result contrasts with the relatively fast decay of subordinate meanings of English homographs (Kellas, Ferraro, & Simpson, 1988; Simpson & Burgess, 1985). Because the decay pattern was similar for both types of Hebrew homographs, the divergence from English should be probably accounted for by language-related factors. One possible source of the different results obtained in Hebrew and in English may be related to the homographic characteristics of the Hebrew orthography. The ubiquity of homography might have shaped the reader's reading strategies. Because ambiguity is so common in reading, the process of semantic and phonologic disambiguation is governed mainly by context. However, the disambiguating context often follows rather than precedes the ambiguous homographs. Therefore, an efficient strategy of processing homographs should require maintaining all the phonologic or semantic alternatives in working memory until the context determines the appropriate one. Note that according to this interpretation the subordinate alternatives do not decay automatically, but remain in memory until disambiguation by context has occurred.

Both phonetic alternatives of heterophonic homographs are automatically activated

Frost (1991) presented additional evidence confirming that both phonologic representations of the ambiguous letter string are automatically activated at some stage after the printed word appears. The aim of this study was to examine directly phonologic and phonetic processing of Hebrew heterophonic homographs. Note that the measurement of semantic facilitation, as used by Frost and Bentin (1992), did not indicate directly whether the presentation of the ambiguous letter string caused the activation of the two phonologic structures related to it, or merely the activation of the two semantic meanings which were accessed directly from the print. To solve this problem, Frost (1991) employed a speech detection task and a task consisting of matching simultaneously presented printed and spoken words. These tasks have been previously shown to detect phonetic and phonologic activation that emerges from the visual presentation of meaningful letter strings (Frost, 1991; Frost & Katz, 1989; Frost, Repp, & Katz, 1988).

The speech detection task is based on an auditory illusion previously reported by Frost et al. (1988). When an amplitude-modulated noise generated from a spoken word is presented simultaneously with the word's printed version, the noise sounds more speechlike than when the print is absent. This auditory illusion suggests that subjects automatically detect correspondences between amplitude envelopes of spoken words and printed stimuli. This speech detection task was employed to examine the processing of Hebrew heterophonic homographs. Subjects were presented with speech-plus-noise and with noise-only trials, and were instructed to detect the speech in the noise. The auditory stimuli were simultaneously presented with printed letter strings that represented two phonological meaningful structures (heterophonic homographs), one dominant and the other subordinate. The bias to falsely detect speech in amplitude-modulated noise when matching print accompanies the auditory presentation occurs only when subjects detect a correspondence between the printed and the spoken information. Therefore, Frost (1991) examined whether subjects detected a correspondence between a printed heterophonic homograph and the masked spoken forms of the *two* phonologic alternatives it represents. The results demonstrated that subjects detected a correspondence between the ambiguous letter string and between the amplitude envelopes of *both* dominant and subordinate phonological alternatives. When the homographs were phonologically disambiguated by adding the vowel marks, similar effects were obtained. Moreover, subjects did not detect any correspondence when the printed pointed alternatives did not correspond to the alternative specified by the noise envelope. These results suggest then, that printed heterophonic homographs automatically activate the two alternative words they represent.

These conclusions were supported by additional experiments employing the matching task. In the matching task subjects are simultaneously presented with a printed word on a computer screen and with a spoken word via headphones. The subjects are instructed to decide as fast as possible whether the stimuli presented in the visual and the auditory modalities are the same or different words. In order for the spoken and the printed forms of words to be matched, they both have to converge at an identical lexical entry. Because the transformation of speech into an orthographic representation is by far less practiced than the transformation of spelling into phonology, the common end result of both print and speech processing in the matching task is presumably a phonological representation in the lexicon (see Frost et al., 1990, for a detailed discussion of the matching task). Frost (1991) presented subjects simultaneously with printed heterophonic homographs and with the spoken forms of the dominant and subordinate alternatives. The subjects were instructed to determine whether the printed words and the spoken words were equivalent. In some of the trials the printed homographs were presented in their pointed form and were therefore disambiguated; that is, the vowel marks unequivocally pointed to either the dominant or the subordinate alternative. In these trials the matching of the visual printed words to the spoken words did not require any ambiguity resolution. In other trials the homographs appeared unpointed, and consequently could be read in two ways. In those trials the outcome of matching the visual words to the spoken words was dependent on the specific phonological alternative generated from the ambiguous consonant string. The aim of the experiment was to compare the decision time for pointed and unpointed print. The results demonstrated that matching the unpointed printed forms of heterophonic homographs to the dominant and subordinate spoken alternatives that were presented auditorily was as fast as matching the pointed unambiguous forms to the respective spoken words. Therefore, these results confirm that both phonologic alternatives were automatically generated from the letter string.

In conclusion, the resolution of phonologic ambiguity in unpointed print is a routine procedure for the Hebrew reader. Our findings suggest that the Hebrew reader develops an orthographic lexicon that serves as an interface to the phonologic lexicon. Each orthographic entry is related to one, two, or more phonologic entries. Lexical decisions in Hebrew are given in reference to this orthographic interface prior to the activation of the phonologic lexicon. However, the activation of an orthographic entry results in the automatic activation of all phonologic entries in the mental lexicon. Semantic activation follows the activation of phonologic entries. Since, in general, while reading, the context disambiguates the phonologically abstract letter string, all phonologic and semantic alternatives remain available for relatively longer periods than in other orthographies such as English. Although all phonologic alternatives are activated following the presentation of the unpointed letters, the more frequent alternative acquires dominance when articulation is required.

Morphologic processing in Hebrew

In the present discussion of Hebrew morphology we will limit ourselves to the processing of roots by the reader. Because the root is the most important determinant of meaning in both spoken and written Hebrew, it has a unique status within the word. Both inflections and derivations in Hebrew modify the root by adding to it prefixes, infixes, and suffixes following specific word patterns. As mentioned in the beginning of this chapter, the root usually specifies a constrained semantic field that constitutes the basic information regarding the meaning of the word. Thus it is fairly reasonable to assume that its extraction from the whole word, whether spoken or written, is a primary process in the analysis of spoken or printed words. We cannot report any data regarding the perception of speech. However, the psychological reality of the status of the root in printed words was examined in several studies.

Morphologic relatedness causes long lasting repetition effects

The preferred technique for investigating morphologic processing in Hebrew was to examine the contribution of morphologic relatedness to pattern of facilitation in the repetition priming task (Bentin & Feldman, 1989; Feldman & Bentin, forthcoming). Bentin and Feldman (1989) examined the effects of morphologic repetition at lag 0 and 15 on lexical decision to the target. Specifically, they compared the effects of pure semantic relatedness, pure morphologic relatedness, and combined semantic and morphologic relatedness, on lexical decisions. In the pure semantic relatedness condition primes and targets consisted of words having different roots but related meanings. In the pure morphologic relatedness condition primes and targets shared the same root but had different meanings (as in the example depicted in Figure 1). Finally, in the combined relatedness condition primes and targets shared both root and meaning. The results showed that semantic relatedness facilitated lexical decisions only at lag 0, whereas pure morphologic relatedness exerted its effect on lexical decisions at lag 0 and 15. Semantic facilitation was greater than morphologic facilitation at lag 0. Facilitation of combined relatedness was as strong as semantic relatedness at lag 0 and similar to pure morphologic relatedness at lag 15. This outcome suggests that semantic activation and morphologic activation have different time courses and arise from two different sources. The presentation of a word containing the root has longer lasting beneficial effects on lexical decisions relative to mere semantic relatedness. Thus, it appears that a previous presentation of the abstract Hebrew root aids lexical processes such as the retrieval of related words and word meanings even at long repetition lags.

Roots are extracted by the reader while processing printed words

In another study, Feldman, Frost, & Dar (forthcoming) examined the ability of skilled readers to detach the phonologic patterns from the roots. This study was based on the segment-shifting task proposed by Feldman (1991). In the segment-shifting task subjects are presented with a printed word in which one segment is underlined. The subjects are required to detach the underlined segment from the word and append it to another word presented underneath. The subjects have to pronounce the second word with the new segment (usually appended to its end) as quickly as possible. The experimental conditions typically consist of underlining a segment that is a suffix morpheme in one word, but not in another (e.g., ER is a suffix morpheme in DRUMMER but not in SUMMER). Is it easier to detach ER from DRUMMER than from SUMMER?

The segment-shifting task was originally employed by Feldman in English and Serbo-

Croatian. These languages are characterized by concatenative morphology where morphologically complex words are constructed from discrete morphemic constituents that are linked linearly: There is a base morpheme to which other elements are appended so as to form a sequence. In languages with concatenative morphology, suffixes and prefixes are regularly appended to the base morpheme in a manner that preserves its phonological and orthographic structure. In contrast to English and Serbo-Croatian, Hebrew is usually characterized by a nonconcatenative morphology. In Hebrew the phonologic word pattern is an infix, not a prefix or a suffix. It is superimposed on the root and changes both its phonologic and its orthographic structure (see the example of "dover" in Figure 1).

Experiments that employed the segment-shifting task in English and Serbo-Croatian yielded straightforward results: It is easier to detach a segment that serves as a morphemic suffix appended to a base morpheme than to detach the same sequence of letters when it is an integral part of a word that cannot be decomposed into morphemic constituents. This outcome suggests that the processing of morphologically complex words in languages with concatenated morphology entails morphemic decomposition. Such decomposition is relatively easy and straightforward when the morphemic constituents are linked linearly. In contrast to English and Serbo-Croatian, the decomposition of Hebrew derivations and inflections into root and word pattern is not as straightforward. This is because the phonemes of the root and the phonemes of the word pattern are intermixed.

Feldman et al. (forthcoming) took advantage of the fact that although formally all words in Hebrew can be defined as containing roots, not all roots are productive. Roots are considered productive if they can be inflected, and other words can be derived from them. A root is considered nonproductive if it cannot be inflected and is therefore contained in only one Hebrew word. Many words in Hebrew form a unique phonemic sequence that does not lend itself to inflections or derivations. Feldman et al. asked whether a specific phonologic word pattern can be detached more easily from words that contain productive roots than from words that contain nonproductive roots.

The experiment was similar to the typical segment-shifting task experiment. Subjects were presented with pointed words and were required to detach the sequence of vowels from the words, to superimpose them on a nonword consonant cluster, and to name it. The results showed that it was easier to detach the vowels from three consonants that were a productive root than to detach them from three letters that were not. In a second experiment similar and even stronger effects were obtained when the word patterns were not merely vowels but consisted of a sequence of vowels and consonants. These results suggest that productive roots have a special status for the Hebrew speaker and reader. Their psychological reality is reflected by their salience relative to the other letters and phonemes constituting the word. It appears that the presentation of a printed word containing a productive root results in the automatic detection of this root, such that the letters of a word are parsed into letters belonging to the root and letters not belonging to it. The important aspect of this morphologic decomposition is that the root letters do not have to appear in adjacent position (as in the second experiment). Even if they are dispersed within the word they are automatically extracted by the reader. We believe that a similar process can be demonstrated in the recognition and understanding of spoken words as well. That is, the phonemes belonging to the root have a unique psychological reality. However, this suggestion requires further investigation.

CONCLUSIONS

The pointed and unpointed Hebrew orthography presents an opportunity to examine reading processes when full or partial phonologic information is conveyed by print. This provides a significant methodological tool for investigating the effects of orthographic depth on visual word recognition, yet avoiding the pitfalls of cross-language designs. Research in reading Hebrew suggests that reading strategies are affected by the presentation or the omission of vowel marks. Efficient reading of unpointed text is based on fast recognition of orthographic clusters that become phonologically and semantically unequivocal given the available context. In contrast, the presentation of vowel marks induces a phonological processing of the printed words, which is often characteristic of shallow orthographies. This suggests that the reader of Hebrew adopts flexible reading strategies that take advantage of all possible phonemic information provided by the print.

REFERENCES

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 340-357.

Bentin, S., (1989). Orthography and phonology in lexical decisions: Evidence from repetition effects at different lags. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 51-72.

Bentin, S. (1992). Phonological awareness, reading, and reading acquisition: A survey and appraisal of current knowledge. *Orthography, phonology, morphology, and meaning* (pp. 193-210). Amsterdam: Elsevier Science Publishers.

Bentin, S., Bargai, N., & Katz, L. (1984). Orthographic and phonemic coding for lexical access: Evidence from Hebrew. *Journal of Experimental Psychology: Learning Memory & Cognition, 10,* 353-368.

Bentin, S., & Feldman, L. B. (1990). The contribution of morphological and semantic relatedness to repetition priming at short and long lags: Evidence from Hebrew. *Quarterly Journal of Experimental Psychology, 42,* 693-711.

Bentin, S., & Frost, R. (1987). Processing lexical ambiguity and visual word recognition in a deep orthography. *Memory & Cognition, 25,* 13-23.

Bentin, S., Hammer, R., & Cahan, S. (1991). The effects of aging and first grade schooling on the development of phonological awareness. *Psychological Science, 2,* 271-274.

Bentin, S., & Leshem, H. (in press). On the interaction between phonological awareness and reading acquisition: It's a two-way street. *Annals of Dyslexia.*

Berman, R. A. (1978). *Modern Hebrew structure.* Tel Aviv: University Publishing Project.

Bertelson, P. (1986). The onset of literacy: Liminal remarks. *Cognition, 24,* 1-30.

Bertelson, P., & de Gelder, B. (1990). The emergence of phonological awareness: Comparative approaches. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception.* Hillsdale, NJ: Erlbaum.

Bradley, L, & Bryant, P. E. (1983). Categorizing sounds and learning to read: A causal connection. *Nature, 301,* 419-421.

Bradley, L, & Bryant, P. E. (1985). *Rhyme and reason in reading and spelling.* Ann Arbor: University of Michigan Press.

Calfee, R., Chapman, R., & Venezky, R. (1972). How a child needs to think to learn to read. In L. W. Gregg (Ed.), *Cognition in Learning and Memory.* New York: Wiley.

Chumbley, J. I., & Balota, D. A. (1984). A word's meaning affects the decision in lexical decision. *Memory & Cognition, 12,* 590-606.

Conrad, R. (1972). Speech and reading. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye.* Cambridge, MA: MIT Press.

Feldman, L. B., & Bentin, S. (submitted). Facilitation due to repetition of disrupted morphemes: evidence from Hebrew.

Feldman, L. B., Frost, R., & Dar, T. (forthcoming). Processing morphemes in different languages.

Frost, R. (1991). Phonetic recoding of print and its effect on the detection of concurrent speech in amplitude modulated noise. *Cognition. 39,* 195-214.

Frost, R. (in press). Prelexical and postlexical strategies in reading: Evidence from a deep and a shallow orthography. *Journal of Experimental Psychology: Learning, Memory and Cognition.*

Frost, R., & Bentin, S. (1992). Processing phonological and semantic ambiguity: Evidence from semantic priming at different SOAs. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 18,* 58-68.

Frost, R., Feldman, L. B., & Katz, L. (1990). Phonological ambiguity and lexical ambiguity: Effects on visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 569-580.

Frost, R., & Katz, L. (1989). Orthographic depth and the interaction of visual and auditory processing in word recognition. *Memory & Cognition, 17,* 302-311.

Frost, R., & Katz, L. (Eds.). (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. *Orthography, phonology, morphology, and meaning* (pp. 67-84). Amsterdam: Elsevier Science Publishers.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 104-114.

Frost, R., Repp, B. H., & Katz, L. (1988). Can speech perception be influenced by a simultaneous presentation of print? *Journal of Memory and Language, 27,* 741-755.

Kellas, G., Ferraro, F. R., & Simpson, G. B. (1988). Lexical ambiguity and the time course of attentional allocation in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 601-609.

Koriat, A. (1984). Reading without vowels: Lexical access in Hebrew. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X.* Hillsdale, NJ: Erlbaum.

Koriat, A. (1985). Lexical access for low and high frequency words in Hebrew. *Memory & Cognition, 13,* 37-44.

Liberman, I. Y., & Shankweiler, D. (1985). Phonology and the problems of learning to read and write. *RASE, 6,* 8-17.

Liberman, I. Y., Shankweiler, D., Fisher, F. W., & Carter, B. J. (1974). Explicit syllable and phoneme segmentation in young children. *Journal of Experimental Child Psychology, 18,* 201-212.

Lundberg, I., Olofsson, A., & Wall, S. (1980). Reading and spelling skills in the first school years, predicted from phonemic awareness skills in kindergarten. *Scandinavian Journal of Psychology, 21,* 159-173.

Mann, V. A. (1984). Longitudinal prediction and prevention of early reading difficulty. Annals of Dyslexia, 34, 117-136.

McCusker, L. X., Hillinger, M. L., & Bias, R. C. (1981). Phonological recoding and reading. *Psychological Bulletin, 88,* 217-245.

Morag, S. (1972). *The vocalization systems of Arabic, Hebrew, and Aramaic.* Mouton & Co.'s- Gravenhage.

Navon, D., & Shimron, Y. (1981). Does word naming involve grapheme to phoneme translation? Evidence from Hebrew. *Journal of Verbal Learning and Verbal Behavior, 20,* 97-109.

Navon, D., & Shimron, Y. (1984). Reading Hebrew: How necessary is the graphemic representation of vowels? In L. Henderson (Ed.), *Orthographies and reading: Perspectives from cognitive psychology, neuropsychology, and linguistics.* London, Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Perfetti, C. A., & Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology, 67,* 461-469.

Shankweiler, D., & Liberman, I. Y. (1976). Exploring the relations between reading and speech. In R. M. Knights & D. J. Bakker (Eds.), *The neuropsychology of learning disorders: Theoretical approaches.* Baltimore, MD: University Park Press.

Simpson, G. B., & Burgess. C. (1985). Activation and selection processes in the recognition of ambiguous words. *Journal of Experimental Psychology: Human Perception & Performance, 11.* 28-39.

FOOTNOTES

*In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 27-44). Amsterdam: Elsevier Science Publishers (1992).

†Department of Psychology, The Hebrew University, Jerusalem.

The Reading Process is Different for Different Orthographies: The Orthographic Depth Hypothesis*

Leonard Katz[†] and Ram Frost[‡]

It has been said that most languages get the orthography they deserve and there is a kernel of truth in that statement. There is generally an underlying rationale of efficiency in matching a language's characteristic phonology and morphology to a written form. Although the final product may turn out to be more suitable for some languages than for others, there are certain basic principles of the fit that can be observed. The attempt to make an efficient match between the written form, on the one hand, and morphology and phonology, on the other, typically determines whether the orthography chosen is a syllabary, a syllabary-cum-logography, or an alphabet. Further, within the group of alphabetic orthographies itself, there are varying degrees of dependence on the strict alphabetic principle: the range of correspondence between grapheme and phoneme varies both in consistency and completeness. The degree of this dependence is to some extent a function of a language's characteristic phonology and morphology, just as was the choice of the kind of orthography itself. We discuss here what this varying dependence on the alphabetic principle may mean for the mental processes involved in reading and writing.

Diversity in writing systems

Although writing systems are, in general terms, systems for communication, what they actually communicate is the spoken language—as opposed to communicating nonverbal ideas and meanings. DeFrancis (1989) reinforced this point with his analysis of so-called pictographic languages, writing systems whose elements are pictures and symbols that do not stand for words. DeFrancis argued that true pictographic systems are not, in principle, effective and showed that existing examples of pictographic systems had been designed only as novelties or playful communication systems. In practice, they were never used to communicate without substantial ancillary aid from spoken language. The example of pictographic writing emphasizes the poverty of written communication that is not based on language. Therefore, because writing systems are systems for representing the spoken language, it is reasonable to suggest that an understanding of the psychological processing involved in using a writing system must include an understanding of the processing of the spoken language. Although the former will not necessarily parallel the latter, it will be constrained by it. A major constraint arises from the spoken language's morphemes which are the smallest units that carry meaning. It is these morphemes of speech that will be the focus of communication, both spoken and written. Word stems are all morphemes and so are their derivational and inflectional affixes; these units of the spoken language must be easily recoverable from the written language.

A large number and variety of writing systems have flourished, evolved and developed, and in many cases, died, over the centuries. Each of the known systems can be categorized as either logographic-phonetic, syllabic, or alphabetic (DeFrancis, 1989). These distinctions are made on the basis of how a script (a set of symbols) relates to the structure of its language. This relationship between a script and its language is what is described by the term orthography (Scheerer,

Work on this paper was supported by National Institute of Child Health and Human Development Grant HD-01994 to Haskins Laboratories. The authors wish to thank I. G. Mattingly for comments on an earlier draft.

1986). The kind of script system and its orthography are typically not wholly the result of accident. It is not accidental that the Chinese languages, for example, have a logographic-phonetic system. In the Chinese orthography, the typical character has two parts to it: a logographic and a phonetic component, the former providing a visually distinctive cue to the semantics and the latter giving the reader a partial guide to the pronunciation. Together, the two components make a combination that specifies completely a unique spoken morpheme. Words may be mono- or polymorphemic.

Chinese morphemes are mainly monosyllabic and, because the variety of possible syllables is limited, there is a high degree of homophony in the language. Such a language is best served by an orthography that distinguishes between the different meanings of morphemes that sound alike. Instead, if the orthography had represented the spoken form alone (e.g., only the phonetic component in the printed word), the reader would not be able to determine the intended meaning of each homophone except, possibly, from the word or sentential context (many polymorphemic words are unique compounds of homophonous morphemes)—but not without the additional cognitive cost needed to resolve the ambiguity. The homophony problem is more of a problem for a reader than for a listener because the listener has more nonverbal contextual information available to assist in determining word meanings.

But a pure logography would not suffice either, for Chinese. If the orthography had no phonetic component, a reader would have to remember, without a phonetic cue, a pronunciation for each of several thousand logographs. This would have effectively limited the number of printed characters that a reader could remember and name to an unacceptably small number. Instead, in modern Mandarin, it is necessary to remember only a small number of phonetic components together with a smaller number of semantic signs. DeFrancis gives the number of these as 895 and 214, respectively (DeFrancis, 1989). A phonetic and a semantic component are paired to produce a character; the effective set consists of 4300 characters, the approximate number considered necessary for full literacy.

In contrast to Chinese, spoken Japanese is polysyllabic and is composed of regular syllable-like components, called moras. Because the number of syllables is small (fewer than 113), it is feasible to represent them by means of a syllabary. The Japanese orthography called kana was such a system, adapted from Chinese characters. However, because there is a good deal of homophony in Japanese, the use of a syllabary alone would not have been without problems and a logography also came into use. That logography is still in use today and is routinely mixed with the use of the kana, the syllabaries being used primarily for morphological affixes and grammatical function words, foreign loan-words, and words not already covered by the Chinese.

Indo-European languages have less homophony and more polysyllabic morphemes than Chinese and Japanese. In addition, the structure of the Indo-European syllable itself is generally more complex than Chinese or Japanese, containing a larger number of phonologically permissible clusters. English is said to have at least 8000 syllables in its phonology, compared to fewer than 1300 for Chinese (DeFrancis 1989). Eight-thousand is far too large a number for an effective syllabary. For English, an alphabet, representing phonemes, is more efficient for learning to read and write. Similarly, the Semitic languages (which include Arabic and Hebrew) would be less suitably represented by a syllabary than by an alphabet. Like Indo-European languages, they too have complex syllable structures. Historically, a consonantal alphabet that developed for West Semitic was the alphabet from which we trace the evolution of modern alphabets.

One of the characteristics of Semitic languages that may have led to the invention of the alphabet is the Semitic triconsonantal root: Semitic words that are related by derivation or inflection have a common core (usually three consonants). Although the vowels in each of the different relatives may be quite different and although there may be additional consonants in a prefix or suffix, there remains an invariant series of three consonants, the root. This core has a strong linguistic salience—it represents a morphological communality of meaning among its family members (see Frost & Bentin, this volume). One can speculate that several early attempts were made at developing a writing system but only an alphabetic system could have captured this communality of morphology efficiently. Syllabic representations would not be optimal: in Hebrew, morpheme boundaries fail to coincide with syllable boundaries (a condition that is true of Indo-European languages as well). Therefore, syllabic representations would not be appropriate for representing morphological units.

The causes of diversity in alphabetic orthographies

Even among the various alphabetic writing systems themselves there are major differences in the degree to which they mirror the phonemic structure of their respective spoken languages. Again, the reason for the differences is largely accounted for by the particular phonological and morphological characteristics of each language. For example, standard written Hebrew is an orthography in which all diacritics (or *points*) are omitted. These diacritics represent nearly all of the vowels and are also used to disambiguate some of the consonants. Nevertheless, writing without diacritics is usually sufficient to indicate the exact intended (i.e., spoken) word if it is supported by a phrasal or sentential context. Thus, although the printed root may be insufficient to allow an unequivocal identification when presented in isolation, when it is presented in a normal context—even a printed one—the combined sources of information are enough for word identification.

In strong contrast to the Hebrew orthography is the Serbo-Croatian. Serbo-Croatian is a major language of the Balkan Peninsula. Its present alphabet was introduced in the early nineteenth century following the principle, "Spell a word like it sounds and speak it the way it is spelled." Each letter represents only one phoneme and each phoneme is represented by only one letter. Moreover, no phoneme in the spoken word is ever excluded in the spelled word. The relation between letters and phonemes is isomorphic and exhaustive. To this day, the Serbo-Croatian spelling system follows the phonemic structure of spoken words. So regular is the relation between speech and writing that the writings of people with different regional pronunciations will show different spellings, mirroring the minor differences in their spoken language. This simple writing system works well for Serbo-Croatian because morphemic variations in the language due to inflection and derivation do not often produce alterations in the phonemic structure of word stems; word stems largely remain intact.

A different state of affairs exists in English. English is somewhere between Hebrew and Serbo-Croatian in the directness with which its phonology is represented in its spelling; there is a large amount of regular phonologic change among words in the same derivational family. Examples of this are the contrasts between the derivational relatives HEAL and HEALTH, between STEAL and STEALTH, etc. Chomsky and Halle (1968) argue that English spelling represents a morphophonemic invariance common to these word pairs, an abstract phonological communality that is below their surface difference in pronunciation. In reading English aloud, the reader must either remember the pronunciation of such a word as a whole or remember the appropriate context-dependent rule for pronunciation. An alternative writing system might have spelled English in the same way that Serbo-Croatian is spelled: with an isomorphic relation between letter and phoneme. However, that method would rob printed English of the advantageous common spelling for words with common morphology. The words HEAL and HEALTH, for example, might then be spelled HEEL and HELTH, disguising their common meaning. The printed form HEEL would also suffer from a double meaning as a consequence of its being homophonic with the word meaning, "part of a foot." English spelling represents a compromise between the attempt to maintain a consistent letter-phoneme relation and the attempt to represent morphological communality among words even at the cost of inconsistency in the letter-phoneme relation.

Thus, alphabetic writing systems reflect the spoken forms of their respective languages with different degrees of consistency and completeness between letter and phoneme. Some of the differences in writing systems have purely political, cultural, or economic causes. But many differences have been motivated by two factors that are purely linguistic. The first has to do with how complex the spoken language is in the relation between phonology and morphology—only a phonologically complex language can have a deep alphabetic orthography. For example, Serbo-Croatian is not phonologically complex. All morphologically related words have a common phonologically invariant core. Two words that are morphologically related will share a common word stem that will necessarily sound the same in both words. Both instances of that common stem will, of course, be spelled the same. Thus, when evaluated by the characteristic of phonological complexity, a language that is not complex can be written (and generally *will* be written) in a shallow orthography, an orthography that tracks the phonology. Secondly, if a language is one that is phonologically complex then the orthography has the option of representing either morphological invariance (a deep orthography) or following grapheme-phoneme invariance (a shallow orthography). As we suggested above, English qualifies as quite complex, phonologically. In principle, it could have been written either as a shallow or a deep orthography. The advantage to English in choosing a deep orthography is in the consistent spelling of morphemic invariances. However, that choice having been made, there are then different pronunciations of the same spelling on occasion (e.g., **HEAL-HEALTH**) and, inadvertently, identical pronunciations for different spellings (e.g., **PEEL-DEAL**).

A different situation exists in Hebrew. Hebrew's phonology is complex; morphemes may undergo considerable sound change under either inflectional or derivational change. On the other hand, because of the pervasiveness of the triconsonantal root in Hebrew, a great deal of morphological constancy exists. Therefore, there was an historical choice, so to speak, for the evolution of the Hebrew orthography: It could have opted for either morphemic or phonemic invariance but, unlike Serbo-Croatian, it could not have contained both in a single orthography because of its phonological complexity. Hebrew initially evolved as an orthography in which the morphology was preserved at the expense of phonological completeness. Vowels were omitted thereby emphasizing the morphologically based consonantal invariance in a given family of word roots. Vowel points were added to the script at a later stage in the orthography's development only because the language was no longer being spoken as a primary language and it was feared that its pronunciation would become corrupted unless vowels were included in the script. Nowadays, the orthography used by adults is the unpointed one, which is graphemically incomplete and somewhat inconsistent to the reader because it omits nearly all of the vowels and makes some of the consonants ambiguous.

In summary, all alphabetic orthographies can be classified according to the transparency of their letter-to-phoneme correspondence, a factor that has been referred to as orthographic depth (Liberman, Liberman, Mattingly, & Shankweiler, 1980). An orthography in which the letters are isomorphic to phonemes in the spoken word (completely and consistently), is orthographically shallow. An orthography in which the letter-phoneme relation is substantially equivocal is said to be deep (e.g., some letters have more than one sound and some phonemes can be written in more than one way or are not represented in the orthography). Shallow orthographies are characteristic of languages in which morphemic relatives have consistent pronunciations.

Differences among alphabetic orthographies in processing printed words: The orthographic depth hypothesis

Our discussion to this point has made the standard argument that there are differences among alphabetic orthographies in orthographic depth and that these differences are a result of differences in their languages' phonology and morphology. In this section, we propose that the differences in orthographic depth lead to processing differences for naming and lexical decision. This proposal is referred to as the orthographic depth hypothesis (ODH). It states that shallow orthographies are more easily able to support a word recognition process that involves the language's phonology. In contrast, deep orthographies encourage a reader to process printed words by referring to their morphology via the printed word's visual-orthographic structure.

We would like to make two points, each independent of the other. The first states that, because shallow orthographies are optimized for assembling phonology from a word's component letters, phonology is more easily available to the reader prelexically than is the case for a deep orthography. The second states that the easier it is to obtain prelexical phonology, the more likely it will be used for both pronunciation and lexical access. Both statements together suggest that the use of assembled phonology should be more prevalent when reading a shallow orthography than when reading a deep orthography. Because shallow orthographies have relatively simple, consistent, and complete connections between letter and phoneme, it is easier for readers to recover more of a printed word's phonology prelexically by assembling it from letter-phoneme correspondences.

Suggested by the above is our assumption that there will always be at least some dependence on phonological coding for the process of reading in any orthography. That is, the processing of (at least) some words will include assembled phonology (at least in part). This assumption can be easily motivated for alphabetic orthographies. The assembling of phonology has a certain precedence in a reader's experience; instruction in reading typically means instruction in decoding, i.e., learning how to use letter-phoneme correspondences. It is well established that beginning readers find it easier to learn to read in shallow orthographies, where those correspondences are most consistent (see, for example, Cossu, Shankweiler, Liberman, Katz, &

Tola, 1988). Even in learning to read Hebrew, instruction is typically given in the shallow pointed orthography instead of the deep unpointed one (the transition to the unpointed form beginning in the third grade). In any orthography, after learning to read by using assembled phonology routines, skilled readers may continue its use to the extent that the cost of doing so is low. This will be particularly true when the orthography is shallow. However, given the experimental evidence, some of which we discuss later, it seems certain that assembled phonology is not used *exclusively*. More likely, a mix of both assembled phonology and visual-orthographic codings are nearly always involved, even in shallow orthographies.

Two versions of the orthographic depth hypothesis

Two versions of the orthographic depth hypothesis (ODH) exist in the current literature. What can be called the *strong ODH* states that phonological representations derived from assembled phonology alone are sufficient for naming and lexical decision in shallow orthographies. Thus, according to the strong ODH, rapid naming in shallow orthographies is a result of only this prelexical analytic process and does not involve pronunciation obtained from memory, i.e., the lexicon. However, we submit that the strong form of the ODH is patently untenable when applied to the orthographies that have typically been used in research on word perception. It is insufficient to account for pronunciation even in a shallow orthography like Serbo-Croatian. This is so because Serbo-Croatian does not represent syllable stress and, even though stress is often predictable, it is not always predictable. Because the final syllable is never stressed, stress is completely predictable for two-syllable words but for words of more than two, it is not. However, one- and two-syllable words make up a large part of normal running text so much or most of the words a reader encounters can be pronounced by means of a prelexical subword analysis. But, of course, many words will be greater than two syllables in length and these can be pronounced correctly only by reference to lexically stored information. In addition, there are some exceptions to the rule that a letter must represent only one phoneme; some final consonant voicing changes occur in speech that are not mirrored in the conventional spelling (these changes are predictable, however). Thus, Serbo-Croatian, although it should be considered an essentially shallow orthography, is not the perfect paradigm of a shallow orthography. We should not expect a strong ODH to make sense for such an orthography.

We support the *weak ODH*. In this version, the phonology needed for the pronunciation of printed words comes not only from prelexical letter-phonology correspondences but also from stored lexical phonology, that is to say, from memory. The latter is the result of a visual-orthographic addressing of lexicon, i.e., a search process that matches the spelling of a whole word or morpheme with its stored phonology. The degree to which a prelexical process is active in naming is a function of an orthography's depth; prelexical analytic processes will be more functional (less costly) in shallow orthographies. However, whether or not these prelexical processes actually dominate orthographic processing for any particular orthography is a question of the demands the two kinds of processes make on the reader's processing resources, a question we discuss further below. We proposed (and supported) this weak form of the ODH in Katz and Feldman (1983) and Frost, Katz, and Bentin (1987); further details are given later in this chapter.

With regard to word recognition (as in lexical decision), some of our colleagues have argued that Serbo-Croatian necessarily involves prelexical (i.e., assembled) phonology (Feldman & Turvey, 1983; Lukatela & Turvey, 1990a). Others have made a similar claim for the obligatory involvement of prelexical phonology in English (Van Orden, Pennington & Stone, 1990; Perfetti, Bell, & Delaney, 1988). However, these researchers have not argued for the *exclusive* involvement of assembled phonology. Logically, assembled prelexical phonological information, without syllable stress information, is sufficient to identify the great majority of words in the English lexicon. However, irregularly spelled words, foreign borrowings, etc., would pose problems for an exclusively phonological mechanism, and, therefore, such a view seems less plausible. Finally, note that we are confining this discussion to the problems of naming and lexical decision; it is an entirely different question to ask whether phonological representations are necessary for *postlexical* processes like syntactic parsing and text comprehension.

Evidence on the questions of phonological recoding and the weak ODH

We discuss next the evidence for the hypotheses that the lexicon is addressed by assembled phonology, presumably in combination with visual-orthographic addressing, and that the specific mix of the two types of codes depends on orthographic depth. We show why single-language studies, in general, are not suitable for testing the weak ODH and mention the few exceptions. Experiments that directly compare orthographies with each other provide the most direct evidence. We will argue that these cross-language comparisons are absolutely critical to an investigation of orthographic depth effects.

It is important to realize that, in the controversy over whether visual-orthographic recoding or phonological recoding is used in word perception, there is little direct evidence of visual-orthographic effects. Instead, the burden of proof is placed on assembled phonology; if no effect of phonology is found, then visual-orthographic coding is said to win, by default. The assumption is not unreasonable because a visual-orthographic representation is obviously available in principle and seems to be the only alternative to assembled phonology. However, it should be kept in mind that, because of this, the experimental evidence hinges on the sensitivity of the experimental task and its dependent measures to phonology. If they fail to indicate the presence of phonology, it may not be because phonology is not operative.

In fact, several experiments have demonstrated that phonological recoding effects can be found even in deep orthographies. In Hebrew, Frost has shown that, if available, the full phonology given by the pointed orthography is preferred for naming; this seems to suggest that even if word recognition does not normally proceed via assembled phonology in Hebrew, the recognition process is prepared to default to its use (Frost, in press). In English, Perfetti and his associates and Van Orden and his associates have presented strong evidence for phonological recoding in lexical access (Perfetti, Bell, & Delaney, 1988, Perfetti, Zhang, & Berent, 1992; Van Orden, Pennington, & Stone 1990; Van Orden, Stone, Garlington, Markson, Pinnt, Simonfy, & Brichetto, 1992).

However, it is one thing to find the active presence of phonological recoding, another to determine the conditions under which phonological recoding occurs, and yet another to determine the degree to which naming or lexical access is dependent on it. Is assembled phonology obligatory or is it rarely used; if neither of these extremes, is it the more preferred or the less preferred code? In this vein, Seidenberg (Seidenberg, 1985; Seidenberg, 1992) has argued that word frequency is the primary factor that determines whether or not assembled phonology is used to access the lexicon. His argument is that in any orthography, whether deep or shallow, frequently seen words will become familiar visual-orthographic patterns and, therefore, rapid visual access will occur before the (presumably) slower phonological code can be assembled from the print. Low frequency words, being less familiar, produce visual-orthographic representations that are less functional; lexical activation builds up more slowly. This gives time for phonological recoding to contact the lexicon first. However, we cannot presently answer questions that are concerned with the relative importance of word frequency to orthographic depth or concerned with the relative dominance of the two kinds of representation, visual-orthographic and phonological. But we can meaningfully address the question of whether or not the relative amount of assembled phonological coding decreases with increasing orthographic depth: the orthographic depth hypothesis.

Comparisons across orthographies

Cross-language experimentation, in which different languages are directly compared, are the critical methodology for studying the orthographic depth hypothesis. Single-language experiments are adequate for testing only the strong form of the ODH, in which shallow orthographies are said to never use lexically stored information for naming—but, as we showed, this is a claim that can be rejected on logical grounds. Single-language experiments are not without interest, however; they can be useful in indicating how easy it is to find effects of phonological coding. This may suggest—but only suggest weakly—what the dominant representation is for an orthography. If it is easy to find effects of phonological coding in Serbo-Croatian and difficult to find those effects in Hebrew, using more or less similar experimental techniques, we may suspect that phonological coding is the dominant (preferred) code in Serbo-Croatian but not in Hebrew. However, such experiments can not rule out the additional use of the alternate type of representation for either orthography. In this vein, we know that it is difficult to find effects of phonological coding in English using standard lexical decision paradigms but, nevertheless, phonological effects can be found using Perfetti's backward masking paradigm, which is apparently more sensitive (Perfetti, 1988). Finally, however, an accurate answer to the question of which type of representation is dominant for a particular orthography can not be given by the current experimental paradigms; the results from these experiments may only reflect the adequacy of the paradigms in capturing the true word recognition process.

The weak form of the ODH proposes that (1) both orthographic information and prelexically assembled phonological information are used for lexical access and (2) the degree to which one kind of information predominates is a function of the structural relationship between orthography and the lexical entry. The Serbo-Croatian orthography, with its simple and consistent letter-phoneme relationships, makes it easy for the reader to learn and maintain the use of assembled phonology. This assembled phonology must address, presumably the same abstract phonology addressed by a listener's spoken language lexicon (although it may be only a subset of the full spoken phonology, because of the absence of stress information, at the least). In contrast, the Hebrew orthography, because it lacks most of the vowels and has many ambiguous consonants, is incapable of providing enough assembled phonology that will consistently identify a unique word in the phonological lexicon (only the consonants can be assembled); therefore, there are fewer benefits in generating phonological information by assembling it from grapheme-phoneme correspondences. These are lessons that the developing reader can learn tacitly, lessons that may lead, eventually, to different dominant modes of printed word processing for Serbo-Croatian and Hebrew readers. For languages that are in between these two extremes, the relative balance of assembled phonology to orthographic representation should reflect the relative efficacy of the two kinds of information in that orthography; some letters or letter-sequences may be simple and consistent and the assembled phonology derived from these may be used along with orthographic information. It is not possible to make a more precise statement without an understanding of the details of the lexical recognition process and the processing resources that are required. For example, for a visual-orthographically coded word to be recognized, a mental representation of that word must have been created previously and stored in the reader's memory. We do not know how to compare the resources needed to create a new orthographic representation with the resources needed to generate assembled phonology; which is more demanding? Neither can we automatically assume that it is easier to access lexicon via a visual-orthographic representation.

Additional complications arise when we try to be more specific about the phonological nature of the information in the lexicon itself; what, exactly, is the information that is represented in lexicon that is, presumably, addressed by assembled phonology? Alphabets mainly represent phonemes but are words in the spoken lexicon to be represented as phoneme sequences? If so, why do syllabic orthographies work at all, since the printed units of syllabaries map onto syllables, not phonemes? In fact, there are several different theoretical descriptions that have been proposed by speech researchers for the lexical representation of a word. Perhaps, the spoken lexicon contains multiple phonological descriptions of a single word, e.g., phonetic, phonemic, syllabic, gestural, etc. The phonology produced by reading may be a subset of the information constituting the spoken lexicon or it may even be different in kind (although related). We do not propose to discuss this problem in detail here, but only wish to point out that there is a companion to the question of how phonological information is used in reading, namely, the question of the nature of the phonological information that is used in spoken word recognition.

Evidence supporting the orthographic depth hypothesis

Some early evidence that there is a relationship between orthographic depth and lexical access was obtained by Katz and Feldman (1981). They compared lexical decision times in Serbo-Croatian and English for printed stimuli that were divided with a slash character. The stimuli were divided either at a syllable boundary (e.g., WA/TER) or one character to either the left or right of the boundary (e.g., W/ATER or WAT/ER). If word recognition involves recoding of the stimulus to a phonological form, and if that phonology includes the syllable as a unit, then division at a syllable boundary—which preserves the syllable units—should be less disruptive than division off the boundary. Pseudowords were similarly divided. Lexical decisions to words and pseudowords that were irregularly divided were slower than lexical decisions to their regularly divided counterparts and the disruptive effect of irregular division was stronger for Serbo-Croatian. The data, then, were consistent with a model of word recognition that assumes the operation of at least some phonological recoding of print prior to lexical access and, further, assumes that phonological recoding is a more consistent determiner of access in shallower orthographies (e.g., Serbo-Croatian) than deeper ones (e.g., English).

In a second study, Katz and Feldman (1983) made a direct test of a second prediction of the orthographic depth hypothesis: that pronouncing a word (naming) depends more on assembled phonology in Serbo-Croatian than in English. A lexical decision experiment and a naming experiment were run in both languages; stimulus words had the same (or similar) meanings in both languages. The subjects were native speakers who were tested in their native countries (Serbia or the United States), in order to avoid subjects who were fluent in both languages. This was done because the experience of a bilingual speaker might have affected his or her strategy for reading.

Each test stimulus (the target), whether it was a word or a nonword, was always preceded by the brief (600 ms) presentation of a real word. On half of those trials when the target was a word, this predecessor was semantically related to the target (e.g., MUSIC-JAZZ); on the other half, it was unrelated. Words also preceded the nonword targets. It is well established, that preceding a stimulus with a semantically related word will facilitate and speed a lexical decision response to the target. Thus, it was expected that reaction time to the target JAZZ would be faster for those subjects who saw it preceded by the word MUSIC than for those subjects who saw it preceded by the word GLASS, which has no strong semantic relation to the target. The critical fact for this kind of experimental technique is that the facilitating effect of MUSIC can only occur by activating the semantic link between it and JAZZ and this linkage necessarily must be within the lexicon. Thus, to the extent that there is facilitation in the subject's recognition of JAZZ, it indicates that recognition is being assisted by activity within the lexicon. Such lexical activity may facilitate whether the lexicon is addressed by orthography or by phonology.

For naming, however, the prediction is different. Naming can be accomplished largely without accessing the lexicon by means of subword letter-to-phonology recoding. Of course, in neither English or Serbo-Croatian can the process be entirely without reference to lexical memory, because the stress of polysyllabic words is not specified in the orthography. Nevertheless, the process of naming can, in principle, be carried out substantially without reference to the lexicon. Thus, if naming in Serbo-Croatian is more dependent on phonological recoding (and less dependent on lexical look-up) than English, naming in Serbo-Croatian ought not to be affected by the semantic priming manipulation, which is necessarily lexical in its locus of operation. Results supported this suggestion: target words that were preceded by semantically related words (e.g., MUSIC-JAZZ) were pronounced faster than target words that were preceded by unrelated words (e.g., GLASS-JAZZ) in the case of English but not in the case of Serbo-Croatian. In contrast, there were equivalent strong effects of semantic priming for lexical decisions, in both languages.

A three-way comparison of Hebrew, Serbo-Croatian, and English increased the range of orthographic depth examined in a single study (Frost, Katz, & Bentin 1987). Necessary to the success of any cross-language experiment is a set of stimulus words that have equivalent critical characteristics in all the languages under consideration: for example, equivalent frequencies of occurrence. Subjectively estimated frequencies of occurrence were obtained and equivalent sets of stimulus words were used in all three test languages. The important comparison, over the three orthographies, was the relation between naming and lexical decision reaction times for words versus nonwords. The importance of this comparison was based on the rationale that in shallow orthographies the lexicon plays only a minor role in the naming process compared to its role in the lexical decision process. The opposite assumption, i.e., that even in shallow orthographies, a skilled reader always employs the orthographic route to accessing the lexicon, predicts that readers in all three orthographies should perform similarly. On the other hand, if the orthographic depth hypothesis is correct, the greatest difference between naming and lexical decision reaction times should be in Serbo-Croatian, which has the shallowest orthography while Hebrew should show the greatest similarity. Results were in line with the orthographic depth hypothesis; naming times were considerably faster than lexical decision times in Serbo-Croatian but, in Hebrew, lexical decision and naming looked quite similar. In Hebrew, it took as long to name a word as to recognize it: a suggestion that naming was accomplished postlexically. In addition, in Serbo-Croatian, the faster responding for naming versus lexical decision was even greater for pseudowords than for words. In these comparisons, English was intermediate. Thus, the results support the hypothesis that the shallower the orthography, the greater the amount of phonological recoding that is carried out in naming. Subsequent experiments in this study, which maximized the potential for lexical processing by semantically priming target words and by varying the relative number of pseudowords further supported this interpretation.

In all the experiments we have discussed to this point, the experimental paradigms used have been naming and lexical decision tasks. These tasks have a disadvantage as methodologies because the phonological variation that is used to affect the subject's response (e.g., the consistency of the grapheme - phoneme relation) is obtained through manipulating the orthography (e.g., different alphabets) and not by manipulating the putative phonology directly. The experimenter never observes any phonologic recoding; its presence is only inferred. Thus, one can not be certain that the differences that are observed are true effects of phonological recoding or, instead, are only the result of orthographic effects which happen to be correlated with phonology. Frost and Katz (1989) addressed this issue by introducing a paradigm in which subjects had to compare a spoken word and a printed word. This paradigm requires subjects to perceive and use phonology in their task processing. Subjects were required to simultaneously read and listen to two words presented by computer and judge whether or not they were the same (i.e., represented the same lexical item). In order to make the comparison, the subject had to mentally place both spoken and printed stimuli into a common representation. This could have been done, in principle, in several ways, although only two possibilities seemed reasonable. The spoken word could have been imagined as a spelled (printed) word or subjects could have generated the phonology of the printed word. The evidence indicated that subjects chose the latter. This was not surprising: subjects have had far more practice reading than spelling. After converting the printed stimulus to a phonological representation, both phonological representations could then have been compared in order to determine if they matched. Over a list of 144 or more trials, subjects made the judgment "Same" or "Different" about each pair of printed and spoken words on each trial. There were three conditions: clear speech and clear print, degraded speech (noise added) and clear print, and clear speech and degraded print (visual noise added). Serbo-Croatian and English native speakers were tested on comparable materials. The effects of degrading were marked; when either the print or the spoken word was degraded, performance declined sharply. However, the difference in latency between the slower responses to print or speech that had been degraded compared to clear print or speech was four times greater in the orthographically deep English than the shallower Serbo-Croatian; degradation had a much stronger deleterious effect in English.

An interactive activation network model can be extended easily to account for these results. The model contains parallel orthographic and phonologic systems that are each multilevel with lateral connections between the two systems at every level. In particular, the sets of graphemic and phonemic nodes are connected together in a manner that reflects the correspondences in a particular orthography: mainly isomorphic connections in a shallow orthography and more complex connections in a deep orthography. The simple isomorphic connections in a shallow orthography should enable subjects to use the printed graphemes to activate their corresponding (unambiguous) phonological nodes, supplementing weaker activation generated more directly by degraded speech. This higher, aggregated, activation should reach threshold fast compared to a network representing a deep orthography with its weaker grapheme-phoneme connections.

Evidence against the orthographic depth hypothesis

We mentioned above the study by Seidenberg (1985) who studied word naming in Chinese and English. In English, he found no difference between regularly spelled words and exception words as long as their word frequency was high, suggesting that phonologic representations play no role in naming frequent words. Differences were found, however, for low frequency words, exception words having the longer latencies. In Cantonese, an analogous pattern was found: There was no significant latency difference between phonograms (compound characters that contain both a phonetic component and a logographic signific) and non-phonograms (characters that contain no phonetic)—as long as they were high frequency. For low frequency items, phonograms were named faster. However, what seemed to drive naming latency most strongly in both languages was word frequency. The results suggest that the effect of frequency, an effect that was similar in both orthographies, may be of overriding importance in determining which kind of lexical access code is successful; differences between orthographies that can affect the coding of phonology may be irrelevant when frequent words are compared.

Besner and Hildebrandt (1987) capitalized on the fact that Japanese is written in three script systems. One of the scripts is a logography that is derived from the Chinese but one in which the characters are pronounced differently. Two of these scripts are essentially syllabic orthographies, katakana and hiragana. Historically, their graphemes evolved from separate sources but they both address the same phonology, similar in this regard to the dual Cyrillic and Roman alphabets of Serbo-Croatian. Unlike Cyrillic and Roman, however, the Japanese scripts are rarely used to write the same words; instead, they "specialize," being used for mutually exclusive vocabularies. In Japanese, then, the pronunciations of those words that are logographic must be recalled lexically. However, those words that are normally printed in katakana and those words that are normally printed in hiragana can, in principle, be pronounced via grapheme-syllable correspondences. In a simple but direct experiment that compared only the katakana and hiragana scripts, subjects named words that were printed either in their normal script or in the other script. When printed in the normal script, naming times were 47 to 65 milliseconds faster than when printed in their atypical script. Besner and Hildebrandt interpreted the results to mean that subjects were not using grapheme-syllable correspondences in order to pronounce the normally printed stimuii because changing the visual-orthographic form had been detrimental; if subjects had been assembling the phonology for naming, they would not have been slowed by the change in grapheme-syllable script system. Thus, there is reason to suspect that Japanese readers always adopt a visual-orthographic mode for naming no matter the depth of the script system they are reading: the deep logography or the shallow syllabaries. Besner and Smith (1992) offer further details.

Additional evidence against the orthographic depth hypothesis was reported by Baluch and Besner (1991). They studied naming in Persian, an orthography which offers a comparison between words that omit the vowels, like all words in Hebrew (opaque words) and other words that are spelled with a relatively full phonological specification, like Serbo-Croatian (transparent words). The difference lies in the representation of the vowels; opaque words have one or more of certain specific vowels that can be written as diacritics but, instead, are typically omitted from the spelling (as in Hebrew) while transparent words contain only those vowels that are never omitted. The authors found that semantic priming equally facilitated both transparent words and opaque words; the weak orthographic depth hypothesis would predict less facilitation for the transparent words, which can be pronounced largely via assembled phonology, needing lexical information only for syllable stress. Differences between opaque and transparent words did appear when pseudowords were included in the list of words to be pronounced; then, only the opaque words were facilitated. The inclusion of pseudowords presumably biased the subject toward the use of addressed phonology as the default because the pseudowords would have had no addressed phonology and grapheme-phoneme correspondence rules were therefore an effective alternative. Apparently, when the recognition process is biased in this way, transparent words are no longer processed via visual-orthographic coding and they are no longer facilitated by semantic priming. Their results suggest that in normal reading, where there are no pseudowords, subjects may always use the direct lexical route, without the use of assembled phonology. Baluch and Besner note that in all the studies in which subjects used phonological recoding, pseudowords had been included in the stimulus list of words, perhaps biasing subjects toward an atypical processing strategy (e.g., (Katz et al., 1983; Frost et al., 1987; etc.).

A similar point was made by Tabossi and Laghi (1992). In a clever series of experiments, they showed that semantic priming effects in naming, which are indicative of lexical processing, disappeared or were attenuated when pseudowords were introduced. The implication is, then, that visual-orthographic coding was the preferred strategy for their subjects. The authors interpret their results to suggest that assembled phonology is produced only under artificial conditions such as when pseudowords are present. Because their experiments used a shallow orthography (Italian), the authors suggest that all orthographies, shallow and deep, use the same mechanism for processing print, i.e., the visual-orthographic route.

However, the alternative explanation, the ODH, is not directly addressed by their study (with one exception, discussed shortly). As we suggest above, no standard experiment on a single script system will be able to test the claim of the ODH that the amount of lexical involvement is greater in shallow than in deep orthographies. The ODH is a statement about relationships *among* orthographies; it does not categorically disallow the use of either assembled phonology or visual-orthographic processing for any orthography (except for the strong form of the hypothesis, which is clearly unacceptable on rational grounds). Thus, it is still not known if some phonological processing is occurring in Italian but is not being observed because of special characteristics of the experiment itself (e.g., the task, stimuli, etc.). As in all situations in which the experimenter is pressing the null hypothesis ("Can we show that *no* phonological processing is occurring?"), the better the set of alternative models for comparison, the more convincing the outcome. In the case of the ODH, the more convincing argument against it would be to show that manipulations that affect semantic facilitation cause identical effects (do not cause differential effects) between Italian and some orthography that is deeper than Italian. Tabossi and Laghi (1992) do, in fact, make this test; when they do, they find evidence that is consistent with the ODH. Semantically primed words were named faster than controls in English but not in Italian, suggesting that naming involved the lexicon more in the deeper English than in Italian. However, both word lists contained pseudowords which would tend to increase the amount of phonological processing for both.

A similar demonstration was made in the Frost, Katz, and Bentin (1987) study. In Experiment 3, the authors explicitly examined the same hypothesis that put forward by Baluch and Besner (1991). In this experiment the ratio of words to pseudowords in the stimulus list was manipulated and its effect on naming was measured in Hebrew, English, and Serbo-Croatian. The results showed marked differences in the pseudoword ratio effect in the three different orthographies. Whereas in Serbo-Croatian the inclusion of pseudowords had almost no effect on naming latencies (consistent with the notion that assembled phonology is the preferred strategy for that orthography), much larger effects were found in English and Hebrew. The point raised by Baluch and Besner is indeed important; pseudowords can, in fact, affect naming strategies. However, this issue has no direct relevance to the ODH. The ODH suggests that the relative effect of pseudoword inclusion should be different in deep and in shallow orthographies. The results of both Frost et al. (1987), and Tabossi and Laghi (1992) are compatible with this notion.

Sebastián-Gallés (1991) presented evidence that was said to be inconsistent with theories that propose different mechanisms for shallow and deep orthographies. Spanish subjects pronounced pseudowords that had been derived from real words; each pseudoword was orthographically similar to its counterpart, except for one or two letters. In some pseudowords, the correct pronunciation of a critical letter (*c* or *g*) changed from its pronunciation in the real word according to grapheme-phoneme correspondence rules, because of a change in the following vowel. In other pseudowords, there was no such change in the vowel (and, therefore, no change in the pronunciation from the real word model). Subjects pronounced about 26% of the change pseudowords contrary to the correspondence rules while only about 10% of the no-change pseudowords were pronounced in that way. Sebastián-Gallés interpreted this result to mean that subjects were using a lexical strategy for pronouncing pseudowords. But this interpretation is warranted only if the theory being tested allows *no* lexical involvement at all in naming: The author was attacking the strong form of the ODH. A closer look at the evidence suggests that the data are, in fact, consistent with the weak ODH. Seventy-four percent of the change pseudowords were pronounced in accordance with spelling-to-sound correspondence rules and only 26% were pronounced "lexically." Thus, a mix of the two processes may have been at work In a second experiment, comparing lexical decision and naming times, Sebastián-Gallés found a moderate correlation (.455) for latencies between the two tasks. When the latency on each task was correlated with word frequency (presumed to be an index of lexical involvement), the correlation was of greater magnitude for lexical decision (-.497) than for naming (-.298). This result is consistent with a continuity of lexical involvement, naming being under weaker lexical control than lexical decision. A final experiment in this series showed semantic priming for naming under conditions where it is usually not found in shallow orthographies, viz., when pseudowords are included in the list of stimuli. Such results do suggest more lexical involvement in naming for a shallow orthography than other research has suggested. Nevertheless, in the author's conclusion, Sebastián-Gallés interprets the sum of the results to mean that "... lexical access from print in Spanish involves the use of orthographic information during at least some of the processing time," a statement that is consistent with the weak ODH.

Concluding remarks

Our working hypothesis has been that all alphabetic orthographies make some use of assembled phonology for word recognition. The proposal that a mixture of prelexical and visual-orthographic information is used for word recognition is consistent with the weak form of the ODH. The approach we suggest is in line with those models that contain dual phonological and orthographic representations, such as the dual route models (see Paap, Noel, & Johansen, 1992) and network models (see Se.denberg, 1992). The question of just how prevalent the use of phonology is, relative to the use of visual-orthographic coding, for any given orthography is an open one and is not addressed by the ODH itself. It could even be the case that the predominant lexical access code for frequent words in the shallow Serbo-Croatian is actually visual-orthographic or, on the other hand, that the predominant code in Hebrew is based on the partial phonological information that can be assembled from the unpointed letters (although either possibility seems unlikely from the present evidence). The ODH does not specify what degree of orthographic depth determines predominance for, say, visual-orthographic coding. Of course, orthographic depth will not be the only determiner of dominance; the reader's experience should play a major role as well.

Which of the codes is dominant might be determined by a mechanism like the following. Assume that the processing system is capable of using either code: a dual-code model. Suppose that a phonological representation is the *default* code for any given word but processing of a word via its phonological code can be replaced by processing via its visual-orthographic representation when the word has been experienced by the reader a sufficient number of times: a word frequency criterion. The premise that phonology is the default code is based on the fact that it is typically the code of instruction and the beginning reader receives much practice in its use. (One piece of evidence supporting the notion that the default code is phonological is that even adult readers of Hebrew prefer to use phonological information if it is available; Frost, in press). The criterion word frequency that is required in order to replace processing by assembled phonology with processing by visual-orthographic representations should be a function of the costs involved—in part, the cost for assembling the phonological representation and using it to access the lexicon. A higher replacement criterion will obtain in a shallow orthography where assembled phonology is easy to generate than in a deeper orthography.

What other factors affect the cost? Besides the ease of assembling phonology, a second factor is the ease with which phonology can be used in the lexical search process itself. Likewise, the ease of generating a visual-orthographic representation that is suitable for lexical search, the ease with which *that* information can address the lexicon, and, of course, the cost involved in establishing the visual-orthographically coded lexical representation in the first place, all need to be evaluated in establishing the criterion. The tradeoff between the generation of information that can be used for lexical access and the access process itself is important. Visual-orthographic codes may be costly for both the beginning and skilled readers to generate, particularly if a word is morphologically complex (it may require decomposition, in that case). However, the search process based on a visual-orthographic representation may be rapid for the skilled reader once he or she has a well-established visual-orthographic representation in lexical memory. Phonological codes may be difficult to generate but, once obtained, may be a fairly natural (i.e., low cost) way of addressing lexicon; after all, our primary lexicon, the speech lexicon, is based on phonology. Although the phonological lexicon may have taken years for a child to develop, it is thereafter available free to the reading process. However, it is obvious that we know very little about this tradeoff in terms of processing costs. The answer to the question about which of the two representations, phonological or visual-orthographic, is dominant depends on knowing more than we presently do about the perceptual and cognitive resources involved in word recognition. This question is discussed in some detail by Seidenberg, (1992). Nevertheless, we repeat that even in the absence of a fuller understanding of how the word recognition process draws on these resources, it is still a plausible (and testable) hypothesis that word recognition in shallow orthoгraphies will depend more on phonological representations simply because such information is available at less cost in those orthographies. Much of the evidence we have presented here is consistent with that hypothesis.

There has been substantial progress over the past decade in understanding the mechanisms behind naming and recognizing printed words. Importantly, research has involved an increasing variety of languages and writing systems, forcing theory to be general enough to encompass this wider scope. The vitality of this research is great and even shows signs of increasing in its pace. We hope that the main import of this article will be to clarify some issues in this area on the differential effects of writing systems on word perception. If there is presently significant (although perhaps not universal) agreement among researchers that visual-orthographic and assembled phonological representations may both play roles in word perception, then the next phase of research activity must include ways of assessing the conditions under which they are active, the relative contributions of each, and the mechanisms of their action. This requires a switch from research designs that address qualitative questions (e.g., "Is the lexicon accessed phonologically or not?") to designs that address the relative balance of phonological and visual-orthographic coding.

The best candidate for a heuristic framework for this research may be network modeling which offers a natural way of simulating the relationships between orthography and phonology, orthography and morphology, and phonology and morphology (at one level) and between these coded representations and the lexicon (at another). Likewise, differences between and within orthographies concerning the consistency, regularity, and frequency of these relationships can be implemented as initial constraints on such networks. Seidenberg (1992) suggests that network architectures offer ways of modeling how the various general cognitive resources involved are adapted to the processing of printed words: how the system assembles itself under the constraints of language, orthography, and memory. Implicit in this characterization is the additional possibility of modeling the historical evolution of a writing system. While the resources of memory, perceptual discrimination, and the like, may be constant, languages and their orthographies have not been immutable and their histories of change are well known in many cases. The orthographic depth hypothesis itself is a statement about a part of this larger issue of the fit between writing systems and human capabilities. The hypothesis embodies the assumption of covariant learning (Van Orden et al., 1991). That is, the structure and operation of the network should reflect the contingencies among phonology, morphology, and orthography that exist for printed words and, therefore, the contingencies that will be experienced by a reader. Each orthography, shallow or deep, defines its own pattern of contingencies. Additional progress in this area should come from requiring our ideas about the differences among orthographies to be made precise enough to be modeled.

REFERENCES

Baluch, B., & Besner, D. (1991). Visual word recognition: Evidence for strategic control of lexical and nonlexical routines in oral reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 644-652.

Besner, D., & Hilderbrandt, N. (1987). Orthographic and phonological codes in the oral reading of Japanese kana. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 335-343.

Besner, D., & Smith, M. C. (1992). Basic processes in reading: Is the orthographic depth hypothesis sinking? In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 45-66). Amsterdam: Elsevier Science Publishers.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper & Row.

Cossu, G., Shankweiler, D., Liberman, I. Y., Katz, L., & Tola, G. (1988). Awareness of phonological segments and reading ability in Italian children. *Applied Psycholinguistics, 9,* 1-16.

DeFrancis, J. (1989). *Visible Speech: The diverse oneness of writing systems.* Honolulu: University of Hawaii Press.

Feldman, L. B., & Turvey, M. T. (1983). Word recognition in Serbo-Croation is phonologically analytic. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 288-298.

Frost, R. (in press). Prelexical and postlexical strategies in reading: Evidence from a deep and a shallow orthography. *Journal of Experimental Psychology: Learning, Memory and Cognition.*

Frost, R., & Katz, L. (1989). Orthographic depth and the interaction of visual and auditory processing in word recognition. *Memory & Cognition, 17,* 302-310.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 104-115.

Katz, L., & Feldman, L. B. (1981). Linguistic coding in word recognition: Comparisons between a deep and a shallow orthography. In A. M. Lesgold &. C. A. Perfetti (Eds.), *Interactive processes in reading.* Hillsdale, NJ: Erlbaum.

Katz, L., & Feldman, L. B. (1983). Relation between pronunciation and recognition of printed words in deep and shallow orthographies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 157-166.

Liberman, I. Y., Liberman, A. M., Mattingly, I. G., & Shankweiler, D. L (1980). Orthography and the beginning reader. In J. F. Kavanagh &. R. L. Vanezky (Eds.). *Orthography, reading and dyslexia* (pp. 137-153). Baltimore: University Park Press.

Lukatela, G., & Turvey, M. T. (1990a). Phonemic similarity effects and prelexical phonology. *European Journal of Cognitive Psychology, 18,* 128-152.

Paap, K. R., Noel, R. W., & Johansen, L. S. (1992). Dual-route Models of Print to Sound: Red Herrings and Real Horses. In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 293-318). Amsterdam: Elsevier Science Publishers.

Perfetti, C. A., Bell, L. C., & Delaney, S. M. (1988). Automatic (prelexical) phonetic activation in silent word reading: Evidence from backward masking. *Memory and Language, 27,* 59-70.

Perfetti, C. A., Zhang, S., & Berent, I. (1992). Reading in English and Chinese: Evidence for a "universal" phonological principle. In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 277-248). Amsterdam: Elsevier Science Publishers

Scheerer, E. (1986). Orthography and lexical access. In G. Augst (Ed.), *New trends in graphemics and orthography* (pp. 262-286). Berlin: De Gruyter.

Sebastián-Gallés, N. (1991). Reading by analogy in a shallow orthography. *Journal of Experimental Psychology: Human Perception and Performance, 17(2)*, 471-477.

Seidenberg, M. S. (1985). The time-course of phonological code activation in two writing systems. *Cognition, 19*, 1-30.

Seidenberg, M. S. (1992). Beyond orthographic depth in reading: Equitable division of labor. In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 85-118). Amsterdam: Elsevier Science Publishers.

Seidenberg, M., Waters, G., Barnes, M. A., & Tannenhaus, M. K. (1984). When does irregular spelling and pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior, 23*, 383-404.

Tabossi, P., & Laghi, L. (1992). Semantic priming in the pronunciation of words in two writing systems: Italian and English. *Memory & Cognition, 20*, 315-328.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review 97*, 488-522.

Van Orden, G. C., Stone, G. O., Garlington, K. L., Markson, L. R., Pinnt, G. S., Simonfy, C. M., & Brichetto, T. (1992). "Assembled" phonology and reading: A case study on how theoretical perspective shapes empirical investigation. In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 249-292). Amsterdam: Elsevier Science Publishers.

FOOTNOTES

[*] In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 67-84). Amsterdam: Elsevier Science Publishers (1992).

[†] Also Department of Psychology, University of Connecticut, Storrs.

[‡] Department of Psychology, Hebrew University.

An Examination of "The Simple View of Reading"*

Lois G. Dreyer[†] and Leonard Katz[‡]

This study assesses the generality and predictive validity of Gough and Tunmer's (1986) model of the reading process, The Simple View, which holds that reading comprehension can be predicted by just two of its central components: decoding ability and linguistic or listening comprehension. The Simple View model posits the relationships between these variables as R = D x L rather than R = D + L, such that there could be no reading comprehension where either decoding or listening comprehension equals zero. Subjects for this study were 137 English-speaking third grade students followed longitudinally. We found mixed support for the Simple View model itself but strong support for its components: whether configured as sum (D + L) or product (D x L), decoding and linguistic comprehension are essential factors in reading comprehension.

Theoretical models of reading can have specific implications for reading instruction (Beck & McKeown, 1986; Singer, 1985). A useful distinction between a *theory*, which explains a phenomenon and a *model*, which specifies the interrelationships between a particular theory's variables, mechanisms and constructs, is provided by Singer and Ruddell (1985). By determining which factors are central to the reading process, models can guide teachers in making instructional decisions.

The purpose of the current study was to assess the generality and predictive validity of "The Simple View of Reading," a model proposed by Gough and Tunmer (1986), with support provided by Hoover and Gough (1990). In short, The Simple View holds that reading comprehension ability can be predicted by two components: decoding, defined as efficient word recognition, and linguistic comprehension, that is, using lexical, or word level information to achieve sentence and discourse interpretations (Hoover & Gough, 1990). Decoding and linguistic comprehension are both considered necessary for success in reading while neither of the two components is individually sufficient. The relationships, then, between reading comprehension (R), decoding (D) and linguistic or listening comprehension (L) were hypothesized in The Simple View model to be the following: R = D × L rather than R = D + L such that there could be no reading comprehension where either decoding or listening comprehension equals zero. Reading comprehension in this model is proposed to be the mathematical product of a child's decoding ability and listening comprehension.

We intended to examine how well The Simple View model predicts reading comprehension for a student population that was substantially different from the Spanish-English bilingual population studied by Hoover and Gough. In contrast to Hoover and Gough (1990), our subjects were monolingual English-speaking third graders who were receiving a uniform instructional program in reading. Should the product model provide a good explanation of reading comprehension at third grade level we also intended to assess the ability of the model to account for reading comprehension two years later, when the students were in the fifth grade. If the product model has

Preparation of this manuscript was supported by NICHD Grant #HD-01994 to Haskins Laboratories.

We wish to thank Karin Shields for assistance with data collection. Gratitude is also expressed to Dr. Robert Lichtenfeld and the faculty and students of the Katonah-Lewisboro Schools for their generous cooperation.

predictive as well as concurrent validity this would be a strong test of the model's adequacy.

Method

There were 166 third grade students in four elementary schools in a largely homogeneous middle class suburban school district. The reading curriculum was uniform across the school district and reflected a strong code emphasis in the early grades. In October of their third grade year, 137 of the students who had been in the school district for at least one full school year and who were monolingual (English-speaking) received an individually administered decoding test developed for use in a study of memory factors in decoding ability (Dreyer, 1989; Dreyer & Bryant, submitted). This was a 60-item test of decoding low frequency phonetically regular single syllable real words varying in length and orthographic complexity. None of the items were likely to be known as sight words. Following the Hoover and Gough (1990) study, the current investigation was a secondary analysis on an existing data set.

One week after the decoding test was given, the school district administered its annual achievement battery, the Educational Records Bureau Comprehensive Testing Program II, Level 2, Form C (Educational Testing Service, 1987). Scores on the third grade reading comprehension and listening comprehension subtests of the battery were obtained from school records. Two years later reading comprehension scores at fifth grade level were also obtained from school records. These were available for 121 of the original 137 subjects.

Results and Discussion

Pearson product moment correlations for all measures are shown in Table 1. As can be seen, in this student population, decoding and listening comprehension are highly related to reading comprehension at both third and fifth grade levels. As Hoover and Gough observed in their study, the relation of listening comprehension to reading comprehension increased from third to fifth grade level. However, unlike Hoover and Gough, in our population the relation of decoding to reading comprehension did not decrease, but rather remained stable.

Hoover and Gough (1990) assess the Simple View model (R = D × L) by making and evaluating three predictions. *Prediction One* is that the product of decoding (D) and listening comprehension (L) will account for significant variance in reading comprehension (R) over and above the contribution of the linear combination or sum of D and L. The linear formula is $R = a + b_1D + b_2L$, a standard regression formula. The model that includes both the linear and the product is $R = a + b_1D + b_2L + b_3[D \times L]$. Table 2 presents the squared multiple correlation coefficient, R^2, together with significance tests for the contribution of the product over the linear.

Table 1. *Intercorrelations among measures.*

Measure	2	3	4	5
1. Comp3	.69***	.38***	.62***	.62***
2. Comp5		.46***	.62***	.63***
3. LstComp			.24**	.84***
4. Decoding				.69***
5. Product				

Comp3 = Reading comprehension at third grade
Comp5 = Reading comprehension at fifth grade
LstComp = Listening comprehension at third grade
Product = Product of decoding and listening comprehension indices

**$p < .01$
***$p < .001$

Table 2. *Summary of regression analyses.*

Variable	R square	*R* square change	*F*	df
		Grade 3		
Linear	.439		52.4 ***	2,134
Product	.453	.014	3.4 ns	1,133
Product	.379		82.3 ***	1,135
Linear	.453	.074	18.0 ***	2,133
		Grade 5		
Linear	.469		52.2 ***	2,118
Product	.488	.019	4.1 *	1,117
Product	.401		79.5 ***	1,119
Linear	.488	.081	18.5 ***	2,117

Linear = Linear combination of the decoding and listening comprehension indices.
Product = Product of the decoding and listening comprehension indices.

*$p < .05$
***$p < .001$

Hierarchical multiple regressions were performed, following the analyses of Hoover and Gough. With reading comprehension for third graders as the dependent variable, decoding and listening comprehension scores were entered into the regression; together they accounted for 43.9% of reading comprehension variance. Adding the product vector (D × L) to the two independent variables (D) and (L) increased the proportion of variance accounted for by 1.4%. This increment was about the same size as that found by Hoover and Gough; however, it was not statistically significant in this sample. Also following Hoover and Gough, as a second test of the importance of the product of decoding and listening comprehension (D × L), reading comprehension was regressed first on the product alone: R square was in this case equal to 37.9%. Adding the linear combination of decoding and listening comprehension (D + L) to this regression raised the proportion of variance accounted for by 7.4%, a significant increase. Thus, with these analyses, there is better evidence for a model based on the sum of decoding and listening comprehension (R = D + L) than a model based primarily on the product of these two variables.

An additional set of regressions was performed using third grade decoding and listening comprehension as predictors, as before, but with fifth grade reading comprehension as the outcome variable. The use of fifth grade reading comprehension as a criterion provides a particularly strong test of the product model because any spurious contributions to correlations among variables that are measured simultaneously (like decoding, listening comprehension and reading comprehension) are weakened when a variable (i.e. reading comprehension) is measured two years later. Here, the product gives a significant increment of 1.9% over the linear. However, the reverse order showed that the linear accounted for more, over and above the product (8.1%). Thus, the test of Prediction One did not provide very strong evidence for the superiority of the product of decoding and listening comprehension over the linear combination of these two factors as a model of the reading process.

Prediction Two. If reading comprehension is the mathematical product of D and L, then the correlation between decoding and listening comprehension should change as reading comprehension changes (see Hoover and Gough (1990) Figure 1 and Table 3). Specifically, the correlation should be high and positive when reading comprehension is strong and should decline to zero when reading comprehension is zero. This is tested by taking successive subsamples of the data such that all subjects are included in the first sample, the best readers (as measured by comprehension) are excluded from the next sample, additional good readers are removed for the third sample, and so on, until only the poorest readers are left. We again used as criteria both third grade and fifth grade reading comprehension. As can be seen in Tables 3 and 4 the correlations for the four overlapping samples do, in fact, decrease for third grade reading comprehension, from a high of .241 to a low of -.019. The first value is significantly different from zero while the second value is not. Although these results are in line with the prediction of The Simple View, they appear to be artifactual.

When the sampling procedure was reversed, beginning with a subsample of the 42 readers with the highest third grade comprehension scores and successively adding more and more poorer readers, we found that the pattern of correlation reversed, instead of remaining similar to the first series of correlations. For fifth grade comprehension the picture is even less clear, with the correlation pattern varying unsystematically. Thus the results of our analysis again provided mixed evidence for The Simple View.

Prediction Three. If the product model holds, then the regression of reading comprehension on listening comprehension should maintain the same intercept as decoding ability decreases but the slope of the function should decrease (again see Figure 1 in Hoover and Gough, 1990.) For this analysis the data were divided into three groups on the basis of decoding score. The results are presented in Table 5.

When third grade reading comprehension was regressed on listening comprehension (L) for each of these three groups, we found that as Hoover and Gough predicted, the intercept remained fairly constant and the slope decreased (from .5342 to .2925) as decoding ability decreased. However, when fifth grade reading comprehension was the criterion, there was no such monotonic change in slope and even the intercepts varied nonmonotonically. Again, we feel these results are mixed with regard to support for The Simple View of Reading.

Table 3. *Descriptive statistics and correlations between decoding and listening comprehension for successive sample reductions based on decreasing reading comprehension skill at Grade 3.*

Quartile	n	Correlation	Descriptive statistics (mean and standard deviation)			
		r	D	L	P	R
		Series 1				
1 - 4	137	.241 **	28.82	12.36	363.74	25.78
			7.32	4.22	164.10	7.07
1 - 3	116	.198 *	28.01	11.79	336.33	24.23
			7.60	7.60	155.07	6.52
	65	.094 ns	25.38	10.85	278.43	19.77
			8.39	3.99	140.36	5.31
1	33	-.019 ns	22.12	10.18	224.61	15.67
			8.59	3.91	117.08	4.26
		Series 2				
4	42	-.153 n.s.	33.05	14.95	492.36	33.00
			3.44	3.48	119.90	1.9
3 - 4	72	.138 n.s.	31.92	13.74	440.75	31.24
			4.33	3.97	145.42	2.68
2 - 4	104	.176 n.s.	30.94	13.06	407.88	29.01
			5.38	4.10	152.08	4.13
1 - 4	137	.241 **	28.82	12.36	363.74	25.78
			7.32	4.22	164.10	7.07

D = Decoding index
L = Listening comprehension index
P = Product of decoding and listening comprehension
R = Reading comprehension
* $p < .05$
** $p < .01$

Table 4. *Descriptive statistics and correlations between decoding and listening comprehension for successive sample reductions based on decreasing reading comprehension skill at Grade 5.*

Quartile	n	Correlation	Descriptive statistics (mean and standard deviation)			
		r	D	L	P	R
		Series 1				
1 - 4	121	.298 **	29.55	12.37	373.17	26.36
			5.84	4.36	159.29	6.50
1 - 3	92	.174 n.s.	28.23	11.52	329.80	24.82
			6.04	4.39	146.82	6.52
1 - 2	63	.199 n.s.	25.38	10.78	297.65	22.51
			8.39	4.34	143.29	6.03
1	30	.239 n.s.	23.77	10.17	247.87	19.90
			5.96	4.53	126.42	5.79
		Series 2				
4	29	.365 *	33.76	15.07	510.76	31.28
			1.94	3.00	113.53	3.14
3 - 4	58	.038 n.s.	32.19	14.10	455.21	30.55
			4.03	3.70	133.90	3.87
2 - 4	91	.134 n.s.	31.46	13.10	414.48	28.50
			4.38	4.08	147.38	5.19
1 - 4	121	.298 **	29.55	12.37	373.17	26.36
			5.84	4.36	159.29	6.50

D = Decoding index
L = Listening comprehension index
P = Product of decoding and listening comprehension
R = Reading comprehension
* $p < .05$
** $p < .01$

Table 5. *Regressions of reading comprehension at third and fifth grade on listening comprehension by decoding ability.*

Decoding level %	n	Intercept	*R* square
		Grade 3	
75 - 100	26	23.526	.211 *
50 - 74	51	23.333	.123 **
0 - 49	53	17.718	.034 n.s.
		Grade 5	
75 - 100	37	11.052	.508 **
50 - 74	34	18.513	.083 n.s.
0 - 49	50	8.573	.088 n.s.

* $p < .05$
**$p < .01$

A limitation in this study should be acknowledged. There was a ceiling effect on our decoding measure such that approximately 8% of the subject population performed at ceiling level. Had the decoding measure included nonwords and multisyllabic words, the results might have more closely replicated the findings of Hoover and Gough.

It is intriguing that such a complex activity as reading could be strongly predicted by just two of its central components. Whether configured as sum or product, it is clear that decoding and linguistic comprehension are essential factors in reading comprehension, as Hoover and Gough suggest. The general theory underlying The Simple View of Reading is therefore clearly supported by our results. In testing the specifics of the model itself, however, we found, as others have (e.g. Stanovich, Cunningham and Feeman, 1984; P. B. Gough, personal communication, May 29, 1991) that the sum of decoding and listening comprehension accounts for so much variance in reading comprehension that there may be little room left for improvement by the product.

What are the implications for reading instruction? We believe that our findings confirm the importance of word recognition skills in reading. For those children who do not develop decoding skills for themselves through exposure to print, explicit systematic instruction and opportunities for practice in meaningful contexts would seem to be essential to reading progress and should be an integral part of a rich classroom reading and language program.

REFERENCES

Beck, I. L., & McKeown, M. G. (1986). Application of theories of reading to instruction. In N. L. Stein (Ed.), *Literacy in American schools: Learning to read and write* (pp. 63-83). Chicago, IL: University of Chicago Press.

Dreyer, L. G. (1989). *The relationship of children's phonological memory to decoding and reading ability.* Unpublished doctoral dissertation, Columbia University.

Dreyer, L. G., & Bryant, N. D. (submitted). Phonological memory as a component of decoding ability in reading.

Educational Records Bureau Comprehensive Testing Program II (1987). Princeton, NJ: Educational Testing Service.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7,* 6-10.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2,* 127-160.

Singer, H. (1985). Models of reading have direct implications for instruction: The affirmative position. In J. A. Niles (Ed.), *Issues in literacy: A research perspective* (pp. 402-413). Rochester, NY: National Reading Conference.

Singer, H., & Ruddell, R. B. (1985). Introduction to models of reading. In H. Singer & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (3rd ed., pp. 620-629). Newark, DE: International Reading Association.

Stanovich, K. E., Cunningham, A. E., & Feeman, D. J. (1984). Intelligence, cognitive skills, and early reading progress. *Reading Research Quarterly, 19,* 278-303.

FOOTNOTES

[*] In C. K. Kinzer & D. J. Leu (Eds.), *Literacy research, theory, and practice: Views from many perspectives,* 41st Yearbook of the National Reading Conference (pp. 169-175). Chicago, IL: National Reading Conference.

[†] Also Southern Connecticut State University.

[‡] Also University of Connecticut.

Phonological Awareness, Reading, and Reading Acquisition: A Survey and Appraisal of Current Knowledge*

Shlomo Bentin†

Phonetic production and perception are part of the natural endowment of the human race. As soon as infants can be tested, they show an ability to distinguish between phonetic categories (e.g., Kuhl, 1987; Kuhl & Meltzoff, 1982; Molfeese, & Molfeese, 1979) and very early in life they are able to use phonetic elements and a few rules of combination to form phonologic structures that represent words. Children's phonological perception ability is, in fact, admirable. Even though the several phonetic gestures that are included in a phonological structure are co-articulated and therefore their acoustic effects overlap, very young children are able to decipher the phonetic code and distinguish between words on the basis of single phonemes (Eimas, 1975; Eimas, Miller, & Jusczyk, 1987; Eimas, Sequeland, Jusczyk, & Vigorito, 1971; Morse, 1972). Moreover, the deciphering of the phonetic code requires very little attention and effort. These findings lead several investigators to propose that the perception of speech is accomplished by a precognitive process controlled by a distinct biological module which is specialized to recover the coarticulated gestures from the acoustic stream and provide the cognitive system with unequivocal phonological information (Liberman & Mattingly, 1989; Mattingly & Liberman, 1990).

In contrast to their well developed phonological ability young children cannot reflect on or intentionally manipulate structural features of spoken language. Most four-to-five year-old children will not be able, for example, to tell what a word's first phoneme is, or how the word ends. Putting it differently, young children do not have the metalinguistic ability that would enable them to manipulate sub-word phonological elements (Bruce, 1964; for a recent review and an alternative perspective see Goswami & Bryant, 1990). This metalinguistic ability has been labeled *phonological awareness* (Liberman, 1973; Mattingly, 1972). The study of phonological awareness is important because the last two decades of research have provided ample evidence for its intimate relationship with reading acquisition and skill. In the present chapter I examine the nature of phonological awareness, its acquisition and development, and its role in reading acquisition.

Forms and levels of phonological awareness

By definition, awareness should be an a'l-or-none aptitude. In support of this view, studies in our laboratory (Leshem, unpublished doctoral dissertation), as well as in others (Calfee, Chapman, & Venezky, 1972; Stanovitch, Cunningham, & Cramer, 1984) revealed that the distribution of children's performance on tests of phonemic segmentation is bimodal: on a particular test, individual scores were either very high or very low. Additional support to this view was provided by several authors who have shown that pre-school children as well as illiterate adults can learn initial consonant deletion within a single session if they are provided with corrective feedback (Content, Kolinsky, Morais, & Bertelson, 1986; Morais, Content, Bertelson, Cary, & Kolinsky, 1988). Other authors, however, postulated that the development of explicit representation of phonemic structures could well be gradual (Content et al., 1986). This view was

This study was supported by the Israel Foundations Trustees and by NICHHD Grant HD-01994 to Haskins Laboratories.

based on results showing that children's performance on different tests of phonological awareness varied considerably (e.g., Stanovitch Cunningham, & Cramer, 1984). For example, preschool children are relatively successful in rhyme detection tasks (Bradley, & Bryant, 1983; Lenel & Cantor, 1981; Maclean, Bryant, & Bradley, 1987), can accurately count the number of syllables in words (Liberman, Shankweiler, Fisher, & Carter, 1974), but they cannot isolate single phonemes (Liberman, Shankweiler, Liberman, Fowler, & Fisher, 1977; Rosner & Simon, 1971). The "all-or-none" view of awareness and the variability in performing different tests of phonological awareness can be reconciled by assuming that phonological awareness is a heterogenic metalinguistic competence involving abilities that differ in developmental trends and origins. Indeed, several recent reports emphasized the heterogeneous nature of phonological awareness (Bertelson & de Gelder, 1989; Bertelson, de Gelder, Tfouni, & Morais, 1989). In order to understand what the different forms of phonological awareness might be, we should first survey the ways phonological awareness has been assessed.

Because phonological awareness refers to the phonological structure of spoken words, phonological awareness tests require the ability to either detect, isolate, or manipulate sub-word phonological segments (or some combination of the above). Some tests require these aptitudes explicitly. These are, for example, phoneme isolation ("What is the first/last sound in desk ?"; e.g., Bentin, Hammer, & Cahan, 1991; Wallach & Wallach, 1976), phoneme segmentation ("What sounds do you hear in the word hot?"; e.g., Fox & Routh, 1975; Williams, 1980), phoneme counting ("How many sounds do you hear in the word cake?"; Liberman et al., 1974; Yopp, 1985), and specifying a delcted phoneme ("What sound do you hear in cat, that is missing from at?"; Bentin & Leshem, in press; Stanovich et al., 1984). In other tests, correct performance requires sensitivity to sub-word phonological segments, although awareness of those segments is not explicitly tested. Such tests are, for example, the detection and/or production of rhyme ("Does sun rhyme with run?"; e.g., Calfee et al., 1972; Maclean et al., 1987), word-to-word matching ("Do pen and pipe begin the same?"; e.g., Bentin et al., 1991; Wallach & Wallach, 1976), phoneme reversal ("Say on with the first sound last and the last sound first"; Alegria, Pignot, & Morais, 1982), and phoneme deletion ("What would be left if you took out the /t/ from told?"; e.g., Bruce, 1964; Rosner, 1975; Morais, Cary, Alegria, & Bertelson, 1979). Tests also differ in the size of the segment they refer to. Some tests require awareness of single phonemes while others require awareness of sub-syllabic segments such as the word's onset or rime[1] (Kirtley, Bryant, Maclean, & Bradley, 1989; Treiman, 1985) or of syllabic segments (e.g., syllable counting; Liberman et al., 1974). Hence, phonological awareness was tested in many ways and, apparently, the observed level of "phonological awareness" was determined to some extent by the particular tests used.[2] The above survey suggests that tests of phonological awareness may differ along at least three dimensions: 1) operation required (detection, isolation, or manipulation of the phonological segment); 2) manner of testing awareness of phonological codes (indirect or explicit); and 3) size of the relevant phonological segment (syllabic, sub-syllabic, phonemic). Although the above dimensions are not entirely orthogonal (most detection tests, for example, are also indirect), a detailed examination of previous reports shows that the performance on different tests of phonological awareness varied systematically along all three dimensions.

Regardless of their size, detection of phonological segments was better than isolation, while the manipulation of segments was the poorest and latest accomplished task. For example, 29 out of 66 four-year-old children were able to detect the one word (out of three) which did not rhyme with the others, but only 8 where able to produce rhymes to target words (Maclean et al., 1987). Similarly, most studies revealed that children in kindergarten are usually very poor at isolating one phoneme of a word (Bentin et al., 1991; Lundeberg, Frost, & Petersen, 1988) or repeating an utterance after deleting one phoneme (e.g., Bruce, 1964; Rosner & Simon, 1971; Content et al., 1986), but they are more successful when they have to match words or detect oddity among words on the basis of only one phoneme (Content et al., 1986; Stanovich, Cunningham, & Crammer, 1984; Yopp, 1988). Children are more aware of syllabic and subsyllabic segments than they are of phonemic segments. For example, children start detecting rhymes and common phonemic clusters at the onset of a word much before they can match words on the basis of single phonemes (e.g., Bradley & Bryant, 1983; 1985). Similarly, preschool children are considerably more accurate in counting the syllables than the phonemes included in words (Cossu, Shankweiler, Liberman, Katz, & Tola, 1988; Liberman et al., 1974; Treiman & Baron, 1981), and the same is true for more sophisticated manipulations of syllables vs. phonemes such as segmentation (Fox & Routh, 1975; Lundberg et al., 1988) and reversal (Content, Morais, Alegria, & Bertelson, 1982; Mann, 1984). Finally, it appears that children's performance is better when phonological awareness is tested indirectly than in explicit tests. For example, counting the number of phonemes (particularly when tokens or wooden blocks are used) is more accurate than "spelling out" the sounds of a word (Yopp, 1988).

The considerable variation in pre-schoolers' performance on different tests of phonological awareness was mentioned and discussed by several authors. However, most of these authors were concerned primarily with the selection of the tests that were most reliable and best correlated with reading skills (e.g., Golnikoff, 1978; Lewkowicz, 1980; Torneous, 1984). Other authors simply partitioned the different tests into coherent groups (e.g., Content et al., 1986) or established a hierarchy of tests according to relative difficulty (e.g., Roberts, 1975; Stanovich et al., 1984). Only a few authors used this variation to analyze the nature and components of phonological awareness. A notable exception is the factor analysis that was recently performed by Yopp (1988) on the results of kindergarten children in ten tests of phonological awareness. The factor analysis revealed that only two factors accounted for most of the variance. Tests of phonemic segmentation, sound isolation, and phoneme counting had high loadings on Factor 1 and low loadings on Factor 2. Tests requiring the deletion of phonological segments and tests of word matching on the basis of single phonemes had moderate to high loadings on Factor 2 and low loadings on Factor 1. Because the tests that loaded Factor 2 required more steps to completion and placed a greater burden on short-term memory than the tests that loaded Factor 1, Yopp suggested that Factor 2 reflects a *Compound Phonemic Awareness* whereas Factor 1 reflects a *Simple Phonemic Awareness*. Hence, Yopp explained the variation in performance along the "operation" dimension by assuming that different levels of operation vary in the number of steps required for test completion. A stepwise regression analysis of reading scores on phonological awareness showed that both factors were good predictors of reading ability. This, however, is not surprising, since it turns out that simple phonemic awareness is in fact included in compound phonemic ability. Therefore, although the description of the two factors might help explain the variability of phonological awareness measures, it adds very little to the explanation of the relationship between phonological awareness and reading.

One of the most reliable sources of variation in phonological awareness performance is the size of the test-relevant segment. Most studies of phonological awareness showed that most preschool children can segment words into syllables but cannot manipulate or isolate single phonemes (Bruce, 1964; Calfee, 1977; Calfee, Lindamood, & Lindamood, 1973; Fox & Routh, 1975; Hakes, 1980; Liberman et al., 1974; Lundberg et al., 1988; Rosner & Simon, 1981; Treiman & Baron, 1981; Zhurova, 1963). Other studies found that four-year-old children can detect rhymes and can match words on the basis of common subsyllabic segments (e.g., Bradley & Bryant, 1985; Kirtley et al. 1989) but are unable to match words on the basis of single phonemes (Maclean et al., 1987). A possible explanation of the difference between children's ability to detect, count, and manipulate syllables or sub-syllabic clusters and their performance with single phonemes is to assume, as Content et al., (1986) did, that phonological awareness is a gradually developing ability, or that there are "levels" of phonological awareness (e.g., Goswami & Bryant, 1990). However, it is also possible that there is a qualitative distinction between the awareness of single phonemes and the awareness of multiphonemic structures which accounts for the observed difference in performance with the two types of segment. In other words, it is possible that awareness and manipulation of single phonemes and detection and sensitivity to syllabic or intrasyllabic structures are qualitatively different forms of phonological awareness rather than two levels along a continuum of one ability. I will try to defend this qualitative distinction.

As a consequence of the process of coarticulation that characterizes speech production, the sound frequency patterns forming acoustic segments in speech reflect the combined contribution of several complex gestures, each intended to produce a different phone. Moreover, because a phone can be coarticulated with different phonetic contexts, there can be no direct correspondence in segmentation between the acoustic signal and the phonetic message it conveys. Therefore, speech perception cannot be based on a simple translation from a set of auditory representations to a set of perceptual phonetic categories. Consequently, awareness of each of the phonemes conveyed by one acoustic segment probably follows a more basic and automatic process of phonetic deciphering. This is probably why, although phonetic distinctions in speech are easy and natural, awareness of phonetic categories appears much later in ontogenetic development and probably requires more than simple cognitive maturation. This awareness require the ability to break up the coarticulated phonological segments and isolate their individual phonemic constituents.

The above analysis implies that segmentation should be relatively easy when the required phonological units correspond to perceived acoustic segments but difficult when the disentangling of coarticulated phones is required. Coarticulated phonological units usually include a highly resonant nucleus (a vowel) flanked by one or several consonants, together forming a syllable. Therefore, syllabic segmentation can be based on simple auditory perception and might not reflect genuine phonological awareness. This view also suggests that the isolation of stop consonants should be significantly more difficult than the isolation of steady-state vowels because the former have no independent acoustic existence—they are always coarticulated. The latter hypothesis, however, is only partly supported by empirical evidence. Previous studies of initial phoneme isolation (Bentin & Leshem, in press) and initial phoneme deletion (Content et al., 1982; Content et al., 1986) suggested that the performance of preschool children was better with vowels than with consonants; that order of difficulty was reversed, however, when the last (rather than the first) phoneme had to be isolated: Final consonants were easier to isolate than final vowels (Bentin & Leshem, in press). A similar pattern was found when performance with stop consonants was compared to performance with fricatives (Content et al., 1986).

In contrast to the commonly reported failure of pre-literate children to isolate and manipulate single phonemes which are perceived in coarticulated form, most studies demonstrate that children are considerably more successful in detecting and producing rhymes. The sensitivity to rhymes might be taken as evidence for a second form of phonological awareness because it also requires the breaking of coarticulated phonetic clusters. For example, the recognition that the monosyllabic words "beg" and "leg" rhyme involves breaking them into b-eg and l-eg segments, and recognizing that the end segments of each syllable sound alike. Because the same children cannot usually tell that /b/ is the first and /g/ is the last phone in "beg," it is conceivable that rhyme detection and phonemic segmentation require different phonological skills. The most outstanding attempt to explain this difference was made by Peter Bryant, Lynette Bradley and their collaborators at Oxford.

The basic idea advocated by the Oxford group is that there are linguistically valid segments intermediate between single phonemes and syllables. These segments were labeled *onset* and *rime*. The onset is the consonant or string of consonants that precedes the vowel in a syllable, and the rime is the rest of the syllable. For example the onset of the monosyllabic word "black" is /bl/ and its rime is /ack/. Note, that although the onset and the rime are phonologically defined units, the validity of this distinction was based either on observing the nature of errors in speech (MacKay, 1972), or on linguistic constraints on sequences of phonemes (Halle & Vergnaud, 1980). Hence, the validity of this phonological categorization is not based on phonetic considerations and is very different from the distinction between phonetic categories that was discussed above. Nevertheless, awareness of this intrasyllabic segmentation and the ability to manipulate these segments require, as mentioned above, breaking the coarticulated unit of perception, and they may therefore be considered a form of phonological awareness.

Words that rhyme share, by definition, the same rime. Therefore, the reliable demonstrations that four-year-old children can detect and produce rhymes proves that they are aware of the rimes of syllables. Are they also similarly aware of the onsets? That evidence is less compelling. Kirtley et al. (1989) attempted to demonstrate such an awareness, however most of their evidence is based on negative findings and their interpretations are speculative. In their study they used oddity tasks with different word sets. First they replicated the finding that it is easier to find an "odd" word among four when the commonality is based on the initial consonant than on the final consonant. Their interpretation of this phenomenon was that the initial consonant formed the whole onset of the word whereas the final consonant was only a part of the rime. However, in order for this interpretation to hold unequivocally they would have had to show that when the initial consonant was only a part of the onset (such as the /s/ in "string") its detection should be more difficult. Unfortunately such a comparison has not been attempted. Moreover, our own observations (Leshem, unpublished doctoral dissertation) suggest that the opposite is true. Our five-year-old subjects were more successful in isolating the initial consonant in words that began with a CCV string than in words that began with a CV string.

In a second experiment Kirtley et al. (1989) used different oddity combinations aimed at distinguishing between situations in which the odd word could be detected on the basis of a full intrasyllabic segment or required the breaking of such segments. In all the conditions the results supported the prediction that it is easier to detect oddity on the basis of intact intrasyllabic structures. However, the same results could be interpreted solely on the basis of special sensitivity to the rime, without making any assumptions about the onset. Moreover, as in the first experiment, no multi-phonemic segments were used, and so the "onset" was always confounded with a single initial phoneme. Nevertheless, it should be stressed that the authors' interpretation is not counter-intuitive and may be right. At this time, however, all we can say is that there is strong evidence for a particular sensitivity to the rime of syllables which may have been induced by extensive experience with rhymes. Moreover, this form of sensitivity was shown only in detection tasks. To the best of my knowledge, awareness of onset and rimes has never been shown in tests of segmentation or isolation. Hence, it is possible that children who are able to detect rhymes and correctly select the odd word in an oddity tests are sensitive to sub-syllabic units but still unable to point out the phonological segment on which their decision is based. Consequently, it is possible that sensitivity to sub-syllabic segments, either as suggested by the Oxford group or limited to rimes only, reflects a qualitatively different form of phonological awareness than sensitivity to single phonemes. Later in this chapter I discuss the relevance of both forms of phonological awareness to reading.

In conclusion, this section shows that there are, in fact, only two forms of phonological awareness: One which is demonstrated by the ability to isolate segments and manipulate single phonemes, and one demonstrated by sensitivity to the rime and perhaps to the onset of syllables The first requires explicit knowledge about the phonemic segment and, therefore I will label it "phonemic awareness"; the second is reflected indirectly in the detection of oddity and commonality between words on the basis of subsyllabic segments. I will label this second form "early phonological awareness." Other tests vary along different dimensions but do not reflect any separate ability.

Factors influencing the development of phonological awareness

Clearly, phonological awareness is not an innate aptitude; it is probably triggered by some experience. Therefore, the questions of "how" and "when" this skill appears have frequently been raised and have been the subject of much controversy. Some authors claimed that phonological awareness is triggered, or at least considerably enhanced, by exposure to the alphabet (e.g., Bertelson et al., 1985; Bertelson & de Gelder, 1990). Others proposed that phonological awareness develops a long time before children learn to read, through experiences which at the time have nothing to do with reading (e.g., Bryant & Bradley, 1985). As we will see, however, these are not mutually exclusive theories, because each of the proponents is actually talking about of a different form of phonological awareness.

There is ample evidence that learning to read affects phonological awareness skills. For example, using consonant addition and deletion tasks, Read, Zhang, Nie, and Ding (1986) found well-developed phonological awareness in Chinese subjects who learned to read a recently developed Chinese alphabetic system (Pinyin) but not among subjects who read only the logographic system (Kanji). The mean percentage of correct performance was 83% in the former group but only 21% in the later. Along the same lines, Mann (1986) reported that first-graders in Japan who learned how to read a syllabary (Kana) were good at manipulating syllables but significantly inferior to American first-graders in manipulating phonemes. Equivalent results were found with Belgian children in the first grade; those who learned to read according to the "analytic" (segmental) method performed better on tests of phonemic segmentation than those who learned to read by the "global" (holistic) method (Alegria, Pignot, & Morais 1982). However, the strongest support for the view that in the absence of reading acquisition phonemic segmentation skills do not develop spontaneously is provided by a series of studies by Morais and his colleagues showing that illiterate adults perform very poorly on tests of phoneme deletion, although they may manipulate phonology at syllabic and word levels (Morais et al. 1979, 1986, 1987). Similar results were also found with semi-literate adults (Read & Ruyter, 1985) and with the reading disabled (Bryne & Ledez, 1983). The ability of the illiterate subjects to manipulate multi-phonemic units as opposed to single phonemes is congruent with findings in preschool children.

As reviewed in the previous section, there is no doubt that some three-year-old children and most four-year-old children recognize and play with rhymes (e.g., Chukovsky, 1963). Formal testing has shown that when either detection through oddity or the production of rhyme and alliteration was involved, many three- and four-year-old children could make judgments about the component sounds (particularly rimes) in words that they heard or uttered (Maclean et al., 1987). The significant ability of pre-literate children to detect rhymes as well as to perform above chance in oddity tests based on sensitivity to subsyllabic segments (Bradley & Bryant, 1985; Kirtley et al., 1989) lead to the conclusion that phonological awareness exist before reading acquisition (Maclean et al., 1987). Note, however, that the apparent disagreement between the two views stems from a different definition of phonological awareness. Those who found signs of phonological awareness in three and four-year-old children refer primarily to what we called the early form of phonological awareness—the one which focuses on subsyllabic segments and is tested indirectly. In contrast, the defenders of the alternative view, (i.e., that phonological awareness is triggered by the exposure to the alphabetic principle), refer to phonemic awareness—the one reflected in the ability to explicitly manipulate single phonemes deciphered from the coarticulated unit. This fact was recently recognized by both parties (Bertelson et al., 1989; Goswami & Bryant, 1990).

Having resolved the above controversy, we are still left with several important questions. What are the factors that affect the development of the two forms of phonological awareness? Is the early form a precursor of the later form? Does the early form develop spontaneously or is it the result of explicit or implicit instruction? Is exposure to the alphabetic principle the only factor influencing the development of phonemic awareness? Can the development of phonemic awareness be accelerated and achieved prior to exposure to the alphabetic principle? The available literature may provide answers to some of these questions.

The impressive studies reported by Bradley and Bryant (1985) prove beyond any reasonable doubt that explicit training with sound categorization improves performance on oddity tests based on rime and onset. In other words, the early form of phonological awareness can be significantly improved in kindergarten by explicit training. This, however, does not prove that this metaphonological ability cannot occur spontaneously. A direct answer to this question requires a rigid control of children's pre-test experience with rhymes. Obviously, it is practically impossible to control children's experience in life, and so we are forced to address this question only indirectly. For example, in a longitudinal study by Maclean et al. (1987), young children's performance on different tests of rhyme and alliteration detection as well as their knowledge of nursery rhymes, was related to their socio-economical background and their parents' education. Although it is a rough estimate, it would not be completely wrong to assume that children of middle-class highly educated parents had more opportunities to be exposed to nursery poems and other forms of rhymes than chidden coming of lower-class poorly educated parents. Therefore, a significant difference between the performance of the two groups might indicate that experience with rhymes is a critical trigger of the early phonological awareness. The results of this comparison suggested that at the earliest age tested (3 years old) children coming from the "privileged" homes were more successful in the detection of alliteration than the other children, but there were no differences in the detection of rhymes, and both groups were equally knowledgeable about nursery rhymes. Moreover, even the small difference did not last. There was no sign of influence of family background after the initial tests. On the basis of these results, and considering that illiterate adults who showed no phonemic awareness were nevertheless sensitive to rhyme judgments and vowel deletion (e.g., Bertelson et al., 1989), we may safely conclude that the early phonological awareness, which does not require awareness of single phonemes, can be easily triggered without explicit instruction and may develop independently of reading acquisition.

In contrast to their sensitivity to syllabic and sub-syllabic phonological segments, evidence from studies with illiterates suggests that the ability to isolate and manipulate single phonemes that are coarticulated in speech (i.e., phonemic awareness) does not develop spontaneously (Morais et al., 1979; Morais, Bertelson, Cary, & Alegria, 1986). These authors proposed that learning to read an alphabetic orthography provides most children (and adults) with the opportunity to develop full phonemic awareness. In contrast to speech, where individual phonemes are coarticulated, in writing the phonemes are represented by clearly defined orthographic segments, the letters. Assuming that children learn about these letter-sound correspondence when they learn to read, it seems likely that during the acquisition of reading skills they become explicitly aware that words are formed of the sounds which the letters represent. Indeed, most studies revealed a significant gap between the phonemic segmentation skills of first-graders and of kindergarten children. For example, Liberman et al. (1974) found that none of the pre-kindergartners and only 17% of the kindergartners tested were able to parse words into phonemes, while 70% of the first-graders tested succeeded in doing so.

A caveat about interpreting developmental studies of phonological awareness, and particularly the striking improvement in phonemic segmentation ability during the first grade, is that all such studies share the serious problem of the possible confounding of differences in the extent or method of reading acquisition with other age-related variables that may have influenced phonological awareness (e.g., the amounts of informal linguistic experience and general cognitive development). In addition, the comparison of illiterate and ex-illiterate adults may be compromised, for example, because the choice to join a literacy program in adulthood was probably not arbitrary. Therefore, before definite claims about a causal relationship between reading acquisition and the emergence of phonemic awareness could be made, it was still necessary to isolate the effect of reading acquisition on the appearance and development of awareness of individual phonemic segments.

Owing to the impossibility of experimenting with elementary school attendance, previous attempts to control for general age-related effects on phonological awareness were based on comparisons between the youngest and the oldest children within one grade level (Bowey & Francis, 1991), or between the oldest children in the kindergarten and the youngest children in the first grade (Bowey & Francis, 1991; Morrison, 1988). Although suggestive, this approach suffers from a serious shortcoming of selection, because the cutoff date for school admission is never strictly imposed. Moreover, the exceptions are not random: Intellectually advanced children who are slightly younger than the official school age are often admitted, while children who are somewhat older than the cutoff point but insufficiently developed may be held back an additional year (Cahan & Davis, 1987; Cahan & Cohen, 1989). This creates a situation of "missing" children in each grade, particularly among children at the extreme ages. Such selective misplacement usually leads to overestimation of the schooling effect (Cahan & Cohen, 1989).

In a recent study Bentin, Hammer, & Cahan (1991) proposed a solution to this problem. Rather then comparing empirically obtained data from children at the extreme ages in each grade, the authors predicted these data on the basis of the best fitting regression of test scores on chronological age, across the entire legal age range in each grade, with the exclusion of the selection-tainted birth dates near the cutoff point. The separate effects of schooling and one year of age were estimated by means of a regression discontinuity design (Cook & Campbell, 1979) involving the regressions of phonemic segmentation scores on chronological age. The effect of age was reflected by the slope of the within-grade regressions, whereas the effect of schooling was reflected in the discontinuity between the two regression lines. The results of this analysis are presented in Figure 1.

As evident in Figure 1, the percentage of correct responses on the phonemic segmentation battery was higher in school children (76%, SD=14%) than in the kindergarten group (35%, SD=23%) ($t(674)=29.12$, $p<.0001$). However, this difference reflected the combined effects of age and schooling. The comparison of the independent schooling and age effects revealed that, although both effects were statistically significant, the effect of schooling (reading acquisition) was four times as large as the effect of one year of chronological maturation.

The results of the Bentin et al. (1991) study pointed to schooling (learning to read) as a major factor affecting the development of phonological awareness. This is not to say, however, that exposure to an alphabetic orthography is the only way to trigger phonemic awareness. There is ample evidence for the efficiency of tuition in metaphonological skills outside the context of reading acquisition (e.g., Lundberg et al., 1988). Significant improvement in phonemic segmentation skills were obtained using different training methods such as the use of visual aids to represent phonemes (Elkonin, 1973; Lindamood & Lindamood, 1969), the designing of speech-correction games played with puppets that impersonated human speakers (Content et al., 1982), simply using corrective information during testing in successive blocks (Content et al., 1986), and designing speech-sound oriented group games (Olofsson & Lundberg, 1983). In all these studies, however, the experimental groups were selected from a normal population of children.

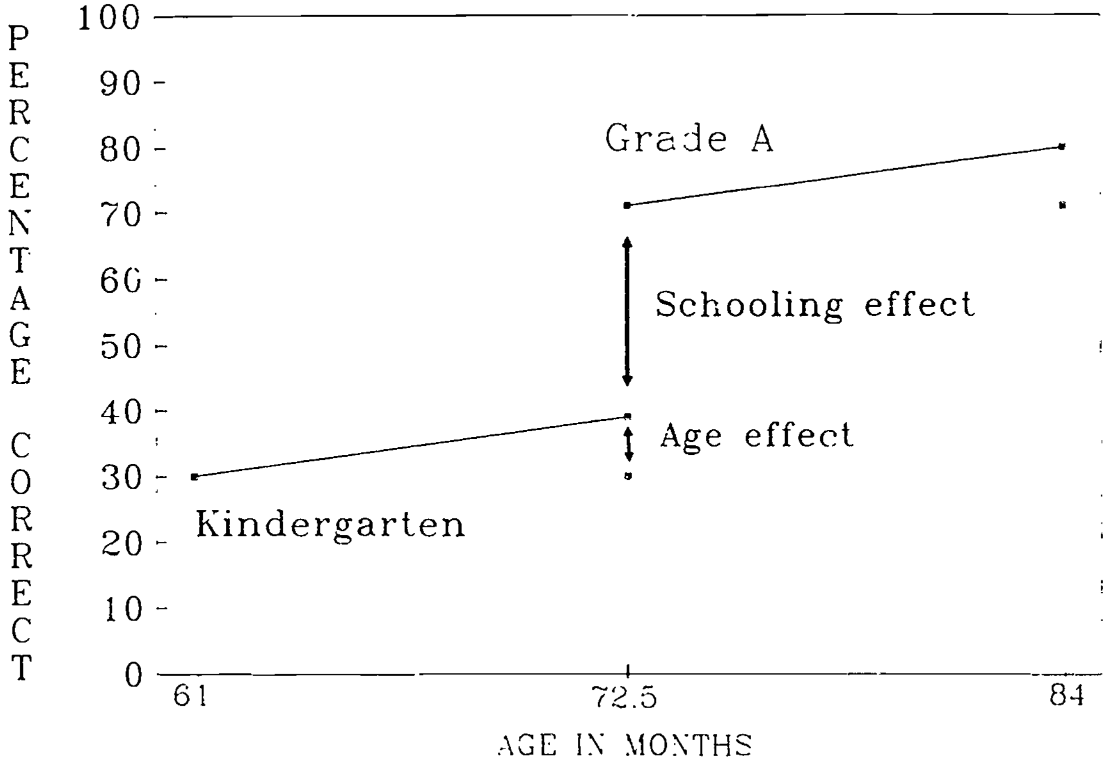


Figure 1. **Schooling and age effects on the development of phonological awareness in kindergarten and Grade A children.**

Moreover, in many of the previous studies the control groups were not trained for other language abilities (but see Ball & Blachman, 1991). Therefore, the specific effect of training in phonological awareness may have been confounded with the positive effects that training in general linguistic skills may have on reading acquisition and might have been limited to linguistically well-developed children. In a recent study we rectified these problems, finding that training in segmentation skills significantly improved phonemic segmentation ability in five-year-old children who were initially at the lower end of the distribution of scores on a battery of phonemic segmentation tests (Bentin & Leshem, in press). Moreover, in that study we found that children who had been trained in phonemic segmentation were able to apply their newly acquired metaphonological skills in other tests of phonological awareness.

In conclusion, the presently available evidence suggests that early phonological awareness, i.e., the ability to detect and produce rhymes and the sensitivity to subsyllabic segments, develops differently from phonemic awareness (i.e., the ability to isolate and manipulate individual phonemes in speech). The former appears to emerge almost automatically and instantaneously in the great majority of children when they are first exposed to nursery rhymes or other forms of phonological word games and develops independently of reading instruction. The latter, on the other hand, is triggered in most children when they come to understand the alphabetic principle during the acquisition of reading in an alphabetic orthography. However, phonemic awareness can be also be triggered and full phonemic awareness can be developed in pre-readers by explicit training of phonemic segmentation skills. There is no direct evidence for interdependence between the two forms of phonological awareness. It is conceivable, however, that well-developed awareness of rhymes and subsyllabic segments is necessary for a smooth acquisition of phonemic awareness during reading instruction. In other words, it is possible that a well-developed early phonological awareness is a prerequisite for the emergence of phonemic awareness without explicit instruction. Indirect evidence for this hypothesis is reviewed in the next section.

Reading and phonological awareness: It's a two-way street

Although studying the development of metaphonological skills is important in its own right, the significance of phonological awareness is considerably enhanced by its well-established relationship with the acquisition of reading skills. Many studies have demonstrated that children's performance in various phonological awareness tests highly correlates with their reading skill in the early school grades in English (Bradley & Bryant, 1985; Calfee et al., 1973; Fox & Routh, 1975; Liberman et al., 1977; Rosner & Simon, 1971; Treiman & Baron, 1981; Tunmer & Nesdale, 1986), as well as in other languages such as Italian (Cossu et al., 1988), Swedish (Lundberg, Olofsson, & Wall, 1980), Spanish (de Manrique & Gramigna, 1984), French (Bertelson, 1987), and Hebrew (Bentin & Leshem, in press). Correlative studies were applied in developing tools for predicting success in reading (Blachman, 1984; Juel, Griffith, & Gough, 1986; Lundberg et al., 1980; Mann, 1984; Share, Jorm, MacLean, & Matthews, 1984); however, they tell us very little about the nature of the relationship. A high positive correlation might exist between two independent skills if they are similarly affected by a third factor. On the other hand, it is also possible that the correlation reflects a causal relationship, as, for example, when one skill is a pre-requisite or trigger for the second.

Theoretical considerations suggest that phonological awareness and the acquisition of the alphabetic principle are directly interdependent, and that the positive correlation might reflect mutual influence and even causal relations between these two skills. The alphabet is the latest and probably the most advanced form of writing (DeFrancis, 1989). One of its most important virtues is that, like speech, it uses a relatively small set of well-defined symbols (the letters) that can be combined in a practically infinite number of ways to represent all the possible words in a language. The representation of words by orthographic patterns is efficient only because the basic units of writing, the letters, are mapped onto the basic units of speech, the phones. Thus, words are not represented in writing by arbitrary and holistically distinguished patterns but rather, the combination of letters that represents a particular word is fully determined by the sequence of phonemes of which the word is composed. Hence, in order to understand a written word the reader must be able to decipher the phonological unit from its written form. Even assuming that a fluent reader may form direct associations between some written patterns and their meanings, and use these associations to access the semantic information directly, the ability to decipher phonology from writing is a prerequisite for reading and understanding written words at the first encounter, and needs to be mastered before efficient reading can occur. This is the essence of the alphabetic principle, and this is the reason why reading and writing require a reasonable awareness of the internal phonological structure of spoken words. (For detailed discussion of these considerations see, for example, Ehri, 1979; Leong, 1986; Liberman, 1989; Liberman & Liberman, 1990; Liberman, Shankweiler, & Liberman, 1989; Rozin & Gleitman, 1977.)

The above account for the reading process implies that, regardless of the particular teaching method adopted by the teacher, in the process of learning to read children learn the basic mapping rules from the domain of letters to the range of phonemes. Obviously, the acquisition of mapping rules requires explicit knowledge of the members of the domain and of the range. The items in the domain (the letters) are explicitly taught by the teacher. On the other hand, the members of the range (the phonemes) are not explicitly taught in the classroom. When children start learning to read they are expected to be aware of the phonological structure of spoken words, or at least to become aware of it very quickly. Indeed, as reviewed in the previous section, most children become aware of the phonemic structure of spoken words fairly easily, as a consequence of exposure to the alphabet, which leads to the understanding of the alphabetic principle. Unfortunately, for a significant proportion of children mere exposure to the alphabet is not sufficient, and they consequently develop a reading disability. Several studies have demonstrated that these children may be helped by explicit training in phonological awareness in parallel to reading acquisition (Perfetti, Beck, Bell, & Hughes, 1987; Wallach & Wallach, 1976; Williams, 1980) or preferably during kindergarten (Ball & Blachman, 1991; Bentin & Leshem, in press; Bradley & Bryant, 1983, 1985; Lundberg, Frost, & Peterson, 1988; Vellutino & Scanlon, 1984). A survey of these studies may shed additional light on the metaphonological prerequisites of reading acquisition.

In their initial longitudinal study, Bradley and Bryant (1985) trained four- and five-year-old children to categorize words on the basis of initial sound, and to be aware of that common sound. Some of the children were also given experience with plastic letters. Children in control groups were trained for conceptual categorization or received no training whatsoever. When they reached school, the reading, spelling and mathematical ability of the children in the four groups were compared. The results of these comparisons showed that the children who had been trained to categorize words on the basis of initial phonemes were better in reading and spelling than the children in the control groups, whereas the mathematical skills of all four groups were equal. It was also found that the reading and spelling performance of children who were given experience with plastic letters in addition to phonemic categorization surpassed that of children who were trained only in sound categorization. Finally, a follow-up of this study (Bradley, 1989) revealed that the advantage gained by the experimental groups was maintained five years later: At the age of 13 years, their reading performance was still better than that of the control groups. It is important to note that in this early study the children were trained to make phonemic distinctions, because in more recent publications the Oxford group seems to be convinced that the form of phonological awareness important for reading acquisition is sensitivity to the onset and rime of syllables (Goswami & Bryant, 1990; Maclean et al., 1987).

Although a substantial positive correlation was found between the early phonological awareness and reading acquisition, I doubt that awareness of subsyllabic segments alone is sufficient for understanding the alphabetic principle First, although several letter-strings frequently appear together (for example */ing/*), many do not. In fact, in reading, as in speech, the distinction among words is frequently based on one letter. Therefore, I am in greater agreement with the following conclusion drawn by the same group: "... a major step in learning to read may take place when the child learns to break the rime into its constituent sounds by detaching... the preceding vowel from the final consonant" (Kirtley et al., 1989). In fact, there is ample evidence that in training phonemic segmentation facilitates reading acquisition (e.g., Ball & Blachman, 1991; Cunningham, 1988, in press; Lundberg et al., 1988) but not a single study in which training *only* in rhyming skills facilitated reading acquisition. In this context it is interesting to mention our own training study (Bentin & Leshem, in press), because the structure of Hebrew orthography, in which vowels are represented by diacritical marks appended to the consonants (see Frost & Bentin, this volume), would be the ideal orthography to make use of onset and rimes rather than phonemes in reading.

In our study we trained four groups of five-year-old children selected from the lower end of the distribution of scores on a phonemic segmentation test-battery. Group I was trained in phonemic segmentation; group II was trained in phonemic segmentation and also in recognizing letters of the alphabet and relating them to their sound; group III was trained in general linguistic abilities such as vocabulary enhancement, sentence comprehension, etc.; group IV received no training. Training, in groups of four children, lasted for 10 weeks with two 1/2-hour sessions per week. A year later, the reading performance of these children was assessed and compared with the performance of children who were comparable to the four training groups, except for being at the higher end of the distribution of scores of the initial phonemic awareness battery (Group V). The children were tested after four months and nine months of reading instruction. Each test consisted of lists of items that the children were instructed to read aloud. Two lists included words and two lists nonwords. The lists included an equal number of monosyllabic and disyllabic items. Table 1 presents the percentage of correctly read words in each group, for each stimulus type.

As evident in Table 1, reading skills in the first grade were significantly correlated with the phonemic segmentation skills that were assessed in the kindergarten before training, and were influenced by training segmentation skills. Because there are no standardized reading tests in Hebrew (except for reading comprehension) it is difficult to interpret the absolute scores. Note, however, that these tests were constructed in collaboration with the teachers in the respective schools and were designed to reflect the expected level of reading at each testing time.

Table 1. *Percentage of correctly read items (SEm) of each stimulus type after 4 months and 9 months of reading instruction. Note that different tests were given each time, to correspond with the respective reading level.*

STIMULUS TYPE	FIRST READING TEST				
	GROUP I	GROUP II	GROUP III	GROUP IV	GROUP V
WORDS					
One Syllable	85.5 (4.7)	76.7 (8.5)	63.4 (8.5)	59.9 (6.3)	94.5 (2.0)
Two Syllables	72.9 (4.9)	66.1 (7.9)	46.2 (7.8)	46.3 (7.4)	87.1 (3.7)
NONWORDS					
One Syllable	64.7 (7.9)	43.4 (9.7)	26.1 (7.0)	28.3 (8.2)	75.0 (5.2)
Two Syllables	58.4 (7.8)	45.3 (9.7)	25.5 (7.2)	24.4 (8.5)	64.9 (7.7)
	SECOND READING TEST				
WORDS					
One Syllable	66.6 (6.0)	70.7 (5.7)	35.2 (6.1)	38.6 (11.)	68.2 (4.8)
Two Syllables	63.9 (5.6)	66.2 (7.8)	21.8 (4.2)	29.6 (10.)	64.4 (4.5)
NONWORDS					
One Syllable	63.4 (6.3)	59.6 (7.7)	27.0 (5.3)	28.8 (11.)	67.4 (6.8)
Two Syllables	52.7 (5.8)	47.5 (9.7)	16.3 (3.5)	19.6 (8.6)	57.5 (7.5)

Therefore, it is suggestive to observe that the reading performance of children who were initially low in phonemic awareness and received no training in phonemic segmentation was about 40%, which according to school standards means failure. In contrast, children from the same population who received training and improved their phonemic awareness scored around 70%, almost as well as children who were initially high in phonemic awareness. These data are particularly important because we tested children who learn to read an orthography in which, because of its specific characteristics, the basic segment usually used by teachers for reading instruction is a consonant-vowel combination.

In conclusion, the evidence relating reading acquisition to phonological awareness is robust. It suggests that the alphabetic principle requires the ability to isolate and manipulate single phonemes in coarticulated speech. The major factor that triggers this ability is exposure to the alphabet. However, phonemic awareness cannot be triggered by the alphabet unless the early form of phonological awareness is well developed. Children who do not meet this prerequisite must be explicitly trained for phonemic segmentation. Our data show that training phonemic segmentation in kindergarten for a relatively short period is effective in inducing the metaphonological skills required for easy acquisition of reading. With younger children, however, or with children who are language-delayed the training program should probably begin with the establishment or improvement of sensitivity to rhymes and the ability to detect the onset and rime of the syllables.

REFERENCES

Alegria, J., Pignot, E., & Morais, J. (1982). Phonetic analysis of speech and memory codes in beginning readers. *Memory & Cognition, 10*, 451-556.

Ball, E. W., & Blachman, B. A. (1991). Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly, 26*, 49-65.

Bentin, S., Hammer, R., & Cahan, S. (1991). The effects of aging and first year schooling on the development of phonological awareness. *Psychological Science, 2*, 271-274.

Bentin, S., & Leshem, H. (in press). On the interaction of phonological awareness and reading acquisition: It's a two-way street. *Annals of Dyslexia.*

Bertelson, P., & De Gelder, B. (1990). The emergence of phonological awareness: Comparative approaches. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception.* Hillsdale, NJ: Erlbaum.

Bertelson, P., de Gelder, B., Tfouni, L. V., & Morais, J. (1989). Metaphonological abilities of adult illiterates: New evidence of heterogeneity. *European Journal of Cognitive Psychology, 1*, 239-250.

Bertelson, P., Morais, J., Alegria, J., & Content, A. (1985). Phonetic analysis capacity and learning to read. *Nature, 313*, 73-74.

Blachman, B. A. (1984). Language analysis skills and early reading acquisition. In G. Wallach & K. Butler (Eds.), *Language learning disabilities in school-age children* (pp. 271-287). Baltimore, MD: Williams and Wilkins.

Bowey, J. A., & Francis, J. (1991). Phonological analysis as a function of age and exposure to reading instruction. *Applied Psycholinguistics, 12*, 91-121.

Bradley, L. (1989). Predicting learning disabilities. In J. J. Dumont & H. Nakken (Eds.), *Learning disabilities: Cognitive, social and remedial aspects* (pp. 1-17). Amsterdam: Swets & Zeitlinger.

Bradley, L., & Bryant, P. (1983). Categorizing sounds and learning to read: A causal connection. *Nature, 301*, 419-421.

Bradley, L., & Bryant, P. (1985). *Rhyme and reason in reading and spelling.* Ann Arbor: University of Michigan Press.

Bryant, P., & Bradley, L. (1985). *Children's reading problems.* Oxford: Blackwell.

Bruce, D. J. (1964). The analysis of word sounds. *British Journal of Educational Psychology, 34*, 158-170.

Byrne, B., & Ledez, J. (1983). Phonological awareness in reading disabled adults. *Australian Journal of Psychology.* 35, 185-197.

Cahan, S., & Cohen, N. (1989). Age vs. schooling effects on intelligence development. *Child Development, 60*, 1236-1249.

Cahan, S., & Davis, D. (1987). A between-grades-level approach to the investigation of the absolute effects of schooling on achievement. *American Educational Research Journal 24*, 1-13.

Calfee, R. C., Chapman, R., & Venezky, R. (1972). How a child needs to think to learn to read. In L. W. Gregg (Ed.), *Cognition in learning and memory.* New York: John Wiley & Sons.

Calfee, R. C., Lindamood, P., & Lindamood, C. (1973). Acoustic-phonetic skill and reading—kindergarten through twelfth grade. *Journal of Educational Psychology, 64*, 293-298.

Chukovsky, K. (1963). *From two to five.* Berkeley: University of California Press.

Content, A., Kolinsky, R., Morais, J., & Bertelson, P. (1986). Phonetic segmentation in pre-readers: Effects of corrective information. *Journal of Experimental Child Psychology, 42*, 49-72.

Content, A., Morais, J., Alegria, J., & Bertelson, P. (1982). Accelerating the development of segmentation skills in kindergartens. *Cahiers de Psychologie Cognitive, 2*, 259-269.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design analysis issues for field settings.* Boston: Rand McNally.

Cossu, G., Shankweiler, D., Liberman, I. Y., Katz, L., & Tola, G. (1988). Awareness of phonological segments and reading ability in Italian children. *Applied Psycholinguistics, 9*, 1-16.

Cunningham, A. E. (1988). *A developmental study of instruction in phonemic awareness.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 1988.

Cunningham, A. E. (in press). Explicit vs. implicit instruction in phonemic awareness. *Journal of Experimental Child Psychology.*

DeFrancis, J. (1989). *Visible speech: The diverse oneness of writing systems.* Honolulu: University of Hawaii Press.

Eimas, P. D. (1975). Speech perception in infancy. In L.B. Cohen & P. Salapatek (Eds.). *Infant perception: From sensation to cognition* (Vol. 2). New York: Academic Press.

Eimas, P. D., Sequeland, E. R., Jusczyk, P. W., & Vigorito, T. 1971). Speech perception in infants. *Science, 171*, 303-306.

Eimas, P. D., Miller, J. L., & Jusczyk, P. W. (1987). On infant speech perception and the acquisition of language. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 161-195). Cambridge University Press.

Ehri. L. C. (1979). Linguistic insight: Threshold of reading acquisition. In T. G. Waller & G. E. MacKinnon (Eds.), *Reading research: Advances in theory and practice.*(Vol 1, pp. 63-114). New York: Academic Press.

Elkonin, D. B. (1963). The psychology of mastering the elements of reading. In B. Simon & J. Simon (Eds.), *Educational psychology in the U.S.S.R.* London: Routledge & Kagan Paul.

Fox, B., & Routh D. K. (1975). Analyzing spoken language into words, syllables and phonemes: A developmental study. *Journal of Psycholinguistic Research, 4*, 331-342.

Golnikoff, R. M. (1978). Phonemic awareness skills and reading achievement. In F. B. Murrary & J. J. Pikulski (Eds.), *The acquisition of reading: Cognitive, linguistic, and perceptual prerequisites.* Baltimore: University Park Press.

Goswami, U., & Bryant, P. (1990). *Phonological skills and learning to read.* East Sussex: Erlbaum.

Halle, M., & Vergnaud, J. (1980). Three dimensional phonology. *Journal of Linguistic Research, 1*, 83-105.

Juel, C., Griffith, P., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in the first and second grade. *Journal of Educational Psychology, 78*, 243- 255.

Kirtley, C., Bryant, P., MacLean, M., & Bradley, L. (1989). Rhyme, rime, and the onset of reading. *Journal of Experimental Child Psychology, 48*, 224-245.

Kuhl, P. K. (1987). Perception of speech and sound in early infancy. In A. Salapateck & L. Cohen (Eds.), *Handbook of infant perception: Vol. II, From perception to cognition* (pp. 274-282). New York: Academic Press.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*, 1138-1141.

Lenel, J C., & Cantor, J. H. (1981). Rhyme recognition and phonemic perception in young children. *Journal of Psycholinguistic Research, 10*, 57-68.

Leong, C. K. (1986). The role of language awareness in reading proficiency. In G. T. Pavlidis & D. F. Fisher (Eds.), *Dyslexia: Its neuropsychology and treatment* (pp. 131-148). London: Wiley.

Lewkowicz, N. K. (1980). Phonemic awareness training: What to teach and how to teach it. *Journal of Educational Psychology, 72,* 686-700.

Liberman, A. M. (1989). Reading is hard just because listening is easy. In C. von Euler (Ed.), *Wenner-Gren International Symposium Series: Brain and reading* (pp. 197-205). Hampshire, England: Macmillan.

Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science, 243,* 489-494.

Liberman, I. Y. (1973). Segmentation of the spoken word and reading acquisition. *Bulletin of the Orton Society, 23,* 65-77.

Liberman, I. Y. & Liberman, A. M. (1990). Whole word vs. code emphasis: Underlying assumptions and their implications for reading instruction. *Bulletin of the Orton Society, 40,* 51-76.

Liberman, I. Y., Shankweiler, D. Fischer, F. W., & Carter, B. J. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology, 18,* 201-212.

Liberman, I. Y., Shankweiler, D., & Liberman, A. M. (1989). The alphabetic principle and learning to read. In D. Shankweiler & I. Y. Liberman (Eds.), *Phonology and reading disability: Solving the reading puzzle* (pp. 1-34). Ann Arbor: the University of Michigan Press.

Liberman, I. Y., Shankweiler, D., Liberman, A. M., Fowler, C., & Fisher, F. W. (1977). Phonetic segmentation and recording in the beginning reader. In A. S. Reber & D. L. Scarborough (Eds.), *Toward a psychology of reading* (pp. 207-226). Hillsdale, NJ: Erlbaum.

Lindamood, C. H., & Lindamood, P. C. (1969). *The A.D.D. program: Auditory discrimination in depth.* Boston: Teaching Research Corporation.

Lundberg, I., Frost, J., & Peterson, O-P. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly, 23,* 263-284.

Lundberg, I., Olofsson, A., & Wall, S. (1980). Reading and spelling skills in the first school years predicted from phonemic awareness skills in kindergarten. *Scandinavian Journal of Psychology, 21,* 159-173.

MacKay, D. G. (1972). The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology, 3,* 210-227.

Maclean, M., Bryant, P., & Bradley, L. (1987). Rhymes, nursery rhymes, and reading in early childhood. *Merrill-Palmer Quarterly, 33,* 255-281.

Mann, V. A. (1984). Longitudinal prediction and prevention of early reading difficulty. *Annals of Dyslexia, 34,* 117-136.

Mann, V. A. (1986). Phonological awareness: The role of reading experience. *Cognition, 24,* 65-92.

de Manrique, A. M. B., & Gramigna, S., (1984). La segmentacion fonologica y silabica en ninos de preescolar y primer grado. *Lectura y Vida* 5, 4-13.

Mattingly, I. G. (1972). Reading, the linguistic process, and linguistic awareness. In J. F. Kavenagh & I. G. Mattingly (Eds.), *Language by ear and by eye.* Cambridge Mass: MIT Press.

Mattingly, I. G. & Liberman, A. M. (1990). Speech and other auditory modules. In G. M. Edelman, W. E. Gall, & W. M. Cowan (Eds.), *Signal and sense: Local and global order in perceptual maps* (pp. 501-519). New York: Wiley.

Molfeese, D. L., & Molfeese, D. L. (1979). VOT distinctions in infants: Learned or innate? In H. A. Whitaker, & H. Whitaker (Eds.), *Studies in neurolinguistics* (Vol 4). New York: Academic Press.

Morais, J., Content, A., Bertelson, P., Cary, L., & Kolinsky, R. (1988). Is there a sensitive period for the acquisition of segmental analysis? *Cognitive Neuropsychology, 5,* 347- 352.

Morais, J., Bertelson, P., Cary, L., & Alegria, J. (1986). Literacy training and speech segmentation. *Cognition, 24,* 45-64.

Morais J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition, 7,* 323-331.

Morse, P. A. (1972). The discrimination of speech and non-speech stimuli in early infancy. *Journal of Experimental Child Psychology, 14,* 447-492.

Morrison, F. J. (1988). Development of phonemic awareness: A natural experiment. Paper presented at the *Annual Meeting of the Psychonomic Society,* Chicago.

Olofsson, A., & Lundberg, I. (1983). Can phonemic awareness be trained in kindergarten? *Scandinavian Journal of Psychology, 24,* 35-44.

Perfetti, C. A., Beck, I., Bell, L. C., & Hughes, C. 1987. Phonemic knowledge and to read are reciprocal: A longitudinal study of first grade children. *Merrill-Palmer Quarterly, 33,* 283-319.

Read, C. A., & Ruyter, L. (1985). Reading and spelling in adults of low literacy. *Remedial and Special Education, 6,* 43-52.

Read, C. A., Zhang, Y., Nie, H., & Ding, B. (1986). The ability to manipulate speech sounds depends on knowing alphabetic reading. *Cognition, 24,* 31-44.

Roberts, T. (1975). Skills of analysis and synthesis in the early stages of reading. *British Journal of Educational Psychology, 45,* 3-9.

Rosner, J. & Simon, D. P. (1971). The auditory analysis test: An initial report. *Journal of Learning Disabilities, 4,* 384-392.

Rozin, P., & Gleitman, L. (1977). The structure of the acquisition of reading: II. The reading process and the acquisition of the alphabetic principle. In A. S. Reber and D. L. Scarborough (Eds.), *Toward a psychology of reading* (pp. 55-141). Hillsdale, NJ: Erlbaum.

Share, D. L., Jorm, A. F., MacLean, R., & Matthews, R. (1984). Sources of individual differences in reading achievement. *Journal of Educational Psychology, 76,* 1309-1324.

Stanovitch, K. E., Cunningham, A., & Cramer, B. (1984). Assessing phonological awareness in kindergarten children: Issues of task comparability. *Journal of Experimental Child Psychology, 38,* 175-190.

Torneous, M. (1984). Phonological awareness and reading: Chicken and egg problem? *Journal of Educational Psychology, 76,* 1346-1358.

Treiman, R. (1985). Onset and rimes as units of spoken syllables: Evidence from children. *Journal of Experimental Child Psychology, 39,* 161-181.

Tunmer, W. E., & Nesdale, A. R. (1986). Phonemic segmentation skill and beginning reading. *Journal of Educational Psychology, 77,* 417-427.

Vellutino, F. R., & Scanlon, D. (1987). Phonological coding, phonological awareness, and reading ability: Evidence from a longitudinal and experimental study. *Merrill-Palmer Quarterly, 33,* 321-363.

Wallach, M., & Wallach, L. (1976). *Teaching all children to read.* Chicago: University of Chicago Press.

Williams, J. P. (1980). Teaching decoding with an emphasis on phoneme analysis and phoneme blending. *Journal of Educational Psychology, 72,* 1-15.

Yopp, H. K. (1988). The validity and reliability of phonemic awareness tests. *Reading Research Quarterly, 23,* 159-177.

Zhurova, L. E. (1963). The development of analysis of words into their sounds by preschool children. *Soviet Psychology and Psychiatry, 2,* 17-27.

FOOTNOTES

*In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 193-210). Amsterdam: Elsevier Science Publishers (1992).

†Department of Psychology and School of Education, Hebrew University, Jerusalem, Israel.

[1] The "onset" and the "rime" are, respectively the consonant (or consonants) that precede the vowel, and the rest of the syllable (Halle & Vergnaud, 1980; MacKay, 1972). I will elaborate on these segments and their relationship to phonological awareness later in the text.

[2] An additional test that was often used is blending, i.e. the ability to form a word by synthesizing syllables or phonemes uttered by the experimenter (e.g., Fox & Routh, 1975; Lundberg, Frost, & Petersen, 1988). I think, however, that although this test requires the manipulation of phonological units, it does not require explicit deciphering of the phonological code, and therefore it examines a skill that is basically different from phonological awareness.

Morphological Analysis in Word Recognition*

Laurie B. Feldman[†] and Darinka Andjelković[‡]

One of the most refined techniques for investigating morphological processing in word recognition is the variant of the lexical decision task known as repetition priming (Stanners, Neiser, Hernon, & Hall 1979). It provides a primary source of evidence, according to Henderson (1989), of facilitation between words formed from the same morpheme (i.e., morphological relatives). Generally, target (second presentation) decision latencies and error rates are reduced in the context of morphologically-related primes (first presentation). Words related to the target (e.g., HEALS) can be forms that are unaffixed (e.g., HEAL), inflected (e.g., HEALED) or derived (e.g., HEALER) in either the same or different modalities (e.g., print or speech) and they can be separated by as many as fifty intervening items. Effects of morphological relatedness have been observed in the lexical decision task across a variety of languages including Serbo-Croatian (Feldman & Fowler, 1987) and Hebrew (Bentin & Feldman, 1990), English (Fowler, Napps, & Feldman, 1985; Feldman, 1991) and American Sign Language (Hanson & Feldman, 1991; see also Emmorey, 1989).

In this chapter, we review evidence of morphological processing in word recognition. In the first two sections, studies of morphology that employ the repetition priming techniques are described. In section one, morphological and orthographic similarity effects are contrasted because alternative accounts of morphological effects in word recognition often minimize the role of the morpheme unit and focus on orthographic and phonological similarity of morphologically-related words. In section two, morphological effects are differentiated from effects due to semantic association, although both are based on word meaning. The repetition priming procedure is a viable tool for studying how morphological relationships among words are represented in the lexicon; however, a confounding episodic contribution to the pattern of facilitation can also occur (e.g., Feustel, Shriffrin, & Salasoo, 1983). It is important, therefore, to differentiate morphological effects from episodic and other types of facilitation and to provide converging evidence of morphological analysis from other word recognition tasks. Morphological effects have also been observed in sentence verification and oral reading tasks. For example, facilitation due to shared morphemes (and/or shared syntactic structure based on the ordering of subject, object and verb constituents) in prime and target sentences have been obtained. In section three, morphological and syntactic facilitation effects are examined in a sentence comprehension task. Finally, it is useful to ask whether analysis of a word's constituent structure at the level of the morpheme is linked to analysis of that word at other linguistic levels. For example, the association between morphological and phonological processing in beginning readers has been examined by comparing performances on such tasks. In section four, a common underlying skill for morphological and phonological analysis is revealed.

1. Distinguishing between orthographic and morphological effects

The standard morphological formation processes in English typically entail prefixation and suffixation to a base morpheme. As a consequence, forms with a common base morpheme generally share orthographic and phonological as well as morphological structure. For regular forms, there is structural transparency (Henderson, 1989) in

The research reported here was conducted at the Laboratory of Experimental Psychology at the University of Belgrade and at Hebrew University in Jerusalem and was supported by funds from National Institute Of Child Health and Development Grant HD-01994 to Haskins Laboratories and by funds from the University of Belgrade.

that related forms are structured around the same base morpheme (e.g., COMPUTE- COMPUTER). Moreover, to the extent that compositionality is present, related words will commonly also have similar meanings. Covariation of morphological and orthographic structure in related forms invites an orthographic account of the morphological effects observed in tasks such repetition priming.

In the repetition priming task, first (prime) and second (target) presentations are separated by an average of ten and sometimes as many as 50 intervening items. Target latency as a function of type of prime is examined. Changes in spelling or pronunciation tend not to significantly diminish the effect of morphological relatedness in this task so that prime-target pairs such as SLEPT-SLEEP or HEALTH-HEAL produce facilitation equivalent to SLEEP-SLEEP or HEALED-HEAL respectively (Fowler et al., 1985; Feldman & Moskovljević, 1987). Similarly, formal similarity of morphologically-unrelated prime and target (viz., phonologically and orthographically but not morphologically similar pairs such as DIET and DIE) does not result in priming at these long lags (Feldman & Moskovljević, 1987; Hanson & Wilkenfeld, 1985; Napps, 1989; Napps & Fowler, 1987). Despite an absence of effects due exclusively to spelling/pronunciation and orthographic form in the repetition priming literature, there persists a tendency to try to interpret morphological effects as effects of orthographic structure. For example, Seidenberg (1987) following Adams (1981) suggested that patterns of high and low bigram frequency could account for morphological patterning because transitional probabilities of letter sequences that straddle a (syllabic or) morphological boundary tend to be low (bigram troughs) relative to probabilities of sequences internal to a unit (cf. Rapp, 1992).

The similar response patterns for regular and irregular forms described above provide evidence against an orthographic account of morphological facilitation in repetition priming. Other morphological effects inconsistent with an orthographic account are based on (1) manipulations of alphabet and on the (2) the absence of an effect when a target is preceded by an unrelated word with a similar orthographic and morphological structure. These findings will be reviewed in the remainder of section one.

Many readers of Serbo-Croatian are fluent in both the Roman and Cyrillic alphabets. For such readers, this situation has been exploited in the study of morphological processing (Feldman & Moskovljević, 1987; Experiment 1). The rationale was that if the facilitation observed in repetition priming arises in a relatively early stage of processing and represents repeated analysis of the same orthographic pattern, then the facilitation due to morphological relatedness should differ when successive presentations of the target word alternated alphabet versus when they preserved alphabet. In that study (Feldman & Moskovljević, 1987), lags ranged from 7 to 13 items and alphabet was manipulated between subjects so that a one alphabet (preserved) and a two alphabet (alternated) condition existed. Because it is possible that the time course of activation of visual form varies with lag (Monsell, 1985; Ratcliff, Hockley, & McKoon 1985), a second study was conducted in which alphabet was manipulated within subjects and a more expanded range of lags (3 to 20 intervening items) was included (Feldman, 1992). In both, words and pseudowords were presented twice, with a lag of intervening items. Subjects who were students at the University of Belgrade were instructed to perform a lexical decision to each letter string as it appeared. In the alternated alphabet condition, prime and target were transcribed in different alphabets (e.g., НОГА-NOGOM). In the preserved alphabet condition, prime and target were in the same alphabet (e.g., NOGA-NOGOM). Decision latencies to targets that were preceded by primes where target and prime either alternated or preserved alphabet were compared in an attempt to find evidence for facilitation based on repetitions of specific visual patterns.

Results of the two alphabet alternation experiments are summarized in Table 1. In neither experiment was the effect of alphabet (preserved/alternated) significant. Moreover, in the latter experiments, it was the case that for words, neither the effect of lag nor the interaction of alphabet by lag was significant. Stated generally, significant target facilitation occurred when primes appeared in either the same alphabet or in a different alphabet from the target and target facilitation was no greater in the alphabet preserved condition than in the alternated condition. Obviously words presented and represented in the same alphabet are more visually similar than are the Roman and Cyrillic alternatives of a word. Yet, in the repetition priming task where several items intervened between first and second presentations, no significant increment to facilitation was observed on the alphabet preserved trials relative to the alphabet alternating trials. This finding was observed at lags as short as 3 and as long as 20 items.

Table 1. *Mean decision latencies (ms) and errors for words in the alphabet preserved and alphabet alternated conditions of the repetition priming task .*

First		Repetition: Alphabet		
Presentation	Lag	Alternated	Preserved	Difference
(Feldman and Moskovljević, 1987; Experiment 1.)				
642	10		552	90
678	10	588		90
(Feldman, 1992: Experiment 1a)				
words				
651	10		601	50
12.7		6.6		6.1
			592	59
			7.3	5.4
	20	607		44
		4.5		8.2
			595	56
			7.3	5.4
(Feldman, 1992: Experiment 1b)				
words				
628	3	562		66
10.8		8.3		2.5
			562	66
			5.9	4.9
	10	567		61
		7.9		2.9
			573	55
			7.9	2.9

A second strategy for differentiating morphological and orthographic effects entailed comparing morphologically-related primes to unrelated primes that share orthographic structure. In Serbo-Croatian, it is possible to identify pairs of unrelated words that are formed around homographic base morphemes. For example, the word "BOR" in nominative singular, meaning "pine," is masculine in gender while the word "BORA," meaning wrinkle, in nominative singular is feminine. They have homographic base morphemes spelled BOR but, because of gender differences, require different sets of inflectional affixes. In a repetition priming task (Feldman & Andjelković, 1991; Experiment 3), targets (e.g., BOROVI from BOR) were preceded an average of ten items earlier in the list by an identical repetition (e.g., BOROVI), by a morphologically-related form (e.g., BOR) or by a morphologically-unrelated but homographic form (e.g., BORAMA from BORA). An analysis of variance revealed a significant effect of prime type in both reaction time and errors. Results are summarized in Table 2. Target decision latencies were 589 ms in the identity condition, 617 ms in the morphological condition and 656 ms in the orthographic condition. Decision latencies in the no prime condition were 661 ms. Target error rates were 3.3% in the identity, 4.4% in the morphological and 12.6% in the orthographic condition. Error rates in the no prime condition were 16.7%. The effect of prime type was significant with both subjects (*F1*) and item (*F2*) as random effect variables, with both reaction time and errors as dependent measures. Post-hoc tests indicated that the morphological and orthographic prime conditions were significantly different.

Table 2. *Mean reaction times and errors for targets (e.g., BOROVI) following identity, morphological and orthographic primes and for first presentations of the target in repetition priming.*

Prime Type	Example	Latency	Errors
Identity	borovi	589	3.3
Morphological	bor	617	4.4
Orthographic	borama	656	16.7
First Presentation		661	16.7

The orthographic and no prime conditions did not differ with either the reaction time or the error measure. That is, no facilitation with orthographically similar but morphologically unrelated words was observed when an average of ten items intervened between first and second presentations in a repetition priming task. These results are consistent with an earlier experiment conducted with Serbo-Croatian materials (Feldman & Moskovljević, 1987; Experiment 2) in which decision latencies for target words such as STAN meaning "APARTMENT" were reduced by the prior presentation of a derivationally-related prime such as the word STANČIĆ meaning "LITTLE APARTMENT." By contrast, the word STANICA did not reduce target latencies. Note that this word is morphologically unrelated but orthographically similar to the target, is composed of one morpheme and means "BUS STATION." In contrast to Feldman and Moskovljević (1987), in the present study, orthographic primes, as morphological primes, were morphologically complex forms consisting of a base morpheme and an inflectional affix. Nevertheless, orthographic primes were equivalent to the no prime condition. Collectively, these studies refute the hypothesis that orthographic similarity underlies morphonological facilitation in repetition priming.

In summary, relative to the no prime condition, both morphological relatives and identical repetitions facilitated target recognition. The orthographic prime condition was not significantly different from the no prime condition. Finally, and most important to the present discussion of morphological effects in repetition priming, target latencies and errors following morphological primes and orthographic primes at long lags were significantly different both in the analysis by subjects and in the analysis by items.

Orthographic and morphological primes also differentially influenced target latency when presented immediately in succession. In a traditional immediate priming task, morphological primes facilitate and orthographic primes may inhibit. However, the orthographic effect is sensitive to the density of the orthographic neighborhood of the prime (Forster, Davis, Schoknecht, & Carter, 1987) as well as the relative frequency of prime and target and the presence or absence of a mask (Segui & Grainger, 1990). For example, without a mask, lower frequency orthographic primes tend to inhibit whereas with a mask, it is the higher frequency prime that shows inhibition.

A recent report with French materials is consistent with this characterization of morphological as contrasted with orthographic primes (Grainger, Colé & Segui, 1991). In that study, masked primes consisted of morphological, orthographic or unrelated derivations of the target. Decision latencies for targets were fastest in the morphological condition (619 ms), (numerically but not statistically) slowest in the orthographic condition (653 ms) and intermediate (639 ms) in the unrelated condition. In those materials, however, orthographic primes (e.g., AFFIRMÉ-REFORMÉ) tended to be less similar to the target than were morphological primes (e.g., DEFORMÉ-REFORMÉ) and this might account for the marginally significant inhibition in the orthographic condition. Nevertheless, the critical point is that morphological primes showed facilitation whereas orthographic primes showed weak inhibition, at best, and conservatively, no difference from the unrelated condition.

A study recently completed in Serbo-Croatian replicates the difference between orthographic and morphological primes presented in immediate succession with an unmasked prime (Feldman & Andjelković, 1991). In a series of two experiments, targets consisting of morphologically-complex forms were preceded by either a morphological relative, an unrelated word formed from a homographic base morpheme or an orthographically and morphologically unrelated word. For example, the target BOROVI was preceded by (1) BOR which is inflectionally related, (2) BORI which is not related morphologically although it is orthographically similar because its base morpheme is homographic and by (3) KRV which is unrelated along both morphological and orthographic dimensions. In one experiment (Experiment 2), primes were of higher frequency than targets. In a second (Experiment 3), primes were of lower frequency than targets. In both experiments, primes without masks were presented to university students in Belgrade for 250 ms and followed by a blank for 50 ms after which the target appeared for 1000 ms. Latencies greater than 2SD from the mean were treated as errors. Results are summarized in Tables 3 and 4.

Table 3. *Mean reaction times (and percent errors) for targets (e.g., borovi) following morphological, orthographic and unrelated primes. Primes were higher in relative frequency than their targets.*

Prime Type	Example	Latency	Errors
Morphological	bor	684	24
Orthographic	borovi	754	45
Unrelated	krv	738	38

Table 4. *Mean reaction times (and percent errors) for targets (e.g.. bori) following morphological, orthographic and unrelated primes. Primes were lower in relative frequency than their targets.*

Prime Type	Example	Latency	Errors
Morphological	borovi	687	15
Orthographic	bor	743	46
Unrelated	krv	746	30

Morphological and unrelated primes differed significantly at short lags in both experiments and this outcome replicates the morphological facilitation observed at longer lags with similar materials. When primes were less frequent than targets, significant orthographic inhibition ($F1$ and $F2$) was evident in the error measure but not in the reaction time measure. This finding is consistent with the results of Segui and Grainger (1990) using morphologically simple materials and unmasked primes. When primes were more frequent than targets, orthographic inhibition was not statistically significant although the reaction time pattern was quite similar to the pattern obtained when primes were less frequent that their targets.

In Italian, inhibitory effects between homographic base morphemes (e.g., FINA-FINIRE which mean "thin" in feminine singular and "to finish," respectively) have been reported in both a double lexical decision (viz., are both letter strings words?) and in a lexical decision task where prime and target are presented successively (Laudanna, Badecker, & Caramazza, 1989). Results were interpreted as evidence of inhibitory connections between homographic base morphemes in the lexicon. By this account, a differential effect of the relative frequencies of prime and target is not anticipated although it could be accommodated. More problematic is the failure to observe inhibition among homographic forms concurrent with facilitation among morphologically-related forms at longer intervals between presentations viz., at the lags incorporated into the repetition priming task. If inhibition reflects a principle of lexical organization then a justification for its sensitivity to lag is needed. Regardless of whether orthographic primes slow or impair accuracy to targets relative to unrelated primes and whether homographic base morphemes relative to other types of orthographic controls pose a special problem for the representation of morphology, it is useful to focus on the reliable facilitation obtained when items are morphologically-related as compared to either unrelated or orthographic conditions. This is true both in repetition priming with average lags of ten items and in successive priming with or without a mask.

In summary, morphologically-related words that undergo changes in spelling and/or pronunciation so that the base morpheme is partially obscured produce the same pattern of facilitation as do related forms that are structurally transparent. This finding has been observed in Serbo-Croatian (Feldman & Fowler, 1987) as well as English (Fowler et al., 1985; see also Kelliher & Henderson, 1990; Nagy, Anderson, Schommer, Scott, & Stillman, 1989) and presents a challenge for an orthographic account of facilitation in the repetition priming task. In addition, for morphologically-related prime-target pairs, the effect of presenting repetitions in the same alphabet was no different than the effect of alternating alphabet (Feldman & Moskovljević, 1987; Feldman, 1992). This outcome suggests that the basis of facilitation must be sufficiently abstract to tolerate changes in visual form introduced by manipulations of alphabet. Finally, when homographic base morphemes appeared in prime and target facilitation was observed among forms that shared a base morpheme but not among unrelated forms. In fact, the contrast between homographic and no prime conditions sometimes revealed marginal inhibition. Evidently, morphological effects cannot be described in terms of shared orthographic structure.

2. Distinguishing between semantic and morphological effects

Related forms, by definition, share a base morpheme. Because morphemes are generally defined as units of meaning, it is plausible that morphological facilitation reflects the semantic similarity of prime and target. Linguists distinguish between two types of morphological relatives, inflections and derivations, and these forms differ with respect to the productivity of rules and the predictability of their meaning from a semantic analysis of the base and its constituents (Aronoff, 1976). Whereas inflections rarely produce new shades of meaning, derivations are much less constrained semantically and historically often change meaning once formed (e.g., TERRIFIC and TERROR). For example (from Henderson, 1985), the prefix UN typically modifies the base adjective or verb in a predictable manner (e.g., UNCLEAR, UNDRESS) but some forms are derived from obsolete or rare bases (e.g., UNKEMPT) (Lakoff, 1971; Jackendoff, 1975). Moreover, forms such as UN+base+ABLE are semantically ambiguous insofar as the prefix can modify either the base verb (V) or the (V+ABLE) adjective (e.g., UNDOABLE may mean UNDO+ABLE or UN+DOABLE). Inconsistencies of semantic composition are obvious in semantic comparisons of base-derivation pairs such as DISCUSS-DISCUSSION, CONGREGATE-CONGREGATION and PROFESS-PROFESSION and point to the unpredictability inherent in a semantic analysis of some complex forms from their constituents.

Nevertheless, morphologically-related words tend to have similar meanings. Moreover, because semantic facilitation has been so thoroughly studied (see Neely, 1991, for a review), it is important to contrast facilitation due to morphological relatedness and to other types of semantic relatedness. Semantic contributions to the pattern of facilitation in repetition priming were explored with derivational relatives in English and in Hebrew. In a repetition priming study with English materials (Feldman, 1991), semantic overlap was first assessed by a separate group of subjects who scaled the items for semantic distance. (Due to the constraints of English, word class changes and other aspects of semantic predictability were not well controlled.) For each target, a morphological relative close and remote in meaning was selected. Small but statistically equivalent effects of the prior presentation of the same morpheme in a related word were observed for derivational relatives that were semantically close (36 ms) and semantically remote (30 ms) whereas identical presentations produced robust facilitation (93 ms). Evidently, extent of semantic overlap did not influence the pattern of facilitation in repetition priming.

In a second study (by C.A. Fowler, reported in Feldman, 1991), a target (e.g., HOT) was preceded at least ten items earlier in the list by the same item (e.g., HOT) or by a strong antonym (e.g., COLD). No facilitation was observed in the antonym condition relative to the identity and no prime (i.e., initial) conditions. In the sense that these items were highly predictable and semantically constrained, it is difficult to argue that semantic similarity underlies facilitation in repetition priming.

A third study (Bentin & Feldman, 1990) compared facilitation by associative and morphological primes at long and at short lags. Materials were Hebrew words; prime conditions consisted of morphological, semantic or both morphological and semantic relatives of the target. For example, the word meaning "LIBRARY" was preceded by the word for "NUMBER" which is formed from the same root or base morpheme, by the word for "LIBRARIAN" which is also formed from the same root and by the word for "READING" which is semantically but not morphologically related. Magnitude of morphological facilitation did not change over lags whereas that for semantic facilitation did. Moreover, when the prime immediately preceded the target, semantic and semantic-plus-morphological primes showed greater facilitation than did morphological primes but when an average of ten items intervened, morphological and semantic-plus-morphological showed equivalent facilitation. Evidently the patterns of associative and morphological facilitation are differentially affected by lag (see Table 5).

In conclusion, similarity of meaning between morphologically-related prime and target does not affect the pattern of facilitation in repetition priming. Morphological relatives closely related and remotely related semantically both produced the same pattern of facilitation at long lags. Moreover, closely related antonym pairs produced no effect under conditions where morphological relatedness effects were observed. Patterns of morphological facilitation in this task are, therefore, not easily described in terms of semantic similarity. In addition, morphological effects occurred at separations between prime and target that considerably exceed those at which semantic/associative priming has been demonstrated (Bentin & Feldman, 1990; see also Dannenbring & Briand, 1982; Emmorey, 1989; Henderson, Wallis & Knight, 1984; Napps, 1989). Evidently, morphological and semantic facilitation reflect different underlying mechanisms.

Table 5. *Mean reaction times and errors for targets (e.g., ספריה meaning "library") following morphological, semantic plus morphological, semantic and no prime in repetition priming (Bentin & Feldman, 1990).*

			Lag			
			lag 0		lag 15	
Prime Type	Example	Meaning	RT	Errors	RT	Errors
Morphological	מספר	number	589	1.8	587	1.0
Semantic/Morphological	ספרן	librarian	559	1.0	583	2.6
Semantic	קריאה	reading	563	2.1	611	1.1
Filler			606	1.2	606	1.2

3. Morphological effects in sentence contexts

The study of word recognition is sometimes represented as the interface between the domains of perception and language processing. It is the case, however, that the status of linguistic codes in word recognition is problematic (Henderson, 1989). In the first two sections of this chapter, we showed that morphological effects could be experimentally differentiated from effects of (1) shared orthographic and (2) shared semantic structure. In the next two sections, the relation between morphology and other types of linguistic patterning are examined. In section three, patterns of facilitation due to morphological and syntactic similarity between prime and target sentences are explored. In section four, the association between morphological and phonological analysis is examined in beginning readers.

Effects of morphological relatedness have been observed in sentence contexts as well as in isolated words. In one study (Feldman & Andjelković, 1990), students from the University of Belgrade were presented with pairs of sentences that either shared the same syntactic structure (subject verb (SV) or verb object (VO)) and/or shared the same base morphemes. In one experiment, subjects were required to read the sentence aloud and onset to vocalization was measured. A similar task has been reported to show effects of syntactic structure (Bock, 1986; 1990). In two related experiments, subjects were required to judge whether the sentence made sense and latency and errors were measured. For example, subjects saw target sentences such as VODIČI PLIVAJU which means "The guides swim" and has a subject (S) verb (V) structure. Across subjects, that sentence was preceded by four types of prime sentences. These included (1) Morphologically unrelated and structurally dissimilar primes consisting of sentences such as ČITA KNJIGU which means "He reads a book." This sentence has a verb (V) object (O) structure (which is grammatical in Serbo-Croatian because pronouns need not be specified outside the verb). Also included were (2) morphologically unrelated but structurally similar primes consisting of sentences such as ŽENA ČITA which has a SV structure and means "The woman reads." and (3) morphologically related and structurally similar primes consisting of sentences such as VODIĆ PLIVA which has a SV structure and means "The guide swims." Finally, there were (4) morphologically related but structurally dissimilar primes consisting of sentences such as VODI PLIVAČA which has a VO structure and means "He guides the swimmer." Primes and targets were constructed so that the same order of the two base morphemes (i.e., VOD- and PLIV-) was maintained over all prime and target sentences in which they appeared.

Foil sentences were morphologically related or unrelated and had either the same or a different constituent structure. Morphologically related foils contained the same base morphemes in illegal combinations such as verbal affixes on nouns and nominal affixes on verbs. Morphologically unrelated foils were composed of legal morphological combinations but were semantically anomalous. The primes for foil sentences were always semantically and syntactically acceptable.

In the oral reading task, the prime and target members of a pair were presented in different alphabets. Primes were always presented in Roman and targets in Cyrillic. In this way, the visual similarity of successive presentations of a morpheme was reduced. Significant effects of morphological relatedness were observed in the oral reading task. Effects of structural similarity were absent (see Table 6). While this outcome can be interpreted as evidence of facilitation due to repetition of base morphemes in prime and target sentences, an alternative account of morphological effects in this task focuses on the repetition of the initial syllable (e.g., VOD-) in all related sentences but not in unrelated sentences. Therefore, it is important to replicate the morphological effect in a task where an advantage based on repetition of the first syllable effect seems unlikely to occur.

In the verification task, subjects had to decide whether each target sentence made sense. Therefore, latencies presumably reflect more than just processing of the first syllable. All prime sentences were meaningful as were half of the target sentences. Primes and targets were printed in the same alphabet. Otherwise, materials were identical to those of the previous experiment. Results indicated that latencies were significantly longer following morphologically unrelated primes than following morphologically related primes (see Table 6). In addition, latencies were longer following structurally dissimilar sentences than following structurally similar sentences in both the morphologically related and the unrelated conditions. The interaction was significant by *F1* but not by *F2*. As in the previous experiment, the effect of morphology could reflect priming of morphemic units over different sentence structures. Alternatively, it could simply reflect episodic repetitions of the same letter sequence (e.g., VOD-) in sentence initial position. The episodic account seems unlikely in a verification task where the entire sentence must be processed before responding. Moreover, it does not explain the significant difference between morphologically related prime sentences with same and different structures (i.e., 901 vs. 975). In addition, when morphemes were not repeated (i.e., for morphologically unrelated sentences), structurally dissimilar primes and structurally similar primes had significantly different effects on target latencies. This outcome indicates that the verification task is sensitive to sentence structure defined over different surface forms. In summary, when the experimental task requires that the entire sentence be processed, facilitation can arise between sentences with similar structures.

Table 6. *Oral reading and verification times (errors in parentheses) for target sentences primed by morphologically and structurally similar sentences (From Feldman & Andjelković, 1990).*

target: VODIČI PLIVAJU (SV)

	Sentence structure	
	Same structure	Different structure
Morphology		
Related	VODIĆ PLIVA (SV)	VODI PLIVAČA (VO)
Unrelated	ŽENA ČITA (SV)	ČITA KNJIGU (VO)
Task		
Oral reading:		
Related	676 (5)	690 (8)
Unrelated	718 (9)	730 (9)
Verification:		
Related	901 (3)	975 (16)
Unrelated	973 (12)	1002 (16)

An examination of the materials used in the previous two experiments revealed that morphologically related and structurally related primes consisted of words related by inflection to the target sentence whereas morphologically related but structurally dissimilar sentences generally consisted of words related to the target by derivation. In a final experiment in this study, effects of sentence structure and morphology were again investigated in a sentence verification task. In contrast to the previous experiments, here all critical items consisted of morphologically related prime-target sentences. Moreover, in addition to sentence structure, type of morphological relatedness (viz., inflection/derivation) was manipulated. The essence of stimulus construction entailed identifying base morphemes that could function as part of either a noun or a verb. For example, subjects saw target sentences (constructed around the morphemes PLIV- which means "swim" and VOD- which means "guide") such as PLIVAJU VODIČI which means "The guides swim" and has a VS structure. As in the previous experiments, that sentence was preceded by four types of prime sentences across different groups of subjects. All primes were morphologically related by either inflection or derivation to the target. In inflected sentences, the word class of the base morphemes was preserved over prime and target sentences. In derived sentences, the word class of the base morphemes changed in prime and target sentences. In addition, primes and targets varied with respect to similarity of sentence structure. Four combinations of sentence structure and morphology were possible: (1) structurally dissimilar inflectional primes consisting of VO sentences such as PLIVA KA VODIČU which means "He swims toward the guide." (2) structurally dissimilar derivational primes consisting of VO sentences such as VODIŠ PLIVAČA which means "You guide the swimmer." (3) structurally similar inflectional primes consisting of SV sentences such as VODIĆ PLIVA which means "The guide swims" and finally, (4) structurally similar derivational primes consisting of SV sentences such as PLIVAČ VODI which means "The swimmer guides" (see Table 7). Primes and targets were constructed so that the same order of sentence constituents (viz., S,V,O) was preserved throughout a set. The advantage of constructing materials in this way is that repeated base morphemes (i.e., VOD- and PLIV-) do not always appear in the same position in prime and target sentences. For example, both VODIĆ PLIVA and PLIVAČ VODI have subject before verb but the ordering of base morphemes in these sentences differ. The disadvantage is that by preserving the ordering of elements in a pair with similar structure, the effect of changing syntactic role for a particular base morpheme may be lost. Prime and target sentences were printed in different alphabets.

Results indicated that the effect of morphology was significant (by both *F1* and *F2*) for both the latency and the error measures such that derivations produced less facilitation than did inflections. This finding suggests that the effect of sentence structure observed in the previous verification experiment can be attributed to different effects on targets of prime sentences related by inflection and by derivation. Because inflectionally related primes differed only in number from the target whereas derivationally related sentences transformed the base morpheme of the noun into a verb and the base morpheme of the verb into a noun, it was always the case that inflectional sentences were semantically more similar to the target than were derivational sentences.

Table 7. *Verification times (and errors in parentheses) for target sentences primed by morphologically and structurally similar and dissimilar sentences.*

target: PLIVAJU VODIČI (VS)

	Sentence structure	
	Same structure	Different structure
Morphology:		
inflection	VODIĆ PLIVA (SV)	PLIVA KA VODIČU (VO)
derivation	PLIVAČ VODI (SV)	VODIŠ PLIVAČA (VO)
inflection	958 (6)	949 (10)
derivation	1084 (20)	1100 (21)

The effect of sentence structure was not replicated. The failure to obtain an effect of sentence structure for either inflectionally or derivationally related primes indicates that facilitation based on repetition of the initial syllable is not in itself a plausible account of facilitation. The absence of a sentence structure effect most likely reflects the way in which structure was manipulated in the present experiment. Specifically, the order of (subject, object and verb) constituents ways not always preserved across structurally related primes and their targets.

Morphological effects occur in sentence contexts as well as in isolated words. In the first two experiments in this series, all related targets started with the same initial syllable (base morpheme). Therefore, effects of morphological relatedness could simply be anticipation effects based on repetition of the first syllable. This account is not plausible in the third experiment, however. In fact, in that experiment, same structure and different structure sentences started with different initial syllables (morphemes) and yet there was no effect of sentence structure of the prime (958 ms vs. 949 ms). In sentences related by inflection, the absence of an effect of structure was not anticipated and needs tc be investigated further. Specifically, in contrast to English, in Serbo-Croatian it is possible to independently manipulate repetition of sentence constituents (subject, object and verb) and the ordering of those constituents.

Even in contexts and tasks that focus on sentence processing it appears that the morphemic constituents of words are analyzed. That is, morphological analysis is not restricted to isolated words in the word recognition task. In these experiments, it is evident that activation among the morphological constituents of prime and target sentences is not necessarily tied to their syntactic role in a sentence nor to the ordering of morphemes. In the next section, associations between morphological processes and other analytic processes are investigated.

4. Associations between phonological and morphological analysis

The beginning reader provides a window through which to evaluate the relation between morphological and phonological analysis in word recognition. In one study (Feldman, Andjelković, & Fowler, in preparation), children between seven and eight years of age who were native speakers of Serbo-Croatian were administered both a morphological and a phonological task. In the morphological task, children were auditorily presented with a source word and a sentence frame and their task was to complete the sentence by adjusting the morphological affixes on the source word to make it fit semantically and syntactically with the sentence frame. Sentence frames were constructed so that depending on the source word, either an inflectional or a derivational substitution was required and frames were paired so that for one source word both an inflectional and a derivational adjustment were necessary. For example, some children were required to fit the source word KUVAR meaning "a cook" (agent in nominative singular) into the sentence frame MAMA MI POMAŽE DA ____ which means "Mother helps me to ___." This sentence requires the first person singular verb form KUVAM which is related by derivation to the source word KUVAR. For other children, the infinitive KUVATI was presented as the source word for the same sentence frame. Here the source word and the response are related by inflection. Still other groups of children viewed the same source words (i.e., KUVAR, KUVATI) on different sentence frames. For example, OVAJ RESTORAN IMA DOBROG ___ which means "This restaurant has a good___" requires the accusative singular form of the agent KUVARA. This response is inflectionally related to KUVAR and derivationally related to KUVATI. All subjects were tested on all four combinations of sentence frame and source word and across subjects the same base morpheme appeared in all conditions. This design minimized effects due to sentence frame and to the morphological complexity and the familiarity of the source word as well as the correct response.

Thirty six sentences were constructed. Each contained between four and six words. The target word was always in final position and varied with respect to word class. Sentences were read aloud by the experimenter. The source word was presented both before and after the sentence frame and all source word-sentence frame combinations required at least one morphological substitution. Forty children randomly selected from an urban elementary school in Belgrade, Yugoslavia, were tested individually by a native speaker of Serbo-Croatian and the subjects' responses were transcribed by that experimenter. Results indicated that inflectional responses tended to be correct more frequently than were derivational responses. Mean errors were 1.0 and 3.5 respectively out of a maximum of 18.

Subsequent to the sentence completion task, all subjects participated in a phoneme deletion task (Rosner & Simon, 1971). Subjects heard words and pronounced them aloud without the designated phoneme. All words became orthographically legal but meaningless sequences after phoneme deletion. The position of the deleted letter was balanced across words and, in the source word, it always constituted part of a cluster. Responses were transcribed by the same adult native speaker of Serbo-Croatian. Performance on the phoneme deletion task was correlated with performance on the inflectional and derivational conditions (summed over sentence frame) of the sentence completion task.

Results indicated a significant correlation between phonemic deletion and each morpheme condition r =.37 for inflections and r =.52 for derivations, respectively. Finally, to each child was administered verbal and nonverbal intelligence tests and these served as covariant controls. Verbal intelligence accounted for 33% of the variance and nonverbal intelligence accounted for 25% of the variance on derivational morphology score. The contribution of intelligence to the inflectional score was not significant but this outcome may reflect the near perfect performance and consequent lack of variability on the inflectional task. Most important, results revealed a significant relationship between phonological and derivational performance even when effects of intelligence were partialled out. That is, with controls for verbal and nonverbal intelligence, performance on a phonological task was still a significant indicator of performance on a (derivational) morphology task. It accounted for a significant 14% of the variance.

Evidently, the ability to explicitly manipulate phonemic segments is associated with the ability to complete sentences with a syntactically correct form. This relationship is independent of intelligence and can be interpreted as evidence of a general linguistic style of analysis that is not tied to particular units. This outcome has also been observed for learners of English which conveys syntactic information through fixed word order in contrast to Serbo-Croatian where word order is relatively free to vary (Fowler, 1988; 1990).

It is often claimed that metalinguistic skill is the single most important factor in learning to read (e.g., Tunmer, 1988). While there is ample compelling evidence that reading an alphabetic orthography requires phoneme awareness, evidence that awareness of linguistic units above the level of the phoneme makes an independent contribution to reading skill is sparse (but see Fowler, 1988; 1990). In the present study, the awareness of morphemes in young readers is associated with phoneme awareness and the relationship cannot be explained by general intelligence or by vocabulary knowledge.

5. Morphological effects reflect linguistic analysis

It is sometimes claimed that three related skills underlie the language user's command of morphology (Tyler & Nagy, 1989). Primary is an appreciation of the *internal structure* of a word such that the presence of a shared component among morphological relatives is recognized, either explicitly or implicitly. Experimental evidence for morphological analysis of a word's structure comes primarily from patterns of facilitation in a priming task in which the same base morpheme is repeated. Recognition by skilled readers is facilitated when the same morphological components recur and the basis of this facilitation can be neither semantic nor orthographic in origin.

Once words can be analyzed with respect to morphological components then it is reasonable to ask whether skilled readers are sensitive to the constraints on *combinations* of morphemes or to the *syntactic implications* of appending particular affixes to a base morpheme. The design of the foils in the sentence verification and oral reading tasks forced adult subjects to attend to these dimensions because some sentences were composed of illegal morphological combinations. Although the foils were not analyzed, morphological analysis was evident in the facilitation to target sentences composed from the same morphological constituents as their prime sentences. Here, effects of visual similarity were eliminated by presenting members of a pair in contrasting alphabets. Interestingly, equivalent facilitation occurred when, across prime and target sentences, a base morpheme changed word class (derivational relatives) and when it did not (inflectional relatives) even though they differed with respect to semantic similarity. It has been reported that skills of morphological analysis emerge before combinatory or syntactic skills (Tyler & Nagy, 1989). In the sentence completion task, however, seven and eight year olds were able to produce the appropriate inflectional and derivation affixes for a variety of syntactic contexts. In order to respond accurately in this task, subjects had to segment the base morpheme from its source word as well as generate a syntactically appropriate affix. Sometimes this entailed forming a verb from a noun or a noun from a verb and it always required the addition of an inflectional affix. Evidently, the metalinguistic demands of this task allowed the children to utilize their knowledge of morphemes and how they combine in particular syntactic contexts.

In conclusion, evidence for morphological analysis in word recognition is not tied exclusively to the repetition priming task although that task has allowed a differentiation between morphological analysis and effects of orthographic or semantic structure. Morphological effects are also evident in a task where constituents of a sentence are experimentally manipulated. Although the mechanism of syntactic effects in the verification task is not clear, it is certain that facilitation occurs when the constituents of words are repeated. This finding is surprising because the task fosters analysis at a level more abstract than the morpheme or the combination of morphemes that comprise the words of a sentence. Evidently, readers cannot refrain from engaging in analysis at the level of the morpheme.

It has recently been demonstrated that time to recognize a target word is influenced by its frequency relative to the other words in its orthographic neighborhood (Grainger, 1990). Similarly for a morphologically simple word, the frequency of words that are inflectionally and derivationally related to it influences the pattern of reaction times in lexical decision (Katz, Rexer, & Lukatela, 1991; Nagy, Anderson, Schommer, Scott, & Stallman, 1989; Taft & Forster, 1975) and it is not necessary that the shared base morpheme of those words be explicitly represented in the surface form (Kelliher & Henderson, 1990). These findings suggest that recognition of any particular word is influenced by properties of other words that are related along some dimension. This similarity is often captured in terms of organizational properties of the lexicon that are distributed rather than tied to one lexical entry. Perhaps the illusion of a shared orthographic or semantic component among morphological relatives has misguided the investigation of morphological processing and undermined our understanding of the status of the morpheme as an abstract linguistic unit.

It has been observed that children who are good at phonological analysis also tend to be good at morphological analysis and that their performance cannot be attributed to either verbal intelligence (and good vocabulary) or to general intelligence. This finding helps to elucidate the value of morphological analysis. It is well established that good and poor beginning readers differ in their ability to grasp the phonological structure of a word in a variety of experimental tasks. Phonological analysis helps the beginning reader map unfamiliar written words into a spoken form which may be familiar even when the written form is not. The essence of this process is an explicit appreciation of the linguistic components of a written word and how they map onto phonemes. It is important because it underlies the ability to read unfamiliar words and combinatorial productivity of the writing system in general. Morphological analysis may serve a similar function. Insofar as words can be constructed from and decomposed into meaningful components and those components can be recombined into new words, morphological analysis enhances the productivity of the reader.

REFERENCES

Aronoff, M. (1976). *Word formation in generative grammar.* Cambridge, MA: MIT Press.

Adams, M. J. (1981). What good is orthographic redundancy? In O. Tzeng & H. Singer (Eds.), *Perception of print: Reading research in experimental psychology.* Hillsdale, NJ: Erlbaum.

Bentin, S., & Feldman, L. B. (1990). The contribution of morphological and semantic relatedness to repetition priming at short and long lags: Evidence from Hebrew. *Quarterly Journal of Experimental Psychology, 42A*, 693-711.

Bock, J. K. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12(4)*, 575-586.

Bock, J. K., & Loebell, H. (1990). Framing sentences. *Cognition, 35*, 1-39.

Dannenbring, G. L., & Briand, K. (1982). Semantic priming and the word repetition effect in a lexical decision task. *Canadian Journal of Psychology, 36*, 435-444.

Emmorey, K. D. (1985). Auditory morphological priming in the lexicon. *Language and Cognitive Processes, 4*, 73-92.

Feldman, L. B. (1992). Bi-alphabetism and the design of a reading mechanism. In D. M. Willows, R. S. Kruk, & E. Corcos (Eds.), *Visual processes in reading and reading disabilities*. Hillsdale, NJ: Erlbaum.

Feldman, L. B. (1991). Morphological relationships revealed through the repetition priming task. In M. Noonan, P. Downing, & S. Lima (Eds.), *Linguistics and literacy*. (239-254). Amsterdam/Philadelphia: John Benjamins.

Feldman, L. B., & Andjelković, D. (1990, November). *Morphemic and syntactic facilitation in sentence contexts*. Paper presented at the Psychonomic Society, New Orleans, LA.

Feldman, L. B., & Andjelković, D. (1991, November). *Morphological priming at long and short lags in visual word recognition*. Paper presented at the Psychonomic Society, San Francisco, CA.

Feldman, L. B., Andjelković, D. & Fowler, A. (ms in preparation). Morphological and phonological analysis in beginning readers.

Feldman, L. B., & Fowler, C. A. (1987). The inflected noun system in Serbo-Croatian: Lexical representation of morphological structure. *Memory & Cognition, 15*, 1-12.

Feldman, L. B., & Moskovljević, J. (1987). Repetition priming is not purely episodic in origin. *Journal of Experimental Psychology: Learning, Memory & Cognition, 13*, 573-581.

Feustel, T. C., Shriffrin, R. M., & Salasoo, A. (1983). Episodic and lexical contributions to the repetition priming effect in word identification. *Journal of Experimental Psychology: General, 112*, 309-346.

Forster, K. I., & Davies, C., Schoknecht, C., & Carter, R. (1987). Masked priming with graphemically related forms: Repetition or partial activation? *Quarterly Journal of Experimental Psychology, 39A*, 211-251.

Fowler, C. A., Napps, S. E., & Feldman, L. B. (1985). Relations among regular and irregular morphologically related words in the lexicon as revealed by repetition priming. *Memory & Cognition, 13*, 241-255.

Fowler, A. (1988). Grammaticality judgments and reading skill in grade 2. *Annals of Dyslexia, 38*, 73-84.

Fowler, A. (1990). Factors contributing to performance on phonological awareness tasks. *Haskins Laboratories Status Report on Speech Research, SR-103/104*, 137-152.

Grainger, J. (1990). Word frequency and lexical neighborhood effects in lexical decision and naming. *Memory & Language, 29*, 28-244.

Grainger, J., Colé, P., & Segui, J. (1991). Masked morphological priming in visual word recognition. *Memory & Language, 30*, 370-384.

Hanson, V. L., & Feldman. L. B. (1989). Language specificity in lexical organization: Evidence from deaf signers' lexical organization of ASL and English. *Memory & Cognition, 17*, 292-301.

Hanson, V. L., & Wilkenfeld, D. (1985). Morphophonology and lexical organization in deaf readers. *Language and Speech, 28*, 269-280.

Henderson, L. (1985). Toward a psychology of morphemes. In A. W. Ellis (Ed.), *Progress in the psychology of language* (pp. 15-72). London: Erlbaum.

Henderson, L. (1989). On the mental representation of morphology and its diagnosis by measures of visual access speed. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 357-391). Cambridge, MA: MIT Press.

Henderson, L., Wallis, J., & Knight, D. (1984). Morphemic structure and lexical access. In H. Bouma & D. Bouwhuis (Eds.), *Attention and performance* X (pp. 211-224). Hillsdale, NJ: Erlbaum.

Jackendoff, R. (1975). Morphological and semantic regularities in the lexicon. *Language, 51*, 639-671.

Katz, L., Rexer, K., & Lukatela, G. (1991). The processing of inflected words. *Psychological Research, 53*, 25-31.

Kelliher, S., & Henderson, L. (1990). Morphologically based frequency effects in the recognition of irregularly inflected verbs. *British Journal of Psychology, 81*, 527-539.

Lakoff, G. (1971). *Syntactic irregularity*. New York: Rinehart & Winston.

Laudanna, A., Badecker, W., & Caramazza, A. (1989). Priming homographic stems. *Journal of Memory & Language, 28*, 531-546.

Monsell, S. (1985). Repetition and the lexicon. In A. W. Ellis (Ed.), *Progress in the psychology of language* (pp. 147-195). London: Erlbaum.

Nagy, W., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24*, 263-282.

Napps, S. E. (1989). Morphemic relationships in the lexicon: Are they distinct from semantic and formal relationships? *Memory & Cognition, 17*, 729-739.

Napps, S. E., & Fowler, C. A. (1987). Formal relationships among words and the organization of the mental lexicon. *Journal of Psycholinguistic Research, 16*, 257-272.

Rapp, B. C. (1992). The nature of sublexical orthographic organization: The bigram trough hypothesis examined. *Journal of Memory and Language, 31*, 33-53.

Ratcliff, R., Hockley, W., & McKoon, G. (1985). Components of activation: Repetition and priming effects in lexical decision and recognition. *Journal of Experimental Psychology: General, 114*, 435-450.

Rosner, J., & Simon, D. P. (1971). The auditory analysis test: An initial report. *Journal of Learning Disabilities. 4*, 384-392.

Seidenberg, M. (1987). Sublexical structures in visual word recognition: Access units or orthographic redundancy? In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading*. (pp. 245-263). Hillsdale, NJ: Erlbaum.

Segui, J., & Grainger, J. (1990). Priming word recognition with orthographic neighbors: Effects of relative prime-target frequency. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 65-76.

Stanners, R. F., Neiser, J. J., Hernon, W. P., & Hall, R. (1979). Memory representation for morphologically related words. *Journal of Verbal Learning and Verbal Behavior, 18*, 399-412.

Tunmer, W. (1988). Metalinguistic abilities and beginning reading. *Reading Research Quarterly, 23*, 134-158.

Tyler, A., & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory & Language, 28*, 649-667.

FOOTNOTES

*In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 343-360). Amsterdam: Elsevier Science Publishers (1992).

†Also State University of New York at Albany.

‡University of Belgrade.

Haskins Laboratories Status Report on Speech Research
1992, SR-111/112, 193-204

Can Theories of Word Recognition Remain Stubbornly Nonphonological?*

Claudia Carello,[†] M. T. Turvey,[†] and Georgije Lukatela[‡]

The issue of how readers get from the printed word to its lexical representation is a hotly contested one (see Carr & Pollatsek, 1985; Humphreys & Evett, 1985; Van Orden, Pennington, & Stone, 1990, for reviews). Candidate routes are the visual and the phonological. In the visual route, lexical entries are said to be accessed directly on the basis of orthographic properties. The phonological route requires that lexical access be mediated by the recoding of graphemes into their corresponding phonemes. Considerable experimental data have been offered in support of both types of routes. The bulk of research on word identification using English language materials has been taken to implicate the dominance of a visual access route with, perhaps, an optional but not preferred phonological route (e.g., Coltheart, Besner, Jonasson, & Davelaar, 1979; Humphreys & Evett, 1985). Data on word identification using Serbo-Croatian language materials point unequivocally to a nonoptional phonological access route (e.g., Lukatela & Turvey, 1990a, b; Lukatela, Carello, & Turvey, 1990).

We assume that the basic mechanism of written language processing is the same for all languages. Different data patterns among languages, therefore, are to be taken as evidence of how that mechanism can be fine-tuned by the structure of a particular language. We will use some differences and similarities between Serbo-Croatian, English, and Hebrew to elucidate possible features of a written language processing mechanism that would allow such patterns to arise.

Given the nature of the data that have been obtained with Serbo-Croatian, such a mechanism must allow for automatic prelexical phonology. Therefore, we must begin with the assumption that all writing systems are phonological—they provide a system for transcribing phonologically any possible word of the language (A. M. Liberman, in press; Mattingly, 1985). The variety of orthographies do this in more or less straightforward ways, resulting in their being phonologically shallow or deep (I. Y. Liberman, A. M. Liberman, Mattingly, & Shankweiler, 1980). How orthographic depth has been interpreted mechanistically will be addressed. We will ultimately claim that the stubborn rejection of phonology in the prevailing theories of reading cannot be sustained within a consistent theory of language processing that accommodates all of the facts, not just those that are convenient (nor, we might add, just those obtained with English).

WHY WRITING SYSTEMS MUST BE PHONOLOGICAL

As will be shown in later sections, the evidence from experiments in Serbo-Croatian overwhelmingly favors phonological mediation. But our argument begins at a more fundamental level—why that should be an expected outcome. The theoretical backdrop concerns the relation of reading to speech. The basic fact is that phonological structures are the raw materials on which syntactic processes normally work in comprehending speech. These processes are well in place by the time one has to commence the less natural task of learning to read. Given that reading, in contrast to speaking and understanding, is something that must be learned explicitly, how might that be accomplished?

We wish to dedicate this chapter to our dear friend and colleague Alvin Liberman who has provided the occasional gentle reminder that reading and speech must be intimately connected. Preparation of this manuscript was supported in part by National Institutes of Child Health and Human Development Grants HD-01994 and HD-08495 to Haskins Laboratories and the University of Belgrade, respectively.

> The seemingly sensible strategy for the reader is to use the optical shapes to access phonological structures early in the reading process. Once the reader has done that, he has put the hard part of reading behind him, for everything else will be done automatically by language processes that he commands by virtue of his humanity (A. M. Liberman, 1991, pp. 242-243).

The alternative to this seemingly sensible strategy is that readers concoct nonlinguistic processes that bar them from the ordinary language processor for as long as possible. The effect is that the route that they take to the lexicon and the base representations that they find there are kept—for whatever reason—"stubbornly nonphonological" (A. M. Liberman, 1991, p. 242).

We suspect, instead, that it is reading theorists rather than readers who remain stubbornly nonphonological, both in denying the plausibility of taking advantage of extant phonological processes and overlooking the phonological basis of writing systems. The ultimate constraint on an orthography is that it permit any *possible* utterance in the language to be transcribed—it must respect the allowable phonological forms of the spoken language (determined by its articulatory gestures and their combinations). We note three aspects of how orthographies accomplish this openness (see Mattingly, 1985, 1992). First, orthographies transcribe linguistic units rather than acoustic or phonetic properties which are too context-sensitive. Second, the linguistic units that are transcribed seem to be words, irrespective of how the graphemic units are framed (cf. Wang, 1981). Third, words are transcribed by exploiting the phonological structure of the language, not by using word-specific symbols. This last point is critical for productivity; it allows a systematic way to transcribe novel utterances.

The morphological and phonological structure of a language determine what form this phonological exploitation takes. Alphabetic (or, more generally, segmental) writing systems are more appropriate for languages that "have fairly elaborate syllable structures, large and rather inefficiently exploited inventories of morphemes, and little homophony" (I. Liberman et al., 1980. p. 149). Syllabic writing systems, in contrast, are more appropriate for languages with a small number of syllables (usually with a regular CVCV... structure without consonant clusters). It should be noted that Chinese, despite its reputation in the folklore of orthographies, reflects phonological constraints as well. Its mislabeling as pictographic or ideographic has more to do with socio-cultural agendas than with what is represented by its orthography, namely, syllables (Mattingly, 1992). Logograms are a secondary accompaniment as they must be in a productive, complete writing system (DeFrancis, 1989; Mattingly, 1992; Wang, 1981).

THE ORTHOGRAPHIC DEPTH HYPOTHESIS

Serbo-Croatian, English and Hebrew differ in how straightforwardly their orthographies transcribe the sounds of the spoken language. In Serbo-Croatian, a grapheme such as G is pronounced /g/ regardless of the context in which it appears. There are no irregular pronunciations, silent letters, doubled letters, and so on. In English, G might be pronounced /g/, /j/, or /zh/, or not pronounced at all, depending on whatever else is in the letter string. In Hebrew, vowels are not even represented in 90% of the written material that adults encounter. Homographs are common; the pronunciation of an isolated word depends on which vowels are elected by a reader. This kind of difference has been referred to as orthographic depth[1] (Frost, Katz, & Bentin, 1987; I. Liberman et al., 1980; Lukatela, Popadić, Ognjenović, & Turvey, 1980; Sebastián-Gallés, 1991).

Orthographic depth has been considered relevant to reading because it seems to imply that getting from script to sound is more or less dependable for different languages and, therefore, should be more or less apparent in reading processes. While more or less dependable is fairly well agreed upon, more or less apparent has been subject to some interpretation which we would like to clarify in the context of our recently developed network formulation of Serbo-Croatian word recognition. The theme is that the easier it is to "get to" the sound from the spelling the more likely the reader is to do so in ordinary reading, using that as a basis for getting to other things as well—such as the lexicon. To some, this suggests that readers of a phonologically shallow orthography will access the lexicon phonologically whereas readers of a deep orthography will access the lexicon visually. The reasoning concerns how efficiently articulatory codes are provided. If the translation is complex and takes a long time, they won't be used (e.g., Frost et al., (1987).

Although not formulated with orthographic depth in mind, dual route theories are consistent with this reasoning. A phonological route to the lexicon is not used in English, it is argued, because the irregularity of script-to-sound makes the translation take too long; visual access, in contrast, is achieved rapidly. A phonological influence will be felt only for those letter strings for which visual access is slowed. This would include nonwords and, perhaps, low frequency words. Under a dual route interpretation, a phonological route to the lexicon would be impossible in Hebrew because the letter strings are so ambiguous.

More recently we have tried to take care in referring to *how apparent* the involvement of phonology is in accessing the lexicon as opposed to *whether or not* it is involved. We are trying to finesse two issues here. One issue has to do with the ease of demonstrating phonological involvement in Serbo-Croatian due to particular methodological advantages (versus processing differences between Serbo-Croatian and deeper orthographies). As noted in detail elsewhere, Serbo-Croatian is not only shallow, it is shallow in two largely distinct but partially overlapping scripts. The nature of the overlap is such that some letters are pronounced the same in the two alphabets while others are pronounced differently depending on which alphabet the reader uses. This allows the construction of letter strings in which a host of properties (semantics, syntax, frequency, associative relatedness) can be controlled experimentally while distinguishing graphemic from phonemic similarity. If experiments in English or Hebrew could be similarly contrived, they might, in principle, show unequivocal phonological involvement as well.

The second issue concerns the mechanistic interpretation of orthographic depth. The tradition has been to couch it in terms of discrete grapheme-phoneme correspondence rules or GPCs. More recent models, such as those pioneered by McClelland and Rumelhart (1986) and envisioned by Van Orden et al. (1990), could accommodate (in principle) orthographic depth with respect to the strength and number of connections in a parallel distributed network. Distinctions between these two kinds of approaches will be considered in some detail before turning to experimental demonstrations of phonological involvement in word recognition.

MECHANISTIC INTERPRETATIONS OF ORTHOGRAPHIC DEPTH

Grapheme-phoneme correspondence rules constitute the more familiar treatment of how one gets from spelling to sound (e.g., Coltheart, 1977, 1978). They specify how particular letters or clusters of letters are to be pronounced. And they do so discretely; the rules do not vary in strength. Thus, under this treatment, a shallow orthography is one that has relatively few rules and whose words can be relied upon to follow them (e.g., in Serbo-Croatian, the rules are defined at the level of the individual grapheme because their pronunciation is not changed by being combined with different combinations of graphemes). A deep orthography may have numerous rules or exceptions to its rules (e.g., in English, when a word ends in E, the E is silent and the preceding vowel is long; CAVE obeys this but HAVE does not) or, perhaps, application of its rules is simply inadequate to allow a reader to settle on a single pronunciation (e.g., in Hebrew, the standard printed form omits vowel marks so that a particular letter string can be pronounced as different words depending on which vowels are elected). However reliable they are, GPCs more or less embody what is orthographically legal in a given language. (This fact is responsible for the easy link between GPCs and pseudowords: GPCs may be useful at least insofar as they permit one to pronounce a novel letter string.)

Under the discrete symbol, rule-based characterization of assembling phonology, it is possible to consider that readers of different kinds of orthographies are engaging in different kinds of processes. Those for whom GPCs are reliable would be well-served to try them since a straightforward translation might be faster than a lexical search for a visual match to the orthographic pattern. But those for whom GPCs are unreliable or inadequate might be forced to be visual readers since using the rules would be slow and error prone. Visual readers, then, would be engaged in a search for a word-specific match in the lexicon.[2] Novel letter strings might allow them to apply GPCs but don't require it; a pronunciation can be generated by (visual) analogy to a real word (e.g., Glushko, 1979; Kay & Marcel, 1981).

This interpretation of orthographic depth means that there is only one kind of processing. Indeed, parallel distributed networks destroy the basis for considering "routes" to the lexicon as if they were independent pathways with no mingling of activation. At bottom, they allow us to reject the dual route model altogether, including its logic for inferring nonphonology.

THE LOGIC OF INFERRING NONPHONOLOGY IN THE ABSENCE OF A DUAL ROUTE THEORY

In the logic of dual-route theory, lexical context effects are thought to undermine the case for assembled phonology. Since rules are discrete, not graded, things like frequency should not matter: K should be pronounced /k/ whether it appears in KICK or KALE. If the higher frequency word is pronounced faster, it must be because its visual form is more familiar. But in an interactive network, the lower threshold of high frequency word units means that they would be activated sooner by the pattern of excitation arising through the phoneme unit level. Indeed, with communication between levels, many lexical properties can be expected to influence pronunciation. Automatic involvement of the lexicon is inevitable given the bi-alphabetic nature of Serbo-Croatian. The presence of phonologically ambiguous letters generates activity along two letter-phoneme connections. If there are, say, two phonologically ambiguous letters in a four letter word, four pronunciations of that letter string are assembled. Each gives rise to some activation at the word unit level depending on how closely the phonemes match the word unit (with respect to number and order) and on the word units' frequencies. Activation of certain word units is strengthened by interaction between word and phoneme levels and continues until above-threshold activation of a single word unit emerges. Of course, this interactive processing is not limited to phonologically ambiguous words; it is characteristic of the ordinary language processor. Phonologically unique letter strings generate a single code but it partially activates a number of word units. Although a single word unit emerges quickly from interaction between word and phoneme levels, interactive processes nonetheless provide the opportunity for lexical influences on pronunciations assembled on the basis of prelexical phonology. That is to say, phonological codes are assembled by the prelexical phonological connections but a single pronunciation is settled on out of the global pattern of activation. Lexical involvement does not contravene prelexical phonology (Carello, Lukatela, & Turvey, under review).

Relatedly, we argue that the distinction between assembled and accessed phonology has been cut too sharply, as if phonological information came only from one source or the other. To arrive at the lexicon phonologically does not mean that the assembled code carries every phonological nuance. Linguistic features such as stress and prosody must be derived from the phonological representation in the lexicon which has itself been accessed via prelexical phonological connections. For example, the stress pattern of Serbo-Croatian words—which is not marked in the orthography—can be rising or falling and long or short. In addition, although the first syllable is usually stressed, occasionally the second syllable is stressed instead. A correct pronunciation requires information about the stress pattern which can only be had at the word unit level. But information about the stress pattern is only made available once the word unit has been activated by the phonological code. That is to say, the existence of accessed phonology does not contravene prelexical phonology (see Lukatela & Turvey, 1990, for experimental ramifications of differing stress patterns between contexts and targets).

THE LOGIC OF INFERRING PHONOLOGY

The case for prelexical phonology is, at the very least, not undercut by the existence of lexical context effects. But what kind of evidence would make the case for prelexical phonology? If we invert the logic that has been established by those advocating primacy of the word-specific visual route, we can look for several things. Phonological influences should be observed on acceptance latencies (which are, by and large, faster than rejection latencies), especially on high frequency words. Such evidence would support the claim that the influence is felt in ordinary word recognition, not just for letter strings that have no lexical entries (cf. Coltheart, Davelaar, Jonasson, & Besner, 1977; Kay & Marcel, 1981). Phonological effects should be apparent in both naming and lexical decision. Naming is important because it is supposed to be free of post-lexical influences (Balota & Chumbley, 1985; West & Stanovich, 1982). The effect ought to depend on the number of constituents (letters, syllables) in a word. A phonologically analytic process would reflect the burden of decoding details of the orthographic structure (Green & Shallice, 1976). Phonological effects should persist in the face of experimental conditions that discourage the use of prelexical phonology. Strategic insensitivity would suggest that the phonological route is nonoptional (cf. Hawkins, Reicher, Rogers, & Peterson, 1976). Finally, phonological effects must appear over and above effects due to graphemic similarity. Orthographically similar rhyming items should behave the same as orthographically dissimilar rhyming items but different from orthographically similar nonrhyming items (cf. Evett & Humphreys, 1981).

THE CASE FOR PRELEXICAL PHONOLOGY IN SERBO-CROATIAN

The case for prelexical phonology has been made on each of these points using the Serbo-Croatian language. These results can be organized around three general manipulations permitted by exploiting the two alphabets: (1) comparisons between phonologically unique letter strings, composed exclusively of unique and common letters, and phonologically ambiguous letter strings, composed exclusively of common and ambiguous letters; (2) comparisons of phonemically and graphemically similar pairs, written in the same alphabet, and phonemically similar but graphemically dissimilar pairs, with the context and target written in different alphabets; and (3) comparisons of phonologically ambiguous pseudowords in which a mixed interpretation of the letters in a single letter string either is or is not a word.

Manipulations of the first type produce the so-called Phonological Ambiguity Effect—letter strings with more than one phonological interpretation are associated with longer latencies and higher errors than letter strings with only one phonological interpretation *even though they are the same words.* For example, VETAR and BETAP are Roman and Cyrillic, respectively, for "the wind." VETAR has one phonological interpretation, /vetar/, because V and R are uniquely Roman letters and E, T, and A are common. BETAP, in contrast, has four phonological interpretations because B and P can be read (differently) in Roman and Cyrillic. The Cyrillic interpretation of both yields /vetar/. The Phonological Ambiguity Effect occurs in naming and lexical decision (e.g., Lukatela, Feldman et al., 1989; Lukatela, Turvey et al., 1989), is larger for words (independent of frequency) than pseudowords (e.g., Feldman & Turvey, 1983; Lukatela, Feldman, et al., 1989), increases with more phonologically ambiguous letters (Feldman, Kostić, Lukatela, & Turvey, 1983; Feldman & Turvey, 1983), decreases with more unique letters (Lukatela, Feldman, et al., 1989), and persists despite instructions (Lukatela, Savić, Gligorijević, Ognjenović, & Turvey, 1978) or experience favoring one alphabet (Feldman & Turvey, 1983) or discouraging phonological coding (Lukatela, Feldman, et al., 1989).

Manipulations of the second type produce phonemic similarity effects. Naming latencies to the word target PUŽIĆ (/puzhich/) and the pseudoword target PUDIĆ (/pudich/) are facilitated to the same degree by phonemically similar contexts, whether those are graphemically similar (PUTIĆ, /putich/) or dissimilar (ПУТИЋ, /putich/) (Lukatela & Turvey, 1990a). For lexical decision latencies, the direction of the phonemic similarity effect depends on target frequency, the ordinal position of the distinguishing phoneme, and lexicality. Phonemic similarity effects persist even when the context is masked (for both word-pseudoword and pseudoword-word sequences, Lukatela & Turvey, 1990a), and when graphemic similarity is further reduced by writing contexts in lower case and targets in upper case, for example, pasus-ПАСУЉ, /pasus-pasulj/ (Lukatela et al., 1990). Finally, target identification under conditions of backward masking—a target followed by a pseudoword mask which is itself followed by a pattern mask—is enhanced when the pseudoword mask is phonologically similar to the target (Lukatela & Turvey, 1990b). The mask presumably continues activation at the phoneme unit level that had been initiated by the target (Naish, 1980; Perfetti, Bell, & Delaney, 1988).

Manipulations of the third type produce "virtual word" effects. BEMAP and HAPEM both differ from a real word by one letter (BETAP and XAPEM, respectively). B, P, and H have different interpretations in Roman and Cyrillic so that a phonologically analytic processing of each string would produce four codes. For BEMAP none of these is a word, whereas for HAPEM one is a word. HAPEM-type strings produce a much larger false positive error rate: 30% vs. 3%. When a HAPEM-type follows a context associatively related to the virtual word interpretation, false positives increase to 55%, compared to 7% for BEMAP-types following associates of their source words (Lukatela, Feldman, et al., 1989; Lukatela, Turvey, et al., 1989). In naming, the mixed alphabet (virtual word) interpretation of HAPEM-type strings occurred 3-4 times more often than the mixed alphabet interpretation of BEMAP-types. These differences arise even though the two types of pseudowords are equally similar visually to a real word—they differ by one letter. But whereas every code for BEMAP is also one phoneme different from a real word, one code for HAPEM shares all phonemes with a real word. Virtual word effects derive from prelexical phonology.

For Serbo-Croatian, in sum, the requisite patterns of results have been obtained to allow the conclusion of prelexical phonology. Phonological involvement has been demonstrated on "yes" responses, with high frequency words, on words more than pseudowords, in naming as well as lexical decision; it is sensitive to the number of constituents, and persists despite experimental conditions that might discourage it; finally, phonological effects are independent of graphemic effects which, in fact, do not occur. The results from English and Hebrew do not permit quite the same point by point confirmation. But there are what we might consider "existence proofs" for a number of them.

THE CASE FOR PRELEXICAL PHONOLOGY IN ENGLISH

The supporting English results can be organized around four general manipulations, the first three of which exploit the deep orthography: (1) comparisons between pseudohomophones, nonwords that are pronounced the same as real words but spelled differently, and spelling controls, nonwords that differ from the targets by the same number of letters as the pseudohomophone but are pronounced differently; (2) comparisons between homophones, words that are pronounced the same as target words but spelled differently, and spelling controls; (3) comparisons of phonologically consistent pairs in which a given stem receives the same phonological interpretation, and phonologically inconsistent pairs in which a given stem receives different phonological interpretations; and (4) comparisons of phonemically and graphemically similar pairs, phonemically similar and graphemically dissimilar pairs, and graphemically similar but phonemically dissimilar pairs.

Manipulations of the first type provided some of the earliest suggestions of phonological involvement in lexical decision (e.g., Rubenstein, Lewis, & Rubenstein, 1971). But since this effect was on "no" responses which are already slow, delayed rejections of pseudohomophones was soon interpreted as implicating phonological involvement only when the direct route hadn't worked fast enough (see Van Orden et al., 1990, for a rebuttal of the logic behind the so-called "delayed phonology hypothesis"). Not prone to such an indictment are recent experiments showing associative priming by and of pseudohomophones: TABLE facilitated the naming of the pseudohomophone CHARE relative to the spelling control CHARK; the pseudohomophone prime TAYBLE facilitated the naming of CHAIR relative to the spelling control prime TARBLE (Lukatela & Turvey, 1991). In both of these instances, in order for the associative relationship to have had an effect, the lexicon must have been accessed and it must have been accessed through phonology. In four experiments, the graphemic control did not produce a significant effect (and the numerical difference was always in the wrong direction). In contrast, TAYBLE did not differ from TABLE in its effect on naming CHAIR. Moreover, the word targets (and source words of the pseudohomophones) were of relatively high frequency (217 according to the norms of Francis & Kučera, 1982).

Other experiments have demonstrated additional dimensions of equivalency in the processing of pseudohomophones and their real word counterparts (Lukatela & Turvey, in press). Between the presentation and recall of one or five digits, subjects performed a secondary task of naming a visually presented letter string—a pseudohomophone (e.g., FOLE, HOAP) or its lexical counterpart (FOAL, HOPE). If nonwords are named by a slow (resource expensive) process that assembles the letter string's phonology and words are named by a fast (resource inexpensive) process that accesses lexical phonology (see Paap & Noel, 1991), then memory load should interact with lexicality (HOPE vs. HOAP, FOAL vs. FOLE). To the contrary, three experiments found that load interacted only with frequency (HOPE vs. FOAL, HOAP vs. FOLE), suggesting that pseudohomophones and their word counterparts are processed similarly, namely, phonologically. An example of the form of the interaction is shown in Figure 1. In a fourth experiment the associative priming-of-naming task described above was secondary to the memory task.

In elaboration of Lukatela and Turvey's (1991) observations, associative priming (HOPE-DESPAIR, FOAL-HORSE) was equaled by pseudohomophone associative priming (HOAP-DESPAIR, FOLE-HORSE) with memory load affecting both kinds of priming in the same way.

Manipulations of the second type show homophony effects on rejection latencies, this time in semantic categorization tasks: BEATS takes longer to reject as a member of the category VEGETABLE than do other foils (e.g., Meyer & Ruddy, 1973). Finer analyses, however, reveal homophony to be influential on faster yes responses as well: The false positive error rate is higher for homophones than for spelling controls (18.5% vs. 3.0%, Van Orden, 1987) and the false positive "yes" latencies are comparable to the correct "yes" latencies (Van Orden, Johnston, & Hale, 1988). Moreover, although orthographic similarity of homophones (BEATS is more like BEETS than ROWS is like ROSE) matters under unmasked conditions, the orthographic effect disappears when targets are pattern-masked while the homophony effect remains strong (Van Orden, 1987).

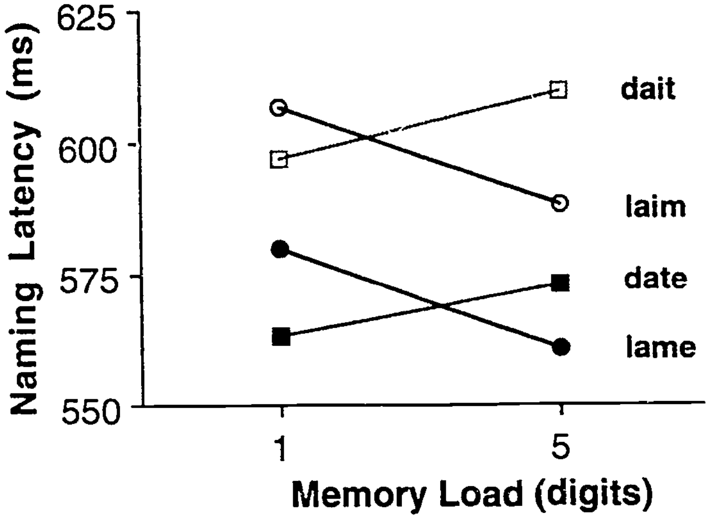


Figure 1. High frequency words and their pseudohomophones (closed and open squares, respectively) are hindered by increased memory load. The opposite pattern is obtained with low frequency words and their pseudohomophones (closed and open circles, respectively). That is, words are more similar to their nonlexical but phonologically identical counterparts than they are to each other.

Van Orden argues that this supports the role of phonological mediation as an early source of constraint on word identification (Van Orden, 1987; Van Orden et al., 1990).

Manipulations of the third type show differences in priming effects between graphemically similar pairs that are also phonologically similar (BRIBE-TRIBE) and graphemically similar pairs that are phonologically dissimilar (TOUCH-COUCH). Generally, phonological consistency is beneficial and phonological inconsistency is detrimental (Hanson & Fowler, 1987; Meyer, Schvaneveldt, & Ruddy, 1974). Where there are priming effects for both types of pairs, the effect with phonologically similar pairs is greater (Hanson & Fowler, 1987). Even the results of Evett and Humphreys (1981), who did not find differences due to consistency when the primes were masked, have been interpreted as supportive of phonological mediation by "noisy phonologic codes" (Van Orden et al., 1990, p. 495). The epithet noisy is applied on the assumption that TOUCH would give rise to a code that had elements of both /tutch/ and /towtch/.[3] Therefore, priming of COUCH by TOUCH is, in fact, a phonological effect. Van Orden (1987; Van Orden et al., 1990) argues further that sometimes noisy codes are sufficient to distinguish words from nonwords (e.g., when the pseudoword foils are illegal nonwords), in which case there would be no advantage for phonologically consistent pairs. Phonological inconsistency will be detrimental when noisy codes must be cleaned up, viz., for foils that are legal nonwords. These are the results reported by Shulman, Hornak, and Sanders (1978) and Hanson and Fowler (1987). Interestingly, detrimental phonological inconsistency effects—those historically taken to demonstrate phonological mediation—are most likely under experimental conditions that ought to discourage phonology were it optional (Van Orden et al., 1990). That is to say, with legal nonword foils, words would be better distinguished by a graphemic code were it an option.

Manipulations of the fourth type have produced inconsistent results. While facilitation for phonemically similar, graphemically dissimilar pairs has been reported (Hillinger, 1980), this has not been replicated, either in lexical decision (Martin & Jensen, 1988) or naming (Peter, Turvey, & Lukatela, 1990). But graphemic priming was not found either. As an important aside, we note that this latter result would appear to be in sharp contradiction of the major expectation from the hypothesized visual, word-specific route. If lexical items are coded visually (more precisely, orthographically), then preceding words that are visually similar to immediately subsequent words should facilitate decisions on the immediately subsequent words. That such visually based facilitation is difficult to obtain (ordinarily investigators have to impose a number of additional manipulations, such as severe forward masking of the prime, to reveal slight effects [e.g., Forster, 1987]) should be taken as prima facie evidence that visual access is neither prominent nor particularly straightforward. Curiously, proponents of the visual, word-specific route have been mute on this fai'ure to prime the lexicon visually.

More reliable than the results from forward phonemic priming are results from masked backward priming (a target followed by a pseudoword mask which is itself followed by a pattern mask): Targets are more likely to be identified when the pseudoword mask is phonemically rather than graphemically similar to it (Naish, 1980; Perfetti et al., 1988). Manipulations of this fourth type can be combined with those of the second type. ROWS is more likely to be recognized as a member of the category FLOWER when followed by a phonemically similar rather than graphemically similar pseudoword mask (Peter & Turvey, 1992).

In sum, the results for English are accumulating to allow the conclusion of prelexical phonology. Phonological involvement has been demonstrated on "yes" responses, with high frequency words, and in naming as well as lexical decision; it has occurred despite experimental conditions that might discourage it; and, finally, phonological effects have been obtained that are over and above graphemic effects which are, in fact, unreliable.

THE CASE FOR PRELEXICAL PHONOLOGY IN HEBREW

Our assertion that the underlying processing is the same across languages requires that there be at least some evidence of prelexical phonology in the deepest orthographies. The phoneme layer still exists even though the letter to phoneme connections might be multiple and very weak. Support for phonological involvement in Hebrew comes from two general manipulations that exploit the fact that vowels are not represented in ordinary text: (1) comparisons of pointed and unpointed letter strings, and (2) comparisons of phonologically ambiguous and unambiguous words.

Manipulations of the first type provided the earliest suggestion of phonological mediation in Hebrew. For some consonant strings, there is only one phonological interpretation with a single lexical entry. Adding the proper vowels redundantly specifies the same pronunciation. Adding certain incorrect vowels specifies other particular pronunciations that are phonotactically legal even though they are without a lexical entry. Adding other incorrect vowels that are allophonic with the correct vowels will specify the correct pronunciation even though that orthographic pattern has no lexical counterpart (it is a pseudohomophone). Navon and Shimron (1981) found that allophonically voweled letter strings (essentially pseudohomophones) did not differ in naming time from ordinary unpointed or correctly pointed letter strings. That is, the correct phonological interpretation accessed its lexical entry even though its orthographic form was novel. Naming was slower when the added vowels specified a pronunciation without a lexical entry. More recently, it has been shown that readers will wait for the vowel marks in a delayed presentation paradigm (consonant string followed at some lag by the diacriticals) even though the orthographic form has only one lexical entry (Frost, 1992). This was true for both high and low frequency words in both lexical decision and naming.

Manipulations of the second type are somewhat similar to manipulations of phonological ambiguity in Serbo-Croatian in that a given letter string can be pronounced in more than one way. In this case, the phonological options come not from choice of alphabet but from choice of vowels to assign to an unpointed letter string. Here we consider only those pronunciations that constitute words (rather than all pronunciations that might be generated by the random assignment of vowels). Consonant strings with three or more phonemic realizations are named more slowly than consonant strings with only one (Bentin, Bargai, & Katz, 1984). When semantic priming contexts are consonant strings with two phonemic realizations and two meanings, one a high frequency word and one a low frequency word, lexical decision is facilitated more by the phonological interpretation associated with the higher frequency word (Fros: & Bentin, 1992). When these same letter strings are pointed (and, therefore, phonologically unambiguous), the amount of facilitation by the low and high frequency versions is the same. Relatedly, contexts with both a high and low frequency meaning but with a single phonological interpretation (like the English word RUN, for example) also produce equivalent facilitation in targets semantically related to either of the two meanings. Taken together, these findings suggest that the ambiguity effect found with heterophonic homographs is phonological rather than semantic in origin (Frost & Bentin, 1992). Delaying the onset of vowel marks after the presentation of ambiguous letter strings with two phonemic realizations slows the naming of words (both high and low frequency) and pseudowords equally (Frost, 1992). This lag effect is larger than that for unambiguous words.

The results for Hebrew suggest at least some involvement of prelexical phonology. It has been demonstrated on "yes" responses, with high frequency words, and in naming as well as lexical decision; it has occurred despite experimental conditions that do not require it; and one phonological effect has been obtained that is over and above a graphemic effect. But the body of data from Hebrew are equivocal, perhaps epitomized by the fact that lexical decision to targets either orthographically or phonemically similar to pseudoword primes are facilitated to the same degree (Bentin et al., 1984).

Nonetheless, the extent of parallel evidence in Serbo-Croatian, English, and Hebrew is impressive. The script-sound relationships in the three languages constitute very different experimental settings. Some of the classes of experiments that we have discussed are not possible in the other language. English and Hebrew have no mixed alphabets; Serbo-Croatian has no phonological inconsistency. For the most part, the differences favor Serbo-Croatian as a vehicle for demonstrating prelexical phonology (Lukatela et al., 1990; Lukatela & Turvey, 1990 a, b). Despite these differences, early nonoptional phonological involvement is apparent in all. Differences that remain are arguably due to differences in covariant learning particularly with respect to letter-phoneme connections.

CONCLUDING REMARKS: THE PRIMACY OF PHONOLOGICAL "DYNAMICS"

We have chosen to build our arguments for reading's natural phonological basis around a hypothesis of prelexical phonology as primary. Roughly interpreted, this hypothesis is that processes intimately connected to those by which speech is produced and perceived constitute the major constraint on the mapping from print to lexicon. The now classic dual-route theory has provided a fairly simple (and empirically fruitful) framework within which to deliberate how a person's knowledge about words might be tapped by letter strings: It is tapped either by the letter strings described in the predicates of the visual system, or by letter strings described in the predicates of the speech system, or both. As the theory tends to go, the visual predicates are more prominent than the speech predicates. Our arguments in this chapter were phrased very much in the context of the dual-route theory, and in reaction to the proposed primacy of visual predicates. The strategy we adopted was chosen because, in many respects, it is the most convenient and the most conducive to communication (relying as it does upon the most conventional understanding). In these final remarks, however, we would like to take a more critical and circumspect stance. We explore the implications of a continuous dynamical perspective on word-recognition processes, the perspective adumbrated in much of the foregoing criticism of the "stubbornly nonphonological" accounts.

Our departure point is an assertion: Learning to read is largely an autonomous process. By this assertion we intend to mean several things. First, reading is achieved by a system capable of attuning to mappings between orthographic and linguistic structures, however arbitrarily complex those mappings might happen to be (that is, it does not require that the mappings be orthogonal or linearly separable). Second, the structures mapped between are characterized by distinguishable features or substructures at many grain sizes; there is, however, no biasing of the system toward any particular grain size. Consequently, attunement may occur to mappings that vary considerably in the sizes of the substructures comprising their domains and codomains. Third, the system's attunement is eventually most pronounced (but not exclusively restricted) to the mappings significant to reading without having to be informed explicitly as to what those particular significant mappings might be. Fourth, the enhanced attunement to reading-significant mappings follows from a generic selection principle: Those mappings are selected that are single-valued, or most nearly so. That is, the more invariant the relation between particular substructures of the orthography and particular linguistic substructures, the more likely is it that that mapping will be selectively enhanced.

In dynamical terms, what are the consequences of invariance—of single-valuedness? An approximate answer, one highlighted by Van Orden et al. (1990), is that resonance or self-consistency is achieved rapidly within the connective matrix binding (the processing units of) the domain's and codomain's substructures. Borrowing from adaptive resonance theory (Grossberg, 1987; Grossberg & Stone, 1986), a resonant mode is achieved when the activity excited in a given layer of processing units from below matches that excited from above. A closely related answer is that the pattern of activity engendered in the network instantiation of the mapping is stable. Consequently, where a mapping deviates from single-valuedness, the time course of achieving resonance is slower and/or the final-state stability is less.

In most languages, if not all, the invariance is greatest between orthography and phonology, roughly speaking, between the spellings of words and the names of words. Patently, the mappings between orthography and the meanings of words, and orthography and the syntactic functions of words, are considerably less consistent. Phonological representations will, therefore, achieve resonance faster, and reach states of stability greater, than other linguistic representations. Again, in terms of adaptive resonance theory, a greater match is achieved, and achieved at a more rapid pace, between the activity patterns in the phonological layer excited by (a) the lexical layer above, and (b) the graphemic layer below. The upshot is that even if many activations of linguistic substructures by orthographic substructures occur concurrently in word recognition, it is the phonological activation that stabilizes earliest, providing a basis for stabilizing the other patterns of linguistic activation (Van Orden et al., 1990).

In these final remarks we have pursued a line of argument constrained by the notions of autonomy and invariant. We have been led to conclude that, in word recognition, the dynamics associated with phonological processes are primary. It will be interesting to see in what directions a theory grounded in dynamics might evolve (along the lines, perhaps, of recent efforts in movement coordination, e.g., Kugler & Turvey, 1987; Schmidt, Beek, Treffner, & Turvey, 1991; Turvey, 1990; Turvey, Schmidt, & Beek, in press) and the kinds of experimental hypotheses to which it might give rise. A benchmark for evaluating such a theory's worth is the dual process theory, which has been the dominant source of stimulation for research on word recognition in recent times. Will a dynamically based theory be as fruitful?

REFERENCES

Balota, D. A., & Chumbley, J. I. (1985). The locus of word frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language, 24*, 89-106.

Bentin, S., Bargai, N., & Katz, L. (1984). Orthographic and phonemic coding for lexical access: Evidence from Hebrew. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 353-368.

Carello, C., Lukatela, G., & Turvey, M. T. (under review). Lexical involvement in naming does not contravene prelexical phonology: Comment on Sebastián-Gallés.

Carr, T. H., & Pollatsek, A. (1985). Recognizing printed words: A look at current models. In D. Besner, T. G. Waller, & G. E. MacKinnon (Eds.), *Reading research: Advances in theory and practice* (Vol. 5, pp. 1-82). Orlando, FL: Academic Press.

Coltheart, M. (1977). Critical notice of Gibson, E. J. and Levin, H., "The psychology of reading." *Quarterly Journal of Psychology, 29*, 157-167.

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing* (pp. 151-216). London: Academic Press.

Coltheart, M., Besner, D., Jonasson, J. T., & Davelaar, E. (1979). Phonological encoding in the lexical decision task. *Quarterly Journal of Experimental Psychology, 31*, 489-507.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535-555). New York: Academic Press.

DeFrancis, J. (1989). *Visible speech: The diverse oneness of writing systems.* Honolulu: University of Hawaii Press.

Evett, L. J., & Humphreys, G. W. (1981). The use of abstract graphemic information in lexical access. *Quarterly Journal of Experimental Psychology, 33*, 325-350.

Feldman, L. B., Kostić, A., Lukatela, G., & Turvey, M. T. (1983). An evaluation of the "basic orthographic syllable structure" in a phonologically shallow orthography. *Psychological Research, 45*, 55-72.

Feldman, L. B., & Turvey, M. T. (1983). Word recognition in Serbo-Croatian is phonologically analytic. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 288-298.

Forster, K. (1987). Form-priming with masked primes: The best match hypothesis. In M. Coltheart (Ed.), *The psychology of reading: Attention and performance XII* (pp. 127-146). Hillsdale, NJ: Lawrence Erlbaum Associates.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage.* Boston: Houghton Mifflin.

Frost, R. (1992). Prelexical and postlexical strategies in reading: Evidence from a deep and a shallow orthography. Paper presented at the Fifth Annual Meeting of the European Society for Cognitive Psychology. Paris, September, 1992.

Frost, R., & Bentin, S. (1992). Processing phonological and semantic ambiguity: Evidence from semantic priming at different SOAs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 58-68.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 104-115.

Glushko, R. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 361-379

Green, D. W., & Shallice, T. (1976). Direct visual access in reading for meaning. *Memory & Cognition, 4*, 753-758.

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science, 11,* 23-63.

Grossberg, S., & Stone, G. (1986). Neural dynamics of word recognition and recall: Priming, learning, and resonance. *Psychological Review, 93,* 46-74.

Hanson, V. L., & Fowler, C. A. (1987). Phonological decoding in word reading: Evidence from hearing and deaf readers. *Memory & Cognition, 15,* 199-207.

Hawkins, H. L., Reicher, G. M., Rogers, M., & Peterson, L. (1976). Flexible coding in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 380-385.

Hillinger, M. L. (1980). Priming effect with phonemically similar words: The encoding bias hypothesis reconsidered. *Memory & Cognition, 8,* 115-123.

Humphreys, G. W., & Evett, L. J. (1985). Are there independent lexical and nonlexical routes in word processing? An evaluation of the dual route theory of reading. *The Behavioral and Brain Sciences, 8,* 689-739.

Kay, J., & Marcel, A. (1981). One process, not two, in reading aloud: Lexical analogies do the work of nonlexical rules. *Quarterly Journal of Experimental Psychology, 33A,* 397-413.

Kugler, P. N., & Turvey, M. T. (1987). *Information, natural law, and the self-assembly of rhythmic movement.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Liberman, A. M. (1991). Observations from the sidelines. *Reading and Writing: An Interdisciplinary Journal, 3,* 429-433.

Liberman, A. M. (in press). The relation of speech to reading and writing. To appear in B. de Gelder & J. Morais (Eds.), *Language and literacy: Comparative approaches.* Cambridge, MA: MIT Press.

Liberman, I. Y., Liberman, A. M., Mattingly, I., & Shankweiler, D. (1980). Orthography and the beginning reader. In J. F. Kavanagh & R. Venezky (Eds.), *Orthography, reading, and dyslexia* (pp. 137-153). Baltimore, MD: University Park Press.

Lukatela, G., Carello, C., & Turvey, M. T. (1990). Phonemic priming by words and pseudowords. *European Journal of Cognitive Psychology, 2,* 375-394.

Lukatela, G., Feldman, L. B., Turvey, M. T., Carello, C., & Katz, L. (1989). Context effects in bi-alphabetical word perception. *Journal of Memory and Language, 28,* 214-236.

Lukatela, G., Popadić, D., Ognjenović, P., & Turvey, M. T. (1980). Lexical decision in a phonologically shallow orthography. *Memory & Cognition, 8,* 124-132.

Lukatela, G., Savić, M., Gligorijević, B., Ognjenović, P., & Turvey, M. T. (1978). Bi-alphabetical lexical decision. *Language and Speech, 21,* 142-165.

Lukatela, G., & Turvey, M. T. (1990a). Phonemic similarity effects and prelexical phonology. *Memory & Cognition, 18,* 128-152.

Lukatela, G., & Turvey, M. T. (1990b). Automatic and prelexical computation of phonology in visual word identification. *European Journal of Cognitive Psychology, 2,* 325-343.

Lukatela, G., & Turvey, M. T. (in press). Similar attentional, frequency, and associative effects for pseudohomophones and words. *Journal of Experimental Psychology: Human Perception and Performance.*

Lukatela, G., & Turvey, M. T. (1991). Phonological access of the lexicon: Evidence from associative priming with pseudohomophones. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 951-966.

Lukatela, G., Turvey, M. T., Feldman, L. B., Carello, C., & Katz, L. (1989). Alphabetic priming in bi-alphabetical word perception. *Journal of Memory and Language, 28,* 237-254.

Martin, R. C., & Jensen, C. R. (1988). Phonological priming in the lexical decision task: A failure to replicate. *Memory & Cognition, 16,* 505-521.

Mattingly, I. (1985). Did orthographies evolve? *RASE, 6,* 18-23.

Mattingly, I. (1992). Linguistic awareness and orthographic form. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning.* Amsterdam, The Netherlands: Elsevier.

McClelland, J. L., & Rumelhart, D. E. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructures of cognition: Vol. 2. Psychological and biological models.* Cambridge, MA: MIT Press.

Meyer, D. E., & Ruddy, M. G. (1973, November). *Lexical memory retrieval based on graphemic and phonemic representation of printed words.* Paper presented at meeting of the Psychonomic Society, St. Louis, MO.

Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. G. (1975). Loci of contextual effects on visual word-recognition. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance V* (pp. 98-118). London: Academic Press.

Naish, P. (1980). The effects of graphemic and phonemic similarity between targets and masks in a backward visual masking paradigm. *Quarterly Journal of Experimental Psychology, 32,* 57-68.

Navon, D., & Shimron, J. (1981). Does word naming involve grapheme-to-phoneme translation? Evidence from Hebrew. *Journal of Verbal Learning and Verbal Behavior, 20,* 97-109.

Paap, K. R., & Noel, R. W. (1991). Dual-route models of print to sound: Still a good horse race. *Psychological Research, 53,* 13-24.

Perfetti, C. A., Bell, L. C., & Delaney, S. (1988). Automatic (prelexical) phonetic activation in silent word reading: Evidence from backward masking. *Journal of Memory and Language, 27,* 59-70.

Peter, M., & Turvey, M. T. (April, 1992). Semantic categorization of backward masking. Paper presented at the Eastern Psychological Association.

Peter, M., Turvey, M. T., & Lukatela, G. (1990). Phonological priming: Failure to replicate in the rapid naming task. *Bulletin of the Psychonomic Society, 28,* 389-392.

Rubenstein, H., Lewis, S. S., & Rubenstein, M. A. (1971). Evidence for phonemic recoding in visual word recognition. *Journal of Verbal Learning and Verbal Behavior, 10,* 645-657.

Schmidt, R. C., Beek, P. J., Treffner, P. J., & Turvey, M. T. (1991). Dynamical substructure of coordinated rhythmic movement. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 635-651.

Sebastián-Gallés, N. (1991). Reading by analogy in a shallow orthography. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 471-477.

Shulman, H. G., Hornak, R., & Sanders, E. (1978). The effect of graphemic, phonetic, and semantic relationships on access to lexical structures. *Memory & Cognition, 6,* 115-123.

Turvey, M. T. (1990). Coordination. *American Psychologist, 45,* 938-953.

Turvey, M. T., Schmidt, R. C., & Beek, P. J. (in press). Fluctuations in inter-limb rhythmic coordinations. In K. Newell (Ed.), *Variability in movement control.* Champaign, IL: Human Kinetics Press.

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound and reading. *Memory & Cognition, 15,* 181-198.

Van Orden, G. C., Johnston, J. C., & Hale, B. L. (1988). Word identification in reading proceeds from spelling to sound to meaning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 371-385.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review, 97,* 488-522.

Wang, W. S-Y. (1981). Language structure and optimal orthography. In O. J. L. Tzeng & H. Singer (Eds.), *Perception of print: Reading research in experimental psychology* (pp. 223-236). Hillsdale, NJ: Erlbaum.

West, R. F., & Stanovich, K. E. (1982). Source of inhibition in experiments on the effect of sentence context on word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 5, 385-399.

FOOTNOTES

*In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 211-226). Amsterdam: Elsevier Science Publishers (1992).

†Also University of Connecticut, Storrs.

‡University of Belgrade.

[1] Orthographic depth, in fact, has a second aspect and that is the relative remoteness of the phonetic representation from the morphophonological representation (I. Liberman et al., 1980). Experimental investigations that deal with orthographic depth tend to focus only on how easily the orthography approximates the phonetic representation.

[2] Van Orden et al. (1990) point out that the way in which the debate has been framed has had the insidious effect of turning psycholinguistics into what it had originally criticized: An account of verbal behavior rooted in specific stimulus-response connections.

[3] This is not unlike what we have proposed for phonologically ambiguous letter strings in Serbo-Croatian—all possible pronunciations of a string are generated before one is settled on through competitive processes.

Haskins Laboratories Status Report on Speech Research
1992, SR-111/112, 205-214

Poor Readers are Not *Easy* to Fool: Comprehension of Adjectives with Exceptional Control Properties[*]

Paul Macaruso,[†] Donald Shankweiler,[‡] Brian Byrne,[†††] and Stephen Crain[‡]

An earlier experiment by Byrne (1981) found that young poor readers tend to act out sentences containing adjectives with object control, like *easy*, as though they were adjectives with subject control, like *eager*. Byrne interpreted this result as evidence that poor readers lag in acquisition of syntactic knowledge underlying this distinction. However, the possibili'y that a processing limitation could have contributed to the poor readers' difficulties with object-control adjectives had not been fully explored. In an effort to tease apart these alternatives, we tested comprehension of object-control adjectives in second-grade good and poor readers using both an act out task and a sentence-picture matching task. Contrary to Byrne's (1981) results, we did not find significant group differences in interpreting object-control adjectives with either task. Reasons for the discrepancy are suggested, and remedies for the pitfalls in designing experiments to assess syntactic knowledge in young children are proposed.

It is becoming increasingly apparent that the difficulties associated with reading disability extend beyond poor word decoding skills. A number of studies have shown that children with reading problems have difficulties in comprehension of certain types of spoken sentences. Relative clauses are perhaps the most frequently mentioned structures that cause problems for poor readers (Bar-Shalom, Crain & Shankweiler, in press; Byrne, 1981; Goldsmith, 1980; Mann, Shankweiler, & Smith, 1984; Smith, Macaruso, Shankweiler & Crain, 1989; Stein, Cairns & Zurif, 1984). Other constructions that have been implicated include passives (Stein, et al., 1984), sentences containing indirect and direct objects (Fletcher, Satz, & Scholes, 1981), certain types of imperative sentences (Macaruso, Bar-Shalom, Crain, & Shankweiler, 1989; Whitehouse, 1983), and sentences containing adjectives with exceptional control properties, such as *easy* (Byrne, 1981; Crain, 1987).

A plausible interpretation of these difficulties with spoken sentences is that young poor readers are often delayed in acquiring certain aspects of syntactic knowledge (e.g., Byrne, 1981; Stein et al., 1984). An alternative explanation is that their difficulties stem from overloaded phonological processing capacities (e.g., Crain & Shankweiler, 1988; Mann et al., 1984; Shankweiler & Crain, 1986; Smith, Mann, & Shankweiler, 1986). A processing limitation may reflect the special memory requirements associated with analysis of complex sentences. In addition, it may reflect the complexity of the task used to assess comprehension.

Accordingly, for some years our research has addressed the problem of disentangling structural and processing contributions to performance on sentence comprehension tasks. For example, Mann et al. (1984) found that young poor readers were significantly worse than good readers in acting out relative clause sentences. In a subsequent

We are grateful to Eva Bar-Shalom, Ruth Fielding-Barnsley, and Suzanne Smith for their help in data collection. We also thank the staff and students of the Armidale (Australia) and Coventry (Connecticut, USA) public schools for their cooperation. F ian Byrne, Stephen Crain, and Donald Shankweiler received financial support under visiting scholar schemes of the University of New England and the University of Connecticut Research Foundation. Stephen Crain and Donald Shankweiler also received support from a Program Project Grant to Haskins Laboratories, New Haven, Connecticut, from the National Institute of Child Health and Human Development. The work of Paul Macaruso was supported by a grant from the James S. McDonnell Foundation to the Neurolinguistics Laboratory, MGH Institute of Health Professions.

study Smith et al. (1989) showed that reducing the syntactically irrelevant complexities of the target sentences and modifying the task to satisfy presuppositions associated with the use of relative clauses resulted in similar high levels of performance by both good and poor readers (see also Crain, Shankweiler, Macaruso, & Bar-Shalom, 1990). Findings such as these suggest that poor readers' difficulties in spoken sentence comprehension do not necessarily reflect deficient structural knowledge, but may instead reflect the processing demands associated with the comprehension task.

Mindful of these findings we were led to reconsider a result in the research literature which is often cited as evidence for a structural deficit in poor readers. One of us (Byrne, 1931) found that young poor readers tend to act out sentences containing adjectives with object control, like *easy*, as though they were adjectives with subject control, like *eager*. For example, they might act out the sentence "The bird is easy to bite" as if it meant "The bird finds it easy to bite (something)." Of course, a subject-control interpretation is correct for "The bird is eager to bite." Because sentences containing object-control adjectives (O-adjectives) and subject-control adjectives (S-adjectives) are short and of equivalent length, Byrne argued that the selective misinterpretation of O-adjective sentences should not be attributed to memory limitations. Instead, difficulties with these sentences were interpreted as evidence of a syntactic delay. Indeed, Chomsky (1969) has proposed that O-adjective sentences are mastered later by normal children than S-adjective sentences because O-adjectives are syntactically more complex. However, the possibility that processing limitations could account for the special difficulty in comprehending O-adjective sentences has not been fully examined.

The purpose of this note is to present some findings that illustrate the practical difficulties in disentangling structural and processing contributions to performance on sentence comprehension tasks. Before accepting the conclusion that young poor readers' misinterpretations of O-adjectives reveal a syntactic delay, we reexamined Byrne's (1881) experiment to identify possible non-syntactic sources of difficulty associated with his act-out task. First, it may be relevant that Byrne did not explicitly test whether his subjects understood the meanings of the lexical items employed in the test sentences. In addition, an act-out task may give rise to interpretive difficulties associated with pragmatic factors. One problem hinges on the abstractness of adjectives like *easy* or *hard*. It may not have been apparent to some of the children how to act out the condition of being "hard to bite," as in "The snake is hard to bite." As pointed out by Hamburger and Crain (1984), pragmatic concerns associated with planning a response may increase memory demands and thus mask syntactic knowledge.

These considerations impelled us to ask whether the poor readers' difficulties with O-adjectives in Byrne's (1981) study may have been due to factors other than syntactic knowledge. We therefore decided to conduct a new experiment to determine if young poor readers have mastered the syntax associated with O-adjectives. In one task we attempted to confirm Byrne's original result by asking good and poor readers to act out O- and S-adjective sentences similar to the ones he used. In the other task, which was administered concurrently with the act-out task, the children's comprehension of O-adjective sentences was assessed using sentence-picture matching, which reduces nonsyntactic processing demands by eliminating the need to plan a response. In addition, the sentence-picture matching task incorporated a control that was missing from the act-out task: the subjects' comprehension of the meanings of the adjectives employed in the test sentences was assessed.

Act-out task. The act-out task was patterned after Byrne (1981), which was based on an earlier study by Cromer (1970). To obtain a generally valid result, the task was administered to two comparable samples of English-speaking children, one from the United States and one from Australia (where Byrne's original study was conducted). Twenty-eight children in the second year of school participated from each country. Each child initially received the Decoding Skills Test (DST) (Richardson & DiBenedetto, 1986) as a measure of reading ability and the Peabody Picture Vocabulary Test - Revised (PPVT) (Dunn & Dunn, 1981) as an measure of vocabulary knowledge, from which an IQ estimate may be obtained.

Characteristics of the subjects are given in Table 1. For the United States sample, the good readers comprised 14 children (five girls, nine boys) with DST scores greater than 69, and the poor readers comprised 14 children (four girls, ten boys) with DST scores less than 56. The DST score is the total number of items read correctly on a list containing 60 words and 60 pseudowords, chosen to give full representation of the syllable types that occur in English. The reader groups did not differ significantly ($p > .05$) in mean age nor in mean PPVT scores.

For the Australian sample, the good readers comprised 14 children (eight girls, six boys) with DST scores greater than 76, and the poor readers comprised 14 children (five girls, nine boys) with DST scores less than 57. The reader groups did not differ significantly in mean age. However, the poor readers' mean PPVT score was significantly lower than that of the good readers. $t(26) = 2.26$, $p < .05$. No poor reader obtained a score less than 85, i.e., one standard deviation below the mean of the normative sample (which is 100). Thus, all subjects displayed vocabulary knowledge within the average range.

Test materials for the act-out task are listed in Appendix 1. The test sentences contained four O-adjectives (*easy*, *fun*, *hard*, *difficult*) and four S-adjectives (*eager*, *glad*, *willing*, *happy*). We used the same set of O- and S-adjectives as Byrne (1981) with one exception: *difficult* replaced *tasty*. There were two sentences for each O-adjective and one for each S-adjective. Each occurrence of an O-adjective was paired with a different verb. Four verbs were used in all: *follow*, *bite*, *touch*, and *kiss*. Byrne had employed only one verb, *bite*, in his test sentences. "Wombat" replaced "squirrel" in the test sentences administered to the Australian children.

Subjects were tested individually. They were first introduced to two hand-puppets, a teddybear and a squirrel (or wombat). The tester then demonstrated two act-out scenes: the teddybear following the squirrel, and the squirrel kissing the teddybear. Subsequently, the children placed the puppets on their hands, and were asked to act out each sentence spoken by the tester. The sentences were presented in the fixed random order shown in Appendix 1.

The results of the act-out task are summarized in Table 2. Overall, the subjects displayed a high level of performance in acting out the O-adjective sentences. Each reader group responded with greater than 75% accuracy, and no group differences were found in either sample ($p > .05$). Thus, we did not confirm the results obtained by Byrne in 1981.

The results were further examined using the classification scheme employed by Byrne in the earlier study (see Table 3). Subjects were categorized according to their performance on O-adjective sentences as follows: *Primitive rule users* interpreted all O-adjectives as if they were S-adjectives (e.g., by having the teddybear do the biting in "The teddybear is easy to bite"). *Intermediates* interpreted some of the O-adjectives as S-adjectives, and *passers* interpreted all O-adjectives correctly.

Table 1. *Characteristics of the good and poor readers.*

	United States				Australia			
	Good		Poor		Good		Poor	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age (months)	92.2	4.6	95.9	8.3	91.5	3.6	93.5	5.6
PPVT	115.4	11.1	108.9	9.4	109.9	7.0	102.3	10.5
DST[a]	97.2	16.1	37.9	12.0	95.9	10.4	28.5	15.8

[a]Maximum DST score = 120

Table 2. *Performance on the act-out task: Number correct.*

	United States				Australia			
	Good		Poor		Good		Poor	
Sentence Type	Mean	SD	Mean	SD	Mean	SD	Mean	SD
O-adjective (8)	7.5	1.4	6.9	2.2	6.3	1.5	7.1	1.3
S-adjective (4)	3.9	0.4	3.3	0.6	3.2	0.8	3.4	0.7

Table 3. *Classification of performances on the act-out task.*

	Byrne (1981)		United States		Australia	
	Good	Poor	Good	Poor	Good	Poor
Primitive	1	4	0	1	0	0
Intermediate	8	14	2	5	11	6
Passer	9	3	12	8	3	8

In the present study only one child (a poor reader in the United States sample) was classified as a primitive rule-user. The remainder of the children interpreted at least some of the O-adjective sentences correctly. More than half of the subjects (55%) were classified as passers, and the remaining subjects fell into the intermediate category. In fact, half of the intermediates produced just one error in acting out the eight O-adjective sentences. If we relax the criterion for "passing" to include 7 out of 8 correct, 77% of the children, including more than half in each group, would be considered as displaying mastery.

Performance on the O-adjective sentences may be contrasted with performance on the putatively less complex S-adjective sentences. As revealed in Table 2, performance on the S-adjective sentences was less than perfect. No reader group difference was found in the Australian sample, but in the United States sample the good readers outperformed the poor readers ($t(26) = 3.00$, $p < .01$). However, we found that these poor readers had difficulty with only one S-adjective sentence: "The teddybear is eager to touch." They were only 43% accurate on this sentence, whereas they were 95% accurate on the remaining S-adjective sentences. For the other three reader groups, more than 65% of the errors with S-adjectives also occurred on this sentence. Apparently, many of the children were uncertain about the control properties of *eager*.

As was the case with S-adjective sentences, performance also varied across the set of O-adjective sentences. For example, 47% of the errors with O-adjectives occurred on the two sentences containing *fun*. These findings suggest that accurate performance on the act-out task is highly dependent on the peculiarities of specific lexical items. The changes in vocabulary introduced into the present study may have had greater implications than we anticipated. The inclusion of a different adjective and different verbs may have made our act-out task easier than Byrne's (1981) original task. In fact, collapsing over groups, the children in the present study had much more success in acting out O-adjective sentences than the children in Byrne's original study (87% correct versus 64% correct). In any case, we believe the modifications made in the present study strengthen the "naturalness" of the task and thus provide a more valid test of the syntactic delay hypothesis (see Shankweiler, Crain, Gorrell, & Tuller, 1989). As we mentioned earlier, another factor associated with accurate performance may be the ability to overcome the pragmatic difficulties intrinsic to the act-out task. Taken together, these factors help to explain why we did not confirm the deficit in poor readers found in Byrne's earlier study.[1]

Sentence-picture matching task. The sentence-picture matching task can be argued to pose fewer processing demands than the act-out task. In sentence-picture matching, the child is required to retain the input string in working memory only long enough to derive an interpretation and match it to an appropriate picture. The demands associated with planning and implementing a response, as in the act-out task, are averted. Based on these observations we chose to assess the syntactic knowledge of good and poor readers with a sentence-picture matching task. Two pictures were created for each test sentence, one depicting the object-control interpretation and the other the subject-control interpretation. For example, the two pictures shown in Figure 1 were presented with the sentence, "The bear is easy to reach." The lower picture shows a boy reaching for a bear that is genuinely easy to reach, which corresponds to the correct object-control interpretation. The foil shows a bear reaching for honey that is easily within reach, which corresponds to the incorrect subject-control interpretation.

Three groups of good and poor readers were tested with the sentence-picture matching task. Two of the groups comprised the same subjects who participated in the act-out task. The third group consisted of 16 good readers and 18 poor readers in the second grade who had participated in a study of relative clause syntax (Smith et al., 1989). Selection criteria for the subjects in the Smith et al. study were essentially the same as in the act-out task.

Test materials for the sentence-picture matching task are listed in Appendix 2. Test sentences 1-8 contained the four O-adjectives: *easy*, *impossible*, *hard*, *difficult*. (Each of these adjectives was used in the act-out task except *impossible*, which replaced *fun*.) Each adjective appeared in two sentences. Test sentences 9-12 contained adjectives that are ambiguous with regard to their control properties. For example, given "The man is too dirty to serve," the child may select a picture of a dirty man not being permitted to serve someone (the subject-control interpretation), or a picture of a dirty man being refused service (the object-control interpretation). The ambiguous sentences allowed us to determine which interpretation the children prefer for adjectives that in the adult grammar have both an object-control and a subject-control interpretation. Sentences 13-16 presented the four O-adjectives in a simplified form (e.g., "It is easy to reach the bear"). These sentences were included as controls to assess the children's comprehension of the specific lexical items and the appropriateness of the pictures. If the children know the meanings of the O-adjectives and the pictures accurately depict these meanings, then they should display few errors on these sentences.

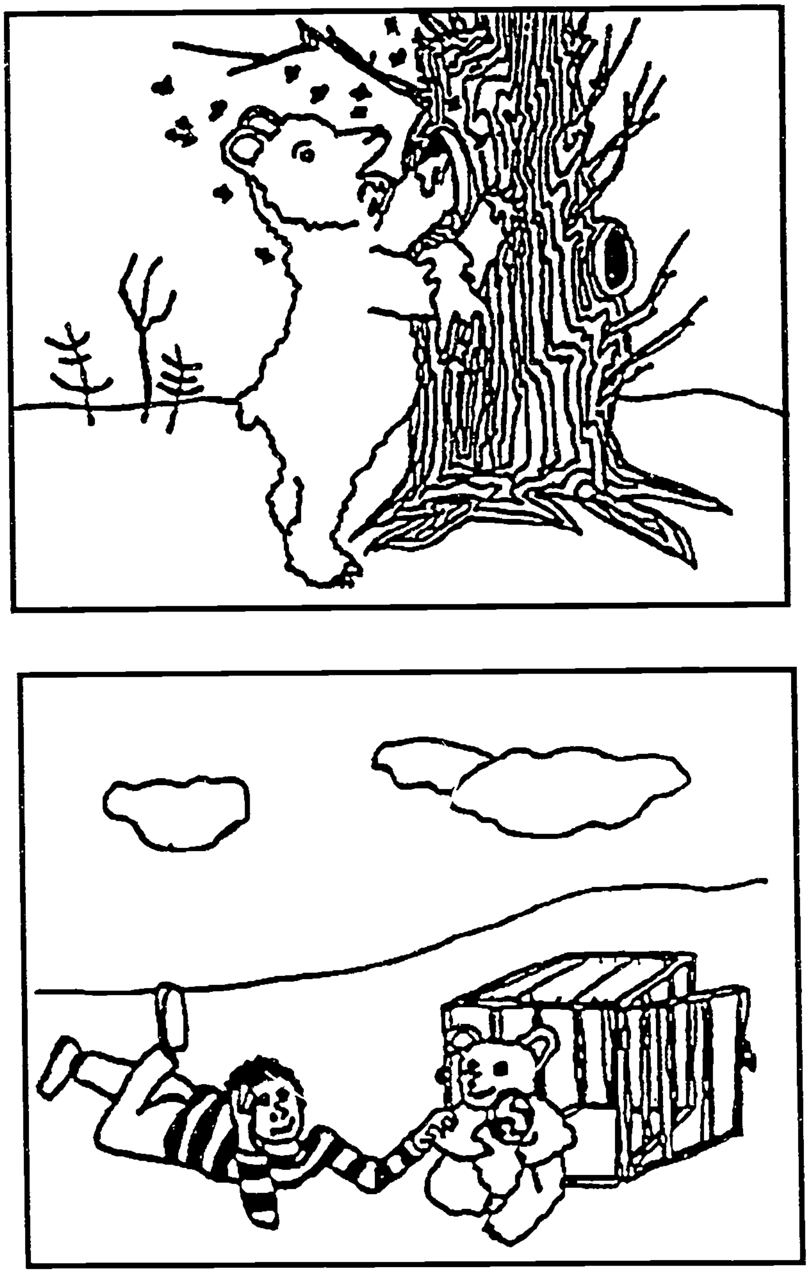


Figure 1. Picture pair for "The bear is easy to reach."

Subjects were tested individually. The tester read each sentence aloud, and the child was asked to point to the picture that best fits the meaning of the sentence. The sentences were presented in the fixed random order shown in Appendix 2.

The mean number of object-control responses for each sentence type and for each reader group is presented in Table 4. No reader group differences were obtained in any of the samples for the eight object-control sentences, for the four control sentences, and for the four ambiguous sentences ($p > .05$).

As evident in Table 4, however, overall performance levels on the O-adjective sentences were less than perfect. Percent correct ranged from a high of 86% to a low of 78%. As in the act-out task, incorrect responses were not uniformly distributed across the sentences. Performance on the two sentences containing the phrase "impossible to jump" accounted for 82% of the errors. For the remaining O-adjective sentences, the subjects responded with 95% accuracy. Unanticipated difficulty with the phrase "impossible to jump" was also apparent in the children's performance on the control sentences. All but one of the errors on these sentences occurred for "It is impossible to jump the frog."

One possible reason the children had difficulty with "impossible to jump" may be because they did not understand the use of *impossible*. However, we consider it more likely that the picture pairs used to test "impossible to jump" were difficult to interpret. One of these picture pairs, used to test "The frog is impossible to jump" and "It is impossible to jump the frog," is presented in Figure 2.

With a new group of subjects we asked whether the picture pairs used with "impossible to jump" are, in fact, ambiguous. In order to rule out limitations in vocabulary knowledge, we decided to pose the question to adults. Twenty-four adults were presented with a shortened version of the sentence-picture matching task in which they were asked to select the correct picture and provide a confidence rating for their choice. Variants of the original test sentences were created by interchanging *impossible* with *hard*, an adjective that did not present much difficulty to the children. Examples of the new sentences are "The frog is hard to jump" and "The boy is impossible to catch." The picture pairs originally presented with *impossible* and the pairs originally presented with *hard* were now tested with each adjective. The crossing of picture pairs and adjectives allowed us to tease apart difficulties associated with the pictures from difficulties associated with *impossible*.

The adults made few errors in selecting the correct picture. However, their confidence ratings clearly indicate that the pictures originally presented with *impossible* were more difficult than the pictures originally presented with *hard*. There were 24 instances of a low confidence rating (or an incorrect response) for the pictures originally presented with *impossible* while only five were given for the pictures originally presented with *hard*. Moreover, ten instances of a low confidence rating (or an incorrect response) occurred when *hard* was presented with the original *impossible* pictures, but only two when *impossible* was presented with the original *hard* pictures. This suggests that the source of difficulty in the children's performance was the original *impossible* pictures and not the adjective itself.

Finally, as revealed in Table 4, no reader group showed a definite preference for either a subject-control or an object-control interpretation for the four ambiguous sentences. Thus, bias did not enter into performance on the O-adjective sentences.

Table 4. *Performance on the sentence-picture matching task: Number object-control responses.*

		United States				Australia				Smith et al (1989)			
		Good		Poor		Good		Poor		Good		Poor	
Sentence Type		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
O-adj.	(8)	6.7	0.9	6.8	1.0	6.9	0.9	6.2	1.6	6.3	0.7	6.6	1.2
Contr.	(4)	3.8	0.4	3.4	0.5	3.6	0.5	3.8	0.4	3.9	0.3	3.7	0.6
Ambig.	(4)	1.8	0.7	2.1	0.8	1.4	1.1	1.6	0.9	2.1	1.0	2.2	1.1

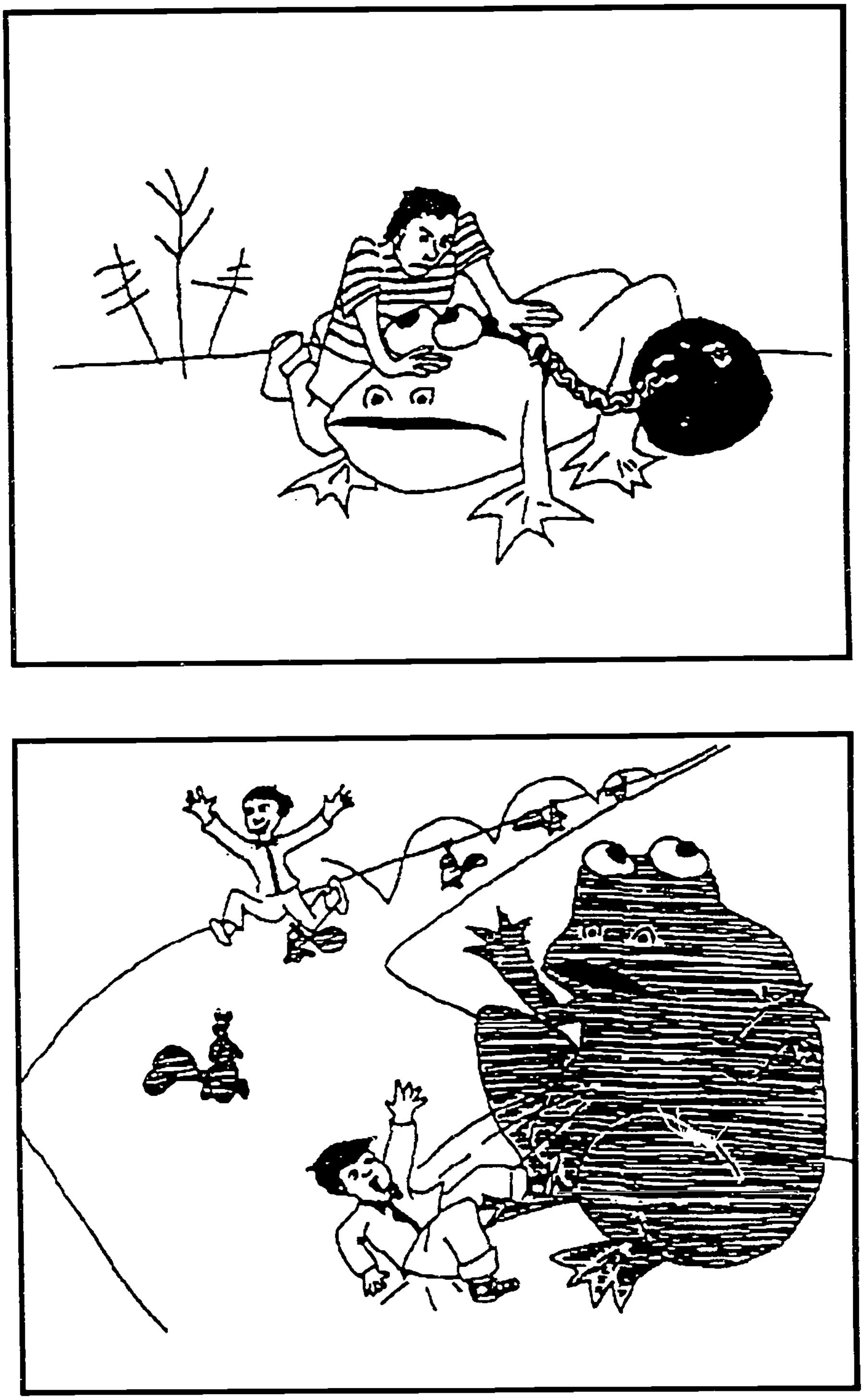


Figure 2. Picture pair for "The frog is impossible to jump."

Discussion

In sum, the results with both tasks support the conclusion that young poor readers are not delayed in acquiring the syntactic knowledge necessary to comprehend sentences containing object-control adjectives like *easy*. The good and poor readers demonstrated high levels of performance with these sentences in both the act-out and sentence-picture matching tasks and no reader group differences were found. Thus, we were unable to confirm the difference obtained by Byrne (1981).

In interpreting these results, we find it useful to distinguish between strong and weak versions of the syntactic delay hypothesis. A strong version would state that young poor readers treat *every* O-adjective as if it were an S-adjective. This version is clearly unsupported by the data. For example, only one poor reader acted out all eight O-adjective sentences incorrectly. Every other child produced at least 3 correct responses in acting out O-adjective sentences. A weak version of the syntactic delay hypothesis would propose that when the control properties of an adjective are unknown, the child reverts to a (default) subject-control interpretation. Even the weak version finds little support in the data. First, the children did not show a bias toward a subject-control over an object-control interpretation when presented with ambiguous adjectives like *dirty* in the sentence-picture matching task. Second, performance on S-adjective sentences in the act-out task was not perfect, which suggests that a default strategy was not followed. For example, many subjects treated the apparently unfamiliar adjective *eager* as if it had the control properties of an O-adjective. Finally, on only 12 occasions in the act-out task did a child respond incorrectly to both presentations of an O-adjective. In contrast, on 34 occasions a child responded incorrectly on one presentation of an O-adjective, but correctly on the other. This pattern of inconsistent responding is also incompatible with the default notion. Thus, the results of this study do not favor either a strong or a weak version of the syntactic delay hypothesis. Instead, they are consistent with a number of recent findings in the literature that argue for intact syntactic abilities in young poor readers in the face of deficiencies in retaining linguistic information.[2]

Our experience in this study highlights some of the practical difficulties that arise in experiments designed to assess syntactic comprehension in young children. First, results from the act-out task show that knowledge of specific lexical items can affect performance on a comprehension task. For example, many children erred in acting out the sentence containing the S-adjective *eager*, presumably because they were unfamiliar with the control properties of this particular adjective. Second, results from the sentence-picture matching task indicate the importance of evaluating the adequacy of the test materials used to assess comprehension. The children displayed poor performance on sentences containing the O-adjective *impossible* not because of failure to comprehend the meaning of the adjective, but because of difficulty in interpreting the pictures presented with this adjective. Thus, children's performance may reflect eccentricities of the particular task used to assess comprehension. This is underlined in comparisons of the children's performance across the two tasks employed in this study. Many of the children who produced errors on O-adjective sentences in the act-out task had no difficulty with the same structures in the sentence-picture matching task. For example, of the six children who had the most difficulty with O-adjective sentences on the act-out task (i.e., had an error rate of 50% or more), three of the children made no errors with O-adjective sentences on the sentence-picture matching task (excluding the two "impossible to jump" sentences).

To conclude, these findings suggest ways that both lexical knowledge and the pragmatic requirements of the experimental task may contribute to success or failure in tests of syntactic comprehension. They teach us to appreciate the need for special attention to controls in developing experimental tasks. It is especially important to ascertain that the subjects comprehend the key lexical items in the relevant contexts. Moreover, the pragmatic requirements of the assessment task need to be carefully scrutinized to insure that they do not create problems of their own.

REFERENCES

Bar-Shalom, E., Crain, S., & Shankweiler, D. (in press). A comparison of comprehension and production abilities of good and poor readers. *Applied Psycholinguistics*.

Byrne, B. (1981). Deficient syntactic control in poor readers: Is a weak phonetic memory code responsible? *Applied Psycholinguistics*, *2*, 201-212.

Chomsky, C. S. (1969). *The acquisition of syntax in children from 5 to 10*. Cambridge, MA: MIT Press.

Crain, S. (1987). On performability: Structure and process in language understanding. *Clinical Linguistics and Phonetics*, *1*, 1-18.

Crain, S., & Shankweiler, D. (1988). Syntactic complexity and reading acquisition. In A. Davidson & G. M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered*. Hillsdale, NJ: Erlbaum.

Crain, S., Shankweiler, D., Macaruso, P., & Bar-Shalom, E. (1990). Working memory and sentence comprehension: Investigations of children with reading disorder. In G. Vallar & T. Shallice (Eds.), *Neuropsychological impairments of short-term memory*. Cambridge, U. K.: Cambridge University Press.

Cromer, R. F. (1970). Children are nice to understand: Surface structure clues for the recovery of a deep structure. *British Journal of Psychology*, *61*, 397-408.

Dunn, L. M., & Dunn, L. M. (1981). *The Peabody Picture Vocabulary Test—Revised*. Circle Pines, MN: American Guidance Service.

Fletcher, J. M., Satz, P., & Scholes, R. J. (1981). Developmental changes in the linguistic performance correlates of reading achievement. *Brain and Language*, *13*, 78-90.

Goldsmith, S. (1980). *Psycholinguistic bases of reading disability: A study in sentence comprehension*. Doctoral dissertation, The City University of New York.

Hamburger, H., & Crain, S. (1984). Acquisition of cognitive compiling. *Cognition, 17*, 85-136.

Macaruso, P., Bar-Shalom, E., Crain, S., & Shankweiler, D. (1989). Comprehension of temporal terms by good and poor readers. *Language and Speech, 32*, 45-67.

Mann, V. A., Shankweiler, D., & Smith, S. T. (1984). The association between comprehension of spoken sentences and early reading ability: The role of phonetic representation. *Journal of Child Language, 11*, 627-643.

Richardson, E., & DiBenedetto, B. (1986). *Decoding Skills Test*. Parkton, MD: York Press.

Shankweiler, D., & Crain, S. (1986). Language mechanisms and reading disorders: A modular approach. *Cognition, 24*, 139-168.

Shankweiler, D., Crain, S., Gorrell, P., & Tuller, B. (1989). Reception of language in Broca's aphasia. *Language and Cognitive Processes, 4*, 1-33.

Smith, S. T., Macaruso, P., Shankweiler, D., & Crain, S. (1989). Syntactic comprehension in young poor readers. *Applied Psycholinguistics, 10*, 429-454.

Smith, S. T., Mann, V. A., & Shankweiler, D. (1986). Spoken sentence comprehension by good and poor readers: A study with the Token Test. *Cortex, 22*, 627-632.

Stein, C. L., Cairns, H. S., & Zurif, E. B. (1984). Sentence comprehension limitations related to syntactic deficits in reading-disabled children. *Applied Psycholinguistics, 5*, 305-322.

Whitehouse, C. C. (1983). Token Test performance by dyslexic adolescents. *Brain and Language, 18*, 224-235.

FOOTNOTES

[*] *Applied Psycholinguistics*, in press.

[†] Neurolinguistics Laboratory, Massachusetts General Hospital.

[‡] Also University of Connecticut, Storrs.

[†††] University of New England, New South Wales, Australia.

[1] Restrictions prohibiting intelligence testing in Australian schools made it infeasible for Byrne (1981) to obtain IQ scores on his subjects. In lieu of a formal test, teachers were asked to match the good and poor readers on numerical skills. As a further control, Byrne excluded from his analysis of performance on O-adjective sentences any subject who failed to score perfectly on the S-adjective sentences. Given this limited set of controls, however, we cannot rule out the possibility that the poor reader group in Byrne's original study contained children with below average intelligence and/or insufficient vocabulary knowledge.

[2] See, in particular, Bar-Shalom et al. (in press), Crain et al. (1990), Macaruso et al. (1989), and Smith et al. (1989).

APPENDIX 1

SENTENCES FOR ACT-OUT TASK

Practice Sentences

1. The squirrel follows the teddybear.
2. The teddybear kisses the squirrel.

Test Sentences

1. The squirrel is happy to follow.
2. The teddybear is easy to touch.
3. The squirrel is fun to bite.
4. The teddybear is hard to kiss.
5. The teddybear is eager to touch.
6. The squirrel is difficult to follow.
7. The teddybear is glad to kiss.
8. The teddybear is easy to bite.
9. The squirrel is fun to kiss.
10. The squirrel is willing to bite.
11. The squirrel is hard to follow.
12. The teddybear is difficult to touch.

APPENDIX 2

SENTENCES FOR SENTENCE-PICTURE MATCHING TASK

Practice Sentences

1. The pig is under the house.
2. The bunny is in the basket.

Test Sentences

1. The kangaroo is easy to reach.
2. The frog is impossible to jump.
3. The rabbit is hard to catch.
4. The porcupine is difficult to chase.
5. The bear is easy to reach.
6. The kangaroo is impossible to jump.
7. The boy is hard to catch.
8. The octopus is difficult to chase.

9. The man is too dirty to serve.
10. The policeman is too nice to shoot.
11. The monster is too nasty to help.
12. The lady is too old to teach.

13. It is easy to reach the kangaroo.
14. It is impossible to jump the frog.
15. It is hard to catch the boy.
16. It is difficult to chase the octopus.

A Review of Daniel Reisberg (Ed.), *Auditory Imagery**

Bruno H. Repp

This book appears to be the first in the psychological literature to be devoted entirely to the topic of auditory imagery. Research on visual imagery has been going on for some time, spearheaded by such authors as Roger Shepard, Stephen Kosslyn, and Ronald Finke. Characteristically, research on the analogous phenomenon in audition has lagged behind, and it is fair to say that even now it is not an area teeming with activity. The purpose of the present volume is evidently to stimulate interest in the topic, as well as to review whatever pertinent findings have been obtained so far. A perusal of the 10 chapters reveals that these findings are still very limited and leads one to wonder whether auditory imagery is going to be as fertile an area of investigation as visual imagery has proved to be.

The relative paucity of empirical data is compensated by the diversity of angles from which the topic is illuminated in this book. Three of the ten chapters deal with musical imagery, two with simple sounds, five with speech. Among the latter, there are discussions of inner speech in the deaf and of auditory hallucinations in schizophrenics. Most of the authors are well-established researchers, though not necessarily in areas primarily concerned with auditory imagery. While some were able to simply summarize their own research, others had the more difficult task of deriving implications for auditory imagery from their ideas and findings on related topics. All contributions, however, are well-written and interesting.

The articles on speech will perhaps be of greater interest to the readers of this journal (i.e., LANGUAGE AND SPEECH) than those on music and other nonspeech sounds. Nevertheless, the issues addressed in the music research are quite pertinent to speech also. Imagery is that facet of memory which retains or regenerates the analog characteristics of the original, modality-specific perceptual experience—it is "surface memory," as it were. Being such a general function, it is equally relevant to speech, music, and environmental sounds.

The order of the chapters is, in editor Reisberg's own words, "somewhat arbitrary," though he does mention the loose organizational principle he had in mind. My summary follows a different but not necessarily better order.

Margaret Jean Intons-Peterson (Chapter 3) acknowledges the theoretical debt of auditory imagery research to visual imagery research; according to her, there are no specific models of auditory imagery as yet, and the models borrowed from vision focus primarily on the relation of imagery to perception. She reviews some of the theoretical concepts as well as the results of several experiments, most of which employed simple stimuli and used reaction time as the dependent variable. In her own research of more than a decade ago, for example, she showed that the time to compare two imagined environmental sounds with respect to their loudness increases with the difference in loudness between them, as assessed by previously having subjects rate the typical loudnesses of these sounds on a scale. This suggests that loudness is a property that is represented in auditory images. Intons-Peterson reviews similar findings suggesting that pitch and timbre are represented literally in images.[1]

The evidence for timbre comes from the work of Robert Crowder who, with Mark Pitt (Chapter 2), reports original data that extend his earlier findings. Despite its narrow focus, this research is significant because it demonstrates, perhaps more convincingly than any other research reported so far, that subjects are able to generate specific auditory images from verbal instructions. Crowder and Pitt asked listeners to make same/different

I am grateful for the support of Haskins Laboratories during the preparation of this review.

judgments about the pitches of two successive tones. When these tones differed in timbre, subjects were slower in making "same pitch" judgments than when the tones had the same timbre as well as pitch. The crucial finding was that the same effect emerged when the first tone in each pair was a sine wave plus a verbal instruction to imagine a particular instrument timbre: Subjects responded faster and/or more accurately to same-pitch pairs when the second tone had the same timbre as the *imagined* timbre of the first tone. In their chapter, Crowder and Pitt report a replication of this finding for plucked versus bowed cello tones. In a subsequent attempt to separate the static and dynamic aspects of this timbre contrast, they used synthetic sounds differing in either spectrum or rise time. They obtained the desired effect with the former variation, but not with the latter. Their tentative conclusion was that dynamic cues to timbre are not represented in imagery, whereas static spectral properties are. Thus, besides providing an elegant experimental demonstration of imagery, these results also suggest that there are aspects of the original perceptual experience that cannot be recreated faithfully.[2]

Some of the evidence for the representation of pitch in auditory imagery comes from Andrea Halpern's research on imagery for songs, which she reviews in Chapter 1. Since songs essentially are a form of speech produced with a particular prescribed rhythmic and intonational pattern, the questions she asked are equally applicable to imagery for memorized stretches of nonmusical speech, such as poems or literary texts. (Psychological research has, unfortunately, neglected these more artistic forms of language use.) Halpern was interested in the temporal layout of imagined songs, and in whether their temporal extent matches that observed in actual singing. She used a probe task similar to one used in studies of the spatial layout of visual images: Subjects were given the initial word of a familiar song and had to decide whether a second word, presented shortly afterwards, occurred in the text of that song. Reaction times for correct responses increased monotonically with the distance between the first and second words in the song, whether or not subjects were instructed to imagine it. In another study, subjects had to make judgments about the relative pitch heights of two monosyllabic words in familiar songs. Again, reaction times increased with the distance between these words. Thus, the subjects (nonmusicians) seemed to scan through a temporal representation of the songs in their heads, as one should expect if auditory imagery is veridical. In further studies, Halpern showed that subjects' imagined tempo was comparable to their preferred tempo when listening to the same song.[3] Halpern also reports some data suggesting that listeners, regardless of musical training, have a long-term memory for the approximate starting pitch of familiar songs and are able to reproduce it by singing, choosing a tone on a keyboard, or giving ratings to presented pitches, though not nearly with the accuracy that possessors of "absolute pitch" might display.

The representation of pitch in musical images is discussed in much more detail by Timothy Hubbard and Keiko Stoeckig (Chapter 9). In fact, their treatment is perhaps somewhat too broad and abstract, lending a certain turgidity to their long chapter. They use the term "qualia" to refer to "a sensory quality or 'raw feel' that makes the experience of imaging similar to the experience of perceiving in a way that abstract representation is not" (p. 199). They also point out that the presence of these qualia has rarely been the focus of the work they review, which includes various models of the mental representation of music (psychoacoustic, rule-based, connectionist), studies of memory for isolated pitches and melodies, theories of representational form, the issue of cognitive penetrability, and questions of methodology. The chapter is valuable in that it provides a broad framework within which to view research on music imagery; however, it also raises the question (in my mind) of whether the subjective "qualia" are really all that important. The more significant question is perhaps how people use their memories and images of auditory properties to accomplish various tasks that are important in their lives. Hubbard and Stoeckig's discussion occasionally gives the impression that their primary aim is the pragmatic one of providing experimental psychologists with grist for their laboratory mills.

The final chapter on music (Chapter 10), by Diana Deutsch and John Pierce, clarifies what I mean by that comment. This unusual and decidedly iconoclastic contribution starts with a series of historical quotes that document the essential and unquestionable role of auditory imagery in composing. The historical survey goes on to bolster the authors' contention that scientists of earlier centuries were well informed about musical phenomena and usually took them into account when theorizing about human auditory capabilities, whereas in this century scientific reductionism has led to a musically uninformed tradition of laboratory psychoacoustics. Deutsch and Pierce argue forcefully against the exclusive focus on low-level explanations in much current auditory research and go so far as to suggest that "most [psychoacoustic] models are a distraction from thought and an inspiration to certitude for the uncertain" (p. 250). They go on to make another controversial point, namely that "good demonstrations are more convincing than experiments," and go on to cite several of these demonstrations, which predictably include Deutsch's well-known octave and scale illusions. These phenomena, however, are of a fairly simple nature, and their relevance to actual music experience is not always clear, as the authors themselves realize. At the end of their chapter, Deutsch and Pierce express the hope that modern music technology will lead to insights that traditional psychoacoustics has failed to provide. Although they come across as favoring the insights of "amateurs at work" (the title of one of their sections) over those of experimental psychologists, their bottom line is that better communication is needed between practicing musicians and ivory tower scientists. The connection with the specific theme of the book becomes very loose by the end of this provocative chapter.

The remaining five chapters deal with various manifestations of linguistic imagery, or "inner speech." Alan Baddeley and Robert Logie (Chapter 8) review some of Baddeley's well-known research on the "phonological loop" of working memory. They speculate that this component may "represent the seat of auditory imagery" (p. 180). Although its function in the temporary storage and maintenance of speech materials has been extensively investigated, the phenomenology of the accompanying imagery, if any, has not been studied. Baddeley and Logie feel that there is an urgent need for such studies. A related question that has been addressed is whether subvocal articulation is involved in phonological working memory. There is evidence that concurrent speech input, as well as speech produced by subjects themselves interferes with some tasks that rely on phonological short-term memory; other tasks, however, seem unaffected. Moreover, there are reports of neurological patients who have normal speech but show a phonological memory deficit, as well as of children who have never learned to speak but show relatively normal phonological function. These findings suggest that phonological processing may be relatively abstract and not necessarily connected with vivid auditory imagery.

This issue is pursued further by Ruth Campbell (Chapter 4) in her discussion of inner speech in deaf individuals. She reviews evidence that individuals without hearing can develop adequate phonological skills, including the ability to read in an alphabetic orthography. Campbell also cites her own, by now well-known finding that lipread words are retained in phonological form, just like auditorily presented speech. This suggests that phonological representations are not necessarily (perhaps even: necessarily not) auditory.

In Chapter 5, David Smith, Daniel Reisberg, and Meg Wilson distinguish between an "inner voice" and an "inner ear" that listens to the inner voice. They examined the effects of concurrent auditory input and/or concurrent articulatory activity on performance on several phonological tasks. In each case, both types of interference were effective, leading the authors to conclude that both processes were involved. (That the articulatory activity also produced auditory input in several instances is a complication that the authors do not discuss.) In a version of Crowder's timbre imaging task, it appeared that only auditory input interfered, so that only the inner ear, but not the inner voice, seemed to be involved—a reasonable conclusion. Some of the other evidence these authors discuss converges with that discussed by Baddeley and Logie (whose phonological store and loop indeed correspond to the inner ear and voice, respectively), and the existence of more abstract phonological processes is also acknowledged by reference to a "lexical ear."

A more critical stance is taken by Donald MacKay (Chapter 6), whose research on internal phonological processes spans more than two decades. With considerably more self-confidence than most other authors in this book, he states right at the outset that inner speech is nonarticulatory and nonauditory. The first claim agrees with the evidence for the abstractness of phonology discussed in other chapters. With regard to the latter claim, MacKay points out that "what seems phenomenally to be auditory often is not" (p. 126), thus revealing a major problem for any phenomenological approach to auditory imagery. MacKay points out (relying on introspection, it seems) that inner speech usually lacks loudness and fundamental frequency variation; he attributes the processing of these qualities to a separate "auditory concept system," distinct from the phonological system. Imagining a concrete voice or words spoken with a specific intonation requires both of these systems. MacKay is critical of some of Reisberg's and Baddeley's conclusions, particularly of the "inner ear" concept: "...the internal listener concept is functionally questionable: The 'double agent' approach to comprehension of internal speech must address the fundamental issue of why speakers must independently 'listen to' the meaning and sound of what they are saying internally when they know all along the meaning and sound of what they are saying" (p. 140). A number of additional related issues are discussed in this chapter, probably the most thought-provoking in the collection.

The last chapter to be mentioned is that by David Smith (Chapter 7) on the auditory hallucinations of schizophrenia. Halpern and MacKay both briefly alluded to the sometimes involuntary and persistent nature of auditory images (such as a "haunting" melody), a phenomenon that is found most dramatically in the illusory voices reported by schizophrenics, which often seem to come from within and seem to speak intelligibly. Smith, revitalizing a neglected theory of these hallucinations, argues that they are a form of inner speech contingent on subvocalization. He cites a variety of reports, some anecdotal, that engagement of the articulators results in a reduction of the vocal hallucinations. The evidence remains merely intriguing but points to a more specific hypothesis about the nature of these hallucinations.

In summary, this is a stimulating collection of articles on a relatively neglected topic. It demonstrates that there are a variety of activities in which auditory imagery plays a role, though some of the most obvious (musical composition and performance) are barely mentioned. In other cases, internal processes rather more abstract than images seem to be involved. The tasks employed in most of this research are fairly distant from real life, and it is not so clear whether further laboratory demonstrations of this kind will contribute any important insights about auditory imagery. It would be worthwhile, perhaps, to look more closely at auditory properties that *cannot* be imagined, such as hinted at by Crowder and Pitt, rather than at those than can. More generally, imagery is perhaps better viewed merely as one end of the memory continuum, the one dealing with analog information, rather than as a special phenomenon on the basis of its subjective "qualia." Its relative neglect in the past may be attributed to experimental psychologists' characteristic preoccupation with discrete, symbolic (and if analog, then visual) processes, and with activities that are technologically rather than culturally significant. I conclude by paraphrasing what I take to be one of Deutsch and Pierce's messages: If you want to learn about auditory imagery, look at what composers do.

REFERENCE

Crowder, R. G. (1973). Representation of speech sounds in precategorical acoustic storage. *Journal of Experimental Psychology, 98,* 14-24.

FOOTNOTES

*Hillsdale, NJ: Lawrence Erlbaum Associates, 1992. 274 pp. $45.00. This review appears in *Language and Speech, 35,* 341-346 (1992).

[1] Intons-Peterson also reports that the time needed to generate the image of a single sound does not depend on its loudness. However, it is not clear why it should; this point is also made by Hubbard and Stoeckig in Chapter 9 (p. 221, Footnote 2). The same comment applies to Intons-Peterson's analogous results for pitch comparisons. Her conclusion that loudness and pitch are sometimes not represented in images may not be justified.

[2] Although the authors do not elaborate on this point, the failure to find evidence for dynamic properties in auditory images is clearly reminiscent of Crowder's well-known finding that stop consonants are poorly retained in precategorical auditory memory (see, e.g., Crowder, 1973). It would be interesting to see a demonstration that stop consonants are likewise difficult to imagine.

[3] It should not be concluded from these data, however, that a memorized song (or stretch of speech) needs to be scanned from beginning to end in order to determine that some word or phrase occurs in it; surely, there must be multiple "access points" in longer structures such as Handel oratorios or plays by Shakespeare. Scanning such as demonstrated by Halpern may be mandatory, however, within small structural units such as sentences or clauses. The size of the units that must be exhaustively scanned would be a worthwhile topic for further investigation, as it bears on the memory representation of large-scale musical and linguistic structures.

Haskins Laboratories Status Report on Speech Research
1992, SR-111/112, 219-226

A Review of Mari Riess Jones and Susan Holleran (Eds.), *Cognitive Bases of Musical Communication**

Bruno H. Repp

This volume is the result of a recent (1988) initiative of the American Psychological Association (APA), the Scientific Conferences Program. The eleventh conference supported by this program, on "Cognitive Bases of Musical Communication," was held at The Ohio State University in April of 1990 and resulted in the present book. Not all of the conference papers have been included: There were 19 invited speakers, while there are only 15 chapters in the book, plus a brief introduction by the editors. Missing in particular is an invited lecture by Jean-Jacques Nattiez, which is only mentioned in the preface. Discussions are not included either. The chapters are brief and of fairly uniform length. Evidently, the editors had to follow strict guidelines from their sponsors. As indicated in the preface, one of the goals was to "bring some of the insights concerning communication via musical events into mainstream psychology" (p. xi). I take this to mean that the book was intended for a nonspecialist readership.

The 15 chapters are grouped into five sections: (1) Communication, Meaning, and Affect in Music; (2) The Influence of Structure on Musical Understanding; (3) Pitch and the Function of Tonality; (4) Acquisition and Representation of Musical Knowledge; (5) Communicating Interpretations Through Performance. Each section is preceded by a brief introduction. The group of authors includes psychologists as well as musicologists and philosophers. Two participants are from England; the others represent the cream of the crop of American cognitive (psycho)musicologists.

In their brief overview, **Mari Riess Jones** and **Susan Holleran** review some historical background and hint at how the influential but disparate theories of Heinrich Schenker and Leonard Meyer reverberate through the writings of some of the present authors. The editors also point out that, in studying the communicative function of music, the problem of multiple interpretations plays a central role.

This problem of indeterminacy is addressed head-on by philosopher **Robert Kraut**. He begins by citing Quine's (1960) controversial claim that even language is indeterminate in that every utterance can be assigned distinct (though closely related) meanings, more like different perspectives on the same event. He then raises the question of whether music can similarly be understood in different ways. Understanding music, he says, "is a matter of experiencing appropriate qualitative states in response to it" (p. 15). But what is the yardstick for appropriateness? Kraut's answer is that, in analogy to language, which is understood only by members of a reference population (viz., those who speak the language), proper understanding of music must be defined with reference to a special population of listeners. He variously defines this population as the one "which is *responsible for the musical event* in question" (p. 20, his italics) or (referring to Beethoven's works) as "Beethoven's

I am grateful to Mari Riess Jones and Patrick Shove for helpful comments on an earlier draft of this review, and to Haskins Laboratories for financial support.

sophisticated peers" (p. 17).[1] Nevertheless, he realizes that even within such a narrow population of specialists, there may still be room for different experiential responses. (In the extreme case, of course, the size of the reference population becomes 1, which is not at all unusual in discourse about music.) Kraut's bottom line is that, the smaller and the more homogeneous the reference population, the more determinate musical events appear to be (though this is by no means proven, except for n=1). He seems to consider the choice of a standpoint along this continuum a matter of personal preference.

This is a thought-provoking essay, though Kraut's oscillations of argument (probably relished by philosophers) made me quite dizzy. A serious problem of the discussion seems to me the abstractness with which the notion of understanding is treated. "Experiences of stability and tension, of metrical groupings, of tonal centers, of variations on harmonic, melodic, or rhythmic structures, and the like" (p. 15) is as far as Kraut gets in defining what musical understanding might actually entail. Thus there *is* a theoretical vocabulary in which musical experiences of a structural kind can be characterized fairly precisely. Moreover, there is a variety of psychological techniques available to assess these experiences indirectly in individuals who have no musical education and thus cannot describe what they perceive (see, e.g., Krumhansl, 1990). Once such techniques are applied, it becomes an empirical question whether and how often contrasting musical experiences can actually be observed. Kraut seems to assume (on logical rather than empirical grounds) that they are common. I find it more likely that the musical experiences of different listeners differ in degree of elaboration. The diversity of musical perception may be often overestimated; for example, there are many basic phenomena of auditory organization that ensure that listeners experience similar grouping structures when listening to music (cf. Bregman, 1990). The difference between the musical expert and the novice is likely to lie in the relative *richness* of the experience, in the ability to focus attention on different levels of detail, and often simply in the ability to verbalize and put technical terms to what is perceived. With regard to *emotional* experience, which Kraut dismisses early on as narrowly confined ("pending further discussion"—p. 12), the situation may be more egalitarian. Emotional experiences in response to music may be just as strong in the novice as in the expert, and they are also likely to be of the same kind, within broad limits (cf. Clynes, 1977, and Sloboda's article, discussed below). Moreover, the kind of unreflective response that can lead to a strong love and enthusiasm for music may constitute a form of musical understanding quite on a par with the more technical understanding evinced by musicologists (see Cook, 1990). Kraut makes no attempt to distinguish different forms of musical understanding; he treats it as if it were a single variable in some abstract logical calculus.

Continuing in a philosophical vein, **Diana Raffman** embarks on a (tentative) definition of musical semantics. With Lerdahl and Jackendoff's (1983) generative grammar as her starting point, she proceeds to argue that the grammar, "to have any explanatory force, must be motivated by an appeal to semantic considerations" (p. 24). She first illustrates this claim with reference to language, where the purpose of a grammar is to explain how language users understand what they hear (or, more often, what they read).[2] She argues that semantic context plays an important role in language understanding. The explanandum of music theory is said to be "the kaleidoscopic sequence of peculiarly musical feelings we experience on hearing a performance" (p. 28). These feelings are probably identical with the musical experiences Kraut referred to, but Raffman is barely more specific than her colleague in defining them. She argues that they may be analogous to contextual-semantic factors in language, and thus may be regarded as semantic themselves. Following Kraut's technique of immediately retreating once an advance has been made, she promptly casts doubt on her own proposal but concludes that, whatever these feelings are, they are what music theory is trying to explain. I understand this as paraphrasing the commonplace observation that music theorists rely on their own intuitions about music in formulating grammatical rules, just as linguists do in constructing their grammars.

I find myself on firmer ground with **John Sloboda's** empirical study of emotional responses to music. Here, at last, is a clear-cut and objectively measurable definition of what we are talking about: crying, shivering, accelerated heart beat. To be sure, these are extreme and correspondingly rare reactions, but they are important because of their salience in long-term memory and because they may motivate an individual's life-long occupation with music. Sloboda reviews several studies that required adults to provide a description of emotional responses to music heard recently or in the distant past, but the most interesting part is his recent attempt to link reports of the above-named physiological reactions to particular structural properties of the music. The fact that music can elicit such plainly observable responses at all is a fact worth contemplating (cf. Clynes, 1977).[3] Sloboda's data must be considered quite preliminary, but they constitute a powerful point of entry into the connection between sound and emotion. Incidentally, that connection does not seem to exhibit the indeterminacy that plagues cognitive analyses of musical structure.

The second part of the book has the curious title, "The Influence of Structure on Musical Understanding," which suggests that the object of understanding is not the structure itself. Indeed, the editors' introduction refers to "musical ideas" that need to be understood, but without specifying their nature. If ideas are not structural themselves, what are they? There is a lot of undefined vocabulary floating around in these initial discussions.

However, there is no such vagueness in **Ray Jackendoff's** article. He defines musical understanding as "the unconscious construction of abstract musical structures," as set forth in his influential book with Fred Lerdahl (Lerdahl & Jackendoff, 1983). Here Jackendoff begins to outline the form that a theory of real-time musical processing might take. He illustrates the gradual development of a structural representation in a hypothetical listener's mind with the help of a concrete musical example and then goes on to discuss how the "processor" (a term that strikes me as ugly and inhuman) might deal with indeterminacy. After considering a serial single-choice model (which continuously makes decisions and backtracks to correct mistakes) and a serial indeterministic model (which delays decisions until they can be made with virtual certainty), Jackendoff argues in favor of a parallel multiple-analysis model which entertains multiple structural hypotheses, even though only one structure may be available to consciousness at any one time. These ideas are analogous to those proposed in psycholinguistics to account for real-time language processing, and given the recent history of cognitive psychology the serial hypotheses seem moribund from the beginning. On the other hand, the parallel hypothesis may be too general to be refutable; it may be more of an appropriate mode of thought on the part of the investigator.

Jackendoff concludes by arguing that the musical processor is modular in Fodor's (1983) sense: It generates structures and expectations autonomously, so that it is unaffected (or indeed enriched) by the familiarity of a musical piece. This explains why surprising musical events retain their interest on repeated listening. Less appealing is Jackendoff's notion of "musical affect" which he ties to the generation of musical structures, among other things. Emotional responses to music tend to be far more differentiated and deserve a richer characterization (as, for example, attempted by Cooke, 1959). Nor is it clear that they have much to do with the structure building discussed here. Jackendoff's "affect" is analogous to a child's feelings while building a Lego construction. Apart from this reservation, this is an extremely lucid and instructive presentation which leads directly to some empirically testable predictions. It is too bad that its cash value, as it were, is reduced by the fact that an expanded and thus even more informative version has already been published *and* reprinted (Jackendoff, 1991, 1992).[4]

Eugene Narmour's subsequent essay covers some basic concepts of his theory of melodic structure, described at length in his book (Narmour, 1990), which appeared soon after the conference and has received much attention in the meantime. Therefore, not much needs to be said about this preview (or postview) in which Narmour illustrates how melodic implications at one level may be enhanced or contradicted by implications at a higher level. In addition to postulating initially a distinction between top-down and bottom-up expectations, which could come into conflict, Narmour sees another conflict-generating mechanism within the bottom-up level, arising from the hierarchical levels of musical structure. What is not clear to me is whether implications at higher levels can be assumed to be as "bottom-up" as those at the foreground level.

Temporal distance and hierarchical abstraction may well weaken the perceptual immediacy of the Gestalt laws on which bottom-up implications are thought to rest. Thus the higher-level implications may actually be top-down effects.[5]

In the final article in this section, **Mari Riess Jones** argues that music can be attended to in different ways that correspond to different time spans in the hierarchical structure (see also Jones & Boltz, 1989). Analytic attending (over short time spans or short "serial integration regions," essentially rhythmic groups) is said to be incompatible with more global ("future-oriented") attending. Because of this flexibility of attentional focus, Jones argues that there are different ways of mentally representing a piece of music. The validity of that conclusion, however, rests on what "attending" is really meant to be. If it just refers to what the listener is *conscious* of, subconscious processing and representation of other hierarchical levels is by no means ruled out (cf. Jackendoff above). Jones's theory may then simply refer to listeners' ability to focus at will on different levels in the structural hierarchy. In a complex structure, it may take time and experience to discover some of the higher levels, but it is difficult to see how the lowest level (the musical foreground) could ever not be represented in the listener's mind, regardless of attentional strategy. I have some difficulty with Jones's claim that attending determines structural representation; it seems to me that, on the contrary, attending to higher levels presupposes that a structural framework has been erected. Without such a framework, attention will remain at the level of primary musical events, which I take to be rhythmic groups or gestures. Jones offers an analogy with the visual inspection of an art object: The viewer may focus on detail by standing close or on global structure by standing back. In the auditory modality, however, there is no "standing back," except in a metaphorical sense. The information always enters at the same time scale, and higher-level regularities must be abstracted from the input.

Proceeding now to the third section on "Pitch and the Function of Tonality," we find a paper by **Diana Deutsch** on the tritone paradox. This work has been presented in several other places, including two recent articles in this journal (Deutsch, 1991; Deutsch, North, & Ray, 1990) and one in a popular science magazine (Deutsch, 1992), so even nonspecialists may have a feeling of *déja vu*. Of course, this does not diminish the importance of the research, which demonstrates striking individual differences in the perception of the relative pitch height of two Shepard tones forming the interval of a tritone. Not only has Deutsch shown that perception of these tones as rising or falling in pitch depends on their pitch class, but that the pattern of this dependency is quite different for listeners from California and from Southern England. For the Californian subjects, there appears to be a relationship to the range of fundamental frequency in speaking (Deutsch, 1991); corresponding evidence for the British subjects has yet to be presented. Deutsch's claim that speakers have a language- or dialect-based pitch template in their heads is provocative, but it stands on three legs only and is in urgent need of additional support.

In the following chapter, **Helen Brown** takes psychologists to task for simplistic approaches to the concept of tonality. Using stimuli from several published experiments as examples, she demonstrates that in many instances tone sequences classified a priori as "atonal" can be shown to have tonal interpretations when notated with enharmonic substitutions. Even more importantly, Brown demonstrates with musical examples and results from earlier experiments of her own that the temporal sequence of tones is a crucial determinant of tonal implications. Her valuable discussion underlines the fact that the psychology of music is an interdisciplinary enterprise which requires the musicologist's analytic acumen as well as the psychologist's methodological skills.

David Butler, Brown's occasional collaborator, continues in a similar vein by reporting some informal experiments which demonstrate that listeners can infer the tonality of a melodic excerpt in both major and minor modes. This is preceded by a discussion of interval frequencies in the two modes, which leads to the prediction that the tonal center might be more difficult to determine in the minor than in the major mode. Butler's observations seem to refute that idea, but the demonstrations are so limited as to reinforce the second half of the conclusion stated at the end of the preceding paragraph. A more extensive and better controlled study is called for to address the hypothesis in a rigorous manner.

The section concludes with **Fred Lerdahl's** "pitch-space journeys" through two Chopin preludes. Extending the theoretical apparatus developed in Lerdahl and Jackendoff (1983), he graphs the melodic/harmonic progression of one prelude as a path in "regional" (i.e., key), chordal/regional, or scale-degree space. Consideration of the unresolved harmonies of the second prelude leads to a "regional prolongational analysis," which represents implicit as well as explicit tonicities. This is sophisticated stuff and, in Lerdahl's own words, "an exercise in theoretical rather than experimental music psychology." It warrants careful study but may be a bit too advanced for the nonspecialist.

The next section, on "Acquisition and Representation of Musical Knowledge," opens with a chapter by **Carol Krumhansl** on "Internal Representations for Music Perception and Performance." Of all the authors in this book, she seems to have taken most seriously the assignment of writing for a general readership. The questions she addresses are very basic and important, but they are dealt with in summary fashion, evidently due to the space constraints. TAs a result, the answers provided are sometimes uncomfortably reductionistic; at other times, they seem too obvious to me. The conclusion that music cognition requires both iconic (surface) and symbolic internal representations is an example of the latter.[6] At the end of the chapter, Krumhansl reports some results from a study of musical memory (Krumhansl, 1991) which show that surface characteristics are retained after a single hearing.

The following two chapters both deal with connectionist computer models of tonal structure, but in quite different ways. **Jamshed Bharucha**'s brief summary of some basic features of his MUSACT model is lucid and readily understandable by the nonexpert. The model employs an "unsupervised" learning algorithm (i.e., without feedback) to construct chord and key representations from chordal inputs. It illustrates how a quasi-neural network can extract systematic relationships from structured input, without being taught these relationships explicitly. Of course, the relationships are implicit in the input, and the demonstration is perhaps less impressive when one thinks of the surface properties of the input as a manifestation of its underlying organization to begin with. That is, as long as the model only recovers the structure we already know, we learn more about the model than about the object of study.

Robert Gjerdingen's approach is more ambitious and reveals the musicologist behind it. He begins with a homage to Leonard Meyer and then discusses briefly some harmonic/melodic schemata in Mozart's music. In contrast to the sober precision of the preceding chapters, Gjerdingen's words dance on the page and entice the reader to join him in his metaphoric exploration of musical phenomena. How welcome these lively stylistic touches are in discourse about music! The innocent reader, having accepted Gjerdingen's invitation to the dance, is whirled through an increasingly complex succession of connectionist modelling efforts dealing with harmonic, rhythmic, and melodic schemata considerably more advanced than those considered by Bharucha. This is very interesting stuff, but difficult to follow in such a condensed presentation. Luckily, more detailed discussions are available elsewhere (e.g., Gjerdingen, 1990). Still, the glimpses provided here are sufficient to reveal a fundamental difference between Bharucha's and Gjerdingen's approaches: Whereas the former, true to his psychological background, is concerned primarily with modelling the knowledge of musically untrained listeners, Gjerdingen's aim is to show that neural networks can be as sophisticated and multifaceted as a master analyst such as Leonard Meyer. However, unless they can go beyond human ingenuity, the project will remain an exercise in artificial intelligence. And does human ingenuity *need* to be exceeded?

We come now to the final section, "Communicating Interpretations through Performance," containing two chapters. The first of these, by **Caroline Palmer**, is a slightly expanded version of a paper published previously (Palmer, 1989). Palmer is concerned with how pianists use expressive timing variations to convey different structural interpretations of the same score. This is an interesting and important issue. In a first example, taken from a Chopin prelude, the timing profile of an expressive performance is compared with that of a deliberately inexpressive performance.[7] The phrase-final lengthening evident in the former is absent in the latter, which suggests that it is part of the performer's strategy to emphasize structural groupings. Palmer's second example, the beginning of Brahms' Intermezzo in A major, op.118, no.2, is more problematic. Pianists were asked to indicate their "phrasing interpretation" by placing slurs in the unmarked score; a single pianist subsequently performed the excerpt according to two alternative markings, and performance analysis revealed timing variations that corresponded to the intended phrasings. Palmer concludes that the different phrasings conveyed different structural interpretations--different "phrase structures." However, the melodic/rhythmic grouping structure is quite unambiguous in this example, and it is difficult to imagine an alternative structural interpretation.[8] The different "phrasing interpretations" are merely different choices of surface articulation within the same underlying structure. One could imagine many alternative examples, however, where different phrasings do disambiguate underlying structural alternatives.

In the second half of her chapter, Palmer reports on a study of errors in piano performance. (It is not clear why the pianists committed so many errors to begin with.) The most frequent type of error, deletions (i.e., omission of a note), is shown to occur almost never at "phrase" boundaries, whereas perseverations (repetition of a note) are more frequent at boundaries than within groups. The definition of boundary location is not always clear, however, and relies in part on the "phrasing interpretation" manipulation described above. Therefore, the error pattern may well be more closely related to surface articulation patterns than to the underlying grouping structure. The distinction between these two semi-independent aspects needs to be addressed more carefully in future work.

In the final chapter, **L. Henry Shaffer** provides a preliminary but intriguing examination of four pianists' multiple performances of a virtually unknown piece by Beethoven. Global timing curves and dynamic traces at the bar level are provided. One pianist, who was provided with a score from which all expression marks had been deleted, produced more variable performances than the others, which suggests (not surprisingly) that expressive marks in the score are not redundant. Shaffer makes a valiant attempt to characterize the different performances but concludes that his descriptions "need to be superseded by a language better suited to the task of analyzing expression" (p. 277). That, of course, is a crucial problem. When Shaffer says that one pianist "misses a sense of wonder in the modulation and a feeling of uplift in the semitone-raised melody" (p. 275), he is talking about his subjective impressions rather than about well-defined objective correlates of these qualities. Nevertheless, Shaffer's descriptive analysis is insightful and instructive. It is embedded in a tentative theoretical framework, according to which music provides an abstract narrative: "...we can think of the musical structure as describing an implicit event, and the gestures of musical expression as corresponding to the emotional gestures of an implicit protagonist who witnesses or participates in the event. Thus, the performer's interpretation can be viewed as helping to define the character of the protagonist" (p. 265). Later he appropriately describes these ideas as having "some heuristic value in opening up the empirical study of expression" (p. 277). This is a promising alternative to musicologically tinged approaches, and I look forward to Shaffer's future explorations in this direction.

To sum up, this book has some strengths and some weaknesses. The weaknesses are in large part a consequence of forcing this group of excellent authors to write short presentations for a general audience, rather than expanding on their latest ideas in depth. As a result, the book holds little attraction for the specialist to whom much of the work described will be familiar from more detailed presentations elsewhere. Clearly, the book has more to offer to the general reader who wants to inform him/herself about what is going on in music psychology. Some of the contributions (Jackendoff, Deutsch, Brown, Bharucha) are admirably suited for that purpose. Others (Kraut, Narmour, Jones, Lerdahl, Gjerdingen) may tax the nonspecialist. Some of the contributions (Raffman, Butler, Palmer, Shaffer) have a very preliminary character, and at least one author (Krumhansl) seems to have bent over backwards to write in a tutorial manner, at the cost of originality. So, what the book ultimately adds up to is a collection of visiting cards from some of the best people in the field. At the very least, it may provide an incentive for nonspecialists or graduate students to delve more deeply into the literature. Through its mixture of authors from different backgrounds and their individual styles, it also illustrates the gap that still separates psychologists from musicologists and philosophers. Thus it reinforces the need for communication and cooperation across traditional disciplinary boundaries--an effort on which the further progress of psychomusicology will crucially depend and to which the conference at Ohio State undoubtedly contributed.

REFERENCES

Bregman, A. S. (1990). *Auditory scene analysis.* Cambridge, MA: MIT Press.

Clynes, M. (1977). *Sentics: The touch of emotions.* New York: Doubleday.

Cook, N. (1990). *Music, imagination, and culture.* Oxford, U.K.: Clarendon Press.

Cooke, D. (1959). *The language of music.* London, U.K.: Oxford University Press.

Deutsch, D. (1992). Paradoxes of musical pitch. *Scientific American, 267,* 88-95.

Deutsch, D. (1991). The tritone paradox: An influence of language on music perception. *Music Perception, 8,* 335-347.

Deutsch, D., North, T., & Ray, L. (1990). The tritone paradox: correlate with the listener's vocal range for speech. *Music Perception, 7,* 371-384.

Farnetani, E., Torsello, C. T., & Cosi, P. (1988). English compound versus non-compound noun phrases in discourse: an acoustic and perceptual study. *Language and Speech, 31,* 157-180.

Fodor, J. A. (1983). *Modularity of mind.* Cambridge, MA: MIT Press.

Gabrielsson, A. (1987). Once again: The theme from Mozart's Piano Sonata in A major (K.331). A comparison of five performances. In A. Gabrielsson (Ed.), *Action and perception in rhythm and music* (pp. 81- 103). Stockholm, Sweden: Royal Swedish Academy of Music (Publication No.55).

Gjerdingen, R. O. (1990). Categorization of musical patterns by self-organizing neuronlike networks. *Music Perception, 7,* 339-370.

Jackendoff, R. (1991). Musical parsing and musical affect. *Music Perception, 8,* 199-230.

Jackendoff, R. (1992). Chapter 7 in *Languages of the mind: Essays on mental representation.* Cambridge, MA: MIT Press.

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review, 96,* 459-491.

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch.* New York: Oxford University Press.

Krumhansl, C. L. (1981). Memory for musical surface. *Memory & Cognition, 19,* 401-411.

Krumhansl, C. L., & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 579-594.

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music.* Cambridge, MA: MIT Press.

Linell, P. (1982).*The written language bias in linguistics.* Linköping, Sweden: University of Linköping.

Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model* Chicago, IL: University of Chicago Press.

Palmer, C. (1989). Structural representations of music performance. In *Proceedings of the Cognitive Science Society* (pp. 349-356). Hillsdale, NJ: Erlbaum.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America, 90,* 2956-2970.

Quine, W. V. (1960). *Word and object.* Cambridge, MA: MIT Press.

FOOTNOTES

*Washington, DC: American Psychological Association, 1992. 284+xiv pp. $40. This review will appear in *Music Perception.*

[1] Both definitions are problematic; the former (which I understand to refer to contemporary performers) because it neglects the fact that musical norms and experiences change continuously throughout history, the latter (which I understand as referring to musicians contemporary with Beethoven) because it refers to an extinct population that, moreover, probably had only a dim appreciation of Beethoven's path-breaking achievements. A more promising definition might have made use of the criterion of production competence: Just as a speaker of a language is identified by his/her ability to converse in that language, it might be stipulated that competence in a musical style (such as Beethoven's) is evidenced by an ability to improvise, compose, or at least play competently in it. This might disqualify some competent listeners, but it would hardly misclassify a novice.

[2] Raffman seems to exhibit a "written language bias" (Linell, 1982) here. In discussing the famous example, "They are frying chickens," she fails to consider that the apparent ambiguity of the written version is commonly resolved by prosody in spoken language (cf. Farnetani, Cosello, & Cosi, 1988; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991) and not just by pragmatic context. An instance of the analogous "written music bias" occurs when she says that the input to the musical grammar is a "mental score," when in fact it is a stream of sound organized by a performer.

[3] Not long ago I found myself repeatedly moved to tears while playing the central variation of Guy Ropartz's *Ouverture, Variations et Final* (a very rewarding piano composition in the style of César Franck), a section full of appoggiaturas and harmonic sequences—exactly the factors that Sloboda's informants reported to be associated with tearful experiences!

[4] I am a little baffled by this cloning of publications. Don't we have enough reading material already? However, I was glad to notice that "processor" has been changed to "parser" in the later version--better, but still a machine-in-the-mind metaphor. I prefer to say that *listeners* (or their brains) do the parsing.

[5] At the end of his paper, Narmour comes close to actually suggesting an experiment to test his theory. This openness to empirical approaches is remarkable in a musicologist (as it is in Jackendoff, a linguist and philosopher) and bespeaks the interdisciplinary cross-fertilization that the cognitive science movement has fostered. However, Narmour should have acknowledged Krumhansl and Shepard (1979) as the inventors of the probe tone method he proposes.

[6] Personally, I feel very uneasy with the notion of "internal representation" which, as Krumhansl says, is so fundamental to cognitive science approaches. Internal representations always end up being something that actually characterizes the perceptual object, and I prefer to think of these "representations" as 'presentations," i.e., as objective properties rather than as internal states. I cannot share the feeling of discovery that many cognitive scientists seem to have when they find that internal representations mirror external reality. They always do.

[7] The timing patterns are shown as deviations from a hypothetical mechanical performance, plotted as a straight line, a convention

that goes back to Gabrielsson (e.g., 1987). However, this reference seems both arbitrary and superfluous to me; it is more informative to plot the raw durations.

[8] In one version, the "phrase" ends with a sixteenth-note following a dotted eighth-note, whereas the next "phrase" begins with a long note on the downbeat. In such a sequence, the sixteenth-note always functions as an anacrusis, and the dotted eighth-note most likely also. A non-legato connection is quite commendable, however, to emphasize the onset of the note on the downbeat. Slurs that cross group boundaries and/or terminate within rhythmic groups are frequently found in musical scores; they almost always have implications for articulation, but they affect the structural interpretation only if there is structural ambiguity.

Diversity and Commonality in Music Performance: An Analysis of Timing Microstructure in Schumann's "Träumerei"*

Bruno H. Repp

This study attempts to characterize the temporal commonalities and differences among distinguished pianists' interpretations of a well-known piece, Robert Schumann's "Träumerei." Intertone onset intervals (IOIs) were measured in 28 recorded performances. These data were subjected to a variety of statistical analyses, including principal components analysis of longer stretches of music and curve fitting to series of IOIs within brief melodic gestures. Global timing patterns reflected the hierarchical grouping structure of the composition, with pronounced *ritardandi* at the ends of major sections and frequent expressive lengthening of accented tones within melodic gestures. Analysis of local timing patterns, particularly of within-gesture *ritardandi*, revealed that they often followed a parabolic timing function. The major variation in these patterns can be modelled by families of parabolas with a single degree of freedom. The grouping structure, which prescribes the location of major tempo changes, and the parabolic timing function, which represents a natural manner of executing such changes, seem to be the two major constraints under which pianists are operating. Within these constraints, there is room for much individual variation, and there are always exceptions to the rules. The striking individuality of two legendary pianists, Alfred Cortot and Vladimir Horowitz, is objectively demonstrated here, as is the relative eccentricity of several other artists.

INTRODUCTION

A. Diversity and commonality in music performance

More than at any earlier period in musical history, the contemporary scene in serious music is dominated by the performer (see Lipman, 1990). Music consumers thrive on a limited repertoire of standard masterworks, primarily from the 19th century, that are offered again and again in different performances, both in live concerts and on recordings. The great musical events of our time are not the premieres of new compositions, but the appearances and reappearances of superstar conductors and instrumentalists. Young musicians compete for career opportunities by entering competitions in which their performances are compared and evaluated by juries (see Cline, 1985; Horowitz, 1990). While ever new renditions of the standard repertoire vie for the attention of record buyers and concert audiences, a spate of reissues of historical recordings on CDs is offering stiff competition. For some of the more popular works, the Schwann catalogue lists dozens of performances; if deleted records, available in libraries and private collections, are counted, they may run into the hundreds. There have never been such ample opportunities to compare different performances of the same music.

Obviously, this remarkable diversity reflects not only clever marketing and the promotion of superstars, but also the ability of music lovers to distinguish and appreciate different performances. The more sophisticated among these listeners detect qualities in the performances of individual artists that lead them to search out these performers' concerts and recordings. They may voice

This research was made possible by the generosity of Haskins Laboratories (Carol A. Fowler, President). Additional support came from NIH (BRSG) grant RR05596 to the Laboratories. I am grateful to the Record Library, School of Music, Yale University, for providing most of the recordings analyzed here, and to its manager, Karl Schrom, for his help. Thanks also to Pat Shove for many stimulating discussions.

opinions about the quality of particular performances by these and other artists, describing them as "brilliant" or "noble" or "thoughtful." Some of the more gifted professional critics excel in characterizing different performances in terms alternatingly scholarly and poetic.

While the attention of listeners and critics thus is mainly drawn to the *differences* among performances, there are also strong *commonalities*, usually taken for granted and hence unnoticed. Though there is a large variety of acceptable performances of a given piece of music, there is an even larger variety of unacceptable performances, which rarely make their way into the concert halls or onto records. Unless they have the mark of inspired iconoclasm (as do some of the performances by the late Glenn Gould; see, e.g., Lipman, 1984), they quickly succumb to the fierce competition of the musical marketplace. Music teachers, however, have to deal with them every day and try their best to mold immature and wayward students into performers that can be listened to with pleasure. Though teachers may differ considerably in their methods and goals, and are rarely very explicit about what these are, they are transmitting the unwritten rules (though see Lussy, 1882) of a performance tradition that goes back to 19th century central Europe, where most of the standard repertoire originated. Despite various changes in performance practices during the last 200 years, most of them of a narrowly technical nature, there are generally accepted *norms* of musical performance, according to which the artist's actions are largely subordinated to the musical structure. The artist's primary task is the *expression* of the musical structure, so it can be grasped and appreciated by the listener, and make an *impression* on him or her. (See Lussy, 1882; Riemann, 1884; Stein, 1962.) This is presumably done by conventional means that are adapted to the hearers' perceptual and cognitive abilities. However, the particular doses in which these techniques are applied (unconsciously, for the most part) vary from artist to artist and account for individual differences in interpretation.

Thus there are two basic aspects of music performance: a *normative* aspect (i.e., commonality) that represents what is expected from a competent performer and is largely shared by different artists, and an *individual* aspect (i.e., diversity) that differentiates performers. The individual aspect may be conceived of as deviations from a single ideal norm; more profitably, however, it may be thought of as individual settings of free parameters in the definition of the normative behavior. That way, it is possible for different artists to meet the norm equally and yet be discriminably different. For example, two pianists may be equally adept in expressing a particular musical structure in their performances, but one may choose a slow tempo and large tempo changes, whereas the other may prefer a faster tempo and smaller deviations from the rhythmic beat. The former artist may then be characterized as "romantic" or "exuberant," while the latter will evoke epithets such as "restrained" and "noble," though both may receive equal acclaim from a sensitive audience.

The commonality-diversity distinction seems obvious enough, but there is little tangible evidence to substantiate it. Although volumes have been written about different performers and their characteristics, these discussions rarely go beyond generalities, and the vocabulary used (such as the adjectives quoted above) are not specifically linked to particular performance properties; they also may be used differently by different writers. There is no consistent terminology, nor a well-developed and generally accepted theory of performance description and evaluation, that would make possible an objective characterization of performance commonalities and differences. Music criticism is an art rather than a science, and the critic's impressions, accurate as they may be, are filtered through an idiosyncratic web of personal experiences, expectations, preferences, and semantic associations. In fact, the belief is widespread that performance differences *cannot* be characterized objectively.

Such a negative conclusion, however, can only be justified if objective performance analysis has been attempted and has failed. A total failure is highly unlikely, however, for there are many physical properties of a musical performance that, without any question, *can* be measured objectively. Whether objective analysis can capture a performance *exhaustively* may remain in doubt; the proper question is how much can be learned from it. Surely, even a partial characterization in terms of verifiable and replicable observations can make a valuable contribution to our understanding of performance commonalities and differences, which remain mostly a mystery to this day. For example, most music lovers would agree that Artur Rubinstein and Vladimir Horowitz were two very different pianists, whose performances of similar repertoire (e.g., Chopin) are instantly distinguishable. But what is it, really, that makes them so different and individual? And do they have anything in common at all? It is all too easy to couch the answers to such questions in terms of "artistic personality," which do not enlighten us at all about the nature of the differences and commonalities. Objective performance analysis has a contribution to make here.

B. Objective performance analysis

The concept of objective performance analysis goes back to Carl Seashore (1936, 1938, 1947) and his collaborators, who pioneered the use of acoustic analysis techniques to derive "performance scores" that show the exact variations of pitch, timing, and intensity produced by an artist on some instrument. A considerable amount of data was collected in Seashore's laboratory, but their analyses remained rudimentary and focused primarily on technical aspects such as pitch accuracy and *vibrato* in singing and string playing, and chord synchrony on the piano. Although some studies compared different performances of the same music, no statistical characterization of commonalities and differences was attempted, nor were the results interpreted with more than a passing reference to musical structure. These limitations reflect the behavioristic approach of American psychology at the time, as well as the unavailability of psychological performance models and advanced statistical methods. Nevertheless, Seashore (1947, p. 77) was able to reach conclusions that reinforce the premises of the present study:

> ...there is a common stock of principles which competent artists tend to observe;... We should not, of course, assume that there is only one way of phrasing a given selection, but, even with such freedom, two artists will reveal many common principles of artistic deviation. Furthermore, insofar as there are consistent differences in their phrasing, these differences may reveal elements of musical individuality.

Contemporary with Seashore's work, similar research was going on in Germany. Hartmann (1932) compared the timing patterns of two famous pianists' performances of a Beethoven sonata movement and provided detailed numerical descriptions of the differences between them. The small size of the sample, however, limits the generality of his conclusions. Although one of the two artists seemed quite eccentric, this impression cannot be substantiated without more extensive comparisons.

The three decades between roughly 1940 and 1970 were barren years for the objective study of music performance. In the last two decades, however, several researchers have taken up the topic again. Their work, with few exceptions, has taken a case study approach; their focus was not on individual differences but on the instantiation of certain principles in (hopefully, representative) performance samples, mostly from pianists. Thus, Shaffer (1980, 1981, 1984, 1989; Shaffer, Clarke, & Todd, 1985) examined the timing of piano performances from the perspective of motor programming and control (see also Povel, 1977). His erstwhile students, Eric Clarke and Neil Todd, went on to make significant contributions of their own. Clarke (1982, 1985) conducted case studies of piano performances of Satie's music and wrote at length on how musical structures are expressed in timing variations (Clarke, 1983, 1988). Todd (1985, 1989; Shaffer & Todd, 1987) developed a computational model of timing at the phrase level, which has been revised and extended to dynamics in his most recent publications (Todd, 1992a, 1992b). Bengtsson and Gabrielsson (1980; Gabrielsson, 1974; Gabrielsson, Bengtsson, & Gabrielsson, 1983) conducted extensive studies of the performance of different rhythm patterns in relatively simple musical contexts, but with attention to individual differences. Gabrielsson (1985, 1988) has written more generally about timing in music performance. A provocative theory of "composer's pulse" in performance timing has been developed by Clynes (1983, 1987).

There are few studies in the literature that employed what one might consider a representative sample of different performances. Gabrielsson (1987) conducted a detailed comparison of five pianists' performances of the first eight measures of Mozart's Piano Sonata in A major, K. 331; the study included measurements of timing, intensities, and articulation (i.e., tone durations). Palmer (1989) studied the same music as performed by six pianists in "musical" and "unmusical" styles, and she analyzed the timing patterns, note asynchronies, and articulation. Palmer also recorded eight pianists playing the first 16 measures of a Brahms Intermezzo and studied timing patterns in relation to intended phrasing. She concluded that "pianists share a common set of expressive timing methods for translating musical intentions into sounded performance" (p. 345). These studies still used relatively small samples and obtained only limited amounts of data from each performer.

The largest amount of performance data was analyzed in Repp's (1990) study, which included 19 complete performances by famous pianists of a Beethoven sonata movement. The analysis was limited, however, in that it concerned primarily timing patterns at the level of quarter-note beats. Also, the focus of the study was a search for Clynes' (1983) elusive "Beethoven pulse" in the timing patterns. Following the example of Bengtsson and Gabrielsson (1980), Repp applied principal components (factor) analysis to these timing data, to determine how many independent timing patterns were instantiated in this sizeable sample of expert performances. Two factors emerged, the first representing primarily phrase-final lengthening, while the second factor captured other types of expressive timing variation. Musical listeners' evaluations of the performances were also obtained, and some relationships between the measured timing patterns and listeners' judgments were found.

The present study continued the general approach taken by Repp (1990), but without the aim of testing Clynes's theory of composer's pulse. Its main purpose was to assemble a large sample of performances of a particular composition by outstanding artists, and to analyze the timing patterns in detail, using various statistical methods. These methods, it was hoped, would make it possible to separate commonalities from differences. The common patterns would reveal how most pianists transmit musical structure and expression through timing variations, and it was of interest to determine whether there would be a single common factor or several. The characterization of individual differences was also of interest, especially with respect to some of the legendary pianists in the sample, who are known for their individuality.

The music chosen for this investigation was a well-known piano piece from the Romantic period, "Träumerei" by Robert Schumann. It was selected because it is a highly expressive piece that permits much freedom in performance parameters, and hence much room for individual differences in interpretation. Also, there are numerous recordings available.

I. THE MUSIC

"Träumerei" ("Rêverie," "Dreaming") is the seventh of the 13 short pieces that constitute Robert Schumann's (1810-1856) "Kinderszenen" ("Scenes from Childhood"), op. 15. This little suite, universally considered one of the masterpieces in its genre, was composed by Schumann in 1838 when he was secretly engaged to Clara Wieck. The pieces were selected from some 30 pieces composed for Clara around that time, and their titles may have been added as an afterthought. They are not intended for children but rather reflect an adult's recollection of childhood (Brendel, 1981; Chissell, 1987).

"Träumerei" occupies a central position in the "Kinderszenen" suite, not only by its location but by its duration and structural role. It serves as a resting and turning point in the cycle, which shows so many intricate thematic connections that it may be considered a set of free variations (Reti, 1951; Traub, 1981). Its key signature, meter, and motivic content also single it out as the hub of the suite (Brendel, 1981; Traub, 1981). However, it is also often performed by itself, both by classical artists (e.g., it was one of Horowitz's favorite encores) and in numerous popular versions and arrangements. Indeed, "Träumerei" is perhaps *the* most popular Romantic piano piece. Its score is shown in Figure 1.

The melodic/rhythmic structure (or grouping structure; see Lerdahl & Jackendoff, 1983) of "Träumerei" is depicted schematically in Figure 2, which also introduces terminology to be used throughout this paper. The layout of the figure corresponds to that of the score in Figure 1, and bars (measures) are numbered. The piece is composed of three 8-bar periods (A, B, A´), the first of which is obligatorily repeated. Each period is subdivided into two 4-bar phrases, which are represented by staff systems in Figure 1 and by large rectangular boxes in Figure 2. (Actually, the beginning and end of each phrase extend slightly beyond the four bars, overlapping with the preceding and following phrases, respectively.) There are two phrase types, a and b, each of which recurs three times with slight variations (indicated by subscripts in Figure 2). Phrase a1 (period A) is repeated literally in period A´. Phrases b2 and b3, which constitute period B, are structurally identical but differ in key, harmony, and some other details.

The melodic events are divided horizontally (roughly, along the dimension of relative pitch) among four registers or voices (S = soprano, A = alto, T = tenor, B = bass). Vertically (along the dimension of metrical distance), the events within phrases are grouped into *melodic gestures*, which are represented by filled boxes in Figure 2. Characteristically, they extend across bar lines (vertical solid and dashed lines in Figure 2). A melodic gesture (MG) is an expressive unit composed of at least two and rarely more than seven successive tones. It is defined here to begin with the onset of its first tone and to end with the onset of its last tone. This seems reasonable, for a pianist cannot influence the time course or dynamics of the last tone once the key has been struck.

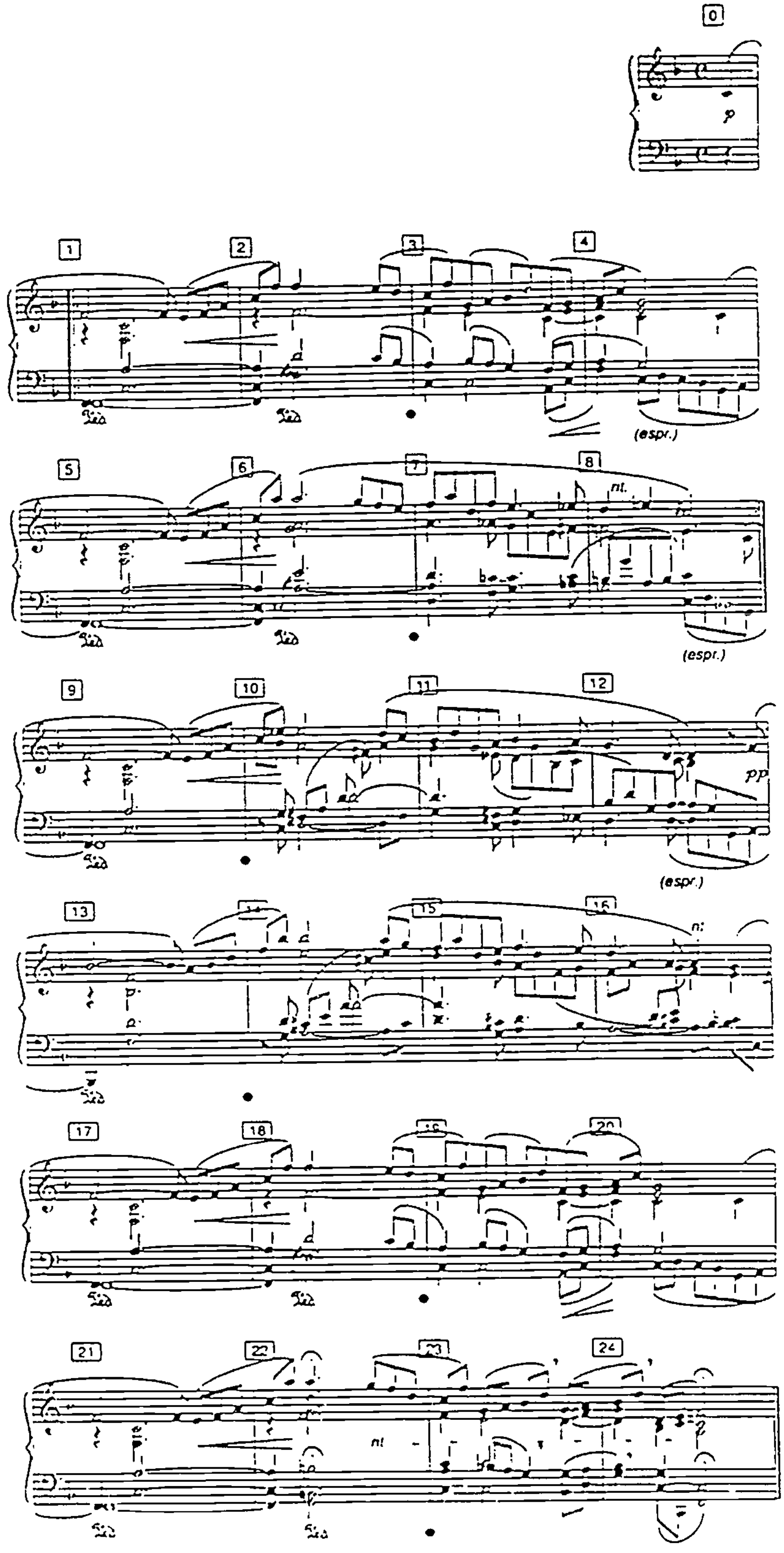


Figure 1. The piano score of "Träumerei," created with MusicProse software following the Clara Schumann (Breitkopf & Härtel) edition (with some deviations in minor details due to software limitations). The layout of the score on the page is intended to highlight the structure of the music.

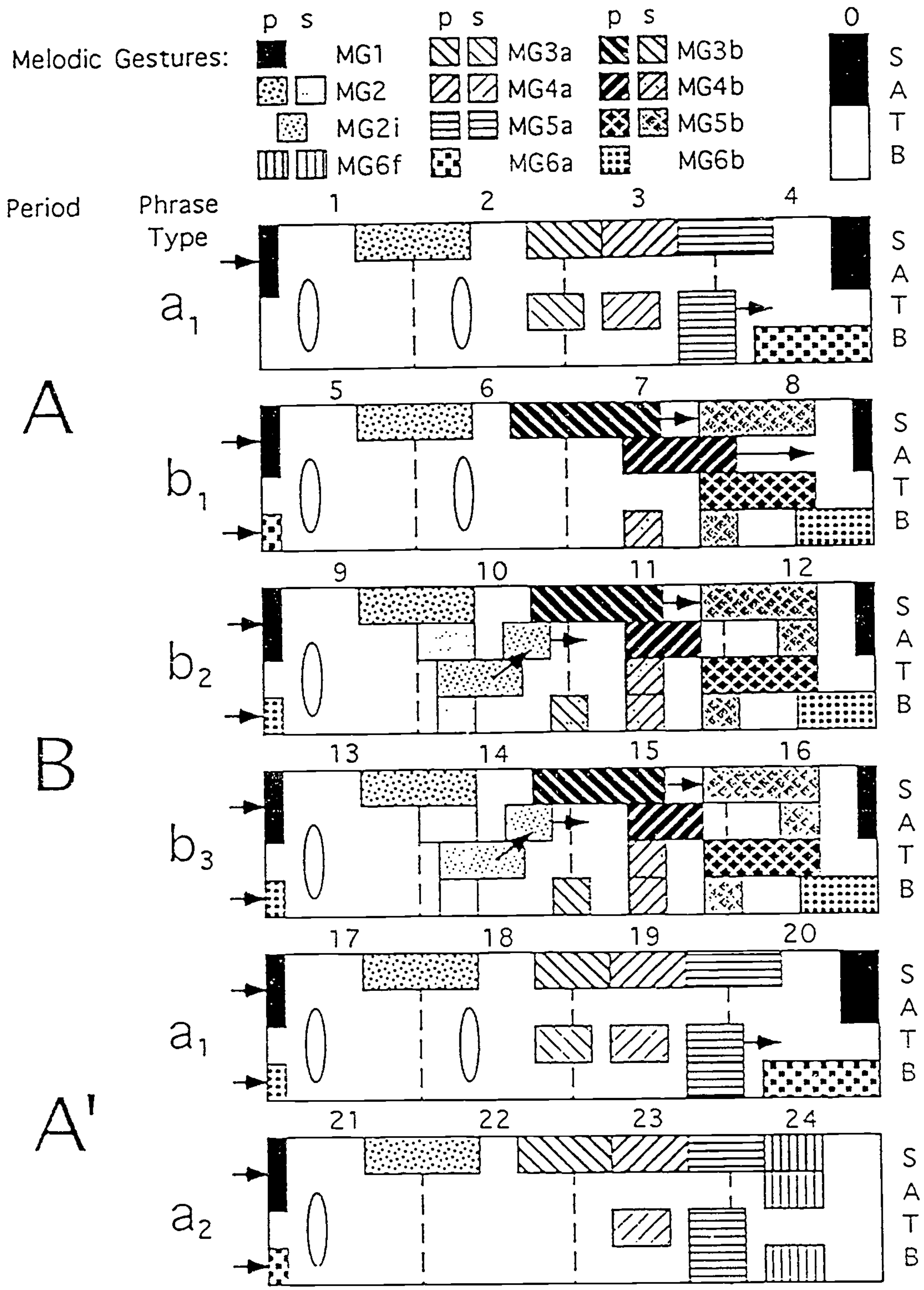


Figure 2. Schematic representation of the melodic/rhythmic structure of "Träumerei." (See text for explanation.)

(In Figure 2, each MG box extends one eighth-note space beyond the metrical onset of the last tone.) Blank spaces represent time spans devoid of MGs; they may contain single tones, sustained tones, or rests. Multivoiced chords having some gestural quality are represented by vertical ellipses in Figure 2. Arrows indicate continuity of a MG across a line break or with a subsequent melodic event. Thus, MG1 and MG6 continue from the end of one line to the beginning of the next, and MG3b and MG5b in the soprano voice are closely linked. When an arrow points into blank space, it points to the onset of a single-tone event that coheres with the MG.

MGs are divided into primary (p) and secondary (s) ones. The latter are usually shorter and accompany primary MGs; they are represented by boxes filled in lighter shades. An exception to this classification is MG2i, a delayed imitation of MG2 divided between the tenor and alto voices (bars 10 and 14). In the following timing analyses, we will essentially be concerned only with the primary MGs, which represent the leading voice(s) in the polyphonic quartet. As can be seen in Figure 2, in phrases of Type a the primary MGs (including the final MG6f in bar 24) are all in the soprano, except for MG6a which is in the bass and overlaps both MG5a and MG1. In Type b phrases, during the second half, the primary MGs cascade down through the four voices, overlapping each other.

II. THE PERFORMANCES

The sample of performances analyzed here includes 28 performances by 24 outstanding pianists, selected according to ready availability (see Table 1). Two artists, Cortot and Horowitz, are represented by three different recordings each.

Table 1. *The artists and their recordings.*

Code	Artist	Recording
ARG	Martha Argerich (1941-)	DG 410 653-2 (CD) [1983]
ARR	Claudio Arrau (1903-1991)	Philips 420-871-2 (CD) [1974]
ASH	Vladimir Ashkenazy (1937-)	London 421 290-2 (CD) [1987]
BRE	Alfred Brendel (1931-)	Philips 9500 964 [1980]
BUN	Stanislav Bunin[a]	DG 427 315-2 (CD) [1988]
CAP	Sylvia Capova[a]	Stradivari SMC-6020 (C) [T] [1987]
CO1	Alfred Cortot (1877-1962)	EMI 3C 153-53793M [1935]
CO2	Alfred Cortot	EMI 3C 153-53794M [1947]
CO3	Alfred Cortot	EMI 3C 153-53795M [1953]
CUR	Clifford Curzon (1907-1982)	London LL-1009 [~1955]
DAV	Fanny Davies (1861-1934)	Pearl GEMM CD 9291 (CD) [1929][b]
DEM	Jorg Demus (1928-)	MHS OR 400 [~1960s]
ESC	Christoph Eschenbach (1940-)	DG 2535 224 [1966]
GIA	Reine Gianoli (1915-1979)	Ades 13.243-2 (CD) [1974]
HO1	Vladimir Horowitz (1904-1989)	RCA LD-7021 [T] [1947]
HO2	Vladimir Horowitz	Columbia MS 6411 [1963]
HO3	Vladimir Horowitz	Columbia MS 6765 [1965] (live) [T]
KAT	Cyprien Katsaris (1951-)	Telefunken 6.42479 AP [1980](live)[T]
KLI	Walter Klien (1928-1991)	Allegretto ACS 8023 (C) [T][c]
KRU	André Krust[a]	MHS 1009 (orig. Erato) [~1960s]
KUB	Antonin Kubalek (1935-)	Dorian DOR-90116 (CD) [1988]
MOI	Benno Moiseiwitsch (1890-1963)	Decca DL 710.048 [~1950s]
NEY	Elly Ney (1882-1968)	Electrola WDLP 561 [~1935]
NOV	Guiomar Novaes (1895-1979)	Vox PL 11.160 (orig. PL8540) [1954]
ORT	Cristina Ortiz (1950-)	MCA Classics MCAC-25234 (C) [1988]
SCH	Artur Schnabel (1882-1951)	Pathé COLH 85 [1947]
SHE	Howard Shelley[a]	CHAN 8814 (CD) [1990]
ZAK	Yakov Zak (1913-1976)	Monitor MC 2039 [~1960]

[a] birthdate not known

[b] originally Pearl CLA 1000

[c] date unknown

Abbreviations: CD = compact disc; C = cassette; < = date of liner notes (recording date probably the same or preceding year). ~ = estimated date; T = "Träumerei" only

Most of the recordings (17) are on long-playing records, 3 are on cassettes, and 8 are on CDs (including a transfer of a 1929 recording by Fanny Davies, a one-time student of Clara Schumann, from a very scratchy original). Most of the performances are of the complete "Kinderszenen," but five are of "Träumerei" only. Two of the latter are from live concerts; all others are studio recordings. Table 1 includes actual or estimated recording dates.

The 24 artists include some of the most renowned pianists of this century as well as some less well-known artists. They can be grouped according to gender (6 female, 18 male), country of origin (5 from Russia; 4 each from Austria and France; 3 from England; 2 each from Germany and Brazil; one each from Czechoslovakia, Argentina, and Chile; one unknown), and approximate age at the time of recording (about equal numbers of young, middle-aged, and old). At least 12 pianists are no longer alive.

III. ANALYSIS METHODS

A. Measurement procedure

All tone interonset intervals (IOIs) were hand-measured by the author using a waveform editing program. Each performance was low-pass filtered at 4.9 kHz and digitized at a 10 kHz sampling rate. The digitized waveform was displayed on the screen of a computer terminal, 2 s at a time. The screen resolution for that display was about 2 ms. A cursor was placed at each tone onset, and a permanent "label" was attached to that point in the waveform file. The differences between the time points of successive labels yielded the IOIs, which were noted down to the nearest millisecond. After some practice, it took about 2 hours to measure one complete performance.

Two problems had to be coped with. One was that the onset of a soft tone was often difficult to detect by eye, particularly in some of the older recordings, which had much surface noise. (DAV was the worst by far.) An auditory method was used in that case: The cursor was moved back in small steps, and the waveform segment up to the cursor was played back until the onset of the tone in question could no longer be heard. This procedure was time-consuming but usually resulted in rather accurate location of tone onsets (cf. Gabrielsson, 1987). The second problem concerned onset asynchronies among simultaneous tones in different voices. Unintended asynchronies, which usually were too small to be detected by eye in the waveform, had to be ignored. Larger asynchronies, which in most cases must have been intended by the artists, were noted (especially in performances by CO1-3, DAV, MOI, and SHE) and measured, but the label from which the IOI was computed was placed at the onset of the major melody tone (usually the one with the highest pitch and the latest onset). The same convention was followed in the case of the written-out *arpeggi* in bars 2, 6, and 18.

B. Data representation

The data were submitted to further analysis in the form of eighth-note IOIs. That is, intervals involving grace note onsets (in bars 2, 6, 8, 16, and 18) were omitted from the data and were considered separately.[1] IOIs longer than a nominal eighth-note were divided into eighth-note intervals of equal duration. Thus, for example, the half-note IOI in bar 2 was represented as four equal eighth-note IOIs. A complete performance thus yielded 254 eighth-note IOIs.

C. Measurement error

The measurement procedure would have been prohibitively time-consuming if maximum accuracy had been aimed for (e.g., by displaying shorter waveform segments on the computer terminal screen). A certain amount of speed-accuracy tradeoff had to be taken into account. The magnitude of the resulting measurement error can be estimated from three performances (BRE, CO2, and the first half of CUR) that, for various reasons, had to be remeasured. Two of them (BRE, CUR) derived from good LPs with a moderate amount of surface noise, whereas CO2 was of older vintage and had special problems related to the pianist's tendency to play the left- and right-hand parts asynchronously. There were several very large discrepancies (in excess of 80 ms) between the two measurements of the CO2 performance, due to a conscious change in the author's criteria for treating asynchronies; these discrepancies were not considered true measurement errors and were omitted from the data presented below, though some smaller inconsistencies of the same kind may still be included.

The measurement error distributions are shown in Table 2. The combined BRE+CUR data represent average accuracy, whereas the CO2 data constitute a worst-case scenario; only DAV was even more difficult to measure. It can be seen that over 90% of the measurement errors in the BRE+CUR data did not exceed 10 ms, which is less than 2% of the average IOI. The largest error was under 35 ms, or about 6% of the average IOI, and the average measurement error was 4.3 ms, less than 1%. In the CO2 data set, about 90% of the errors ranged from 0 to 20 ms, and the largest error was under 50 ms. The average error was 10.4 ms, or about 2%. Measurements of the virtually noiseless CD recordings presumably were even more accurate than suggested by the BRE+CUR data.

Table 2. *Distributions of absolute measurement error in two sets of remeasured data.*

	BRE+CUR (N=412)		CO2 (N=245)	
Range (ms)	Percent	Cum.	Percent	Cum.
0 - 5	77.4	77.4	24.5	24.5
6 - 10	13.8	91.3	40.0	64.5
11 - 15	2.7	93.9	18.8	83.3
16 - 20	1.7	95.6	5.7	89.0
21 - 25	2.2	97.8	3.7	92.7
26 - 30	1.7	99.5	3.3	95.9
31 - 35	0.5	100.0	2.0	98.0
36 - 40			1.2	99.2
41 - 45			0.4	99.6
46 - 50			0.4	100.0

IV. RESULTS AND DISCUSSION

The strategy in presenting the results will be to proceed from more global properties to more detailed aspects.

A. Overall tempo

Perhaps the most obvious dimension along which different performances vary is that of overall tempo. The present set of performances is no exception. Tentative estimates of global tempo were obtained by computing the first quartiles (the 25% point) of the individual eighth-note IOI distributions, multiplying these millisecond values by 2, and dividing them into 60,000. The choice of the first quartile was motivated by the consideration that expressive lengthening of IOIs is both more frequent and more pronounced than shortening, and also by the fact that it gave tempi for two pianists, BRE and DAV, that agreed closely, respectively, with Brendel's (1981) statement of his preferred tempo and with Clara Schumann's (DAV's teacher's) recommended tempo. (Repp, 1993b, will provide a more extended discussion.) Table 3 shows that the tempo range extended from 48 to 79 quarter-notes per minute. Apart from the fact that the three fastest performances are all old recordings, there does not seem to be any systematic relationship between tempo and the time at which the recording was made, nor with pianists' gender, age at the time of recording, or country of origin.

Table 3. *Overall tempi (qpm) estimated from the first quartile of the IOI distribution.*

48	ESC
49	
50	KAT, NEY
51	
52	ZAK
53	CAP
54	KLI
55	
56	CUR
57	BUN, NOV
58	DEM, KUB
59	ARR, KRU, MOI, SCH
60	
61	HO2
62	ARG
63	ASH
64	HO3
65	HO1, SHE
66	CO3
67	BRE, ORT
68	GIA
69	
70	
71	
72	CO1
73	
74	
75	CO2
76	
77	
78	
79	DAV

B. Repeats

The first step in the data analysis was an attempt to eliminate redundancy and reduce random measurement error by averaging across repetitions of the same (or highly similar) musical material. Of course, such averaging is meaningful only if there are no systematic differences in timing microstructure across repeats. The prime target for averaging was the first period, bars 1-8, which was repeated literally, according to Schumann's instructions in the score. All but two pianists (DAV, KRU) observed this repeat.[2]

To compare the first and second repeats for all pianists, a *grand average timing pattern* was first obtained by computing the geometric mean of corresponding IOIs across all 28 performances. (For DAV and KRU, the data of bars 1-8 were simply duplicated for the missing second repeat.) The geometric mean was preferred over the arithmetic mean because it compensated for any tendency of slower performances to show more expressive variability, which would have dominated in an arithmetic average. In this grand average, the two repeats of bars 1-8 were found to have extremely similar timing profiles, with only a slight tendency for the second repeat to be played slower. The correlation between these two averages was 0.987, which indicates that any variations across repeats for individual pianists were either random or idiosyncratic. The correlations between the two repeats in individual performances are shown in Table 4 (column 2); most of them were quite high.

The close similarity of timing patterns across repeats has been noted in virtually every study in which such a comparison was made (see, e.g., Palmer, 1989; Repp, 1990; Seashore, 1938). Thus, it seemed justified to average across the two repeats of bars 1-8 in all further analyses, except as noted.

There were other instances of identical or highly similar musical material in the piece. Their timing patterns may be compared in Figure 3, which plots the grand average eighth-note IOIs. A logarithmic ordinate scale is used here to accommodate the longest values; it also makes the scale comparable to the percentage scale used by authors such as Gabrielsson (1987). The abscissa represents metrical distance (or "score time") in eighth-note steps.[3] Longer notes appear as "plateaus" of multiple eighth-notes; that way, all IOIs are represented on the same proportional scale.

Table 4. *Correlations of the timing profiles for repeated or similar sections (R = repeat; GM = geometric mean).*

	Bars		
Artist	1-8/1R-8R	1-4/17-20	9-12/13-16
ARG	0.781	0.851	0.630
ARR	0.933	0.938	0.850
ASH	0.922	0.951	0.927
BRE	0.953	0.888	0.880
BUN	0.510	0.633	0.632
CAP	0.935	0.922	0.846
CO1	0.859	0.717	0.837
CO2	0.865	0.736	0.851
CO3	0.750	0.741	0.571
CUR	0.937	0.911	0.921
DAV	*	0.863	0.863
DEM	0.938	0.924	0.828
ESC	0.890	0.734	0.802
GIA	0.777	0.809	0.914
HO1	0.918	0.958	0.697
HO2	0.811	0.861	0.822
HO3	0.826	0.738	0.770
KAT	0.825	0.901	0.867
KLI	0.804	0.893	0.861
KRU	*	0.875	0.940
KUB	0.951	0.913	0.888
MOI	0.805	0.855	0.493
NEY	0.920**	0.904	0.872
NOV	0.931	0.906	0.908
ORT	0.677	0.609	0.707
SCH	0.676	0.574	0.748
SHE	0.868	0.831	0.859
ZAK	0.839	0.875	0.909
GM	0.987	0.986	0.950

*no repeat
**bars 1-4 only (see Footnote 2)

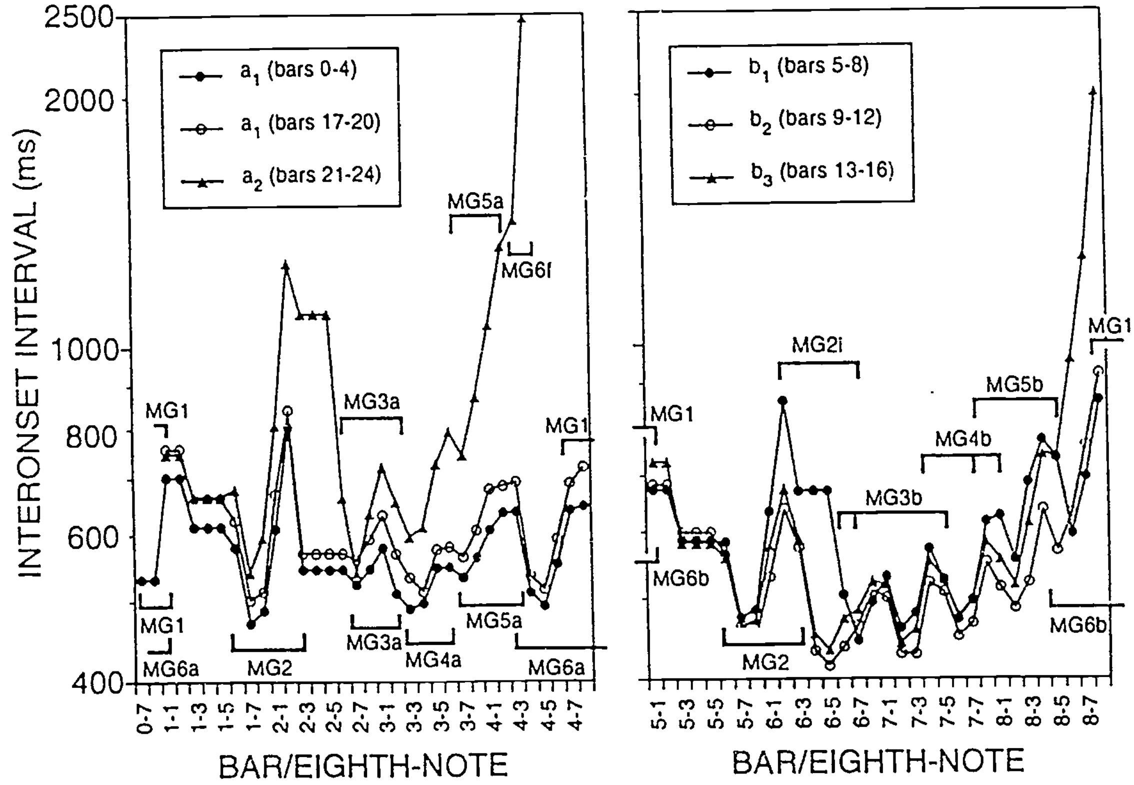


Figure 3. Grand average IOIs (geometric means across all 28 performances). The two panels show the timing profiles for phrases of Type a and b, respectively. Primary melodic gestures (cf. Figure 2) are indicated by brackets. The bar numbers on the abscissa refer to bars 0-8 only; for the later phrases, an appropriate constant must be added.

The left-hand panel of Figure 3 compares the timing profiles for the three Type a phrases. Bars 17-20 (phrase a1) are musically identicalwith bars 1-4 (see Figure 1), and it can be seen that their timing patterns are highly similar (r = 0.986), though bars 17-20 tend to be played at a somewhat slower tempo. The individual correlations for these four bars (Table 4, column 3) are comparable to the correlations between the two repeats of bars 1-8. The timing pattern for bars 21-24 (phrase a2) is initially similar but diverges soon, due to the *fermata* (long hold) in bar 22 and the progressive *ritardando* (slowing of tempo) towards the end of the piece.

The right-hand panel of Figure 3 compares the timing patterns of the three Type b phrases. Once again a striking similarity can be seen, due to the similarity or identity of the melodic structure (see Figure 2). The timing profiles of phrases b2 (bars 9-12) and b3 (bars 13-16) are especially close. Only the (prescribed) *ritardando* in bar 16 is much more pronounced than that in bar 12 (which is not notated). If the final halves of bars 12 and 16 (the *ritardandi*) are excluded, correlations between these two timing patterns are again extremely high (see Table 4, last column). Phrase b1 (bars 5-8) also shows a rather similar profile; differences occur precisely where its musical content deviates from phrases b2 and b3 (especially in bar 6).

The correlations in Table 4 indicate which artists were highly consistent in their timing patterns, and which of them were less consistent. The consistent group is led by ASH, CUR, and NOV and includes many others; the less consistent group includes ARG, BUN, CO3, MOI, ORT, and SCH.[4]

C. Melodic/rhythmic structure and the average timing pattern

Let us examine now the average timing pattern and its relation to the musical structure in greater detail. Although the grand average timing profile is not necessarily an optimal performance timing pattern, it captures features that many individual performances have in common, whereas it suppresses random and idiosyncratic timing deviations that differ from performance to performance.

In Figure 3 it can be seen that there are global temporal trends within phrases, particularly those of Type b. Their timing profiles follow a concave function on which various local peaks are superimposed. Bars 21-24 (phrase a2), too, seem to follow such a curve, though it is grossly distorted by the *fermata* and the final grand *'ardando*. Bars 1-4 and 17-20 (phrase a1) have a flatter global timing shape. The curvilinear trends in Type b phrases are reminiscent of the parabolic curves hypothesized by Todd (1985) as the basic phrasal timing pattern. Todd's model[5], however, operates on the durations of time units equivalent to whole bars, whereas the present timing profiles are across eighth-note units and thus provide much finer resolution. The curves that "cradle" the present timing patterns (i.e., the "bottom lines" of the Type b timing profiles) are, in fact, poorly fit by quadratic functions. However, they do represent the general principle formalized by Todd and observed by many others, that there is a slowing of the tempo at major structural boundaries, in proportion to the importance of the boundaries. Thus the most extreme *ritardando* (prescribed in the score) occurs at the end of the piece; a pronounced slowing down (also prescribed) occurs at the end of the second period (bar 16); next come the end of the first period (bar 8) and the end of phrase b1 (bar 12), which show about the same amount of *ritardando* (not explicit in the score); the end of phrase a1 (bars 4 and 20) shows the smallest, but still quite noticeable slowing down.

The piece begins with a brief melodic gesture (MG1) that ascends from the dominant to the tonic, with accent on the latter (hence the intervening bar line, which indicates accent on the following beat). MG1 recurs in bars 4-5, 8-9, 12-13, 16-17, and 20-21, though the upbeat is shortened in bars 8, 12, and 16 (cf. Figure 1). It thus contains a single IOI of variable nominal length. Its realization was strongly context-dependent: The quarter-note upbeat at the beginning of the piece (bar 0) was short relative to the following IOI, which seemed fairly stable across all phrases. The quarter-note upbeats at the ends of bars 4 and 20, which coincided with MG6a, were relatively longer. The eight-note upbeats in bars 8 and 12, which coincided with MG6b, were even longer. Finally, although the grace-note upbeat of bar 16 is not shown explicitly in Figure 3, the IOI accommodating it was lengthened enormously, due to the major *ritardando* at the end of period B. (More information about the grace-note upbeat is provided later on.)

The chord following MG1 enriches the harmonic and rhythmic texture but is not really part of the melody; however, it may be thought of as a kind of echo of the tonic and thus could be linked with MG1. The chord bisects the inter-gesture interval between MG1 and MG2. Of the resulting two IOIs, the first (nominally shorter) one was consistently longer relative to the second (nominally longer) one, as if pianists tried to equalize the two. This may reflect final (and/or accent-related) lengthening of the tonic in MG1, which is absorbed by the IOI preceding the chord.

MG2 is undoubtedly the most salient melodic gesture of the piece. It comprises five eighth-notes that ascend in increasingly larger pitch steps to a final note which constitutes the apex of the four-bar melodic arch that constitutes each phrase. Variants of the gesture occur in bars 1-2, 5-6, 9-10, 13-14, 17-18, and 21-22. MG2 is unusual because it culminates on the second beat of a bar, negating entirely the customary accent on the first beat. As can be seen in Figure 3, the average temporal pattern of the five IOIs is quite similar in all occurrences of the gesture: The first IOI is close to the proportional duration of the preceding IOI, but the next two IOIs are much shorter. The fourth IOI is longer again, similar to the first, and the fifth IOI is the longest; it is clearly visible as a sharp peak in the timing profile. The duration of that fifth IOI varies with position in the piece: It is least extended in bars 10 and 14, longer in bars 2, 6, and 18, where it accommodates the two grace notes of the left-hand chord (cf. Figure 1), and longest in bar 22 where it leads to the climactic *fermata*. In fact, this eighth-note IOI in bar 22 tended to be lengthened relative to the *fermata* chord itself (the plateau following the peak). The specific *accelerando-ritardando* shape of MG2 will be examined more closely later on.

The IOI following MG2 in bars 2, 6, 18, and 22 (plateau in Figure 3) represents another inter-gesture interval, nominally three or four eighth-notes long. Relative to the last IOI of MG2, this interval was much shorter proprortionally, even in bar 22, where it received the *fermata*. The *fermata* had its maximal effect on the inter-gesture interval, doubling its duration relative to bars 2 and 18, but it also affected the whole preceding MG2. In bars 10 and 14, the inter-gesture IOI is bridged by the inner-voice imitation gesture MG2i, which has a timing pattern that is almost the mirror image of MG2 in that it accelerates during the first three IOIs and slows down only at the end. The first IOI is long because it also represents the last IOI of MG2; the relative length of the second IOI may be transitional, or it may reflect final lengthening of MG2 which was absorbed by MG2i.[6]

The remaining MGs, which descend in a chain from the melodic apex reached by MG2, are patterned differently in phrase Types a and b. The grouping structure of the Type a chain is indicated with slurs in the score (Figure 1). MG3a consists of four (a1) or five (a2) eighth-notes in the soprano voice, with accent on the penultimate; a shorter accompanying gesture occurs in the tenor voice (a1 only). It can be seen in Figure 3 that lengthening on the accented note occurred in performance. MG4a is similar in rhythmic structure to MG3a and exhibits a similar timing pattern, except that the final unstressed eighth-note IOI was lengthened as well, especially in phrase a2, where the final *ritardando* tilted the timing profile. MG5a, reinforced by shorter secondary MGs in the bass voice, is similar to MG4a, except that in phrase a1 it leads into a final half-note that coincides with the onset of MG6a. Its timing pattern is a progressive lengthening, similar to that observed in MG4a. In phrase a2, the final *ritardando* is in full control. The final gesture in phrase a1, MG6a, is in the bass voice and leads back to the low-pitched tonic which doubles the tonic of the next MG1. Its first IOI coincides with the final lengthened IOI of MG5a, and the last two IOIs coincide with the quarter-note upbeat of MG1 and are both lengthened as well. The IOIs in between are considerably shorter. The total timing profile of MG6a is not unlike that of MG2, which it resembles rhythmically. The final gesture in phrase a2, MG6f, is a truncated version of the preceding MGs and shows an enormous slowdown in tempo; note, however, the discontinuity in the *ritardando* between MG5a and MG6f, which is a manifestation of final lengthening in MG5a.

The melodic chain of type b is also divided into four MGs, but they are organized differently. The slurs in the score are not entirely consistent here; what distinguishes the gestures is the fact that they occur in different voices, imitating and overlapping each other. Simultaneously, other voices articulate shorter two-note cadential gestures which, as can be seen in Figure 3, essentially govern the timing pattern. In each MG, *both* IOIs constituting these accompanying motifs were lengthened, the first, unaccented one usually more than the second, accented one. This is explained by the harmonic function of the tone cluster initiating the unaccented IOI, which is that of suspense followed by resolution. In part, however, it may also be due to final lengthening of the preceding, overlapping MG.

In summary, these observations illustrate several general principles of performance timing: (1) Whole phrases tend to show global tempo curves characterized by an initial *accelerando* and a more pronounced final *ritardando* whose degree reflects the degree of finality of the phrase in the hierarchical grouping structure. (2) "Riding" on the global pattern, individual MGs often show a similarly curved local pattern, though accent placement, harmonic factors, and the influence of overlapping MGs in other voices may modulate that pattern in various ways. Gesture-final lengthening is commonly observed.

D. Principal components analysis

An alternative way of capturing commonalities among different performances is offered by the statistical technique of principal components (factor) analysis. This method decomposes the data matrix (N performances by M IOIs) into a number of independent components or factors, each of which resembles a timing profile (i.e., M standardized "factor scores"). The original data are approximated by a weighted sum of these factors; the weights, which differ for each performance, are called "factor loadings" and represent the correlations between the performance timing profiles and the factor score profiles. The degree of approximation (the percentage of the "variance accounted for" or VAF) depends on the number of factors that are considered significant; a common criterion employed here is that they should have "eigenvalues" greater than 1. (Eigenvalue = N times an individual factor's proportion of VAF.) The first factor always accounts for the largest amount of variance and represents a kind of central tendency or most common pattern. Additional factors account for increasingly less variance and thus represent patterns shared by fewer performances. A standard technique for simplifying the factor loadings and thereby increasing the interpretability of the factors is called "Varimax rotation." It increases large factor loadings and reduces small ones without changing the number of factors or the total VAF; however, the VAF is redistributed among the rotated factors.

Principal components analysis thus can reveal whether there is more than one shared timing pattern represented in the sample of 28 performances. If there is essentially only one way of performing the piece, then a single factor should explain most of the variance in the data. If there are several radically different timing patterns, then several orthogonal factors should emerge. Idiosyncratic timing patterns will not lead to a significant factor and will constitute part of the variance not accoun*ted for. Since all timing profiles are standardized and intercorrelated in the course of the analysis, differences in overall tempo are automatically disregarded.

Although it would seem desirable to conduct the analysis on the complete performances, the large *ritardandi* observed universally at the ends of phrases a2, b1, b2, and b3, as well as the large slowdown at the *fermata* in phrase a2, dominate the overall timing pattern and cause high correlations among the performances, with the result that the interesting variation among the shorter IOIs is lost. In fact, a principal components analysis conducted on the complete data yielded only a single factor, indicating that all pianists observed the major *ritardandi*, which is not very surprising. Therefore, it was decided to analyze only the data for bars 0-8 (phrases a1 and b1), which did not include any extreme lengthenings (cf. Figure 3). Because of the parallelism of the timing profiles of these phrases to those of similar type (see Figure 3), such a restricted analysis essentially captures the whole performances minus the major *ritardandi*.

This analysis yielded four significant factors. Together they accounted for 76% of the variance. For individual performances, the VAF ranged from 60% to 90%. The timing profiles representing the first three factors are shown in Figure 4, and the factor loadings of the 28 performances are listed in Table 5.

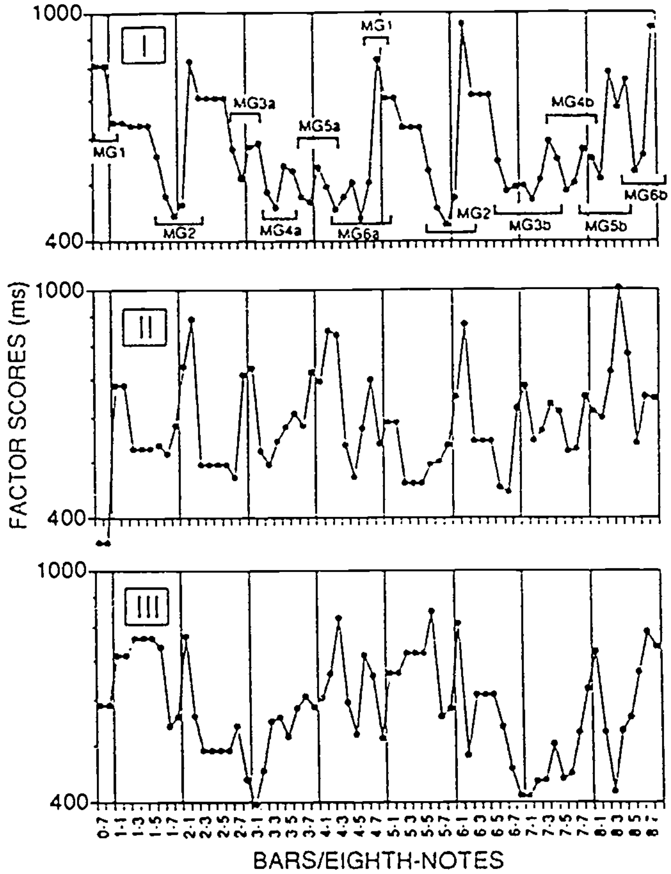


Figure 4. Timing patterns of the first three principal components for bars 0-8. The standardized factor scores were converted into milliseconds by multiplying them with the average within-performance IOI standard deviation and adding this product to the grand mean IOI. Bar lines and MG brackets are added for orientation.

Table 5. *Sorted rotated factor loadings of the 28 performances, bars 0-8. (Loadings smaller than 0.4 are omitted.)*

	I	II	III	IV
SCH	0.844			
ASH	0.760	0.423		
BRE	0.738			
DAV	0.695			
CAP	0.695	0.438	0.411	
SHE	0.692			
KAT	0.685			
KUB	0.672			
ARR	0.664	0.463		
ZAK	0.644	0.460		
ORT	0.637			
CUR	0.615		0.429	
KRU	0.593		0.495	0.445
DEM	0.551	0.495		
HO1		0.888		
HO3		0.886		
HO2		0.837		
ARG		0.770		
NEY		0.665		0.524
ESC	0.482	0.626		
GIA		0.595	0.491	
BUN		0.588		
KLI	0.523	0.570		
NOV	0.509	0.549		0.464
CO3			0.876	
CO1			0.875	
CO2			0.850	
MOI	0.501			0.582

Factor I (VAF = 30%) represents a timing pattern shared (partially) by a large number of performances; half the pianists had their largest loading on this factor, with SCH, ASH, and BRE leading the group. This pattern has the following features: a relatively long initial upbeat (MG1) and pronounced lengthening of the last IOI of MG6, which accompanies the upbeat to the next MG1 (positions 4-8 and 8-8); general lengthening of long IOIs (plateaus in the profile); a relatively long *accelerando* phase in MG2 followed by a pronounced lengthening of its last IOI; and small but regular gestural accent peaks for MG3-MG6, except for special emphasis at the end of MG5b (positions 8-3 to 8-5). This pattern reflects the melodic/rhythmic structure even more clearly than the average timing profile displayed in Figure 3.

Factor II (VAF = 25%) may be called the "Horowitz factor," since the three performances by that artist showed the highest loadings, though ARG, NEY, ESC and a number of other performances shared features of this pattern. Due to the statistical orthogonality of the factors, this timing pattern is radically different from that of Factor I. It is characterized by a very short initial upbeat (essentially reduced to an eighth-note, though it was rarely that extreme in the actual performances; see below); relatively short long IOIs (i.e., low plateaus); no *accelerando* but a pronounced *ritardando* during MG2; pronounced lengthening in MG3a, MG3b, and MG5b, with extreme prolongation of the IOI in position 8-4, which accommodates a grace note.

Factor III (VAF = 15%) is the "Cortot factor"; the three performances by this artist were the only ones that showed substantial loadings. Its timing pattern is most unusual: a relatively short upbeat in MG1; a short final IOI in MG2 (positions 2-2 and 6-2); marked speeding up during MG3 with a following *ritardando* extending through MG4 to the end of MG5; and rushing during the end of MG5b, followed by a pronounced *ritardando* during MG6b.

Factor IV (VAF = 6%) seems less important and will not be discussed in detail, as no artist showed a high loading on it (Table 5).

The three factor timing patterns shown in Figure 4 do not correspond exactly to any real performance, though they are similar to certain performances, as indicated by the factor loadings (Table 5). Most artists' performances are best described by a weighted combination of several factors; thus, for example, the CAP timing profile is a combination of Factors I, II, and III. Such a combination will of course result in an attenuation of extreme features. About one fourth of the variance was not explained by the four factors extracted, and most of that variance was probably nonrandom, representing the artists' individuality.

It had to be asked at this point whether the Horowitz and Cortot factors emerged only because each of these artists was represented by three different performances.[7] The principal components analysis was repeated with only a single performance of each of these pianists included (CO1 and HO2). The three factors that emerged (VAF = 71%) were highly similar to the first three factors of the earlier analysis. The second and third factors still had their highest loadings in HO2 and CO1, respectively. The patterns of factor loadings for the three factors in the two analyses correlated 0.96, 0.98, and 0.95, respectively. Thus, the Horowitz and Cortot factors are not artifacts due to the over-representation of these artists in the sample. They represent two true alternatives to the "standard" pattern of performance timing instantiated by Factor I—alternatives that only Horowitz and Cortot dared to choose in nearly pure form, but that several other pianists incorporated partially into their timing strategies.

E. Detailed analyses

In this section, we examine how melodic gestures and other details of the score were executed by using, as it were, a magnifying glass on the data. While the emphasis has so far been primarily on commonalities, the analyses are now directed increasingly towards uncovering artistic diversity. They reveal the different ways in which a gesture may be shaped by pianists of great authority. They also reveal some unusual, and occasionally questionable, interpretations of notational details in the score. In addition to demonstrating the range of individual differences, these analyses illustrate what temporal patterns are preferred by a majority of the artists, and what patterns are never used, and hence presumably unacceptable.

1. MG1. This gesture consists of merely two tones forming the pitch interval of an ascending fourth. It is followed, however, by a chord whose timing, though not strictly part of the melody, is also of interest. MG1 appears six times: Three times with a nominal quarter-note upbeat (bars 0, 4, and 20) and three times with a nominal eighth-note upbeat (bars 8, 12, 16), one of which (bar 16) is written as a grace note and thus is open to the assignment of an even shorter value. Two of the instances of the gesture are relatively unconstrained because the upbeat is unaccompanied (bars 0 and 16); in the other cases the upbeat coincides with eighth-notes in the bass voice (MG6).

Figure 5 presents the individual data in terms of two IOI ratios, A/(BC) and B/C, where A is the duration of the intragesture IOI (i.e., the upbeat), B is the intergesture IOI preceding the chord (nominally two eighth-notes) and C is the intergesture interval following the chord (nominally three eighth-notes). If the score were played literally, A/(BC) should be either 0.2 or 0.4, depending on whether the upbeat is notated as an eighth-note or a quarter-note, and B/C should be 0.67. In Figure 5, however, the ratios have been normalized with respect to the underlying eighth-note pulse (assumed to be isochronous for this purpose), so that the expected "literal" ratios (effectively, tempo ratios) are equal to 1 in all cases.

Let us consider first the ratio plotte ɔn the abscissa. The top row of panels in Figure 5 shows the three instances in which the MG1 upbeat is nominally a quarter-note. The leftmost panel shows the unconstrained situation at the beginning of the piece (bar 0). It can be seen that very few pianists played this initial upbeat—or the following intergesture interval, as the case may be—"as written" (ratio of 1). The large majority played the upbeat quarter-note *short* relative to the intergesture interval. Several pianists (NEY, CO1, BUN) come very close to playing it as if it were an eighth-note (ratio of 0.5), and one (ARG) does so without question. This may have been a deliberate anticipation of the (notational) shortening of the upbeat in later incarnations of MG1, but it is a true deviation from the written score and is perceived as such.

Later, however, when the quarter-note upbeat occurs together with two eighth-notes in the bass voice (top center and right-hand panels in Figure 5), pianists are more evenly divided between those who shorten the upbeat and those who lengthen it relative to the intergesture interval. While the average ratio is close to 1, there is enormous variation among individual artists, and some who severely shortened the initial, unconstrained upbeat (e.g., ARG, BUN) do not repeat this tendency later on.

The lower row of panels in Figure 5 shows instances where the upbeat is nominally an eighth-note. In two cases (left-hand and center panels) the upbeat is accompanied by a bass-voice eighth-note that marks the approaching end of MG6 and of a whole phrase, so here the overwhelming tendency is to lengthen the upbeat, sometimes to an extent corresponding to a nominal quarter-note (ratio of 2; BUN and ARR in bar 12). In bar 16, where the upbeat is notated as a grace eighth-note, a significant minority of pianists plays the upbeat shorter than an eighth-note, in one case (KAT) shorter than a sixteenth-note (ratio of 0.5). The majority, however, still lengthen it, and at least one (ASH) plays the grace note as if it were a quarter-note (ratio of 2).

There is little indication of clustering or bimodality in these data. Although there are some "outliers," the values of the ratio seem to be rather evenly distributed over a wide range. Correlations of individual timing patterns across the different instances of MG1 are not high, though some consistencies can be seen. Thus ARR and BUN always tend to lengthen the upbeat, whereas ARG, DEM, and CO1-3 tend to shorten it. Some pianists, perhaps to be regarded as "literalists," consistently avoid extremes (e.g., ESC, KRU, MOI, ZAK).

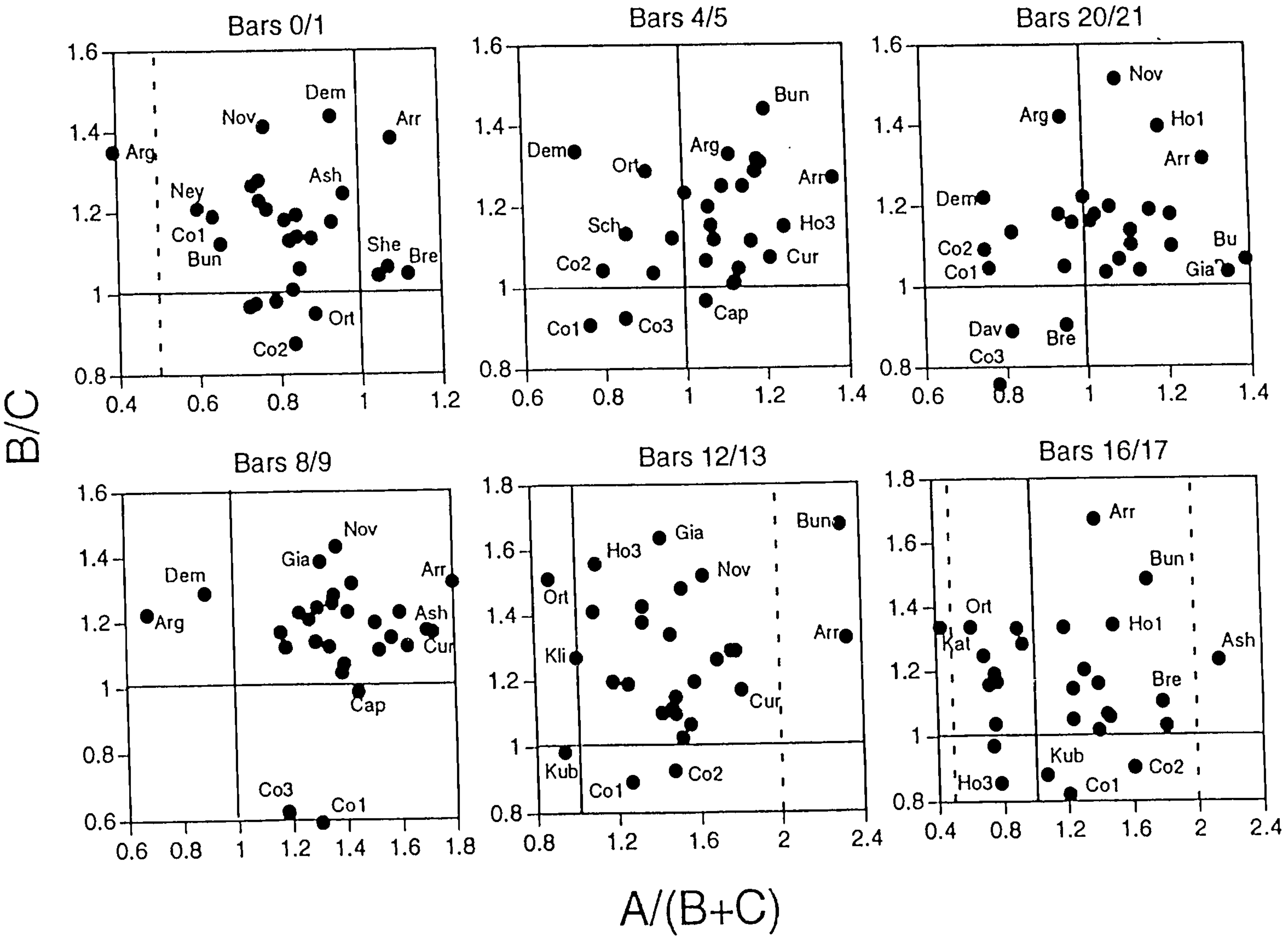


Figure 5. Timing patterns of the six instances of MG1 and the following chord. (See text for explanation of ratios.) Ratios of 1 represent equal underlying beats (i.e., equal tempo). The upper panels plot the three instances where the upbeat is nominally a quarter-note, while the lower panels plot the three instances where the upbeat is nominally an eighth-note (a grace note in bar 16). The dashed lines indicate a doubling or halving of the nominal duration in execution.

With regard to the B/C ratio, plotted on the ordinate, there is more consistency. In all six instances, which are notationally equivalent, there is a strong tendency to lengthen the rest preceding the chord, sometimes to the extent that the onset of the chord occurs in the middle of the intergesture interval (ratio of 1.5). Among those who tend to play the chord very late are ARG, ARR, BUN, and NOV. Nevertheless, there are some pianists who do not show that pattern, most notably Cortot.

Finally, it should be noted that the two ratios are uncorrelated across different artists. Evidently, the timing of MG1 is quite independent of the timing of the chord in the intergesture interval.[8]

2. MG2. This is the signature melodic gesture of the piece, and its manner of execution is crucial to the impression of a performance of "Träumerei." It offers a unique opportunity to investigate the temporal shaping of a gesture, for several reasons: (1) It comprises six notes (i.e., five IOIs)—a sufficient number of degrees of freedom for any constraints on temporal shape to emerge very clearly; (2) it is unidirectional in pitch and uninterrupted by metric accents—the accent that would normally occur on the note following the bar line is clearly suspended in this case; (3) it recurs six times, with slight variations; and (4) pianists are likely to give special attention to its execution.

MG2 occurs in bars 1-2, 5-6, 9-10, 13-14, 17-18, and 21-22. (We will not consider the imitation gesture MG2i in detail.) The versions in bars 1-2 and 17-18 are identical; during the last IOI, two grace notes (a written-out *arpeggio*) occur in the left-hand accompaniment of the melody. In bars 5-6 and 21-22, the penultimate pitch interval is extended from a fourth to a major sixth; in addition, the melody in bars 21-22 leads to a *fermata* and also lacks the grace notes during the last IOI. In bars 9-10 and 13-14, there are no grace notes, the penultimate melodic interval is reduced to a minor third in the first instance, there is a change of key (to B-flat major) in the second instance, and the occurrence of MG2i in the middle voices creates a forward movement that is absent in the other variants. Therefore, timing differences are to be expected reflecting these factors.

The average temporal shapes of the six versions of MG2 are shown in Figure 6. The five IOIs for each version, representing the geometric means across all 28 performances, are plotted here on a linear scale. The five data points for each version have been fitted with a quadratic function (i.e., a parabola), which seems to describe their temporal shape rather well.[9] Each gesture accelerates initially and then slows down at the end. This final *ritardando* is least pronounced in bars 9-10 and 13-14; it is more evident in bars 1-2, 5-6, and 17-18; and in bars 21-22, preceding the *fermata*, it is dramatic. Except in this last instance, there is a tendency for the penultimate IOI to rise slightly above the parabolic trajectory. This was evidently due to including in the average performances such as Cortot's which, as we have seen in Figure 4 (Factor III), showed a pronounced tendency to shorten the last IOI.

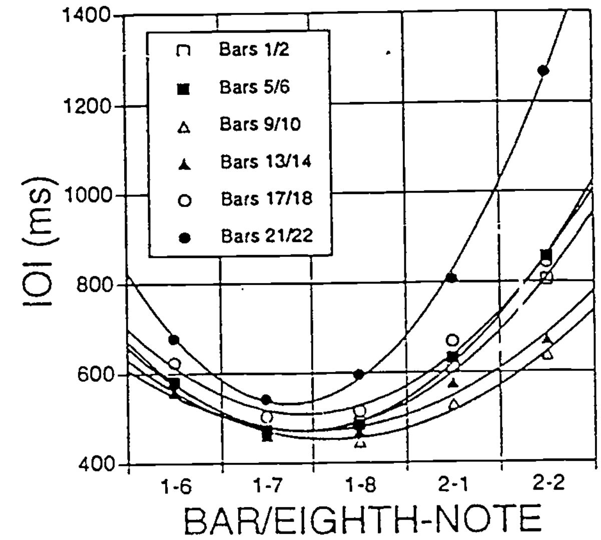


Figure 6. Geometric mean IOIs for the six versions of MG2, with best-fitting quadratic functions. The abscissa labels refer to bars 1 and 2.

The temporal shapes of the six versions of MG2 were examined and fitted with quadratic functions in each of the 28 individual performances (a total of 168 instances). It emerged that most individual artists' timing patterns (87% of all instances) could be described well by parabolas, with fits ranging from good to excellent. What varied between artists and between different instances of MG2 were the curvature and elevation of the parabolic functions, but not so much their goodness of fit (however, see Footnote 9). The data of six individual artists are shown in Figure 7.

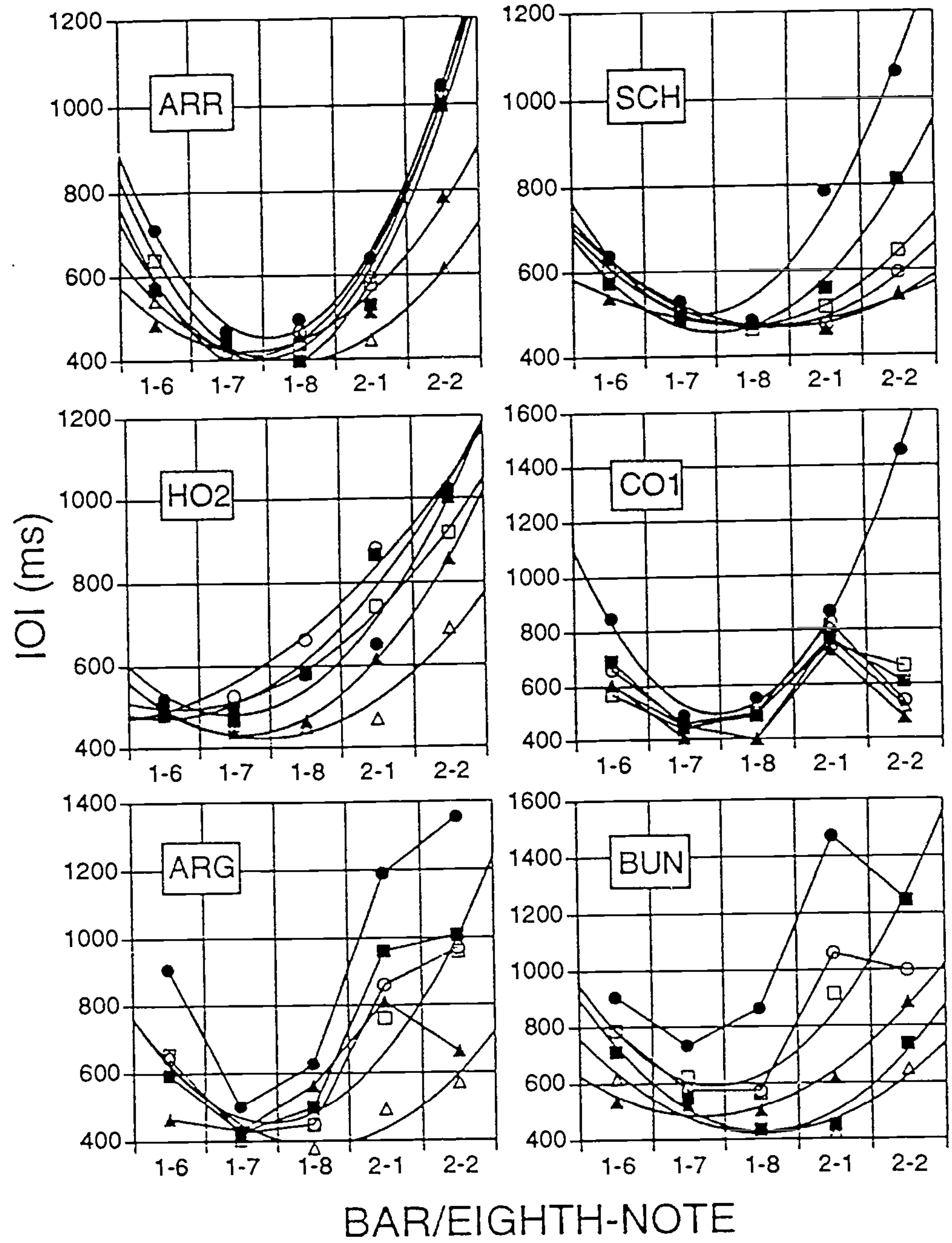


Figure 7. MG2 timing data for six individual artists, with best-fitting quadratic functions. Where the pattern deviated markedly from a parabolic shape, the data points are connected with straight lines. See Figure 6 for the legend of symbols.

The two panels on top illustrate two individual cases that were fit well by parabolas but differed in curvature: The highly modulated timing curves of ARR contrast with the flatter ones of SCH. The center panels show one representative performance of each of the two great individualists, Horowitz and Cortot. Horowitz's curves are fit fairly well by parabolas but generally lack the initial acceleration shown by most other pianists. Cortot shows a truly deviant pattern: In all three performances (which spanned 18 years the last IOI. This seems to indicate that he grouped the penultimate tone with the following long tone, treating it like an upbeat; in fact, the first four IOIs are fit well by a parabola. This idiosyncratic timing shape held only for the first five instances of MG2, however; in bars 21-22, preceding the *fermata*, Cortot produced a beautiful parabolic timing shape in all three performances. Finally, as can be seen in the bottom panels of Figure 7, ARG and BUN, two highly variable pianists, intermittently adopted the alternative timing pattern instantiated by Cortot; one other pianist who did so was CUR. The only other type of deviant timing pattern was shown by ORT who, in bars 9-10 and 13-14 only, lengthened the third IOI above the parabolic trajectory, which led to a W-like shape.

The quadratic functions ($y = C + Lx + Qx^2$) that fit the large majority of individual phrasal shapes are characterized by three parameters: a positive constant C that reflects vertical displacement, related to overall tempo; a generally negative coefficient L that reflects horizontal-vertical displacement of the curve in x-y coordinate space;[10] and a positive coefficient Q that reflects the degree of curvature of the concave (U-shaped) parabola. A question of great theoretical interest was whether there are any constraints among the three parameters; in principle, of course, they are quite independent of each other. The coefficients L and Q were plotted against each other for all 146 individual gestures that followed a parabolic shape. There was a remarkably tight linear relationship between these two parameters, with correlations ranging from 0.93 to 0.98 across the six versions of MG2. The reason for this result is that the location of the minimum of the function is relatively fixed. Therefore, as the curvature of the parabola increases, the negative slope of the tangent at x0 also increases.

If the x coordinate of the minimum is relatively fixed, then the constant C must be correlated with the y intercept of the function and should increase as a function of curvature Q. There were indeed positive linear correlations between C and Q, ranging from 0.79 to 0.93.[11] Thus, both L and C are fairly predictable from Q, which leads to the conclusion that the expressive timing patterns of the large majority of the performances of MG2 can be characterized by a single family of parabolas, varying in Q only. This family of functions is shown in Figure 8; it was generated by increasing Q from 20 to 140 in steps of 20, and by setting L and C according to their average linear regressions on Q. Figure 8 embodies a constraint on the expressive timing shape of MG2 that was obeyed by the majority of pianists. With few exceptions, any individual execution of this phrase can be characterized by a single curvature parameter, with some additional freedom in global tempo (vertical displacement), which is not represented in the figure.

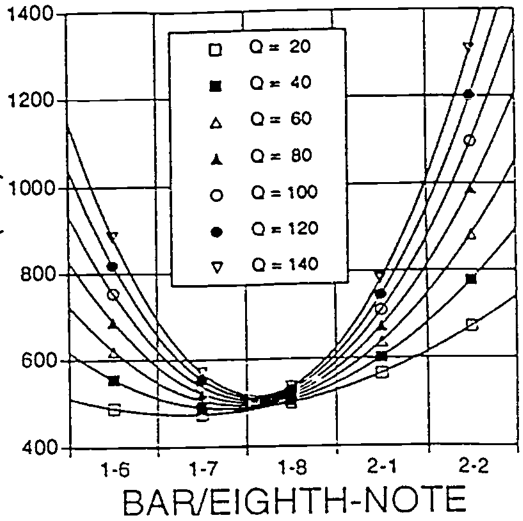


Figure 8. Family of quadratic functions that describes the timing constraint on MG2 observed by the large majority of pianists. The functions vary in curvature (Q), roughly over the observed range. The other coefficients of the quadratic functions were obtained by the average regression equations. Vertical displacement should be considered an additional free parameter.

To give some impression of individual differences in tempo modulation during MG2, Figure 9 plots the average Q coefficient for bars 1-2, 5-6, and 17-18 (which were generally executed similarly; cf. Figure 6) against that for bars 21-22, with individual artists identified. (Cortot is excluded from this plot.) Most pianists employed moderate modulation in the earlier instances of MG2, but gave a very expressive shape to the last instance in bars 21-22, which led to the *fermata*. Some, most notably DEM, used strong modulation in all instances. Horowitz's three performances are at the other extreme. They appear relatively unmodulated despite a substantial *ritardando* because he did not show any initial *accelerando*; therefore, his timing patterns were fit by parabolas with low curvature.

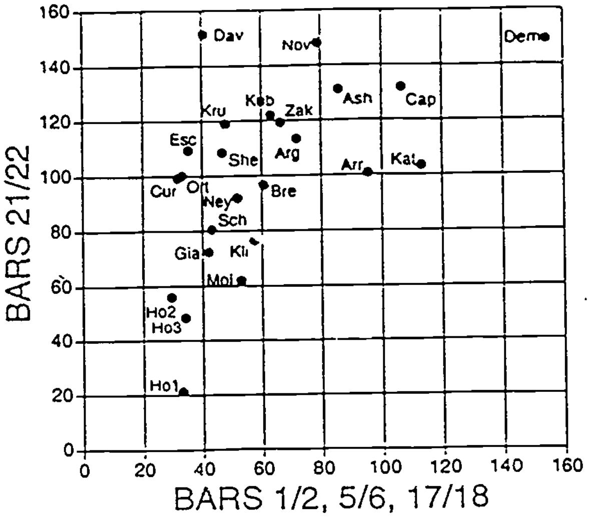


Figure 9. **Individual variations in tempo modulation during MG2. The average curvature (Q) for three instances of the gesture is plotted against the curvature for the last instance (bars 21-22). Only cases following the parabolic timing shape are included.**

This "parabolic constraint" on the timing shape of an expressive gesture is of great theoretical interest, as it supports Todd's (1985) suggestion of the parabola as a basic timing function. It also appears to be consistent with his more recent modelling (Todd, 1992b), even though he proposed that a linear function can account for most of the timing variation; clearly, a quadratic function did even better. The parabolic constraint may thus be understood as an allusion to physical motion (cf. Todd, 1992b), which most performers aim at, and which listeners with some musical education find aesthetically pleasing (Repp, 1992b).

Since MG1, the following inter-gesture interval, and MG2 may form a coherent half-phrase that is planned and executed as a single expressive unit, their timing relationships across the 28 performances were examined (in bars 0-2 only). None of the three coefficients characterizing the parabolic shape of MG2 correlated significantly with either of the two IOI ratios considered in connection with MG1 (Figure 5). However, the total interval preceding the onset of MG2 correlated moderately (0.59-0.72) with all three MG2 coefficients. Thus, the slower the initial tempo, the slower and the more inflected was MG2 (see also Footnote 8).

3. Grace notes accompanying MG2. Only three of the six instances of MG2 have grace notes (really a written-out *arpeggio*) in the left hand during the last IOI; they occur in bars 2, 6, and 18. Since all six instances of MG2 followed a parabolic trajectory, on the average (cf. Figure 6), the grace notes as such were not responsible for the lengthening of the final IOI and the systematic constraints that it followed. Rather, they seemed to be fitted into whatever temporal shape the pianists chose for the primary melodic gesture. We now turn our attention to the execution of these grace notes, which exhibited surprising variability.

In the score, the grace notes are represented as nominal sixteenth-notes. However, since the preceding notes in the left hand are nominally (tied-over) quarter-notes, there is literally no time for the grace notes in the rhythmic scheme; thus they must be "taken away" from the preceding or following note values. The standard way of execution suggested by the spatial placement of the grace notes in the score is to play them within the last IOI of MG2—that is, after the onset of the penultimate melody tone in the soprano voice, but before the onset of the final long tone and the accompanying two-tone chord. However, only 12 of the 24 pianists consistently played the passage according to this conventional interpretation of the notation. (The two repeats of bars 1-8 were examined separately in this analysis.) It seems that this brief passage offered pianists ample opportunity to indulge in "deviations from the score."

Most of the variants are illustrated schematically in Figure 10, which uses horizontal displacement of the critical notes in bar 2 to indicate the actual succession of tones. The standard version is shown in (a). As shown in (b), some pianists delayed the last note in the left hand until after the octave chord in the right hand: ARR (slightly, bar 6 only), DAV (bar 2 only), MOI (bar 6, once in bar 2), NOV (only the first time in bar 2, but very dramatically), and ZAK (consistently). Others (c) played all notes of the left hand before those of the right hand: ARG (except the second time in bar 6), HO1 and HO2 (bar 2). In a variant of that pattern, the right-hand chord was arpeggiated (DEM, consistently) or mysteriously interleaved with the broken chord in the left hand (Horowitz, in bar 6 only, but in all three performances).

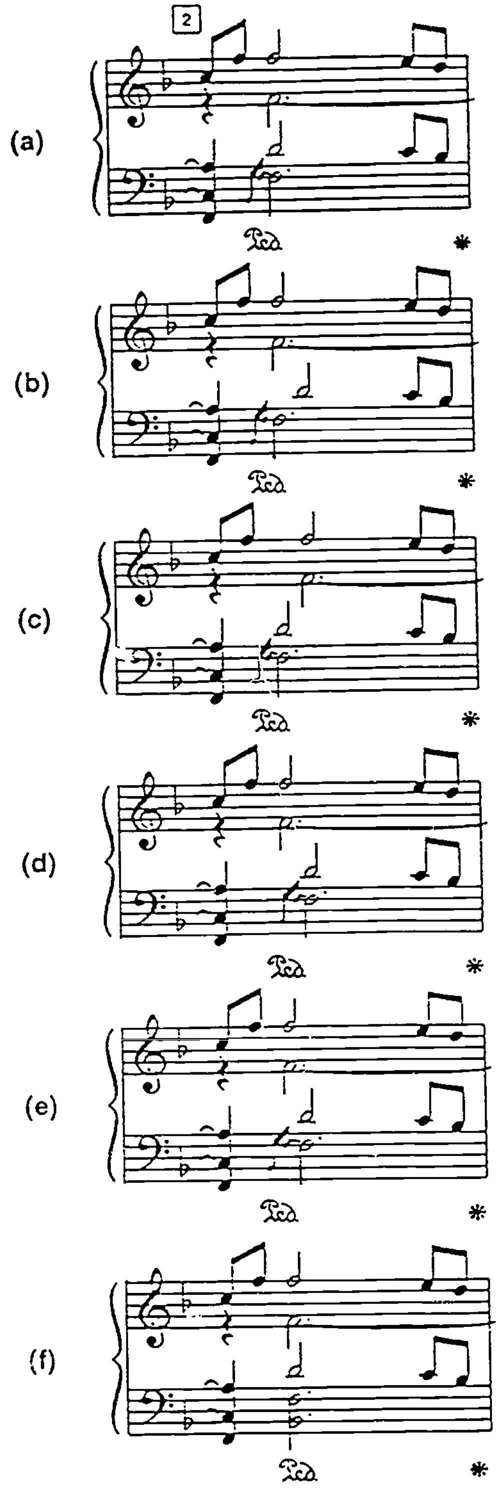


Figure 10. Schematic illustration of six observed manners of execution of the grace notes accompanying MG2. The notation for bars 2 and 18 is shown, but it stands for bar 6 as well. Spatial displacements of the notes symbolize temporal displacements.

In all instances mentioned so far, the onset of the last melody tone of MG2 did follow the two grace notes, so the grace notes were within the last IOI of the gesture. This made it possible to ask about the relative timing of the grace notes within the IOI (see below). In some further cases, however, one or both of the grace notes coincided with the final melody tone. Thus, SCH consistently played the right-hand octave tones simultaneously with the first grace note (d). BUN (bars 2 and 18) and KUB (bars 2 and 6, second repeat only) played the right-hand tones simultaneously with the second grace note (e). Finally, Cortot (bar 6, in all three performances; also bar 18 in CO1) played all tones simultaneously as a single chord (f), as prescribed only in bar 22 of the score. (BUN also did this once in bar 6.)[12]

To investigate the relative timing of the grace notes within the last IOI of MG2, provided they did fall within that IOI (cases a-c in Figure 10), the IOI was divided into three parts (A, B, and C), defined by the respective tone onsets. Two interval ratios were calculated: A/(BC), which indicates the relative delay of the onset of the first grace note, and B/C, which represents the duration of the first grace-note IOI relative to the second. If the first melody tone and the two grace notes were played as a triplet, for example, the two ratios would be 0.5 and 1, respectively; if the grace notes were played as thirty-second notes following a sixteenth-note rest, the ratios would both be 1.

Figure 11 plots A/(B+C) against B/C for all instances where these ratios could be determined. (Bars 2 and 6 are each represented by a single ratio, averaged across the two repeats.) A wide range of ratios was found, but most data points form a cluster in the lower left quadrant, suggesting that the majority of pianists agreed on a particular timing pattern. The central tendency of that cluster is around A/(B+C) = 0.4 and B/C = 0.5 This means that the grace notes were not played evenly: The first grace note started about 40% into the IOI, and the second grace note was about twice as long as the first, which means it started about 60% into the IOI, on the average. Considering the compressed scale of the ordinate in the figure, however, it is clear that there was wide variation in the relative durations of the two grace notes, even within the cluster of relative conformity. (Some of this variation undoubtedly reflects measurement error magnified by ratio calculation.) What was almost always true, however, is that the first grace note started during the first half of the IOI (A/(B+ C) <1), and the second grace note IOI was longer than the first (B/C < 1).

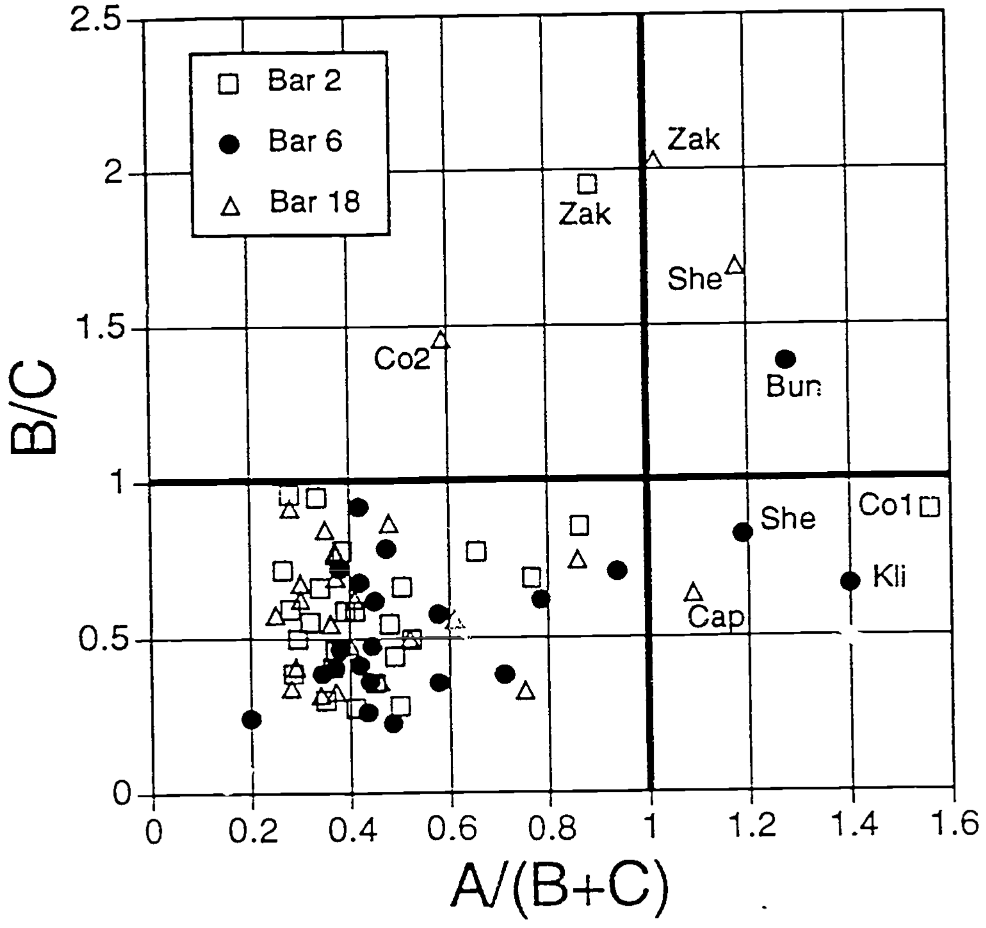


Figure 11. Relative timing of the grace notes within the last IOI of MG2, for cases (a)-(c) in Figure 10. The abscissa shows the ratio of the time before the onset of the first grace note (A) and the remainder of the IOI (BC); the ordinate shows the ratio of the time between the two grace note onsets (B) and the time from the onset of the second grace note to the end of the IOI (C). Ratios for bars 2 and 6 are averaged over the two repeats.

There were some notable exceptions to this pattern, though no individual pianist was consistently deviant. Assuming that these instances are not simply slips of motor control, it would be interesting to know what led individual artists to vary their timing patterns in these ways.

4. MG3a-MG6a. We turn now to a closer examination of the MG chains making up the second half of each phrase. The Type a chain occurs in identical form in bars 2-4 and in bars 18-20 (see Figure 2). It also occurs in varied and abbreviated form in bars 22-24. We will not deal specifically with MG3a and MG3b in bars 22-23, which on the whole were played as in bars 2-3 and 18-19. MG5a and MG6f in bars 23-24, which carried the final large *ritardando*, will be considered separately later on.

As noted earlier, the average timing profiles for bars 2-4 and 18-20 were extremely similar, though bars 18-20 were played at a somewhat slower tempo overall (see Figure 3). In the principal components analysis conducted on bars 0-8, the first factor (see Figure 4, top panel) exhibited a regular sequence of peaks for MG3a-MG6a, which reflect lengthening of accented tones and MG-final IOIs. Detailed examination of the individual timing patterns, however, revealed enormous variety, though not without constraints. No two pianists' interpretations of this MG chain were quite the same, and some artists also changed their patterns significantly between bars 2-4 and bars 18-20. (The two repeats of bars 2-4 will not be considered separately here, though some pianists played even these differently.)

To explore this variability in a reasonably economical way, principal components analysis was used once again, but this time only on the 2 × 18 36 IOIs comprising MG3a-MG6a in bars 2-4 and 18-20, concatenated. From this analysis, six significant factors emerged—more than from the earlier analysis of bars 0-8, where the larger timing deviations in MG1 and MG2 dominated the correlation structure. Thus, there were at least six distinct (i.e., uncorrelated) timing patterns underlying the variation in the data; these patterns are shown in Figure 12. None of them is fully representative of any individual performance, however, and together they account for only 81% of the variance. The factor scores for bars 2-4 and 18-20 were highly similar and are superimposed in the figure. The factor loadings of the individual performances are shown in Table 6.

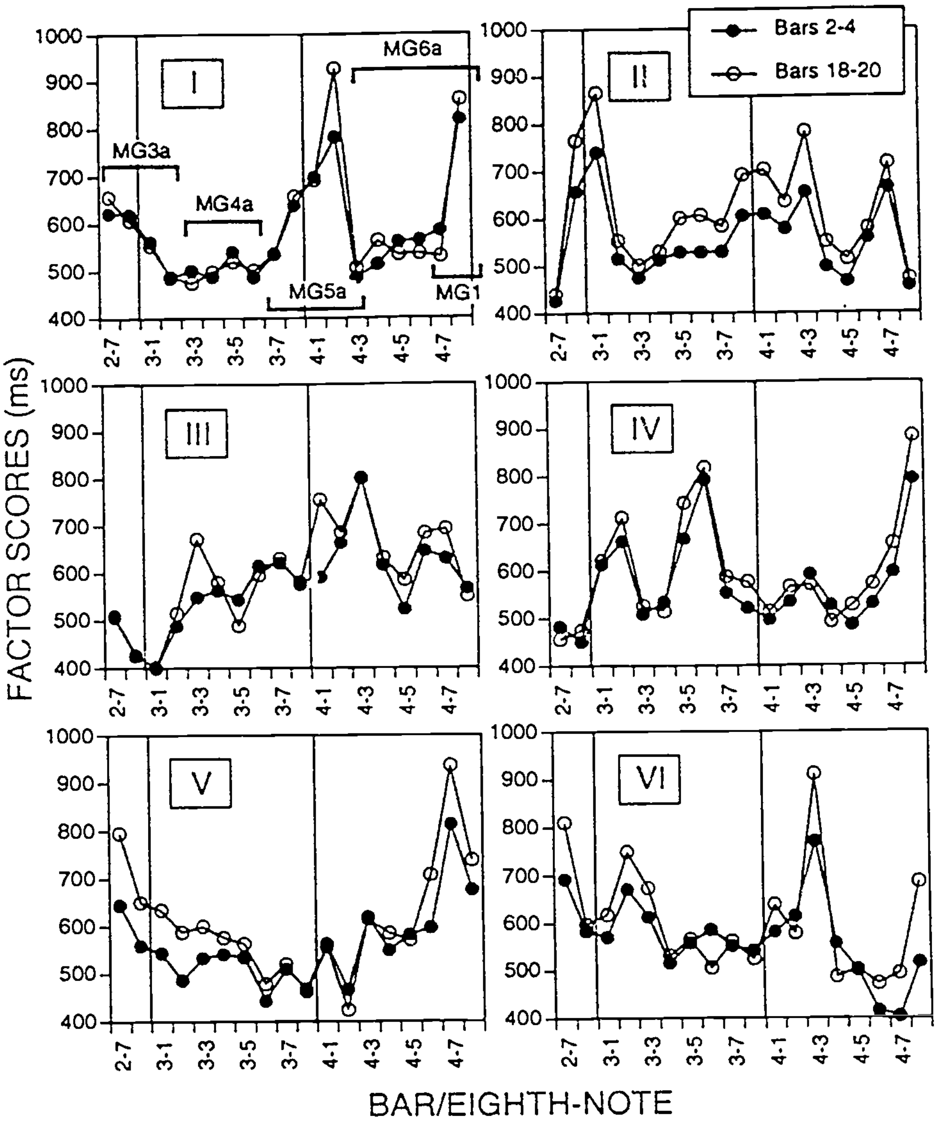


Figure 12. Rescaled factor scores (underlying timing patterns) for the MG chain of Type a in bars 2-4 and 18-20. The abscissa labels refer to bars 2-4.

Table 6. *Sorted rotated factor loadings from the principal components analysis of MG3a-MG6a. (Loadings below 0.4 are omitted.)*

	I	II	III	IV	V	VI
ARR	0.773					
NEY	0.758					
BRE	0.744					
ASH	0.708					
KAT	0.671					0.437
ESC	0.606	0.416				
HO3		0.908				
HO2		0.893				
HO1		0.830				
ARG		0.629	0.402			
CO1			0.873			
CO2			0.832			
CO3			0.784			
ORT		0.446	-0.593			
KUB				0.823		
NOV				0.808		
SCH				0.786		
CUR				0.676	0.541	
MOI					0.864	
SHE	0.564				0.693	
DAV				0.456	0.574	
GIA		0.410	0.535		0.564	
KLI		0.576				0.602
DEM	0.457	0.436				0.565
ZAK	0.427	0.515				0.540
BUN	0.431		0.410	0.491		
KRU	0.414			0.500	0.436	-0.435
CAP				0.455		

Factor I (VAF = 18%) is characterized by acceleration during MG3a; a fast traversal of MG4a; a dramatic *ritardando* within MG5a with maximal lengthening of the final, unaccented note (position 4-2), further augmented in bars 18-20; and a pronounced lengthening of the last IOI in MG6a (position 4-8). The pianists whose interpretations come closest to this pattern are ARR, NEY, BRE, and ASH.

Factor II (VAF = 17%) is the "Horowitz factor." It is characterized by significant lengthening of the accented tone in MG3a as well as of the preceding unaccented tone; slight deceleration during MG4a; final lengthening in MG5a; and a *ritardando* during MG6a, followed by a sudden shortening of the last IOI. Some of these tendencies were exaggerated in bars 18-20 relative to bars 2-4. Horowitz's actual performances (especially HO2 and HO3) resemble this pattern, except that the exaggeration in bars 18-20 is much more dramatic. HO3 differs in that bars 2-4 and 18-20 are executed very similarly, both showing dramatic lengthening in positions 4-2 and 4-3.

Factor III (VAF = 14%) is the "Cortot factor." It is characterized by a fast start; progressive slowing down during MG4a and MG5a, with final lengthening in MG5a; and some shortening of the final IOI in MG6a. Some additional peaks appear in bars 18-20. Cortot's real performances are similar, especially CO1 and CO2, except that the tendencies in bars 18-20 are much more exaggerated. In CO3, there is much lengthening in position 20-2 instead of 20-3 (read 4-2 and 4-3 on the abscissa in Figure 12). Note the sizeable *negative* loading of ORT.

Factor IV (VAF = 14%) shows pronounced final lengthening in both MG3a and MG4a; a "flat" MG5a with slight final lengthening; and a smooth *ritardando* through MG6a, with a very long last IOI. Bars 2-4 and 18-20 are executed very similarly. This beautifully regular pattern comes closest to the actual interpretations of KUB, NOV, and SCH.

Factor V (VAF = 11%) shows a rather flat pattern, with initial acceleration and pronounced final deceleration, but with the final IOI shorter that the penultimate one. This pattern is the least differentiated, as if the whole MG chain were thought of as a single gesture. This pattern comes closest to MOI's performance, though he does show some small local peaks in positions 3-5 and 4-1.

Factor VI (VAF = 7%), finally, begins with a pattern for MG3a and MG4a that is very nearly the opposite of Factor II; MG5a shows striking final lengthening; and MG6a shows only a small lengthening of the final IOI. There is an overall trend to accelerate during the MG chain. This pattern is not representative of any individual performance and occurs only in mixed patterns, such as those of KLI, DEM, and ZAK, all of which exhibit pronounced lengthening in position 4-3.

In summary, the information contained in Figure 12 and Table 6 offers only a moderate amount of data reduction. The timing patterns of individual artists were remarkably diverse. Nevertheless, there are many possible patterns that never occurred, such as lengthening in position 3-4 (the second IOI of MG4a), which clearly was treated as transitional by all pianists. Nor were the patterns random in any way; the majority of the artists produced very similar timing patterns in bars 2-4 and 18-20, and also in the two repeats of bars 2-4 (which were averaged here). The performances with the highest consistency between bars 2-4 and 18-20 were ARR, CAP, HO1, and KUB, whereas the most variable ones were BUN, CO1, CO2, HO2, HO3, and ORT. That Cortot's and Horowitz's within-performance variations were carefully planned is suggested by the fact that they were replicated in different performances (only HO1 retained the pattern of bars 2-4 for bars 18-20). Whether BUN's and ORT's inconsistencies were similarly planned is not known, since only a single performance by each pianist was available for examination.

5. *MG3b-MG6b*. The Type b version of the MG chain occurs three times in the piece: first in bars 6-8, and then in bars 10-12 and 14-16. Because of the substantial *ritardandi* that occur during MG6b, it was decided to treat the last six IOIs of each chain separately. The remaining 3 × 12 = 36 IOIs were subjected to principal components analysis, which yielded six significant factors, just as for the Type a passage. They accounted for 80% of the variance. Their timing profiles are shown in Figure 13, and the matrix of factor loadings is shown in Table 7. As for the Type a chain, the factor patterns are only rarely representative of individual artists' patterns, which more often are a combination of several factor patterns. It may be recalled that the grand average timing profile showed a fairly regular lengthening of pre-accented and accented IOIs in each MG, which really represents final lengthening of the accompanying secondary MG, which ends with an accented tone. The factor profiles essentially represent a varying focus on individual MGs in the chain.

Factor I (VAF = 22%) shows substantial lengthening in MG3b, much less in MG4b, and somewhat more in MG5b. The timing profiles for the three instances of the MG chain are very similar. The pianist most closely associated with this pattern is ARR, and a number of other artists, including Horowitz, show substantial loadings in this factor. (In contrast to the Type a chain, there was no "Horowitz factor" here.)

Factor II (VAF = 21%) is characterized by total absence of lengthening in MG3b and variability during MG4b and MG5b. usually with more pre-accentuation lengthening than accentuation lengthening. Two performances showing this pattern in relatively pure form are CO1 and KRU, and several other performances (including CO2) show substantial loadings. Thus this was again a kind of "Cortot factor," though it was not unique to Cortot. However, CO3 deviates from this pattern and is responsible for a unique factor, after all (Factor VI).

Factor III (VAF = 14%) is characterized by a flat pattern (bars 6-7) or pronounced *accelerando* (bars 10-11 and 14-15) during MG3b, variable lengthening in MG4b, and pronounced lengthening in MG5b. NEY, KAT, and DAV are the best representatives.

The remaining three factors account for much less variance (8% in the case of Factors IV and V, 6% in that of Factor VI), and each of them is largely associated with one particular artist. These three artists, which earlier analyses have already revealed to be somewhat eccentric, are ORT, BUN, and CO3. Thus, ORT put increasing emphasis on MG4b in successive renditions, BUN showed a very atypical peak in position 7-7 and tremendous lengthening in MG4b in bar 7 only, and CO3 showed no lengthening in MG3b and MG4b, and a very variable execution of MG5b.

On the whole, within-performance variability across the three instances of the Type b chain seemed larger than for the Type a chain, whereas between-performance variability seemed somewhat lower. The former observation is not surprising, for the two MG chains of Type a are identical, whereas the three chains of Type b differ harmonically and melodically. No pianist showed nearly identical patterns for all three instances of Type b. Some pianists were markedly more variable than others, however, not to say erratic. BUN leads the list, which also includes ARG, CO3, DAV, and MOI. The relatively most consistent pianists were ASH, DEM, HO2, KLI, NEY, and ZAK. Some pianists (BRE, KRU, SCH) were notable for their relatively flat and uninflected rendering of the whole Type b chain.13

6. *The ritardandi.* As is evident from the grand average timing pattern plotted in Figure 3, major *ritardandi* occurred at the ends of periods and phrases. The largest *ritardando* is in bars 23-24, at the end of the piece. A substantial slowdown occurs also in bar 16, at the end of period B. Less pronounced *ritardandi* occur in bar 8, at the end of period A, and in bar 12, at the end of phrase b2. Thus there are four locations where major *ritardandi* are observed. It may be asked whether the temporal shapes of these *ritardandi* followed any common pattern.

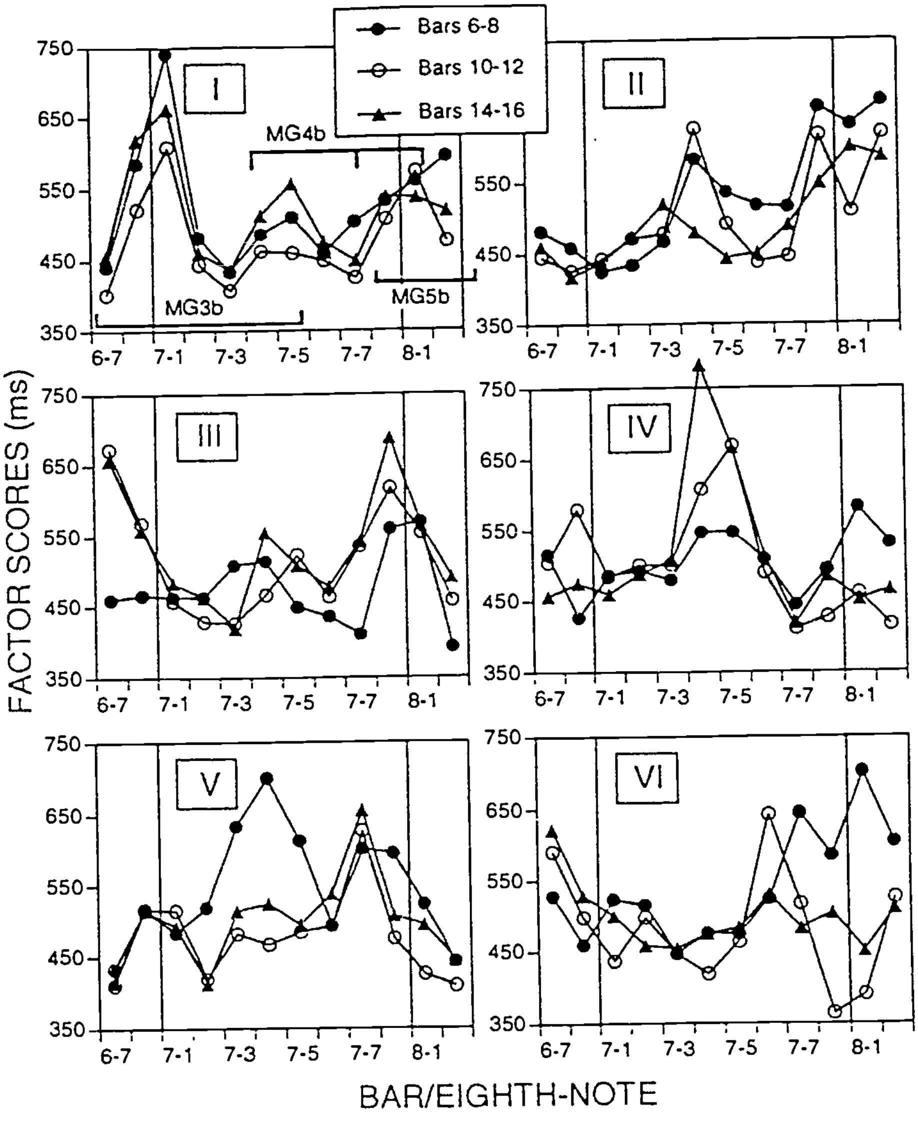


Figure 13. Rescaled factor scores (underlying timing patterns) for MG chain of Type b in bars 6-8, and 10-12, and 14-16. The abscissa labels refer to bars 6-8.

Table 7. *Sorted rotated factor loadings from the principal components analysis of MG3b-MG5b. (Loadings below 0.4 are omitted.)*

	I	II	III	IV	V	VI
ARR	0.923					
HO1	0.788					
ARG	0.745					
HO3	0.740					
KLI	0.715					
MOI	0.701					
ASH	0.648	0.521				
DEM	0.589		0.582			
ESC	0.553			0.490		
GIA	0.531	0.455				
CAP	0.515			0.438	0.450	
CO1		0.860				
KRU		0.829				
BRE		0.733				
CO2		0.699				
NOV	0.418	0.691				
SHE		0.675			0.407	
CUR		0.664	0.488			
SCH		0.629	0.404			
KUB		0.530		0.416		
NEY			0.814			
KAT			0.746			
DAV			0.721			
ZAK	0.407		0.642		0.416	
ORT				0.893		
BUN					0.756	
CO3						0.863
HO2	0.413				0.452	

This question was addressed previously by Sundberg and Verrillo (1980) and Kronman and Sundberg (1987) with regard to the *ritardandi* observed at the ends of performances of "motor music," mostly by Bach. Sundberg and Verrillo plotted an average "retard curve" in terms of inverse IOIs and observed that it could be described in terms of two linear functions: a gradual slowdown followed by a more rapid one. The latter phase often included only three data points, however, and tended to coincide with the last melodic gesture in the music. Kronman and Sundberg abandoned the bilinear model and interpreted the average retard curve as a single continuous function. That function was described well by expressing local tempo (1/IOI) as the square root of the normalized distance from a hypothetical zero point located one beat beyond the onset of the final tone. Kronberg and Sundberg speculated that this function may be an allusion to physical deceleration in natural activities such as walking.

That the average *ritardando* followed a quadratic function is intriguing given the present results concerning the temporal shape of MG2. The *decelerando* in that ascending melodic gesture may be considered an instance of a local *ritardando*, and the principles involved may be quite the same. A square-root (convex) function fitted to reciprocal IOIs implies a quadratic (concave) function fitted to the original IOIs. However, it is difficult to reach strong conclusions from Sundberg's examination of timing patterns averaged across performances of different music. The melodic grouping structure must be taken into account, which was done only to a very limited extent by Sundberg and his colleagues.

This is much easier, of course, when dealing with multiple performances of a single piece of music, as in the present case. Inspection of the grand average timing profile in Figure 3 suggests that *ritardandi* progress smoothly within melodic gestures but are interrupted between gestures. Thus, the final *ritardandi* in bars 12 and 16 (read bar 8 on the abscissa) can be seen to begin with the second IOI. but a "reset" occurs after the fourth IOI, which corresponds to a gestural boundary. The final lengthening of MG5b causes the initial *ritardando* to "overshoot" its trajectory, which is resumed with the onset of MG6b. These *ritardandi* thus are divided into two sequences of 3 and 4 IOIs, respectively. Similarly, in the large final *ritardando* in bars 23-24, which begins as early as position 23-3 or 23-4, two resets occur, corresponding to the ends of MG4a and MG5a, which in this instance are unmistakably indicated in the score by "commas." This grand *ritardando* thus is divided into three groups of IOIs: 3(4), 4, and 2. Because meaningful fitting of quadratic curves requires at least four data points, the following analyses examine one sequence of four coherent IOIs from each of the three *ritardandi*, forming either part of MG6b or of MG5a. The analyses were analogous to those conducted on MG2.

In each case, it was found that a quadratic curve fit the average IOI pattern quite well, perfectly so in the case of MG5a. These curves are shown in Figure 14. Quadratic functions were subsequently fitted to all 28 individual performance timing patterns of each *ritardando* section. Most fits ranged from excellent to acceptable (see Footnote 9). In the case of MG6b (bar 12), there were five clearly deviant cases (ARG, HO3, KUB, and ORT, all of whom shortened the last IOI, and KLI, who showed a rather flat pattern); for MG6b (bar 16) there was none; for MG6a there was one (ARG, who shortened the last IOI).

The coefficients of the quadratic equations describing the acceptable (and better) fits were found to be highly correlated in all instances, just as for MG2. Thus it was again possible to construct families of parabolas that capture a major portion of the individual variation in the shapes of the *ritardandi*. These curves are shown in Figure 15. Taking into account the different time scales on the abscissa, it seems that each *ritardando* has its distinctive range of variation, but all share the property of being largely parabolic in shape. These results for strictly intragestural *ritardandi* support the observations on more heterogeneous materials by Sundberg and Verrillo (1980) and Kronman and Sundberg (1987).

A final analysis was conducted on the last two IOIs of the piece. While it did not make sense to fit any function to them, their correlation could be examined, which was 0.87. Thus, the longer the penultimate IOI, the longer the final IOI.[14] Not only was the relationship quite linear, but the regression line passed almost through the origin, suggesting that the two IOIs were in a constant proportion (1:1.8).

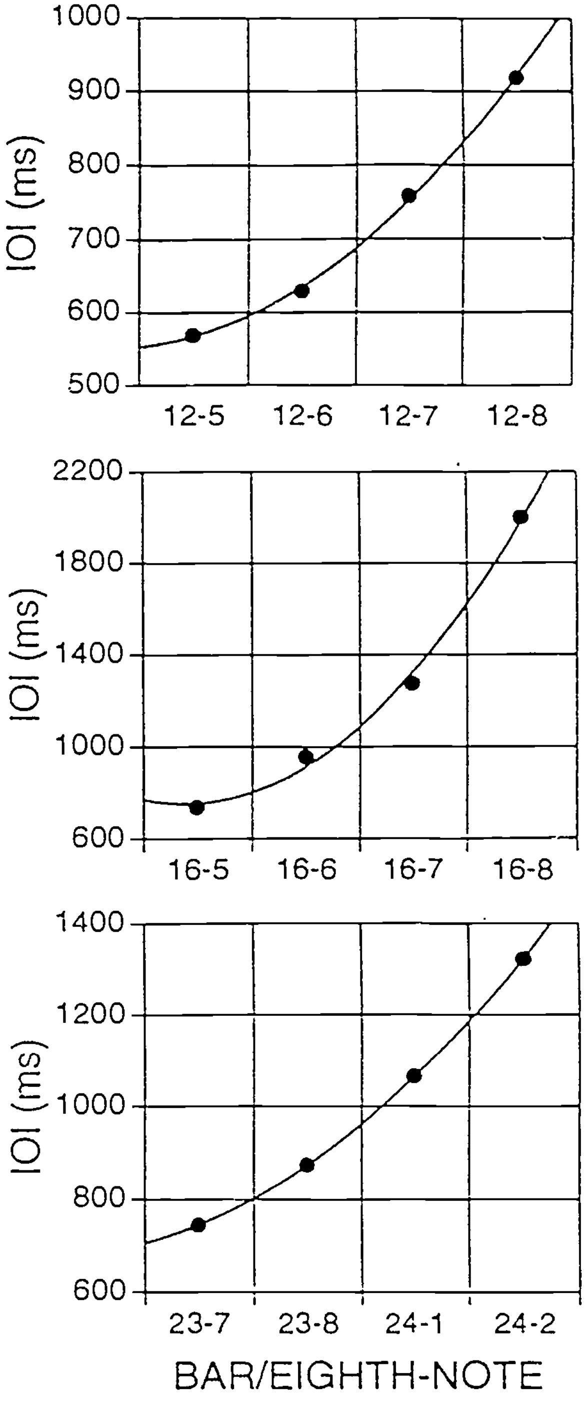


Figure 14. Ritardando functions: Quadratic functions fitted to the grand average (geometric mean) IOIs in MG6b (bars 12 and 16) and in MG5a (bars 23-24).

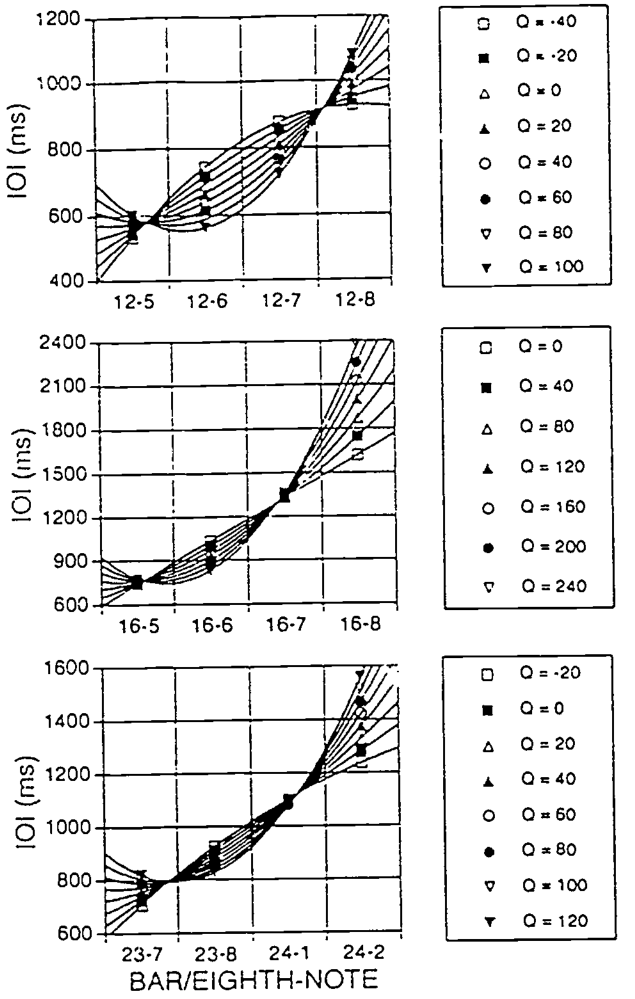


Figure 15. Families of parabolic timing functions for the *ritardandi* in MG6b (bars 12 and 16) and in MG5a (bars 23-24).

V. GENERAL DISCUSSION

The present comprehensive analysis of expressive timing patterns in 28 performances of Schumann's "Träumerei" provides an objective view of the commonalities and differences among great artists' interpretations of one of the masterpieces of the piano literature. At first glance, the differences are perhaps more striking than the commonalities. There is ample material here to support the view that every artist's performance is, in some sense, unique and unlike any other artist's, even if just one physical dimension (timing) is considered. Even the same artist's performances on different occasions, while demonstrably similar, are sufficiently different to be considered distinct and individual events. Yet, there are also significant commonalities that apparently reflect constraints on performance, at least on those performances that have been deemed suitable for commercial distribution. The individual variations among performances largely take place within these constraints, although there are always exceptions, representing conscious or unconscious transgressions of the boundaries established by musical convention.

How should those boundaries be characterized? And are they purely conventional (i.e., arbitrary), or do they represent more general laws of motor behavior and perception that music performance must conform to in order to be naturally expressive?

It is probably futile to attempt to characterize the boundaries of acceptable performance practice. They are manifold and are likely to contract and expand as a function of many factors. It is theoretically more parsimonious to conceive of a performance ideal (norm, prototype) that lies at the center of the hypothetical space enclosed by the boundaries. This ideal may be thought of as a relatively abstract specification that contains free parameters, so that a multiplicity (if not an infinity) of concrete performances can be generated that all more or less satisfy the norm. Because of the enormous complexity of serious music, any concrete performance ideal is very difficult to attain; and, if one were attained, the artist would probably change it the next time, because art thrives on variety. However, the underlying abstract ideal may remain constant, as it embodies generally accepted rules of performance practice. Concrete realizations are conceived, perceived, and judged with reference to the underlying norm, but deviation from the norm (within certain limits of acceptability) can—to some extent, must—be an artistic goal. That is, diversity is as necessary as is commonality: Both uniformity and lack of an aesthetic standard are detrimental to art.

One obvious free parameter is tempo. Even though there may be an "ideal tempo" for a piece of music, this is again an abstraction. There is in fact a *range* of acceptable tempi, and individuals may differ considerably in what they consider "the" ideal tempo. This is illustrated by the performance sample examined here, which represents a very wide range of tempi, nearly all of which seem acceptable to a musical listener (the author must rely on subjective judgment here), even though some sound clearly slow, while others sound fast. To the author, only ESC really sounds "too slow" (though perhaps even that judgment might change if his performance were heard in the context of the whole "Kinderszenen" suite), while that of DAV sounds "too fast," though apparently (Clara) Schumann wanted it that way (see Repp, 1992d). An unusually fast "Träumerei" indeed seems to come across better than an unusually slow one.

The present investigation substantiates two abstract timing constraints embodied in the hypothetical performance ideal. One of them concerns the temporal marking of the melodic/rhythmic structure, the other one has to do with the temporal shaping of individual melodic gestures, particularly of the *ritardandi* within them. The first constraint is by now well known and is implemented in Todd's (1985) model of expressive timing at the phrase level. The principle is that boundaries in the hierarchical grouping structure are generally marked by *ritardandi* whose extent is roughly proportional to the "depth" of the boundary. Thus the most extensive *ritardando* occurs at the end of the piece, where boundaries at all levels coincide; substantial *ritardandi* occur at the ends of major sections, such as 8-bar periods; smaller *ritardandi* occur at the ends of individual phrases and gestures. A performance system such as that devised by Todd may come close to the principles embodied in the performance ideal, though it is not known at present whether a performance strictly following such rules would in fact be perceived as "ideal" by listeners. There is some perceptual evidence, however, that listeners—even those without much musical education—expect phrase-final and gesture-final lengthening to occur (Repp, 1992a).

What is clear from the performances examined here is that variability increases at lower levels of the structural hierarchy. Virtually every performer observes the major *ritardandi* at the ends of major sections, though in different degrees. When the complete performance timing profiles were subjected to principal components analysis, only a single factor emerged, which reflected the qualitative conformity in that regard. When the analysis was restricted to the first 8 bars only, four factors emerged. When the analysis focused on half a phrase, omitting major *ritardandi*, six factors emerged. It seems paradoxical that the number of independent factors increases as the number of data points decreases. What this reflects is the increasing pattern variability at lower levels of the structural hierarchy.

This variability is not likely to reflect a decrease of control over precision in timing, though some of it may. Pianists at the level of accomplishment studied here can control their timing patterns down to a very fine grain. What the variability presumably reflects is a relaxation of the performance constraints imposed by the grouping structure at lower levels of the hierarchy. Not only are the boundaries between individual melodic gestures perhaps less definite than those between larger units, but they are weaker determinants of the timing pattern and compete with other local factors including harmonic progression, melodic pitch contour, and texture. It also seems that the major source of timing variation is not artists' choice of unexpected locations for expressive lengthening (though some of this occurred, too) but varying degrees of emphasis on expected locations. Thus, pianists often omitted lengthening where the detailed grouping structure might have predicted it, whereas they overemphasized other grouping boundaries, as if to compensate. Sometimes this resulted in the creation of hypergestures, which combined two or three elementary melodic motifs. In other words, the lowest level of the grouping hierarchy is structurally flexible; it permits the marking of group boundaries but does not prescribe it. Artists choose from among the possibilities in a manner comparable to varying focus in a spoken sentence.

The other constraint observed in the present data is more novel and more tentative. It is that, within melodic gestures requiring a *ritardando* for whatever reason, this local tempo change is best executed such that successive IOIs follow a parabolic function. This constraint was not only observed in the majority of performances at four different locations in the music (nine, if the different occurrences of MG2 are counted separately), but it also is in agreement with the observations of Kronman and Sundberg (1987) on the shape of final *ritardandi*, as originally described by Sundberg and Verrillo (1980). Sundberg and Verrillo also provided perceptual data suggesting that listeners prefer *ritardandi* corresponding to the original (quadratic) timing curves over other possible timing profiles, and a recent study by this author (Repp, 1992b) on the perceptual evaluation of different timing patterns for MG2 led to similar conclusions.

It is not clear at present whether the parabolic timing constraint can also be applied to melodic gestures that do not include any pronounced *ritardando*; certainly other factors, such as accent location, would have to be taken into account. The larger the number of successive tones in a melodic gesture, the stronger the constraint is likely to be; a minimum of five tones (4 IOIs) is required. It also remains to be seen whether a similar timing constraint operates at higher levels of the grouping hierarchy. Todd (1985) postulated a parabolic timing function as the prototype for within-group timing at the phrase level, though apparently more for the sake of convenience than for any stringent theoretical or empirical reason. Kronman and Sundberg (1987) hint at a grounding of the parabolic timing constraint in elementary principles of (loco)motion, but without elaborating on this hypothesis. Todd (1992b) has claimed that a piecewise-linear function is sufficient to describe local timing changes during a performance, but his own (rather limited) data suggest that a quadratic function provides a better fit. At this point, the general hypothesis may be stated that a parabolic timing profile is in some sense "natural" for both performer and listener, and probably not for purely conventional reasons. A better understanding of the origins of this constraint may elucidate the meaning of the motion metaphor that is often applied to music and its performance (cf. Todd, 1992a, 1992b; Truslit, 1938).

The present performance analyses revealed no clustering of individual artists according to sex, age, or national origin. Their individuality apparently transcended these other factors, whose relevance to music performance may be questioned in any case. There is no doubt, however, that in this sample of highly accomplished artists, perhaps precisely because of their level of artistry, some performances seemed unusual and even deviant. Subjective impressions will have to suffice for the time being: Long before objective measurements confirmed the actual deviations in their timing patterns, the performances by Argerich, Bunin, and Cortot, and to a lesser degree those by Horowitz and Ortiz, struck the author as eccentric and distorted. Interestingly, some of these pianists (especially Bunin) also turned out to be inconsistent within their own performance, as if they had no fixed concept and were exploring possibilities. Other artists, however, were remarkably consistent, for example Arrau. The author's subjective impressions were not only highly reliable on relistening, but they also correspond to what professional critics have to say about some of these pianists' performances. Certainly, Horowitz and Cortot are two of the most unusual pianists of this century, and their greatness may lie precisely in their individuality, which challenges the listener. However, in the context of listening to 28 different performances in sequence, which heightens one's sensitivity to differences and perhaps increases one's reliance on an internalized performance ideal, the unusual performances do not fare so well. The author's favorites are the performances by Curzon, Brendel, and Ashkenazy, which usually were in the middle field in the various analyses reported above, and hence were mentioned only rarely. The author's aesthetic ideal seemed to correspond to the central tendency of the sample examined here. Whether this ideal represents a stable representation of traditional performance norms or whether it is a form of psychophysical adaptation to the range of the performance sample is an intriguing question that warrants further investigation. Also, the absence of the context of the preceding and following pieces of "Kinderszenen" must be acknowledged; interpretations that sound unusual in isolation may sound more convincing in context. Finally, it must be noted that the author's impressions derived not only from the timing patterns of these performances, but also from their intensity patterns (both "horizontal" and "vertical"), their articulation and pedaling, and their overall sound quality. There are many parameters that contribute to the subjective impression of a performance, but the timing pattern is probably the most important one. Nevertheless, future performance analyses will have to consider these other parameters (which are much more difficult to measure in recorded performances) as well as the subjective impressions of more than one experienced listener. Ultimately, we would like to know *why* individual performers play the way they do, and what message they are sending to listeners (cf. Kendall & Carterette, 1990).

The present research illustrates one of two complementary approaches to the objective investigation of musical performance. The other approach is represented by the work of Sundberg and his colleagues on a system of performance rules (e.g., Friberg, 1991; Sundberg, 1988; Sundberg, Friberg, & Frydén, 1989) and by Todd's (1985, 1992a, 1992b) modelling of musical motion. These models are ingenious and important, but they need to be tested on large performance data bases, which do not exist at present. In this study, a very modest beginning was made towards accumulating and organizing data of sufficient depth and variety (though still restricted to a single composition and a single parameter, timing) to present a challenge and testing ground for emerging models of music performance.[15] Soon, of course, much more extensive data bases should become available with the help of technological marvels such as the MIDI-controlled grand piano. There are exciting times ahead for research on music performance, one of the most advanced and culturally significant human skills.

REFERENCES

Bengtsson, I., & Gabrielsson, A. (1980). Methods for analyzing performance of musical rhythm. *Scandinavian Journal of Psychology, 21*, 257-268.

Brendel, A. (1981). Der Interpret muss erwachsen sein. Zu Schumann's "Kinderszenen." *Musica, 6*, 429-433.

Chissell, J. (1987). Liner notes for Claudio Arrau's recording (Philips 420-871-2).

Clarke, E. F. (1982). Timing in the performance of Erik Satie's 'Vexations.' *Acta Psychologica, 50*, 1-19.

Clarke, E. F. (1983). Structure and expression in rhythmic performance. In P. Howell, I. Cross, & R. West (Eds.), *Musical structure and cognition* (pp. 209-236). London: Academic Press.

Clarke, E. F. (1985). Some aspects of rhythm and expression in performances of Erik Satie's "Gnossienne No. 5." *Music Perception, 2*, 299-328.

Clarke, E. F. (1988). Generative principles in music performance. In J. A. Sloboda (Ed.), *Generative processes in music* (pp. 1-26). Oxford: Clarendon Press.

Cline, E. (1985). *Piano competitions: An analysis of their structure, value, and educational implications*. Doctoral dissertation, Indiana University.

Clynes, M. (1983). Expressive microstructure in music, linked to living qualities. In J. Sundberg (Ed.), *Studies of music performance* (pp. 76-181). Stockholm, Sweden: Publication issued by the Royal Swedish Academy of Music No. 39.

Clynes, M. (1987). What can a musician learn about music performance from newly discovered microstructure principles (PM and PAS)? In A. Gabrielsson (Ed.), *Action and perception in rhythm and music* (pp. 201-233). Stockholm, Sweden: Publication issued by the Royal Swedish Academy of Music No. 55.

Friberg, A. (1991). Generative rules for music performance: A formal description of a rule system. *Computer Music Journal, 15*, 56-71.

Gabrielsson, A. (1974). Performance of rhythm patterns. *Scandinavian Journal of Psychology, 15*, 63-72.

Gabrielsson, A. (1985). Interplay between analysis and synthesis in studies of music performance and music experience. *Music Perception, 3*, 59-86.

Gabrielsson, A. (1987). Once again: the theme from Mozart's Piano Sonata in A major (K. 331). In A. Gabrielsson (Ed.), *Action and perception in rhythm and music* (pp. 81-103). Stockholm, Sweden: Publication issued by the Royal Swedish Academy of Music No. 55.

Gabrielsson, A. (1988). Timing in music performance and its relation to music experience. In J.A. Sloboda (Ed.), *Generative processes in music* (pp. 27-51). Oxford: Clarendon Press.

Gabrielsson, A., Bengtsson, I., & Gabrielsson. B. (1983). Performance of musical rhythm in 3/4 and 6/8 meter. *Scandinavian Journal of Psychology, 24*, 193-213.

Hartmann, A. (1932). Untersuchungen über metrisches Verhalten in musikalischen Interpretationsvarianten. *Archiv fur die gesamte Psychologie, 84*, 103-192.

Horowitz, J. (1990). *The ivory trade: Music and the business of music at the Van Cliburn International Piano Competition*. New York: Summit Books.

Kendall, R. A., & Carterette, E. C. (1990). The communication of musical expression. *Music Perception, 8*, 129-163.

Kronman, U., & Sundberg, J. (1987). Is the musical ritard an allusion to physical motion? In A. Gabrielsson (Ed.), *Action and Perception in Rhythm and Music* (pp. 57-68). Stockholm: Publications issued by the Royal Swedish Academy of Music No. 55.

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Lipman, S. (1984). Glenn Gould: His dissent/An obituary (originally published in 1982). In S. Lipman, *The house of music: Art in an era of institutions* (pp. 79-95). Boston: Godine.

Lipman, S. (1990). *Arguing for music/Arguing for culture*. Boston: Godine.

Lussy, M. M. (1882). *Musical expression: Accents, nuances, and tempo in vocal and instrumental music*. London: Novello.

Palmer, C. (1989). Mapping musical thought to musical performance. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 331-346.

Povel, D.-J. (1977). Temporal structure of performed music: some preliminary observations. *Acta Psychologica, 41*, 309-320.

Repp, B. H. (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. *Journal of the Acoustical Society of America, 88*, 622-641.

Repp, B. H. (1992a). Probing the cognitive representation of musical time: Structural constraints on the perception of timing perturbations. *Cognition, 44*, 241-281.

Repp, B. H. (1992b). A constraint on the expressive timing of a melodic gesture: Evidence from performance and aesthetic judgment. *Music Perception, 10*, 221-242.

Repp, B. H. (1993a). Objective performance analysis as a tool for the musical detective. *Journal of the Acoustical Society of America, 93* (forthcoming).

Repp, B. H. (1993b). On determining the global tempo of a temporally modulated music performance.

Réti, R. (1951). *The thematic process in music*. New York: Macmillan.

Riemann, H. (1884). Der Ausdruck in der Musik. In P. Graf Waldersee (Ed.), *Sammlung musikalischer Vorträge* (pp. 43-64). Leipzig: Breitkopf und Härtel.

Seashore, C. E. (1936) (Ed.) *Objective analysis of music performance*. Iowa City, IA: The University Press (University of Iowa Studies in the Psychology of Music, Vol. IV).

Seashore, C. E. (1938). *Psychology of music*. New York: McGraw-Hill. (New York: Dover Publications, 1967.)

Seashore, C. E. (1947). *In search of beauty in music: A scientific approach to musical esthetics*. New York: Ronald Press.

Shaffer, L. H. (1980). Analysing piano performance: a study of concert pianists. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 443-455). Amsterdam: North-Holland.

Shaffer, L. H. (1981). Performances of Chopin, Bach, and Bartok: Studies in motor programming. *Cognitive Psychology, 13*, 326-376.

Shaffer, L. H. (1984). Timing in solo and duet piano performances. *Quarterly Journal of Experimental Psychology, 36A*, 577-595.

Shaffer, L. H. (1989). Cognition and affect in musical performance. *Contemporary Music Review, 4*, 381-389.

Shaffer, L. H., Clarke, E. F., & Todd, N. P. (1985). Metre and rhythm in piano playing *Cognition, 20*, 61-77.

Shaffer, L. H., & Todd, N. P. (1987). The interpretative component in musical performance. In A. Gabrielsson (Ed.), *Action and perception in rhythm and music* (pp. 139-152). Stockholm, Sweden: Publication issued by the Royal Swedish Academy of Music No. 55.

Stein, E. (1962). *Form and performance*. New York: Knopf (New York: Limelight Editions, 1989.)

Sundberg, J. (1988). Computer synthesis of music performance. In J. A. Sloboda (Ed.) *Generative processes in music* (pp. 52-69). Oxford, UK: Clarendon Press.

Sundberg, J., Friberg, A., & Frydén, L. (1989). Rules for automated performance of ensemble music. *Contemporary Music Review*, 3, 89-109.

Sundberg, J., & Verrillo, V. (1980). On the anatomy of the retard: A study of timing in music. *Journal of the Acoustical Society of America*, 68, 772-779.

Todd, N. P. (1985). A model of expressive timing in tonal music. *Music Perception*, 3, 33-58.

Todd, N. P. (1989). A computational model of rubato. *Contemporary Music Review*, 3, 69-88.

Todd, N. P. McA. (1992a). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America*, 91, 3540-3550.

Todd, N. P. McA. (1992b). The kinematics of musical expression. *Journal of the Acoustical Society of America* (in press).

Traub, A. (1981). Die "Kinderszenen" als zyklisches Werk. *Musica*, 6, 424-426.

Truslit, A. (1938). *Gestalt und Bewegung in der Musik*. Berlin-Lichterfelde: Chr. Friedrich Vieweg.

FOOTNOTES

[*] *Journal of the Acoustical Society of America*, 92, 2546-2568 (1993).

[1] An effort is made in this manuscript to distinguish between *notes*, which are graphic symbols, and *tones*, which are sound events. However, when it comes to grace notes, the distinction cannot easily be made, since "grace tone" is not an acceptable term and "grace-note tone" is awkward. It should be understood, then, that "grace note" refers either to the notated symbol or the resulting sound, depending on the context. The same ambiguity holds for "chord," although "tone cluster" is an acceptable term for the sound event that may be used on occasion.

[2] In a third case (NEY), it was discovered that the two repeats were virtually identical from bar 5 on, suggesting duplication by the recording engineers (see Repp, 1993a).

[3] In the designation for a "position," such as 5-8, the two numbers stand for the bar and the eighth-note IOI within it, respectively. In an expression such as "bars 5-8," however, the two numbers stand for the first and last bars in the range referred to.

[4] Most of the performances in the latter group strike the author as mannered. It seems that these pianists deliberately tried to play differently from the norm, but were not willing or able to do so consistently. Perhaps they intended to convey an improvisatory quality.

[5] Todd (1992a, 1992b) has recently revised and extended his model of expressive timing. An application of this model to the present data would be most interesting but exceeds the scope of this paper.

[6] That is to say, there may be an implicit, *expressively modulated* eighth-note pulse going through the longer IOIs. As suggested by the extent of the MG boxes in Figure 2, the first of these pulses "belongs to" the preceding MG, but its actual duration cannot be determined unless it is marked by some tonal event in another voice. MG2i, which breaks up the inter-gesture interval in the soprano voice, may track the implicit pulse induced by the primary MG2 and thus may reveal its final lengthening.

[7] Although these performances were by no means identical, they were more similar than almost any pair of performances by different artists. Cortot's three performances intercorrelated between 0.80 and 0.82, and Horowitz's between 0.81 and 0.92. Only three other values in the 28 × 28 intercorrelation matrix exceeded 0.80: CAP/ZAK (0.86), KAT/SHE (0.82), and CUR/KRU (0.81). On the other hand, some of the lowest correlations were observed between Cortot's and Horowitz's performances (0.15 to 0.44), and also between Cortot and several other artists (ARG, BUN, ESC, KLI, NEY, NOV). Only two other correlations fell below 0.30: BUN/MOI (0.26) and HO1/SCH (0.26). Most correlations were between 0.4 and 0.7. Thus Cortot and Horowitz were rather extreme cases in the present sample, which agrees with their general reputation as highly individual artists.

[8] The possible covariation of adjacent as well as distant IOIs was investigated in a principal components analysis on the large (190 × 190) matrix of intercorrelations among all IOIs, computed across the 28 performances. Its purpose was to "retrieve" the melodic grouping structure through the factors extracted, on the expectation that IOIs would covary more strongly within MGs than between. This expectation was not fulfilled; the analysis yielded a very large number of significant factors, none of which captured much of the variance. They did reflect correlations among some adjacent IOIs and especially between IOIs in structurally analogous positions across phrases of the same type. However, for example, there was no significant relationship between the timing of MG1 and MG2, nor even between the beginning and the end of MG2.

[9] Unfortunately, the software used for curve fitting (DeltaGraph) did not provide a measure of goodness of fit. Therefore, statements about this aspect of the data will remain impressionistic in this paper.

[10] The L coefficient can be understood as follows: If the constant C (without loss of generality) is assumed to be 0, the parabola must pass through the origin, the (0,0) point. L represents the slope of a tangent through the origin. It is zero when the minimum of the parabola is located at the origin. It becomes negative when the minimum of the parabola moves to a positive value along the abscissa, but since the left branch of the parabola must still pass through the origin, this implies a negative value along the ordinate for the minimum.

[11] The correlations varied systematically across the six instances of MG2: The highest correlations (0.91-0.93) were obtained for bars 1-2 and 17-18; lower correlations (0.85-0.86) hold for bars 9-10 and 13-14; and the lowest correlations (both 0.79) were found for bars 5-6 and 21-22. The less tight relationship here must be due to variations in global tempo among the pianists, which may be only weakly related to the degree of tempo modulation in MG2.

[12] For that matter, the *fermata* chord in bar 22, which is clearly notated as requiring simultaneity of all tones, was not played by all pianists in that way: DAV and NOV played a grand *arpeggio*, NEY played a partial arpeggio (two grace notes in the left hand, as in bars 2, 6, and 18), and ASH played only the lowest tone in advance. These variants may have been occasioned by small hands (note that three of the pianists are female), for the chord requires a large span.

[13] One detail skipped in this analysis is the timing of the melody grace note during the last IOI of MG5b (position 8-4). Its onset generally occurred near the middle (40-60%) of the IOI. A few pianists played it a little earlier (CAP, CO2) or later (ARG, ESC, SCH, HO2, ARR). Two artists (CO3, DAV) omitted it altogether.

[14] In a final display of eccentricity, BUN produced a terminal IOI of nearly 5 seconds duration, with ORT not far behind.

[15] The data matrix is available from the author.

Haskins Laboratories Status Report on Speech Research
1992. SR-111/112. 261-264

A Review of *Einführung in die deutsche Phonetik* by Ursula Hirschfeld*

Bruno H. Repp

This video introduction to German phonetics, apparently the first of its kind, was produced by Ursula Hirschfeld who is currently head of the Working Group in Phonetics at the University of Leipzig. This group has been active in applied phonetics for a number of years, with particular attention to the teaching of German as a second language and the development of materials for that purpose. The present video course thus is the outgrowth of extensive practical experience as well as of a solid grounding in phonetic theory. Much of the work was done under rather difficult conditions before the political unification of Germany, and its successful completion reflects the dedication and persistence of Dr. Hirschfeld.

The course is intended for use in teacher training, class instruction, and home study. The 70-minute tape comprises 8 short lessons on the following topics: intonation and stress; unrounded vowels; rounded vowels; diphthongs; plosives; fricatives and fricatives in combination (2 lectures); and nasals. Each lecture features two main protagonists who are seated at a desk facing the viewer: a charming young woman named Christine and a droll life-size puppet (a cross between a university professor and an old Germanic warrior, featuring a movable jaw and a human right hand) named Hermann. Christine speaks the examples while Hermann provides the phonetic explanations (in British English on my tape; German and Polish versions are also available). The explanations are illustrated with tables and graphs which are reproduced in the accompanying booklet, together with all sample utterances, a glossary, and a list of phonetic symbols. Some of the lessons include a display of a schematic vocal tract in motion. All lessons focus briefly on spelling-sound relationships. They further include a brief scene acted out by people in the studio, which uses some of the words practiced previously by Christine. Finally, each lecture concludes with short takes of various passers-by in the streets of different German towns, each of whom speaks some of the target words in a natural and unrehearsed way, presumably elicited by a question from the interviewer. I focus now on each of these components of the didactic action in turn, slicing the pie horizontally, as it were.

It is difficult to find fault with **Christine**. A student of phonetics in real life at the time of recording, she speaks with precision, behaves naturally in front of the camera, and displays a winning smile that makes her fun to watch. She introduces the topic of each lecture in German before Hermann takes over in English, and while her introductory sentence is not likely to be understood by beginners, it does set the scene in the target language. After Hermann introduces some phones or explains some rule, Christine produces the examples, usually first in isolation, then in selected words or short sentences. Each utterance is repeated once, the second time somewhat more casually than the first. There is no time for the student viewer to insert his/her own repetition; while this may be a shortcoming for home study, pauses on the tape would naturally have slackened the pace.

The variety of German being spoken by Christine is not defined explicitly in the booklet. Obviously, it is meant to be representative of

I am grateful for the support of Haskins Laboratories during the preparation of this review and for the comments of several colleagues who viewed the video tape with me. Special thanks to Ursula Hirschfeld for her helpful comments on an earlier draft.

standard German, and indeed it is, in so far as the Northern variety is to be considered as a supraregional standard. In a time of increasing emancipation of regional variants the assumption of such a standard may be questioned, however, and it might have been prudent to specify in the booklet the regional provenance of the dialect spoken. One pronunciation I found surprising was that of the first vowel in *Gläser* (the only example given) as [ɛ:], which seems a "spelling pronunciation" to me; the common rendition would seem to be [e:]. (It certainly is in my own dialect, that of Vienna.) This is confirmed by Christine's production of the same word in a later lecture, where she clearly pronounces it with [e:]. On the other hand, she does pronounce *Mädchen* (which occurs in the fricative lesson) with an [ɛ:], though the vowel is raised slightly in the repetition. The contrast between [ɛ:] and [e:] may well be on its way out as the language continues to change. As Dr. Hirschfeld has pointed out to me in personal communication, however, its inclusion in this course helps focus students' attention on the correct quality for [e:]. The only other peculiarity I noticed is that Christine pronounces the isolated fricative /x/ with some uvular vibration, so that it sounds like a snoring sound.

A special feature of this course are the hand movements with which Christine accompanies some of her articulations. Thus she mimics the F0 contour of different intonations with up-down motions of her right hand (going from right to left on the screen, contrary to the graphs displayed subsequently), marks stressed syllables by bringing the palms of her hands together, moves her fists apart as if stretching a ribbon for tense vowels while moving her parallel open hands downwards for lax vowels, marks glottal stops with downward movements of one hand, etc. (Of course, only one of these movements is executed at a time, to illustrate whatever feature is being discussed.) It is an empirical question to what extent such gestures may be helpful to students. Their inclusion may indicate that Dr. Hirschfeld found this technique useful in her teaching.

In contrast to Christine's distinct and closely observable articulations, **Hermann** moves his bearded lips in an indistinct mumbling motion, suitable for overdubbing of various languages. This works well enough in the English version I viewed; however, in some lectures Hermann introduces other participants in German, speaking in what seems a different voice, which is decidedly strange. (Dr. Hirschfeld has informed me that it is in fact the same bilingual speaker, so I attribute the impression of different voices to prosodic discontinuity.) The English explanations are delivered clearly but cast in a fairly technical language that students not schooled in phonetics may have some difficulty with. In one instance, when explaining the glottal stop, Hermann mispronounces "liaison" as "elision." To illustrate liaison, the example *Mein_Name ist_Schmidt* is used; however, it is not so clear what the alternative pronunciation of *ist_Schmidt* might be. Christine never provides examples of incorrect pronunciations. Also, examples are rarely presented in terms of minimal pairs, and sometimes contrasting segments do not even occur in the same syllabic position, even when that would be phonotactically possible. Apparently, priority was given to keeping the vocabulary within narrow limits (200 relatively common words and names), to enable beginning students to follow along easily.

The terms *gespannt* and *ungespannt* for vowels are rendered by Hermann in English as "taut" versus "loose" at first, but as "tense" versus "lax" later on. (Whether either pair of terms makes sense to a language learner is questionable, since the physiological correlates of these features are complex.) Similarly, for consonants Hermann first speaks of "degrees of tightness," but then reverts to "tense" and "lax." In a somewhat exaggerated illustration of the distinction between tense and lax plosives, Christine articulates /pe, be, te, de, ka, ge/ with a piece of paper in front of her lips. The term "aspiration" is not mentioned by Hermann, and the viewer may wonder how tenseness or tightness causes the paper to move away from Christine's lips. Also, the different vocalic context for /k/ seems awkward; apparently it was chosen to correspond to the German letter name which, however, seems irrelevant in this example. In some of the later lessons, by the way, Hermann (speaking English) refers to letters by their German names, which is jarring.

Two **tables** displayed both on the video screen and in the booklet show the German vowels and consonants in IPA notation, arranged according to their distinctive features. The phonetic symbols are explained in the booklet, but the names of the features are unfortunately rendered in German and thus not readily intelligible to the language learner. (Translations are provided separately in the booklet, however.) While rounded vowels are denoted as *rund* or *gerundet*, there seems to be no German term for "unrounded." Likewise, the term *unbetont* (unstressed), applied to schwa, has no contrasting term in the table. The classification of /j/ as a lax fricative rather than as a glide is debatable, and the manner category *isoliert* (translated as "glottal") for both /l/ and /h/ is new to me.

The animated **vocal tract displays** in three of the lessons are instructive. The accompanying sound is not the output of an articulatory synthesizer but Christine's natural-voice productions that have been aligned with the simulated movements. The display is schematic but adequate. Only in the illustration of nasal articulation, where the air is said to pass through the nose, the absence of a nasal passage in the display is likely to leave the unimaginative viewer puzzled.

A number of tables illustrate **spelling-sound relationships**. Relevant segments in target words are indicated by bolding and are preceded by the IPA symbol and the isolated spelling pattern. In the orthography, long vowels are indicated by underlining and short vowels by a dot underneath; this notation is not explained by Hermann (it is in the glossary in the booklet) but easy to grasp. The fact that double consonants do not indicate a long consonant but rather a short preceding vowel is mentioned in three different lessons. On the other hand, the special nature of the German orthographic symbol *ß*, which permits the preceding vowel to be either long (as in *Maß*) or short (as in *Faß*), is not pointed out. Also, the existence of different /l/ allophones following high and low vowels is not mentioned. Hermann only warns that /l/ never should sound "hard or dull." The allophonic difference may be less pronounced in Northern than in Southern German.

Regarding the pronunciation of the ending *-en* Hermann says that it is not stressed and "in part varies very strongly." He also states that, whereas the nasal is assimilated to the place of articulation of a preceding plosive, it is pronounced as [n] after /t/, /d/, "and the fricatives." Yet, in words such as *laufen* and *lachen*, assimilation is likely to occur, just as with preceding plosives. (The labiodental nasal in *laufen* may be difficult to tell apart from [n], however.) One of the examples given is *Kuchen*, and although Christine does seem to produce a final [n], pronunciation with a final [ŋ] does seem an acceptable variant. Hermann further explains that *-en* retains its vowel (transcribed as [ə] rather than [ɛ]) after vowels, /l/, /r/, and nasals. Yet, reduction is common in most of these contexts, as illustrated by the repeated greetings of *Auf Wiedersehen* (-[ze:n]) at the end of the last lesson.

The first lesson makes use of schematic F0 contours accompanying the orthography, with stress indicated by bolding. The choice of the word *Hallo* to exemplify a trochaic stress pattern is not optimal in view of its long final vowel and its dialectal variation. It is definitely iambic in my Southern dialect, and several of the passers-by saying *Hallo* at the end of the lecture produce it that way. *Halle* (the name of the German city, which is part of the base vocabulary) would have been a better choice. The fall-rise intonation pattern is said to occur in yes-no questions and "very friendly" utterances—a rather vague characterization. The spelling of double /k/ as *ck* would have deserved a special comment.

The phenomenon of final devoicing of plosives could have been treated in more detail. It is presented as a spelling convention rather than as a phonological rule, but whereas the examples for /d/ and /g/ show the consonant in word-final position, the /b/ in the example *gibt* is not word-final. Hermann says that /b,d,g/ "after a vowel as the final part of a word or syllable" are pronounced as [p,t,k]. Viewers might be confused by the reference to syllables (is a word-internal but syllable-final plosive to be devoiced?) and by the fact that the /b/ in the example is neither word- nor syllable-final. However, Hermann merely advises "to pay close attention" and leaves it at that. The same problem occurs in connection with the pronunciation of *-ig* as [-Ix], which is said to apply "at the end of words or syllables."

In discussing the "joining" of plosives and fricatives (the terms "cluster" or "affricate" are not used), Hermann points out that each consonant must be pronounced clearly, which does not jibe with the fact that the /t/ in /ts/ and the /p/ in /pf/ are unaspirated. (By contrast, the /k/ in /kv/ retains its aspiration.) As is illustrated by two of the passers-bye at the end of the lecture, the /p/ in *Pfennig* is often deleted altogether.

The short acted-out **studio scenes** are generally delightful. The emphasis here is on intelligibility rather than spontaneity. Nevertheless, the dialogues are quite natural, and the participants represent both sexes and a wide range of ages. The youngest participant, Tilli (about 4 years old), is very cute in her breakfast scene but mispronounces *für* as [fiɐ]. The sound is sometimes a little distant, so that manual adjustment of the level may be required during playback.

Finally, informal speech is presented in the **street scenes**, which are likewise useful and pleasantly human. The utterances are generally very brief. Again, a healthy variation of ages is represented among the participants, most of whom appear in several lectures. The dialect variation, on the other hand, is relatively limited, undoubtedly due to restrictions on Dr. Hirschfeld's ability to travel outside East Germany. Thus only a single representative of Southern German is included (an elderly man from Vienna), and only a few from West Germany; the majority are from East German towns and exhibit considerable homogeneity of pronunciation.

In summary, while I have done my best to find details to quibble with, there is no question that this is an instructive and entertaining course which should be of great value in supplementing the teaching of German as a foreign language.

FOOTNOTE

*(Video tape with accompanying booklet.) Ismaning, FRG: Max Hueber Verlag, 1992. Running time 70 minutes. DM120. This review will appear in *Language and Speech, 36(1)* (1993).

Haskins Laboratories Status Report on Speech Research
1992, SR-111/112, 265-278

Music as Motion: A Synopsis of Alexander Truslit's (1938) *"Gestaltung und Bewegung in der Musik"**

Bruno H. Repp

Truslit's (1938) monograph, rarely cited nowadays and found in few academic libraries, contains profound insights into the motional character of music and the performer's role in shaping it. His ideas are highly relevant to contemporary attempts to understand the nature of musical motion and its communication in performance, and, although often speculative and supported only by very marginal data, they provide a valuable source of hypotheses for the more extensive and precise empirical inquiries that are feasible now. To bring Truslit's important theoretical contribution to the attention of contemporary researchers and musicians, this paper presents a highly condensed and annotated translation of his book.

TRANSLATOR'S PROLOGUE

Historical consciousness is limited in today's science, particularly in the fast-moving United States. Apart from a few classics that are cited (though perhaps not read) by everyone, literature that goes back more than a few decades is commonly ignored, particularly if it is in a foreign language. Deluged by new publications, few scientists have the time to dig up historical sources, and many are hampered by their lack of foreign language skills. It is also true, of course, that many older works lack the methodological sophistication of contemporary studies. These older authors, however, often made up for this lack of rigor by depth of insight and breadth of view. We can still benefit from their wisdom.

Alexander Truslit's (1938) monograph, "Gestaltung und Bewegung in der Musik" ("Shaping and Motion in Music"),[1] came to my attention through a reference in Gabrielsson (1986)—the only reference to Truslit's work I have encountered. I obtained a copy on interlibrary loan from the University of Iowa, where one of the few copies in the U.S. is located.[2] I was not able to obtain the sound recordings (three 78 rpm discs) that originally accompanied the book.[3] Even so, I found the book extremely stimulating and insightful. Truslit's claims, even though they are largely speculative and subjective, seem highly relevant to contemporary attempts at understanding the motional character of music and its communication in performance.

For those who are interested in these problems but have been unaware of or unable to read Truslit's book, I provide the following synopsis. The text is, with minor exceptions (such as some contractions, added descriptive statements, or added emphasis), a literal translation of statements culled from the approximately 200-page original, which lends itself well to this kind of distillation. German terms are quoted wherever faithful translation seemed difficult. All headings are literal translations of the book's major section headings. My own comments appear as footnotes and in the epilogue.

This excerpt-translation is published with the permission of Chr. Friedrich Vieweg Musikverlag, Munich, Germany. I gratefully acknowledge the generous support of Haskins Laboratories during this work.

Shaping and Motion in Music

by Alexander Truslit

Motion as the Fundamental Element of Music

Musical experience and shaping. To experience music fully, both the listener and the composer or performer must understand its most essential characteristic.[4] This characteristic is the expression of *inner motion*, whose spontaneous manifestation in voice and movement formed the origin of song. This expression is the eternal driving force of music. In music, however, it is *consciously shaped*; inner experience and artistic form merge into an integral process. The artist's motion experience creates the form and gives it content. Creative and re-creative artists use various techniques for shaping their materials. The listener, who is liberated from technical concerns, must nevertheless carry out this shaping process internally (*inneres Mitgestalten*), to realize the full potential of music.

The literature on musical interpretation. There is little scientific literature on the shaping of musical performance (*Vortragsgestaltung*), and what there is offers little of practical value.[5] Authors rarely ask (and never answer) the question in what particular way a *crescendo* or *ritardando* is to be executed; instead they often appeal to the artist's "taste" or "feeling" which, however, cannot form the basis for artistic shaping. In general, the observations of these authors concern surface manifestations (changes in intensity or tempo), but not the underlying force that shapes them. Of greater interest are the ideas of authors (such as Rutz, Sievers, and Becking) who have pointed to a connection between body posture or forms of rhythmic movement and musical performance. The purpose of the present investigation is to ask about the *meaning and inner necessity* of the superficial properties of performance, and to explore more thoroughly the parallelism between musical and bodily processes.

The acoustic elements of shaping. The acoustic elements that an artist manipulates in shaping a performance are pitch, timbre, intensity, and duration; while the first two are of great importance for the composer, the last two are most important for the performer. Music played with tones of equal intensity and exact duration (as notated) sounds lifeless and mechanical. It is variations in duration and intensity that create the iiving shapes in music. Such a "free" realization is by no means unfaithful to the notated score; on the contrary, it is closer to the work's original conception. The artistic shaping of intensity is called *dynamics* (*Dynamik*, 'moving force'); that of durations, *agogics* (*Agogik*, 'conduct').

The lawfulness of musical interpretation. Observation of fine artists shows that these "liberties" in performance are by no means arbitrary. Certain patterns keep recurring and can be captured in "rules" of interpretation. These rules, however, do not reveal the biological significance of the shaping process.

The law of motion in music. As we have noted already, the origin of music lies in inner motion (*innere Bewegtheit*). The composer Carl Maria von Weber has described how, on his travels, the forms of the landscape moving past the coach window would evoke compelling melodic images in his mind. Ethnomusicological research has revealed a relationship between the landscape and the melodic motion of folksongs: The songs of mountain people often show jagged melodic lines, while those of hill dwellers are gently arched, and those of the plains people are monotonous.[6] The inner motion that gives rise to music may be more or less conscious. It is a pure sensation of motion, not necessarily accompanied by emotional experiences.

Just as motion can generate tones, so the tones can elicit a sensation of motion in the listener. Even single tones can accomplish that, if they are dynamically changing. For example, on hearing the tone G4 played on a violin, one observer reported a feeling of moving freely, being suspended in the air; on hearing D6, a feeling of jumping high, quickly climbing up, etc.[7] To have clear experiences of this kind, however, certain prerequisites are necessary; therefore, not everyone makes these observations, even though everyone has the potential for having the corresponding experiences.

The phenomenon of *visual synesthesia* (*Synopsie*) makes it possible to demonstrate these experiences, also for series of connected tones. Some persons who experience the motion character of music and who have a pronounced visual sense can translate the motion from the auditory to the visual domain by drawing "synoptic pictures." On listening to the same music performed in the same way, a synoptic person usually has the same visual image, even if days or years intervene, and even if the music is only imagined. The color of these images usually reflects sonic (timbre, intensity, and register) and harmonic factors, whereas their form depends on pitch, intensity, and duration of the tones—that is, on the melodic-rhythmic motion in the music.[8]

Tonal motion and emotional experience. Musical motion is often just playful, like a bird gliding through the air; however, it can also have a deeper significance—the expression of inner motion. We know from recent research that every experience is accompanied by an inner process of motion, which is often barely perceptible.[9] Therefore, a certain musical motion should elicit in the listener a corresponding motion experience. An experiment was conducted in which two subjects (not musicians, but experienced in psychological observation) were presented with single tones played with different forms of motion on the violin and were asked to describe their experiences.[10] There was considerable agreement between the two subjects in reporting impressions such as "powerful," "sad," "commanding," etc. in response to particular tones. These impressions were determined by the totality of each sound (pitch, timbre, volume, duration), but primarily by its underlying form of motion. This is the element that music has in common with all kinds of other experiences and living things. Moreover, the specific form of the motion reflects the specific kind of the experience. Straight motion (i.e., a tone of constant intensity) can elicit only a very limited range of experiences, or a direct awareness of the tone as such. All more differentiated experiences are associated with *curvilinear* forms of motion.

A survey of the subjects' responses shows that, although they do not refer directly to experiences of motion, they reveal states of tension (e.g., feelings of power, excitement, elevation, contraction, etc.) which tend to result from inner motion. Often the inner motion cannot be described well in words and can be captured precisely only by the musical motion itself.

Historical references to motion in music. Music is tonal motion (*tönende Bewegung*); it was thus from the beginning and will always be that way. However, when we listen to how music is "made" today, we find that tonal motion is often difficult to sense. The music of primitive people has an elementary, vital motion that is often missing in our music of today. There is also evidence that in antiquity and in the middle ages there was a closer connection between motion and music. For example, in ancient Greece the conductor of a chorus would outline the melodic motion with his hand. Early notational systems (e.g., neumes) were a fairly direct representation of the melodic motion. The later development of exact representation of pitches on a staff, however, led to a loss of this graphic immediacy and of a sense for the connection between the separate notes. Modern notation encloses the living music like a rigid and cold armor. Yet it is not difficult to cut through this armor and to make the imprisoned shape come gloriously alive.[11]

Music as a Biologically Conditioned Form of Motion

The concept of motion as it functions in music. The term "motion" (*Bewegung*) has been used by music theorists in many different ways, often to refer just to the tempo. However, we need a broader definition that encompasses all of music and that accounts equally for the experiences of the shaping performer and of the receiving listener. The novelist Jean Paul came close to the truth when he said: "Music is an invisible dance, just as dance is inaudible music." The motion experience elicited by music is of an inner nature and affects the whole being. Its only outward manifestation are subtle tensions of the muscles. These sensations easily pass unnoticed because they are overshadowed by more intense acoustic, visual, and motoric sensations that accompany music. The performer, for example, "experiences" the technical movements he makes in playing his instrument. Yet, if the execution is proper, these movements arise from inner motion in adaptation to the instrument and thus are organically integrated with musical motion.

Musical motion is internal and encompasses the whole human being. It is not only an emotion (*Gemüts-Bewegung*) but also a true motion sensation. It must be distinguished from acoustic vibrations, from sympathetic resonance, from technical movements in playing an instrument, from the sequence of tones (which is only the outward manifestation of the inner process), and from conducting movements (though they merge partially with it). Musical motion can be likened to an invisible, imaginary dance which is free from all physical constraints. Absolute freedom in movement is the special privilege of music. Musical motion is as differentiated and manifold as life itself, and each musical work has its own motion sequence (*Bewegungsablauf*). What all motions have in common is that they communicate from one inner being to another.

From a scientific viewpoint, the experience of musical motion may be counted among the *vestibular sensations*.[12] These sensations arise from movement of the whole body, not of individual limbs. In the emotions, vestibular sensations are encountered in pure form, although little attention has been paid to them. In our experiments, they appear as various patterns of tension in the whole body. We may conclude that even the more differentiated musical experiences are mediated by vestibular sensations. Although they are mostly of a passive nature, they occasionally may have a more active character which then may be transmitted to the limbs.

The mechanical elements of motion. Every motion starts with an *impulse* or energy that gives it direction and velocity. These properties are modified by resistances in the environment and in the moving body itself (friction, mass, etc.). Every movement needs space and time. By observing the movement trajectory in space and time, we can infer the energy that gave rise to it.

The acoustical elements of motion. Movements usually cause sound. The dynamic development of the sound (i.e., its agogics) provides information about the movement trajectory. The closer the moving object, the louder the sound. The faster the movement, the higher the "pitch"; changes in timbre will occur also. Thus, the acoustic dimensions that convey the movement of objects are the very elements that are involved in musical performance, viz., dynamics and agogics.

Since a movement can be recognized by its dynamo-agogics (*Dynamo-Agogik*), it follows that it can also be *represented* by the same dynamo-agogics, so that any movement can be expressed acoustically, hence musically. *Provided the sound has the dynamo-agogic devel-*

opment corresponding to a natural movement, it will evoke the impression of this movement in us. This impression will be strongest when changes in pitch are involved; this is not essential, however. When the speed of a movement does not change much, its nature is conveyed by its dynamo-agogics alone.

The acoustic elements of motion and the shaping of sound. Thus we come to realize that musical dynamics and agogics are nothing but the expression of movement processes (*Bewegungsvorgänge*). Musical shaping is the shaping of movement. Every *crescendo* and *descrescendo*, every *accelerando* and *decelerando*, is nothing but the manifestation of changing motion energies, regardless of whether they are intended as pure movement or as expression of emotion. Therefore it is not sufficient to execute a *crescendo*, for example, by increasing the intensity of the tones in some arbitrary fashion. The dynamic development must arise as expression of a natural movement, in which case the appropriate agogics will also appear, so that the tone sequence assumes a living, true, and eloquent expression.

Just as space and time are inseparable in a real movement, so dynamics and agogics are both necessary for musical motion; neither is sufficient by itself. The function of agogics is to "guide" the dynamics; this function can only be fulfilled when both result from the same movement. Dynamics and agogics thus must be mutually attuned and without contradiction. In music, the motion processes should occur in a natural manner, otherwise they seem strange and do not communicate anything. Motion is perceived as natural only when it obeys the natural laws of movement. Therefore, the dynamic-agogic shaping of music cannot be applied from the outside, but only through the inner execution (*Mitvollzug*) of the appropriate movement. This applies equally when the object of the shaping is an emotion rather than a pure movement.

The biological and psycho-physiological foundations of the law of motion, and of music experience and shaping. Our daily experience from birth on connects sensations of movement, vision, and hearing; thus associations are formed. We also find an expression of acoustic motion (mostly as emotion) in the intonation of spoken language. However, there is also a more direct path from sound to the sensation of motion, through the *vestibulum*. This is evident in fish, for example, which have no basilar membrane but only a vestibulum: They react to sound with movement (e.g., flight). In humans, the vestibulum serves to maintain equilibrium and is closely connected with the system of motoric muscles. An uninhibited body reacts to sound with certain movements. In animals, P. Tullio (1929) has elicited particular movements by acoustically stimulating the exposed labyrinth; the nature of the movement depended on the type of sound.[13] That humans do not often react to sound with overt movements may be attributed to cultural inhibitions. However, an inner movement reaction will often occur, especially when listening to music. Indirect (associative) and direct (vestibular) effects work in concert and form the psycho-physiological basis of musical experience.

Through its connections to the muscular system, the vestibulum is also the natural organ for the motional shaping of music. It controls both dynamics and agogics. However, the vestibular-muscular connection can be disturbed: Voluntary impulses, education, inhibitions, etc. may work against the physiological link.

The Comprehension and Identification (Erfassen und Feststellen) of Musical Motion

The basic forms of motion in music. Let us imagine a simple scale, from C4 to C5 and back to C4. The most obvious motion trajectory (*Bewegungsbahn*) for it would perhaps be a straight line; however, such movements rarely occur in nature. Also, an ascending line for the ascending part and a descending line for the descending part would be unnatural because of the sharp angle on top. In nature, everything moves in *curves*.

If we try to find a suitable curvilinear trajectory for our scale, we note right away that a downward movement is unthinkable for the rising part of the scale. It is in the nature of our functional organization that the general direction of the movement must agree with the (pitch) direction of the tone sequence. Therefore, a rising-falling scale can only have a rising-falling movement.

There are three basic possibilities of an up-down movement, which differ in manner and inner content (*Gehalt*). They are illustrated in Figure 1:

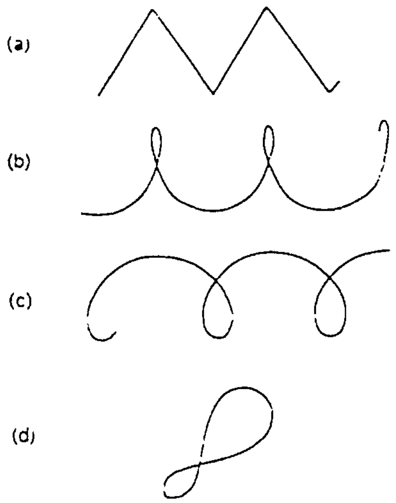


Figure 1. **One artificial and three natural forms of movement (after Truslit's Plate 2): (a) straight; (b) open; (c) closed; (d) winding. The coordinates are spatial (left-right and up-down).**

(1) The *open* movement begins calmly, accelerates on the way up, makes a narrow counter-clockwise loop, and decelerates on the way down.

(2) The *closed* movement begins rapidly, decelerates as it reaches the top, then accelerates on the way down, making a larger clockwise loop if it continues into another movement.

(3) The *winding (gewunden)* movement ascends diagonally into a large counter-clockwise loop and descends fairly vertically, making a smaller clockwise loop at the bottom, which leads it back to its origin.

These three basic forms of curving motion have innumerable variants. The winding movement in particular can vary in the angle and direction of its axis, as well as in the relative magnitude of its two loops.

The acoustic manifestations of the different movement forms, and their examination through measurement. To illustrate these movements, an oboist was asked to play a scale that ascended and descended twice over an octave. Once he played it without inner motion, with *crescendo* and *decrescendo* applied from the outside, as it were. The other times he played it with a lively inner motion, according to the three forms of movement. He had not practiced these forms beforehand but was merely shown the motion curves, with explanations of hcw each motion was to be executed.[14] Even though he had not mastered these movements completely, their differences were nevertheless expressed clearly.

This is evident on listening to the recording. In contrast to the first (straight) example, in which the tones seem to be stationary and unrelated to each other, the other examples exhibit a lively pull, an impulsive forward motion. The open scale scurries along, taking the upper loops in flight; the closed scale moves forward energetically and broadly; the winding scale winds around itself with even greater energy and wide sweep. Moreover, they differ in their speed, with the open scale being fastest and the winding one slowest. These differences were not intended as such; rather, they result necessarily from the peculiarities of the different movements. In each movemer', the descent is somewhat faster than the ascent, which is also quite natural. In fact, all the details of agogics and dynamics follow from the specific characteristics of each movement.[15] Normally, volume will increase with pitch, but the opposite can also occur. Furthermore, there are timbre differences among the movements, with a thinner, more transparent sound in the open movement, a fuller sound in the closed movement, and an even fuller, radiant sound in the winding movement.

It might be asked whether some of these differences are imaginary. To prove their objective existence, measurements were conducted on the sound examples. Using a "film gramophone," the sound wave was recorded on film and then projected onto a wall to magnify it. Thus the durations of tones could be measured to the nearest millisecond, and relative amplitudes were measured in millimeters.[16] The results of these measurements are shown in Figure 2.[17]

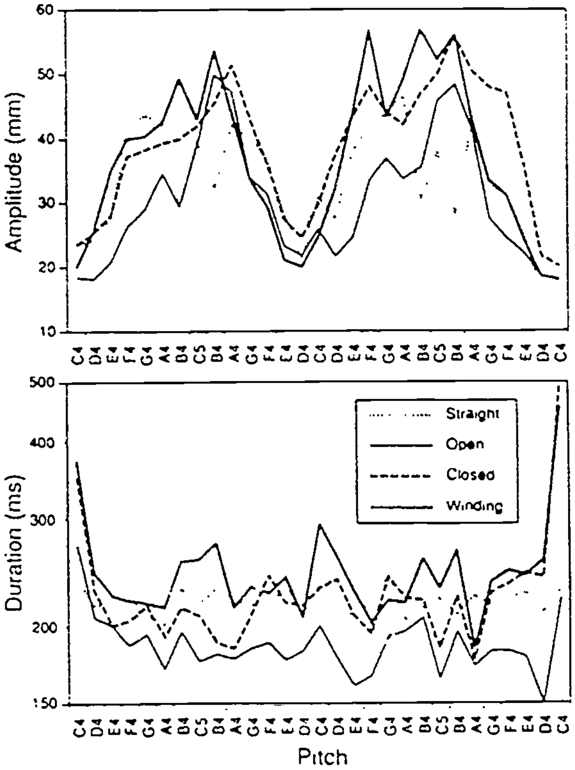


Figure 2. **Relative amplitudes (upper panel) and durations (lower panel) of the tones of a twice ascending and descending scale (C4 to C5), as played "straight" and with three forms of inner movement by an oboist. (Data from Truslit's table on pp. 88-89.)**

It is evident that the open motion proceeds faster than the closed one, which in turn is faster than the winding one. There is also a tendency, most noticeable in open motion, for the scale to be played faster the second time. This speeding up with repetition is a characteristic of natural movement; it is most pronounced at the beginning of a tone series. There is also a tendency for the descending scale to be faster than the ascending scale, especially in the first playing.[18] The unnatural temporal precision of the "straight" scale may be noted. Such mechanical exactness suppresses the inner life and restricts the musical content to the sonic appearance; it has no expressive effect on the listener. In the various forms of motion, on the other hand, if the agogic subtleties occur in the right manner, the tones also seem to arrive with great temporal precision, but with a liveliness that contrasts starkly with metronomic playing. The precision here derives from entirely different sources.[19] Maximum precision is found rarely, but characteristically in the greatest masters of musical performance.

The amplitude curves in Figure 2 show a moderate increase of volume in the open movement, a stronger increase in the closed movement, and an even stronger increase in the winding movement. There are some deviations from the ideal dynamic shapes; however, each movement form is expressed very clearly.[20]

Additional studies confirm these impressions. Thus, in further sound examples, a bassoonist plays a broken chord *staccato* in the three movement forms, and a violinist (somewhat more experienced) plays an ascending series of broken chords in more complex (concatenated) open and closed movement forms.[21]

It should be remembered that dynamics and agogics, although they are measured separately, form an inseparable unit. The amplitude curve depends on the timing of the tones, and a doubling of the tempo, for example, would lead to a much steeper movement trajectory. What amplitude curves (such as shown in Figure 2) cannot represent are the loops and changes in direction of the movement (cf. Figure 1); the loops are stretched out, as it were, along the unidimensional axis of time. When this is taken into account, one can appreciate the complete correspondence of the measured dynamo-agogics with the original movement that shaped the music.

Timbre differences among the movement forms were measured by examining the relative intensities of the partials in bassoon tones (G3, occurring both in an ascending and in a descending *arpeggio*). It appeared that the second harmonic was stronger in closed than in open movement, and strongest in winding movement. It was also noted that intonation tended to be slightly sharp in open motion and slightly flat in winding motion, though it depended on a number of other factors as well. This effect is strongest in the singing voice, and a tone in high *tessitura* may sound impure or wrong when it is mistakenly sung in open rather than in closed or winding motion.

The reception (Erfassen) of the heard motion. Having determined how inner motion manifests itself in simple musical examples, we must now ask whether the listener can consciously apprehend and experience this motion, and whether the motion-based music experience is as naturally given as the motion-based shaping of music.

Sound example No. 10 reproduces an excerpt from Wagner's *Tristan und Isolde* (the end of Scene 1 in Act 2, where Isolde is waving a handkerchief impatiently) whose form of movement is so immediately compelling that it can become conscious on first hearing. If it does not, then perhaps the listener's attention was focused on other musical elements (the sound per se, the rhythm, etc.) or his body was too tense (inhibited) to develop the muscular reactions. In such cases, one should attempt to perform curves in the air by moving both arms in parallel (from the torso) in open, closed, or winding form while listening to the music. Soon it will become apparent that one of these forms fits the music better than the others. How does this come about?

When humans or animals move in a group, there is a strong tendency to move *in phase* with the nearest precursor; this can be seen, for example, in running dogs, jumping children, or migrating birds. Otherwise, there is a feeling of incongruence which arises from the mismatch in the simultaneous experience of one's own movement and that of the precursor. Moving in phase eliminates this incongruence.

Music elicits in us motion sensations comparable to those we experience when we observe another moving person; our own movements, which should adapt to the other's movements, are the arm movements recommended above. Of course, it is possible to execute any arbitrary movement to music; however, if we pay close attention, we find which form generates the least "inner friction." This simple procedure makes it possible to determine *objectively* the forms of movement in music.[22]

It is useful to record every musical form of movement graphically as soon as it has been determined.[23] One often speaks of a "beautiful line" in music without knowing what this really means. A drawn curve with corners or bumps in it does not look good; it does not "sing." Our eyes are extremely sensitive to such small deviations from good form. Our ears, unfortunately, have largely lost (*verlernt*) the ability to attend to the pure motion-determined progress of a melody.

In drawing a curve to represent musical motion, the listener must feel the inner motion and let his hand be guided by it. This is achieved most effectively by people with synoptic abilities (see above). The curve as such is not important, however; it is only an aid to visualizing the motion. The essential thing in music is the experience of natural motion as expression of an event (*Geschehen*). The curve must be experienced as a picture of motion, as the trace of someone moving. It also implies the energy process (*Energieablauf*) necessary for execution of the movement. However, the *character* of the movement is not determined by the visual picture. Depending on whether a small or a large mass is moving, or on the mood in which the movement is carried out, there may be many fine differences in the dynamo-agogic pattern. This "inner differentiation" of the movement form is not visible in the curve and would clutter the picture, were it included. Thus there are many possibilities of individual shaping. By no means does the curve fix the exact form of movement of a musical work.

The original motion (Ur-bewegung) in music, and proof that it can be determined. The assumption seems justified that every inner motion can find only one corresponding motion-based musical expression, and that every motion-based music can contain only one form of movement, namely, the original one that was effective in the composer.[24] In some musical examples, it is fairly obvious what this original motion must have been (e.g., the open movement of waving in the *Tristan* example). In most cases, however, the determination is not so simple and requires a very fine sense for motion (*Bewegungssinn*).

The question may well be raised whether it is possible to find the "correct" motion, and whether it makes any sense to search for it. Although this question seems to be answered in the affirmative by our consideration of the physiological facts and our experiments with synoptic persons, an experiment was nevertheless undertaken. Its purpose was to examine to what extent the form of motion determined from a musical score (rather than from listening to a performance) would agree with the original form of motion that was expressed in the music.

Two subjects (N. and T.) participated, both well versed in the methods described in this book.[25] Subject N. chose a motion curve and sketched it on paper. He/she then carried out the movement with the arms, or sometimes just imagined it, and notated the musical form that came to mind. The resulting notated musical phrases were presented to subject T., who tried to determine the motion in them and drew the corresponding curves. There were 20 examples altogether (see Figure 3). In 13 of these, the recovered motion curves were identical with the original ones, in 6 there were minor differences, and in only one (No. 4) there was a clear disagreement. The experiment thus supports the claim that the original motion pattern can be determined from the musical score.

The significance of the original motion for sonic shaping (klangliches Gestalten). To shape music completely, and to affect the listener as a whole, the artist must shape the work out of the original motion. When music is played with an incorrect motion, there is a disharmony between the tones (whose sequence is a partial expression of the original motion) and the deviant dynamo-agogic shaping, with negative effects on both performer and listener.[26]

The melos as principal carrier of musical motion, and its relation to rhythm and harmony. The *melos* is the principal carrier of motion; a melody without motion is not yet a melos. Melos means "singing," and only a melody filled with motion can "sing."

In the evolution of music lively melodic patterns preceded the development of scales. The rhythms that pervaded them were, like the melodies, lively and fluid, without disruptive accents. Infants, before they even begin to speak, sing or babble with variations in pitch—a first manifestation of melody. In our earlier examples of single tones sung with different forms of motion, we observed a tendency of the tone to either rise or fall in pitch—a tendency towards melody formation.

Beethoven, while composing, would hum or growl up and down in pitch, without singing specific notes. This siren-like up and down is not yet a melody, but it is the original form from which the melody develops. This development (selection of specific tones and intervals, rhythm, metric frame, etc.) may proceed very quickly and unconsciously, so that only the more or less finished music is heard inwardly (just as many processes in ordinary speaking do not reach consciousness).

The same basic motion can give rise to somewhat different musical forms. There may be differences in register, interval size, rhythm, meter, etc. These fine differentiations, however, cause an equally fine detail in the original motion of the emerging musical form. Thus the inner motion expresses itself first in the melody, which then is refined through the addition of rhythmic, metric, harmonic, and other elements. Rhythm and meter must be relatively discreet, so as not to disrupt the melodic motion. In dance music especially, fluid motion may be destroyed by strong accents.

Although "rhythm" originally meant "flowing" or "even motion," the concept has shifted over time to focus on special manifestations such as accent and tone duration, while the element of motion in rhythm has been forgotten.[27] Rhythm occurs to make the motion of even the smallest impulses purposeful and harmonious, and to inject new impulses. Its purpose is *not* to define accents, divisions, groups, or the like; these appear in rhythm without being planned as such.

Rhythm is based on *combination* (*Zusammenfassung*) of similarly structured, recurring events. A running dog automatically combines the movements of its legs into a quaternary rhythm and thus needs only a single impulse instead of four separate ones; that impulse in turn is carried by the principal motion, the rushing ahead. That which has been combined—the content of the rhythm, or rhythmic motive—consists of a chain of smaller impulses, which also have motion character. This form of motion, however, as well as its combination, is different from the motion considered so far, which shapes the melody. Whereas the latter relates to movement of the *whole body* via the labyrinth of the ear, the individual rhythmic impulses as well as their combination relate to movements of *individual limbs* and of the torso via the system of muscles and joints. Thus, overt movements such as foot tapping are easily elicited by rhythmic motion. Rhythmic and melodic motion thus have very different meanings: One affects the limbs, the other the whole person. In its dynamo-agogic consequences, however, rhythmic motion must also follow the law of motion, which is regulated by the vestibulum.

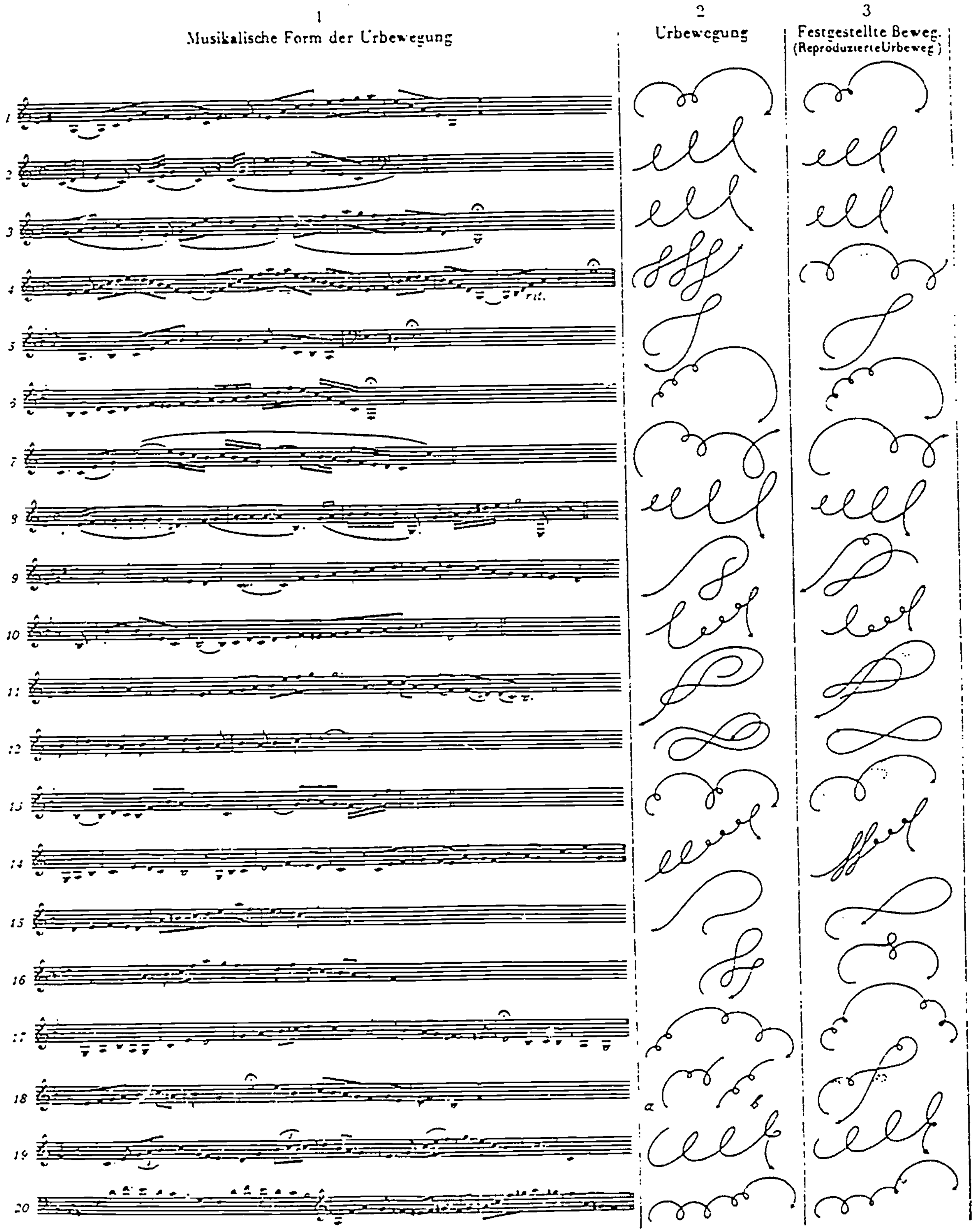


Figure 3. Experiment on the recovery of original motion from notated music. (1) Examples notated by subject N. (2) Original motion curves, as drawn by subject N (3) Motion curves drawn by subject T. (Reproduced from the insert between pages 116 and 117 in Truslit's book.)

Consideration of several recorded examples shows that the rhythm alone (as tapped with the finger, for example) can suggest a motion that is different from that which emerges when the melody is added. The melodic motion carries and absorbs the rhythmic motion, without suppressing its inner pulsing. That the two motions spring from different sources is also evident from encounters with students who have a highly developed rhythmic sense but no feeling for melodic motion, as well as from clinical cases which show that melody production can be impaired while rhythm production remains intact.

Harmonic dissonance can also cause tension in the muscle system, but again in a different way than melody and rhythm.[28] The sharper dissonances, which result in greater tension, tend to lie in those parts of the musical motion curve that must be executed with greater energy and tension. Thus, if the melodic motion is executed correctly, justice is also done to the harmonic relationships.

The determination of the original motion (education of the sense for musical motion through inner listening). When we listen to music that has been shaped by motion, it is easy to hear the expressed motion. However, when we have only the score, how do we arrive at the motion that we need for its sonic shaping?

Consider a single long tone. To determine a motion curve, at least three points are needed; a single tone provides only one, so there is considerable freedom. The motion could go from left to right or from right to left (which, depending on the hand executing the movement, will be towards the body or away from it), and upward or downward. Normally, however, a single long tone will have a curve that rises from left to right, unless some very special expression is intended. If the curve is open, the tone becomes short and pointed; a winding curve gives the tone something disquiet and provocative; a closed curve works best because it results in a quiet and broadly encompassing form.

For a short tone, the closed form is still best, though it begins before the tone sounds; an open form would be appropriate only if the tone is to be understood as a question or scream. However, tones rarely occur in isolation. The accompaniment or the following context usually indicate the original motion, not only for the tone itself but for the whole phrase or section. Incidentally, the motion sense has a tendency to persist in the same form.

For two tones, there is already a better basis for determining the original motion. Thus two tones ascending an octave suggest a closed form rising from left to right and falling slightly at the end, whereas two tones descending one octave result in a closed form rising slightly and then falling from left to right.[29]

To illustrate further the relationship between musical form and motion, let us see how small changes in a musical phrase of about 12 tones change the form of motion. Figure 4 shows nine variations of a simple broken chord and the corresponding motion curves. Example 1 clearly has an open form which rises slowly at first, then faster. In Example 2, these differences are reversed, and the motion is a closed one. A small change on top from C4 to D4 (Example 3), which creates a dissonance and corresponding tension, causes a sharp counter-clockwise turn towards the body, hence a winding motion. Note also that the ascending part of the curve is now rounder, bigger, and fuller of energy than in Example 2, even though the notes are the same. In Example 4, the highest tone is raised but causes little tension because it is a member of the triad; this causes the upper loop of the curve to become bigger and rounder, and the lower loop to adjust correspondingly. When the highest tone rises even further to the sixth scale step (Example 5), it results in harmonic tension, so that the upper loop becomes broad and filled with tension, whereas the lower loop is small.

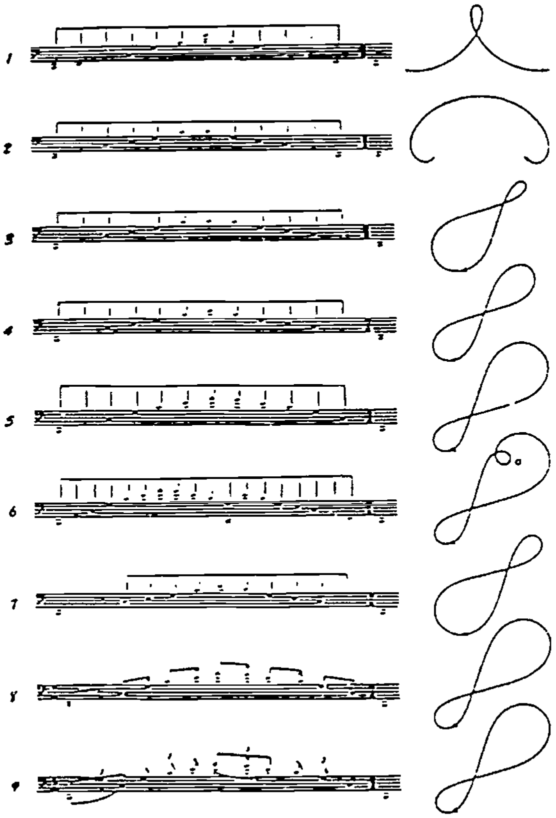


Figure 4. **Nine variations of an *arpeggio* and the corresponding motion curves. (Reproduced from p. 137 of Truslit's book.)**

These examples show how a seemingly insignificant change in the melody can amount to a change in the form of motion. In the first three examples the character of motion has changed. If one tries to perform these examples with other forms of motion (preferably on an instrument), he will see what inhibitions arise and how unfavorably the incorrect curves affect the sound.

Examples 3-5 illustrate how, in addition to the musical "line" visible in the notation, the harmonic development implicit in the tone sequence influences the form of motion. The motions can also differ very much in energy.

In Example 6, the simple "line" is varied by a temporary change of direction, reflected as a small loop in the curve. The overall character of the movement is not affected. The extension of the first tone in Example 7 results in an even broader beginning of the curve. The last two examples (8 and 9) show that neither metric nor rhythmic variations influence the basic form significantly. Similarly, the articulation of the notes (*legato* versus *staccato*) is irrelevant to the basic motion and only of characterizing value.

In all these examples, the motion was not derived from the picture of the notes; rather, it was "intuited" (*empfindungsmäßig "erspürt"*). What was to be demonstrated here is how certain musical moments can determine the course of motion, and how meter, rhythm, phrasing, and articulation align themselves with the musical motion.[30]

We turn now to some examples from the musical literature. With some simple motives, such as the "nature motive" from Wagner's *Das Rheingold*, the form of motion is easily recognized; obviously, it is closed (rising from left to right). A longer simple motive, the "supper motive" from *Parsifal*, has a widely arched closed form, with several small loops where there are dips in the melodic contour as well as harmonic tensions. Each loop, bend, or change in direction is a source of energy that helps sustain the great arch.

It is instructive to compare two examples having rather similar tone sequences but very different forms of motion, caused mainly by the different harmonies: The beginnings of the slow movement of Beethoven's Piano Sonata in c minor, op. 10, No. 1, and of Schubert's *Ave Maria*. The former has a quiet, closed form, whereas the latter has a special winding form, which gives the motion a very different character. Schubert's motive starts with a closed form but then bends over to the left and moves away from the body, whereas Beethoven's motion curve returns to the body.

The initial 8 bars of Brahms' Second Rhapsody, op. 79, show a fine interplay of closed and winding curves. The next 5 bars, however, demand a single closed curve, with some small, imaginary loops. The part for the left hand, however, has a different pattern of two closed curves. It is not unusual to find different motion curves for different voices assigned to the two hands in piano music. This is seen also in the following 7 bars (the second theme), where the right-hand melody has a winding, almost horizontal curve, whereas the left hand has simple closed arches. The 12-bar coda also requires a closed movement.

The graceful Rondo in E-flat major by C.P.E. Bach shows an open movement form. Brahms' famous *Lullaby*, on the other hand, has a horizontal, winding motion over 2 bars at a time. Often it is sung with an open motion, which corresponds to the shaking motion of a cradle rocked with the foot. The original motion, however, corresponds to an old-fashioned cradle suspended from the ceiling, which swings freely and is more effective in putting infants to sleep. This ancient form of rocking is the biologically correct one and probably very important for the development of the vestibular sense of motion.

The correct motion of a musical work cannot be found intellectually; rather, the "guiding hand" in this search is provided by a very fine sense for the (non)correspondence between the inner motion of the music and the movements we have assumed to be "correct." Therefore, it is difficult to *describe* this process. Usually one can only point to outward manifestations of one or the other movement.[31]

The impression that a certain tone has been played too loud or too soft, too long or too short, can only result from a conscious or unconscious *comparison* of what has been heard to a motion-based inner experience of the tonal events. The expressive quality of a tone can only be judged in the context of other tones, and only as expression of the particular local tension (*Spannungsstrecke*) of its motion trajectory. If we do not move along (inwardly) ourselves, then we have "no sense" for irregularities. Even though the vestibular reactions are automatic, they are easily "drowned out" by tonal sensations, metric impulses, etc.

The flux, liveliness, special sound quality, correct tempo, and especially the "comprehensibility" of music result directly from the correct motion. If the presentation of a musical work is "biologically" perfect (i.e., if its inherent dynamo-agogics, resulting from the original motion, are realized), then we do not need any special "interpretation." The personality of the recreating artist comes through automatically in an appropriate way.

The Shaping and Mastery of Musical Motion

Motion-based shaping and its natural mastery. It is commonly believed that performance (*Vortrag*) relies on "inspiration," "intuition," "feeling," or the like. However, these subjective states must not guide the recreating artist, because they do not give access to the original motion. Rather, the original motion may elicit them. Unfortunately, there are few artists today who can shape under the influence of inner motion.

Music teachers, even if they are highly talented, will use in their conscious teaching prescriptions such as "This tone must be the strongest" (but why?), "That tone should be extended" (but by how much?), "This *crescendo* must be natural" (but in what way?), etc. These are hints towards an external application of nuances, which can influence only the exterior of the performance; the inner motion remains obscure. The dynamo-agogic differentiations that express inner motion can be described in rules, but simple application of these rules does not result in living expression. The dynamo-agogics are so subtle that they cannot be grasped intellectually and applied in a calculated fashion. What is important is not the mathematically precise realization of absolute temporal or dynamic values, but the "presence" of motion as a shaping force that realizes *relative* values automatically with absolute precision.

Sonic shaping requires movement of the organs required for singing or playing. These movements have dynamo-agogics corresponding to those of the activating muscles. The sonic events thus correspond to the dynamo-agogic muscular events. If the latter are inhibited, angular, abrupt, or contrived, this will be mirrored in the sonic events; however, if they are holistically experienced, free, and harmonious, so will be the sound.[32]

As our investigations have shown, the inner motion of the performer leads to muscle actions, which in turn lead to sonic events that stimulate the listener's basilar membrane and vestibulum, and thus lead to auditory sensations paired with an inner motion experience and corresponding muscular reactions. All these stages *preserve the same dynamo-agogics*. It is now clear that a person must be *wholly* involved in the shaping and experience of music. Manual dexterity alone does not generate living music; it can dazzle listeners but cannot move them.

The natural development of mastery of movement (education of ear, sense of motion, and the body). Attempts to bring people closer to music by educating their bodies, such as practiced by Dalcroze and his successors, are based on a good idea but suffer from a focus on rhythm. The true musical motion, which is primarily melodic, is easily destroyed by the rhythmic element, which affects the limbs. To get hold of our body musically, we must move it as a whole; its movements must be melodic without losing the rhythm.

These movements are of a different character than the rhythm-oriented ones.[33] They are not gymnastic or dance-like, nor do they represent a translation (*Verdolmetschung*) of the music. They merely open a path to establish the inner link between the original motion of the music and the melodic movements of the body. They magnify and bring to consciousness the subtle muscular reactions to music.[34] Unlike rhythmic movements, they are not swinging or driven by gravity. Rather, they are intensively *guided*, especially in the large curves. The most important muscle is the *latissimus dorsi*, whose pattern of tension follows the corresponding motion curve. The muscular tensions are very different for the (superficially similar) beginnings of a winding and an open curve; they anticipate the further course of the movement. The different forms of movement also affect the diaphragm, which accounts for different sound qualities in singing or blowing a wind instrument. An elastic connection between the limbs and the torso is essential. The body leads the movement, and the arms follow like a single limb.

As the bodily mastery of these movement exercises progresses, so that even the finest musical motions can be experienced with the whole body, the sonic shaping also becomes more perfect, lifelike, and free. There is no need to worry, however, that mastery of the original motion will always and in everyone lead to the same stereotypical sonic expression. The executant's personality, temperament, mood, etc. will introduce differences, while preserving the original form. Nor does the inner motion experience interfere in any way with the playing of music; it only manifests itself as a natural suppleness of the body. Through its effect on the diaphragm, motion-based music making also influences breathing patterns in both performer and listener, and thereby may even have a stimulating and beneficial effect on blood circulation.

Prerequisites for sonic shaping. Before we perform a musical work, we must clarify its original motion. By carrying it out internally, we derive the motion impulses that shape the dynamics and agogics. The performance should not be "thought out"; rather, the musician should concentrate fully on the inner motion experience. There should be no unmotivated tensions or relaxations, which will be perceived as unnatural deviations from the original motion. The inner energy must be maintained across pauses and *fermatas*. The engagement of the body should be intensive but not external. The movement exercises described above are merely preparatory; the exaggeration demanded by them automatically disappears in the actual musical shaping.

The graphic movement curve reproduces only a small part of the actual content of the motion. For example, dynamics and agogics cannot be seen directly in the curve. Only the *experience* of this schematically sketched movement allows these details to be felt. There are many other subtle differentiations to be realized. A natural movement is rarely without a goal, content, or expression. A motion may be filled with passionate impatience or with subdued anger. This essential aspect of musical motion is often difficult to render in words, but it is conveyed by the finely differentiated movement of the tones.

The motion must begin before the first tone is sounded; this avoids unclear or sharp attacks. Successive tones should be regarded as signs of particular motion events, even if they are a series of *staccato* chords. Accents should not be "applied" but should grow out of the special form of motion as sudden condensations of tension.

The sign of a perfectly natural performance is that it *seems* "metronomically accurate." Since the listener moves along with the natural musical motion, the agogic deviations are subjectively compensated, and there are no "double contours." An actually metronomic playing, on the other hand, seems deadpan and stiff. *Naturalness of a* *performance results when its dynamo-agogics lead to the experience of natural movement events.* Once the motional shaping is completely mastered, the consciousness of the form and manner of movement actually recedes, while the finest details of dynamics and agogics are experienced more fully and the motion itself proceeds more precisely and freely.

The relation of musical motion to the other elements of music. It is not necessary to have a virtuoso *technique* to be able to shape according to motion. On the contrary, the "technique" should be born out of the wish to be musically expressive. The meaning of making music is not in the acquisition of a fluent technique, but in the artistic shaping of inner experience. From the very beginning, technical exercises should be guided by motion. Not only can even the simplest forms of musical expression be imbued with meaning and living warmth, but the technique is furthered and differentiated until it dissolves completely in the artistic shaping. Seen in that way, the technical differences among string, wind, and percussion instruments are immaterial; they all serve the same end—the sonic expression of experience. This also applies to conducting, in so far as it indicates the original motions rather than just the beat.

It is very difficult to achieve an unconscious engagement of the sense of motion through technical exercise alone. The easier way to a perfect performance and technique is to develop technique out of motion. The usual "practicing" works against this effort because it involves only individual limbs; it often results in the loss of the natural motion instinct. Many "technical problems" disappear once motion-based shaping and the resulting natural coordination of the muscles are introduced.

The sense for motion also provides the correct *tempo.* It is the movement trajectory that determines the time needed for negotiating it. The beat *(Zählzeit)* is often mistaken for the motion itself. In fact, although an Adagio and a Presto differ widely in their beat count, their motions may be very similar, spanning one bar in one case, but three or four bars in the other. In general, these temporal spans do not vary widely, hence the speed of motion is also approximately the same. If there is a lot "happening," the motion will hold back a little; if there is less, the motion will go faster. To give a performance "pull," rather than speeding up the tempo, it is sufficient to give it motion. Fast tempo, if it is not motivated by motion, creates the impression of rushing, whereas natural motion gives even the fastest piece inner calm. Similarly, accelerations and decelerations will enliven a performance only when they result naturally from the inner motion.

In conveying the melodic motion, a player often must pursue two or more curves simultaneously (e.g., in the Adagio of Beethoven's Piano Sonata op. 27, No. 2). They interweave and form a harmonious whole after having been consciously guided for a while. As far as *rhythm* is concerned, its clarity and precision do not lie in sharp accentuation but in motion-based naturalness. Moreover, as an act of combining, rhythm contains movement elements that tend towards melodic forms. The execution of rhythmic motives must likewise grow out of motion. If they represent a periodic event (such as the "tolling bell" in Chopin's Prelude in b minor), the corresponding activity must be carried out in one's imagination, otherwise it will remain "unexperienced" and not be properly translated into the acoustic domain.

The *meter* should be felt more agogically than dynamically; if it is marked too strongly, it has a disturbing effect. The same applies to *articulation and phrasing*; the onset of the first and the offset of the last tone under a slur are especially important and must not be too sharp or abrupt. *Pauses and fermatas* are given life by motion, which bridges them and thereby determines their correct duration. *Expressive markings* in the score are generally imprecise and more a danger than a help to the performer. The original motion provides a much more precise regulation of dynamo-agogics.[35]

Iconicity (*Bildhaftigkeit*) and *gesture* (*Gebärde*) result from a correspondence between the musical motion and natural forms or activities. *Moods and emotions* give the original motion its special character; these fine details cannot be represented in the graphic curves.

Interpretation (*Auffassung*), a purely intellectual function, must be distinguished from the ability to shape according to motion. However, when the correct motion is found, the "character" of a piece is usually grasped as well. The peculiarities of a given *style*, be it that of a period or of an individual composer, are contained in the original motion, which was formed under their influence. Therefore, a style can be grasped only if the original motion is grasped. (We do not refer here to the external manifestations of style, which are the subject of stylistics and have little significance for shaping.)

The sense of motion also provides a significant aid in *memorizing* music. Sequences of tones are conceived as wholes (*Ganzheiten*), which provides the psychologically correct foundation for memory performance.

Motion-based music experience as "shaping along" (Mitgestalten). The listener must follow the musical motion actively, by "shaping along" inwardly. Even persons with little musical education may be receptive to musical motion, and the conscious experience of motion may generate interest in and enjoyment of music. The ability to experience music as motion constitutes true musicality.

TRANSLATOR'S EPILOGUE

Truslit's book provides no clue about the author's background. Only one other publication is cited, an obscure paper on tonal dissonance (cf. Footnote 28). Most likely, he was a music teacher rather than an academic; his theories seem to derive from an intensive active involvement with music. His systematicity and clarity of thought must be admired, and even though his empirical methods are woefully inadequate by today's standards, the fact that he saw the necessity of objective tests and measurements adds to his credits.

Truslit's ideas are original, coherent, and important. They are delivered in a concise and forceful manner, often reminiscent of James J. Gibson's style, and they antedate by several decades modern developments in psychology and psychomusicology. His assumption that the dynamo-agogics of music constitute auditory information for motion perception presages the tenets of ecological acoustics (see, e.g., Jenkins, 1985). His idea of the transmission of motion information from the musician to the listener via the acoustic medium is essentially a "motor theory," such as proposed by Liberman and his colleagues for speech perception (see, e.g., Liberman & Mattingly, 1985). One of his most intriguing and speculative proposals, that the vestibulum is the organ of musical motion perception, has recently been revived quite independently by Todd (1992a, 1992c). Several authors (Feldman et al., 1992; Repp, 1992; Todd, 1992b) are currently pursuing the question of how to characterize "natural motion" in music performance. There are many affinities between Truslit's ideas and those of Clynes (1977, 1983, 1986, Clynes & Nettheim, 1982) and Gabrielsson (see 1986, 1988), although these latter authors focus more on the rhythmic level, the one Truslit associates with limb movements. The strong biological flavor of Truslit's theory puts him especially in the vicinity of Clynes, whose theories originate in biocybernetics.

As far as I know, no contemporary author has based his or her theories or experiments directly on Truslit's ideas. Perhaps due to the political upheavals of the time, his book seems to have been forgotten as soon as it appeared. I hope to have convinced the reader, however, that Truslit's work deserves to be widely known. It is probably the most coherent and convincing theory of the basis of common music experience—of musical as distinct from musicological listening (cf. Cook, 1990). Approaches towards understanding the latter abound in current music psychology. It is time to give the former its due, and there is no better way to start than by reading what Truslit had to say.

REFERENCES

Clynes, M. (1977). *Sentics The touch of emotions*. New York: Doubleday Anchor.

Clynes, M. (1983). Expressive microstructure in music, linked to living qualities. In J. Sundberg (Ed.), *Studies of music performance* (pp. 76-181). Stockholm: Royal Swedish Academy of Music.

Clynes, M. (1986). Music beyond the score. *Communication & Cognition*, 19, 169-194.

Clynes, M., & Nettheim, N. (1982). The living quality of music: Neurobiologic patterns of communicating feeling. In M. Clynes (Ed.), *Music, mind, and brain: The neuropsychology of music* (pp. 47-82). New York: Plenum Press.

Cook, N. (1990). *Music, imagination, and culture*. Oxford, UK: Clarendon Press.

Feldman, J., Epstein, D., & Richards, W. (1992). Force dynamics of tempo change in music. *Music Perception, 10*. 185-204.

Gabrielsson, A. (1986). Rhythm in music. In J. R. Evans & M. Clynes (Eds.), *Rhythm in psychological, linguistic and musical processes* (pp. 131-167). Springfield, IL: Charles C. Thomas.

Gabrielsson, A. (1988). Timing in music performance and its relations to music experience. In J. A. Sloboda (Ed.), *Generative processes in music* (pp. 27-51). Oxford, U.K.: Clarendon Press.

Hartmann, A. (1932). Untersuchungen uber metrisches Verhalten in musikalischen Interpretationsvarianten. *Archiv fur die gesamte Psychologie, 84*, 103-192.

Jenkins, J. J. (1985). Acoustic information for objects, places. and events. In W. H. Warren & R. E. Shaw (Eds.), *Persistence and change* (pp. 115-138). Hillsdale, NJ: Erlbaum.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1-36.

Repp, B. H. (1992). A constraint on the expressive timing of a melodic gesture: Evidence from performance and aesthetic judgment. *Music Perception, 10*, 221-242.

Seashore, C. E. (1938). *Psychology of Music*. New York: McGraw-Hill. (Reprinted by Dover Publications in 1967.)

Todd, N. P. McA. (1992a). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America, 91*, 3540-3550.

Todd, N. P. McA. (1992b). The kinematics of musical expression. Manuscript submitted for publication.

Todd, N. P. McA. (1992c). The communication of self-motion in musical expression. Unpublished manuscript.

Truslit, A. (1938). *Gestaltung und Bewegung in der Musik*. Berlin-Lichterfelde: Chr. Friedrich Vieweg.

Tullio, P. (1929). *Das Ohr und die Entstehung der Sprache und Schrift*. (Cited by Truslit; no publisher given.)

FOOTNOTES

[*] *Psychology of Music, 21*, 48-72 (1993).

[1] *Gestaltung* (literally, "form-giving activity") is rendered as "shaping" here, *Bewegung* alternately as "motion" or "movement."

[2] Two others, according to the *National Union Catalog*, are at Northwestern University (Evanston, IL) and in the Library of Congress (Washington, DC).

[3] Only recently—too late to take into account in this paper—I learned that the Library of Congress has a complete set of these recordings, the only one currently known to me.

[4] Truslit provides few indications of how broadly he conceived of "music"—whether he included the music of other cultures as well as popular music and the new music of his time. Judging from his musical examples, he probably had in mind primarily the music of the great masters of the 18th and 19th centuries—by no means an exceptional attitude.

[5] Truslit is referring here primarily to the writings of musicians and educators, a number of which he cites on the following pages. He was apparently unaware of the scientific work of Carl Seashore and his associates (see Seashore, 1938) or even of the similar work of Hartmann (1932) in his own country. Undoubtedly, however, and with justification, he would have accused these performance studies of focusing on physical variations as such, rather than on their underlying shaping forces.

[6] A most intriguing observation, the source of which is not revealed.

[7] Standard pitch designations (C4 middle C) are used here. Truslit's "g1" and "d3" thus become G4 and D6, respectively.

[8] Several pages of graphic examples follow, some elicited from experienced synoptic observers, others from three apparently naive subjects who were presented with the same melody played in two different ways. (A more detailed analysis of these acoustic materials, which are among the sound examples accompanying the book, is presented below.) The drawings, some of them quite elaborate and in color, essentially follow the pitch contours of the melodies, but their relative roundness or jaggedness reflects more subtle aspects of motion character.

[9] The names of Palagyi and Klages are mentioned in this connection, but without specific references.

[10] These forms of motion are represented by curves whose nature is not well explained at this point but that suggest *crescendo*, *decrescendo*, and *crescendo-decrescendo*, respectively. The pitch of the tones was also varied. Truslit does not mention whether the violinist (possibly, he himself) was in view. Needless to say, his use of only two subjects makes the results less than conclusive.

[11] A final subsection omitted here, *Formulation of the law of motion*, summarizes the most important claims and raises a few questions for the next section.

[12] This hypothesis, recently (unknowingly) resurrected by Todd (1992a, 1992c), is elaborated below and in an appendix to the book, which is not included in this translation.

[13] These experiments are described in more detail in Truslit's appendix.

[14] The nature of these explanations is not described, nor is it clear whether the curves were presented merely on paper or were also acted out in space.

[15] These details are illustrated schematically in Truslit's Plate 1.

[16] Truslit says they were multiplied by 2 or 3, respectively, though it is not clear why or when either of these factors was employed. The amplitude measured is presumably the maximum excursion for each tone.

[17] This figure was newly drawn from the numerical data provided by Truslit (pp. 88-89). He graphs the same data in his Plate 3, with a separate panel for each movement. The duration variations are difficult to see in his figure, however. He plots the amplitudes as a function of real time rather than of score distance, so that the tones are not evenly spaced and also not vertically aligned for the four movement types. He also superimposes idealized amplitude functions over the data; for this purpose, his representation is in fact preferable.

[18] Again, this seems to be true mainly for the open movement. I have difficulty seeing differences in the other functions.

[19] In contrast to this general statement, Figure 2 does not show much precision in the timing of the three moving scales, which Truslit evidently attributes to the inexperience of the performer. The only systematic timing pattern seems to be a slowing down at the extremes of the scale in the winding movement.

[20] Truslit glosses rather quickly over the amplitude data. He does not mention the slower decrease in amplitude in the closed movement, nor the fact that the amplitude peaks occur beyond the top of the scale (C5). Several functions also have a local amplitude peak during the ascending portion of the scale. The straight scale has a curious amplitude drop near its apex, especially on the seventh scale step (B4).

[21] Numerical data (p. 94) and a graph (Plate 5) are provided for the violinist's performances. These data seem rather more variable to me than Truslit's discussion suggests and therefore are not included here.

[22] At this point follow several pages of discussion of music examples that illustrate different movement forms. Movement curves (more complex than those of Figure 1) are shown as well as some oscillograms, whose amplitude envelope (drawn somewhat impressionistically) is said to indicate the movement form. In addition to the already mentioned excerpt from *Tristan und Isolde* (open form), there are Elisabeth's prayer from Wagner's *Tannhauser* (closed), Schubert's *Ave Maria* (winding), and the aria *Holde Aida* (*Celeste Aida*) from Verdi's *Aida* (winding).

[23] In a footnote, Truslit reminds the reader that these curves are not simply portrayals of the melodic pitch contours. Their precise shapes depend on the dynamic processes resulting from the underlying movements, which do not always coincide with the pitch contour.

[24] It is not quite clear how much detail Truslit means to include in the "original motion." His statements seem less controversial if they are understood as referring merely to the distinction between the three basic types of motion (cf. the preceding paragraph).

[25] The subjects are not further identified. Of subject N. it is said that he/she "had mastered inner motion." Subject T. may have been the author himself.

[26] There follows a discussion of three pairs of recorded musical examples; in each case, one performance represents the correct original motion, according to Truslit, whereas the other reflects an incorrect or degenerate motion. Motion curves, temporal measurements, and oscillograms are provided for some of the examples.

[27] This observation is still largely accurate today, with the significant exception of Alf Gabrielsson's work (see Gabrielsson, 1986).

[28] Truslit's idea of dissonance as "competition of tone sensations" and his description of the resulting tensions as "inner contraction or dilation" due to "nerve excitation," seem esoteric.

[29] These forms constitute about the first and last 60%, respectively, of the first cycle of the closed curve shown in Figure 1c.

[30] The preceding paragraphs have been translated almost completely, to convey these important examples in detail. Due to space limitations, the following discussion of more complex musical examples, while equally instructive, can be reproduced only in rough outline.

[31] There follows a discussion of a more complex example, the first 28 bars of the Prelude to Verdi's *La Traviata*, which is omitted here because it makes extensive reference to sound recordings. Truslit compares in detail two performances by well-known (but unnamed) conductors, one of which shows the correct movement whereas the other does not. The following two, more general paragraphs are embedded in that discussion. Discussion of another sound example, from a Haydn Symphony, concludes the section.

[32] This is illustrated with a sound example, an excerpt from a Chopin Nocturne, played by a pianist first without any instructions, then with the intention of conveying the appropriate motion, which was presented by the experimenter as arm movement and as a drawn curve. A substantial improvement is pointed out.

[33] A series of still photographs illustrates phases in movements corresponding to two of the musical examples, as acted out by a man (presumably, the author).

[34] According to Truslit, the celebrated dancer Isadora Duncan fully recognized the importance of movement in music and, by reproducing its curves, created her unique art in which dance and music fused into a single original motion. He further cites the work of the piano pedagogue Elisabeth Caland.

[35] In a footnote, Truslit argues in favor of performances of Bach's harpsichord works on the modern piano, and of including the dynamic variations that are possible on this instrument, even though they are not "in style."

Haskins Laboratories Status Report on Speech Research
1992, SR-111/112. 279-282

Objective Performance Analysis as a Tool for the Musical Detective*

Bruno H. Repp

The expression and individual character of a musical performance resides in its microstructure, which includes variation in the exact timing and intensities of the tones played. Each performance has a unique microstructure which cannot be reproduced exactly by a human performer, even though repeated performances by the same artist are often highly similar. Objective performance analysis (Seashore, 1936) thus makes it possible to distinguish an identical copy from a repeat performance with a high degree of confidence.

A demonstration of this capability was provided by an unexpected discovery during the analysis of 28 different recorded performances of Robert Schumann's famous piano piece, "Träumerei" (Repp, 1992). The measurements revealed beyond any reasonable doubt that one recording of older vintage, by the German pianist Elly Ney (Electrola WDLP 561), contained two identical sections four measures long.

The first eight measures of "Träumerei" are followed by a repeat sign in the score, which is obeyed in nearly all performances. When the expressive microstructures of the first and second renditions are measured and compared, they are usually found to be quite similar in any given performance, but never identical. Since artists are not machines, there is always a certain amount of uncontrolled variation in timing and intensity, in addition to any variation introduced deliberately by the artist. Such variation was also observed between the two renditions of the first four measures in Elly Ney's recording. The two renditions of measures 5-8, however, were virtually identical, within the limits of measurement error.

The evidence is presented in Figure 1. Figure 1a shows the differences between the first and second renditions in the inter-onset intervals (IOIs) between successive (clusters of) tones, as measured in the digitized acoustic waveform (see Repp, 1992).

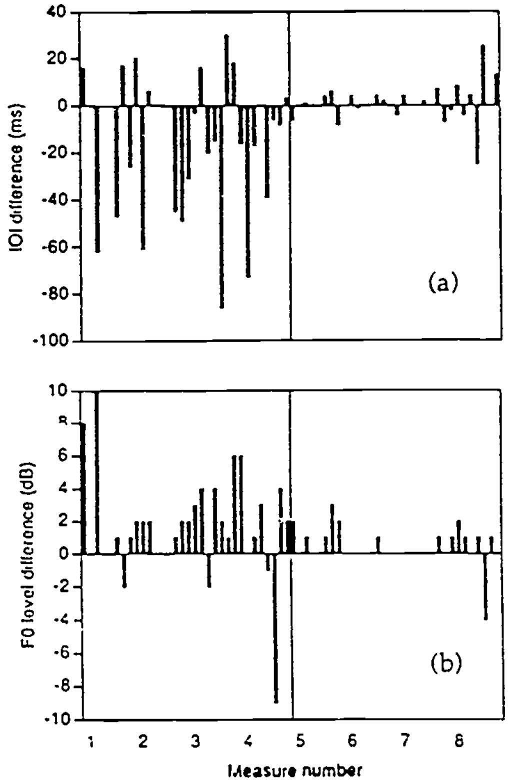


Figure 1. Differences between the first and second renditions in (a) inter-onset intervals (IOIs) and (b) sound levels of the fundamental frequencies (F0) of the melody tones in the spectrum.

This research was made possible by the generosity of Haskins Laboratories and by support from NIH grant RR-05596 to Haskins Laboratories. I am grateful to the record library at Yale University (Karl Schrom, supervisor) for providing the Elly Ney recording, and to Allen Hirson (City University London) for drawing my attention to the potential forensic applications of objective music performance analysis.

Most of these intervals correspond to eighth-notes in the score; any longer IOIs were scaled down to the same level (e.g., the duration of a quarter-note IOI was divided by two). The vertical line divides measures 1-4 from measures 5-8. It is evident that the differences were much smaller in measures 5-8 than in measures 1-4. This was confirmed statistically by a one-way analysis of variance on the absolute difference values [F(1,50) = 23.41, $p <$ .0001].

Three additional considerations suggested that the two renditions of measures 5-8 were in fact identical. First, the differences observed in that section of the music were within the range of measurement error estimated by Repp (1992). The two 25-ms discrepancies in measure 8 occur in adjacent positions and are of opposite sign, which suggests a single larger measurement error in locating a tone onset. This was confirmed by re-examination of the waveforms. Second, the average difference in measures 1-4 was -18.38 ms (s.d. = 31.65 ms), which indicates a slower tempo in the second than in the first rendition, whereas the average difference in measures 5-8 was 1 ms (s.d. = 8.36 ms), suggesting identical tempi. The third and most important argument is that, whereas the differences found in measures 1-4 are quite typical of those found between two renditions of the same music by the same artist (cf. Repp, 1990, 1992), agreement as high as that observed in measures 5-8 has never been encountered by this author. The highest correlation between the IOIs in the two renditions of measures 1-8 for any individual artist was 0.95 (Repp, 1992); for Elly Ney, the correlation was 0.92 for measures 1-4 but 0.998 for measures 5-8.

These arguments are further corroborated by intensity measurements. The particular measure employed here was the rms sound level of the fundamental frequency (F0) in the spectra of successive melody tones, determined by a FFT over a 51.2 ms window whose left edge coincided with tone (cluster) onset, as determined in the waveform (cf. Repp, 1993). Figure 1b shows the differences between the first and second renditions in this measure. Again, the absolute differences in the second half are much smaller than those in the first half [F(1,48) = 16.76, $p <$ .0003].

Although the sound spectra were determined automatically and thus error-free, errors in locating tone onsets, random surface noise from the record, and rounding to the nearest dB value caused measurement error in sound levels. Note that the largest discrepancy, in measure 8, coincides with the single large timing measurement error (Figure 1a). The other discrepancies, too, seem within plausible error margins. The average difference in measures 1-4 was 2.04 dB (s.d. = 3.56 dB), which suggests a somewhat higher dynamic level in the second than in the first rendition, whereas the average difference in measures 5-8 was only 0.5 dB (s.d. = 1.24 dB), suggesting a slightly elevated level. There are not sufficient data available on intensity measurements to judge whether such close agreement as observed here in measures 5-8 could ever be achieved by an artist playing the same music twice; however, it seems highly unlikely.

The essentially independent statistical comparisons of timing and intensity patterns lead to the inescapable conclusion that the second rendition of measures 5-8 is not only much more similar to the first rendition than is the case for measures 1-4, but more similar than is humanly possible. This indicates that measures 5-8 were in fact duplicated by the recording engineers. This was presumably done to cover up some technical problem at the time when the original 78 rpm recordings (dating from the late 1930s) were transferred to LP format. This may be the first time that objective performance analysis has uncovered such "surgery" in a commercial recording.

Of course, the same conclusion could have been reached by simply comparing the raw acoustic waveforms or their amplitude envelopes. The waveforms of the two renditions are shown in Figure 2. Visual comparison fully confirms the conclusion that the first halves of the waveforms are different, whereas the second halves (starting at the arrows) are identical, except for occasional spikes and other small differences caused by pops and surface noise from the record. However, a comparison of waveforms may not be practical if the location and extent of a duplicated portion are unknown. Moreover, if copied music were processed in any way to conceal its identity (e.g., by changing its tempo, filtering the sound, or mixing it with another sound track), waveform comparisons would be much less informative, whereas analysis of musical microstructure, particularly of timing patterns, could still reveal identity (cf. Howard, Hirson, and Lindsey, in press).

A much more difficult question than that of literal identity would be whether two different performances of the same music are by the same artist or by different artists. Analysis of musical microstructure can provide relevant information here, too, provided a data base of many different performances of the same music is available.

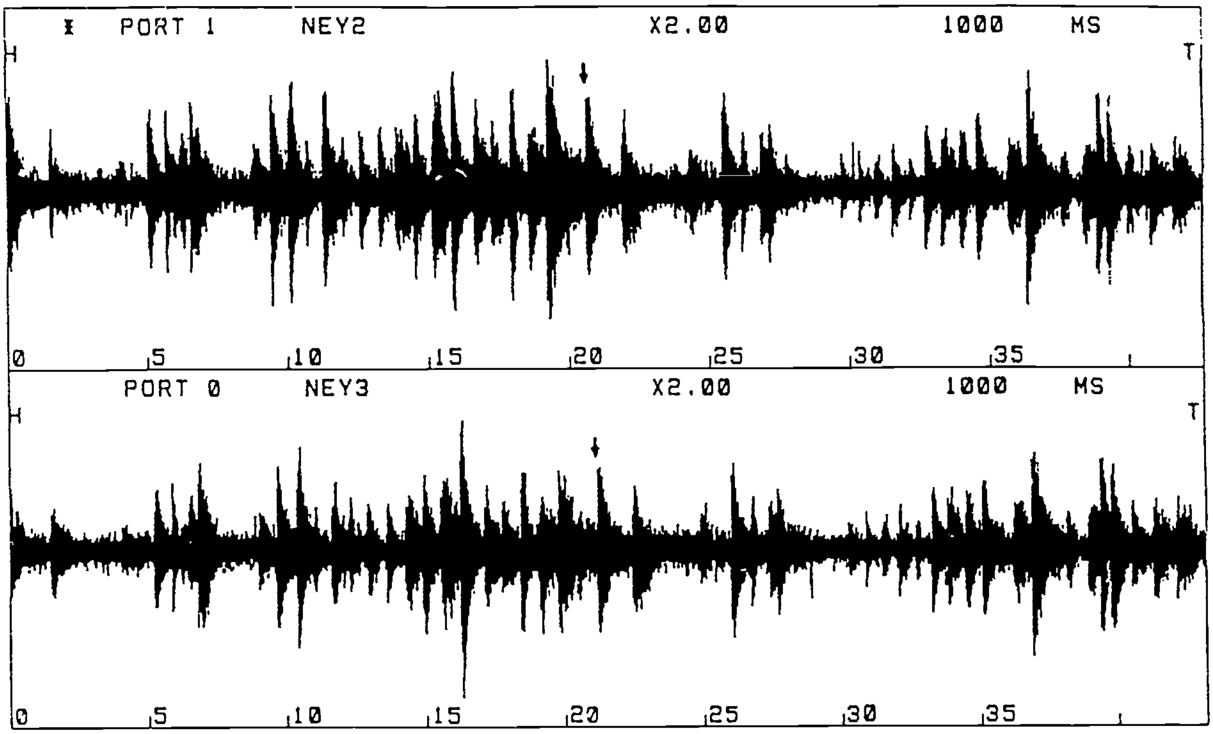


Figure 2. **Acoustic waveforms of the two renditions of measures 1-8 (NEY2 and NEY3, respectively, in this display). The time scale is in seconds. The arrows mark the beginning of measure 5. (The waveforms for measures 5-8 are slightly misaligned due to measures 1-4 being slightly longer overall in the second than in the first rendition.)**

Repp (1992), for example, has observed that different performances by the same artist, recorded many years apart, still tend to be more similar to each other than to almost any other performance by a different artist. Clearly, however, there is a much higher margin of error here, which is inversely related to the size of the available data base.

The most difficult question to pose to a musical detective would be whether two performances of *different* music are by the same artist or by two different artists. This situation is comparable to that encountered in forensic voice recognition, where the speech samples being compared usually differ in content. Although some individual artists are said to have a recognizable individual style that is revealed in all their performances, this is probably too subtle a characteristic to be demonstrated objectively at this stage of the game. Once extensive performance data bases are available, however, the question could be addressed, especially since artists, unlike criminals, are not motivated to disguise their individual characteristics.

REFERENCES

Howard, D. M., Hirson, A., & Lindsey, G. (in press). Acoustic techniques to trace the origins of a musical recording. *Journal of the Forensic Science Society.*

Repp, B. H. (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. *Journal of the Acoustical Society of America, 88,* 622-641.

Repp, B. H. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei." *Journal of the Acoustical Society of America, 92,* 2546-2568.

Repp, B. H. (1993). Some empirical observations on selected acoustic properties of recorded piano tones. *Journal of the Acoustical Society of America*, 93, 000-000.

Seashore, C. E. (1938). The objective recording and analysis of music performance. In C. E. Seashore (Ed.), *Objective analysis of musical performance* (pp. 5-11). Iowa City, IA: The University Press.

FOOTNOTE

Journal of the Acoustical Society of America, 93. 000-000 (1993).

Haskins Laboratories Status Report on Speech Research
1992, SR-111/112, 283-292

Some Empirical Observations on Sound Level Properties of Recorded Piano Tones*

Bruno H. Repp

Preliminary to an attempt at measuring the relative intensities of overlapping tones in acoustically recorded piano music, this study investigated whether the relative peak sound levels of recorded piano tones can be reliably inferred from the levels of their two lowest harmonics, measured in the spectrum near tone onset. Acoustic recordings of single tones were obtained from two computer-controlled mechanical pianos, one upright (Yamaha MX100A Disclavier) and one concert grand (Bösendorfer 290SE), at a range of pitches and hammer velocities. Electronic recordings from a digital piano (Roland RD250S), which were free of mechanical and sound transmission factors, were included for comparison. It was found that, on all three instruments, the levels of the lowest two harmonics (in dB) near tone onset generally increased linearly with the peak root-mean-square (rms) level (in dB) as hammer velocity was varied for any given pitch. The slope of this linear function was fairly constant across mid-range pitches (C2 to C6) for the first harmonic (the fundamental), but increased with pitch for the second harmonic. However, there were two sources of unpredictable variability: On the two mechanical pianos, peak rms level varied considerably across pitches, even though the strings were struck at nominally equal hammer velocities; this was probably due to the combined effects of unevenness in hammer-string interaction, soundboard response, and room acoustics. Moreover, for different pitches at *equal* peak rms levels, the levels of the two lowest harmonics varied substantially, even on the electronic instrument. Because of this variability, the relative levels of the first or second harmonic near onset provide only very rough estimates of the relative peak levels of recorded piano tones.

INTRODUCTION

The present investigation was stimulated by a practical problem. In the course of studying the expressive microstructure of acoustically recorded piano performances (Repp, 1990, 1992), the question arose how to best estimate the relative intensities of individual tones. In piano music there are usually several tones present at the same time, initiated simultaneously or in succession. Their intensities are thus confounded in any overall measure derived from the acoustic waveform, such as the maximum amplitude or the root-mean-square amplitude within some time window. To separate the simultaneous sounds, spectral analysis is necessary. Ideally, one would like to have a technique that automatically separates the spectra of simultaneous complex tones. This is a difficult problem, however, which could not be tackled by this author for a variety of reasons. For the present purpose, therefore, a simpler solution was sought.

It was hypothesized that a reasonable estimate of the relative peak levels of individual tones might be obtained by measuring the relative levels of their fundamental frequencies (first harmonics) in Fourier spectra. The pitch and approximate time of onset of each tone were assumed to be known from the musical score and from acoustic waveform measurements, respectively. Unless the

This research was made possible through the generosity of Haskins Laboratories (Carol A. Fowler, President) and was supported in part by NIH (BRSG) grant RR-05596. I am grateful to Jonathan Berger and Jack Vees (Center for Studies in Music Technology, Yale University) for permitting and helping me to record their Yamaha Disclavier, and to Allan Dudek and Pete Tender (School of Music, The Ohio State University) for providing the Bösendorfer recording, as well as further information and advice. For helpful comments on an earlier version of the manuscript I am indebted to Donald Hall, Caroline Palmer, and an anonymous reviewer.

pitch distance between simultaneous tones is very small, their fundamental frequencies should appear as separate peaks in their combined spectrum. The heights of these peaks should be independent of each other, unless the pitch relationship is such that a higher harmonic of the lower tone coincides with the fundamental of the higher tone (i.e., if they are an octave apart, or a twelfth, and so on). These cases could be dealt with by applying an empirically or theoretically derived correction for additivity of harmonics. In most instances, however, independence of fundamental frequency peaks in the spectrum can be assumed.

Before the problem of simultaneous tones can be addressed, however, it must be determined whether the relative peak levels of *single* tones can be inferred reliably from the relative levels of their fundamentals. Two functions must be determined: one that relates these two quantities for any given pitch, and another that shows how that relationship changes with pitch. It is known from theoretical investigations and measurements of piano string vibration (e.g., Hall & Askenfelt, 1988) that the relative strength of the higher harmonics in the spectrum increases as hammer velocity (and hence peak sound level) increases for a given pitch. There is apparently no information in the literature, however, about whether the increase in the level of any individual harmonic with peak level is linear or nonlinear. It is also known that, as pitch increases, the fundamental becomes increasingly prominent in the spectrum of piano tones, thus accounting for a greater proportion of the overall intensity (see Benade, 1990). Again, however, the precise nature of this functional relationship is apparently not specified in the literature. Because the fundamental can become rather weak at low pitches, due to poor radiation by the soundboard (see Fletcher & Rossing, 1991), the present study not only avoided the extremes of the pitch range but also examined the levels of both the fundamental (first harmonic) and of the second harmonic.

Although the theoretical apparatus may be available to make fairly precise predictions of piano string vibration, which is subject only to variation arising from mechanical factors, piano sounds recorded by a microphone are subject to variation from two additional sources: the radiation characteristics of the soundboard and effects of room acoustics dependent on the position of the microphone. These effects are extremely complex and difficult to predict (Benade, 1990). Commercial piano recordings thus present a rather messy situation from a scientific viewpoint: The specific characteristics of the instrument, the room acoustics, and the microphone placement(s) are all unknown, not to mention any additional sound processing by recording engineers or band limitations and distortion on older recordings. It is impossible to infer the "original" piano sound from such a recording, let alone the hammer velocities that produced the sound; the recorded sounds must be accepted as they come. (After all, they are what the listener hears.) The purpose of this study was to examine whether piano tones recorded under conditions not unlike those of a commercial studio recording would show systematic relationships among the intensity measures of interest, and to assess the magnitude of the unsystematic variation to be reckoned with.

Although the investigation could, in principle, have been conducted on sounds recorded from a conventional instrument played manually, advantage was taken of the availability of two different computer-controlled mechanical pianos on which MIDI or hammer velocity could be specified precisely, so that different strings could be excited with comparable forces. The inclusion of recordings from two different instruments was expected to lend some generality to the conclusions drawn, as well as to provide information about some specific characteristics of each of these instruments. This information might prove useful for investigators in music psychology who plan to use computer-controlled pianos for the generation of materials for perception experiments. For comparison, electronic recordings from an instrument producing synthetic piano sound were included as well. These recordings served as a control or baseline, for they were free of any variability due to mechanical factors and sound transmission (resonance and absorption) effects. Of course, their representativeness depended on the realism of the sound synthesis algorithm; at the very least, however, they provided some information about the output of this proprietary algorithm, which may also be useful to researchers contemplating to use this instrument.

A recent precedent for an empirical investigation of piano acoustics from a psychomusicological perspective is the study by Palmer and Brown (1991). They investigated the maximum amplitude of piano tones as a function of hammer velocity in recordings obtained from a computer-controlled Bösendorfer 290SE concert grand piano located at the Media Lab of the Massachusetts Institute of Technology. Maximum amplitude was measured from digitized waveforms of acoustically recorded tones at several different pitches. Palmer and Brown found that this quantity increased as a linear function of hammer velocity, for velocities between 0.5 and 4 m/s. The slope of the linear function was similar for most pitches tested, but some pitches showed a different slope—an irregularity that remained unexplained. Palmer and Brown further showed that the maximum amplitude of two simultaneous tones was linearly related to the sum of the maximum amplitudes of the individual tones, with a slope somewhat below 1. The rather basic information provided by this empirical study seemed not to be directly available in the literature on piano acoustics, which generally takes an experimental physics approach. The present investigation follows in the footsteps of Palmer and Brown in that it provides some empirical results that may be of interest not only to music acousticians but also to music psychologists using computer-controlled pianos.

I. Methods

A. Instruments. The first instrument recorded from was a Roland RD250S digital piano at Haskins Laboratories. This instrument uses proprietary "structured adaptive" synthesis algorithms, which are said to faithfully recreate the sounds of several different types of piano, including their variations in timbre with pitch and sound level. "Piano 1," which indeed has a fairly realistic—though still recognizably artificial—piano sound, was used for the present recordings (with intermediate "brilliance" setting). The instrument was MIDI-controlled from an IBM-compatible microcomputer using the FORTE sequencing program.

The second instrument was a Yamaha MX100A Disclavier upright piano located at Yale University's Center for Studies in Music Technology. It is a traditional mechanical instrument equipped with optical sensors and solenoids that enable MIDI recording and playback. Control was via a Macintosh IIcx computer using the PERFORMER sequencing program. The translation from MIDI velocities (ranging from 0 to 127) to actual hammer velocities was not known for this instrument.

The third instrument was a Bösendorfer 290SE concert grand piano located in the Music School at The Ohio State University. Like the Yamaha Disclavier, it is a traditional piano equipped with optical sensors and solenoids. Unlike the Yamaha, however, it comes with customized hardware and software that specifies actual hammer velocities (in m/s). In this study, however, the piano was MIDI-controlled from a Macintosh II computer using STUDIO VISION software. The mapping from actual hammer velocities to MIDI velocities had been carried out previously by technicians at The Ohio State University.[1]

B. Recording procedure. The same array of sounds was recorded from each instrument. To keep the study within bounds, extremes of pitch and loudness were avoided. The pitches recorded ranged from C2 to C6 (C4 = middle C, about 262 Hz) in steps of three semitones: C2, E♭2, G♭2, A2, C3, ... C6. Each tone was played once at each of five MIDI velocities: 20, 40, 60, 80, and 100. Each tone lasted several hundreds of milliseconds (well beyond the peak amplitude), and comparable intervals of silence intervened between successive tones.[2] All recordings were made on high quality cassette recorders with Dolby B noise reduction.

The output of the Roland was recorded electronically, going directly from the output jack to the tape recorder, to avoid any effects of sound transduction and room acoustics. The Yamaha upright was recorded with a Sennheiser MD409U3 microphone placed on a chair located about 1 m in front of the keyboard. The lid of the piano was open. The Bösendorfer grand piano was recorded with an AKG 451 cardioid (condenser) microphone placed on the right side, about 1 m from the open lid.[3]

C. Acoustic analysis. Each sequence of recorded tones was played back with Dolby B on a high quality cassette deck and was digitized on a VAX 11/780 computer using the Haskins Laboratories Pulse Code Modulation System (Whalen, Wiley, Rubin, & Cooper, 1990). The sampling rate was 10 kHz, with low-pass filtering at 4.9 kHz but without high-frequency pre-emphasis. It was assumed that frequencies above 5 kHz would contribute little to the peak level of the sounds.

The *peak root-mean-square (rms) level* in dB of each tone was measured by moving a 12.8 ms time window in 6.4 ms steps across each sampled data file, computing the rms level in each window, and picking the maximum value.[4] The resolution was 0.1 dB. Since lower-pitched piano tones have a slower amplitude rise time than higher-pitched tones, the location of the maximum moved closer to tone onset as pitch increased.

The *levels of the first and second harmonics (H1 and H2)* in dB near the onset of each tone were determined (to the nearest 0.5 dB) from a display of the Fourier spectrum of each tone. The spectrum was computed from a 51.2 ms time window whose left edge coincided with the onset of sound energy, as determined visually in magnified waveform displays.[5] Since the amplitude rise was partially or wholly included in this window, and since rise time decreased with pitch, it may have had an indirect effect on H1 and H2 level.

Since only a single token of each pitch-velocity combination had been recorded from each instrument, no estimate was available of the variability in peak rms, H1, and H2 levels for the same key struck repeatedly at the same MIDI velocity. This variability, however, was expected to be small and randomly distributed over all measurements. Hall and Askenfelt (1988) and Palmer and Brown (1991), who did obtain multiple tokens of each tone, commented on the small token variability. As will be seen below, the present functions relating sound level to pitch at different MIDI velocities were strongly parallel, suggesting little uncontrolled within-pitch variability.[6]

It should be emphasized that the absolute dB values shown in the figures below reflect the different recording and playback levels employed for each instrument and hence are not interpretable. Only relative differences are of interest. Due to the different methods of computation, peak rms levels are only about half the magnitude of H1 and H2 levels; only the latter are directly comparable to each other, as they are derived from the same spectra.

II. Results and Discussion

A. Level measures as a function of pitch and MIDI (hammer) velocity

1. Peak rms level. For a constant nominal MIDI velocity, which represents or simulates a constant hammer velocity, peak rms level was expected (perhaps naively) to be constant or slowly changing across different pitches. Figure 1 shows peak rms level as a function of pitch at each of the five MIDI velocities, separately for each instrument. On the Roland, peak rms level was indeed reasonably constant, though at low MIDI velocities there was a tendency for higher-pitched tones to be a few dB more intense than lower-pitched tones. The range of variation at any given MIDI velocity did not exceed 4 dB.

The Yamaha and Bösendorfer data were quite different. There were dramatic variations in peak rms level as a function of pitch. The range of variation at a constant nominal MIDI velocity was as much as 13 dB on each instrument. Moreover, the pattern of variation seemed unsystematic and unrelated between the two instruments. The curves for different MIDI velocities were closely parallel, however, indicating only a small contribution of measurement error or token variability. The observed variation may have been superimposed on a more systematic change in peak rms level with pitch, but any such underlying trend was difficult to discern because of the large variability. If anything, peak rms level tended to decrease with pitch on the Yamaha, while it tended to increase on the Bösendorfer.

In theory, this quasi-random variation in the relative levels of different piano tones could have at least four causes: (a) variation in the force with which the hammers were set into motion, due to imprecise calibration of the electronic components; (b) variation in the mechanical action and in the hammer-string interaction, due to differences in friction, hammer surface, area of hammer-string contact, etc.; (c) variation due to the resonant characteristics of the piano soundboard; and (d) variation due to acoustic frequency absorption or enhancement by surfaces and objects in the room. The relative contributions of these sources are unknown. Fletcher and Rossing (1991, p. 327) display a figure from a study by Lieber (1979) showing similar variation in the relative sound levels of (upright) piano tones when the strings were struck at a constant force. Since Lieber averaged over several different microphone positions and controlled physical force, his data reflect variation mainly due to (b) and (c).[7]

Figure 1 shows that, orthogonal to the variation in relative peak level across pitches, there was a systematic increase in peak rms level with MIDI velocity on each instrument—certainly an expected result. Moreover, the nature of that increase seemed similar at all pitches; hence the parallelism of the functions in each panel of Figure 1. It made sense, therefore, to average across all pitches on each instrument to examine the relationship between *average* peak rms level and MIDI velocity. In each case, this function was mildly nonlinear and negatively accelerated; it was fit well by a quadratic (second-order polynomial) curve. In other words, peak rms level tended to increase faster at low than at high MIDI velocities. The function was steepest and most curved for the Roland and flattest and least curved for the Yamaha, due to differences in dynamic range. The dynamic range (the difference between the average peak rms values at MIDI velocities of 20 and 100) was 23 dB on the Roland, 15 dB on the Yamaha, and 18 dB on the Bösendorfer.[8]

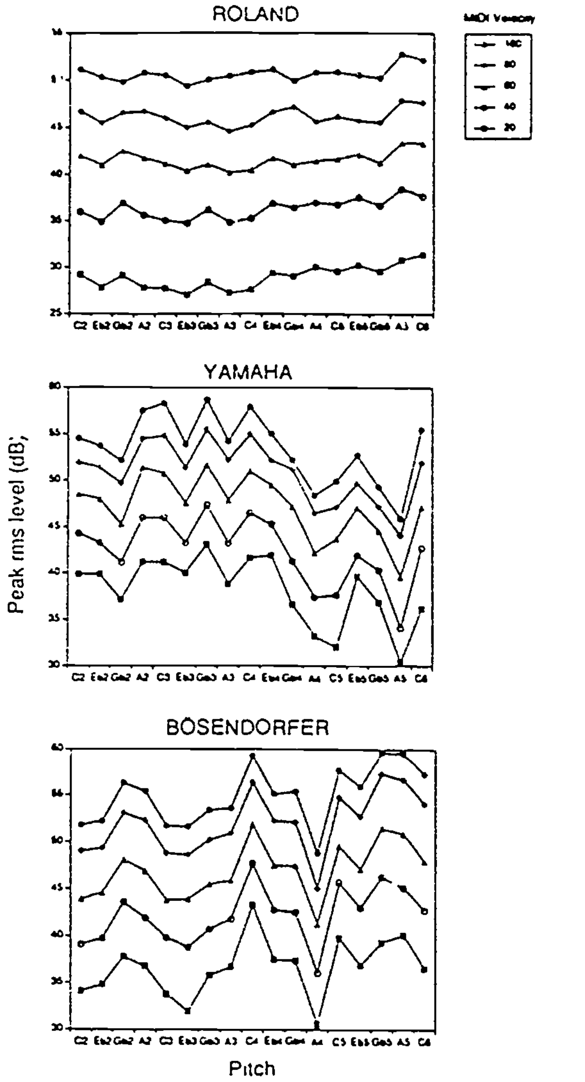


Figure 1. Peak rms level as a function of pitch and MIDI velocity on three instruments.

Since the nominal hammer velocities (in m/s) were known on the Bösendorfer, the function relating average peak rms level and hammer velocity could be examined. It was approximately logarithmic. This is in agreement with the results of Palmer and Brown (1991) who found maximum amplitude (measured on a linear scale, not in dB) to be a linear function of hammer velocity.9

2. *H1 level.* It is known that the relative prominence of the fundamental frequency in the spectrum of piano tones increases with pitch (Hall and Askenfelt, 1988). Therefore, the function relating H1 level to pitch at a constant MIDI velocity was expected to be rising rather than flat. However, given the unsystematic variation in peak rms level across pitches on the two mechanical pianos (Figure 1), comparable irregularities in the .elationship of H1 level to pitch were to be expected for these instruments.

The relevant data are shown in Figure 2. The Roland results indeed show gradually rising functions, with variability comparable in magnitude to that of peak rms level. However, there were a few unexpected (and unexplained) failures of H1 level to increase further at high MIDI velocities (at pitches E♭3, G♭4, and especially A4). Apart from these flukes, however, the functions for different MIDI velocities were parallel. It appears that the Roland synthesis algorithm successfully simulates the expected increase in relative H1 level with pitch.

The Yamaha data were quite different. At pitches below G♭2, there was a precipitous decline in H1 level, evidently due to the "critical frequency" of the soundboard, below which the sound is not radiated effectively (see Fletcher and Rossing, 1991, p. 326). Above G♭2, there was no systematic increase in H1 level with pitch. Instead, there was major variability, similar in magnitude to that observed for peak rms level. (Note that the ordinate scale is compressed in Figure 2 relative to Figure 1.)

The Bösendorfer data did show an overall trend for H1 level to increase with pitch, but with major irregularities, as expected. The different MIDI velocity functions were fairly parallel. The change of H1 level with pitch seemed quite different on the Bösendorfer than on the Yamaha, although on both instruments there were dips in level at pitches E♭3 and A4. This may be a coincidence, however. There was no abrupt fall-off in H1 level at the lowest frequencies, presumably due to the lower critical frequency of the larger soundboard of a grand piano (cf. Suzuki, 1986).

Since the functions in Figure 2 were generally quite parallel, except for a few anomalies in the Roland data, it seemed justified to examine the relationship between average H1 level (averaged across pitches) and MIDI velocity. As for peak rms level, these functions were negatively accelerated (quadratic), more so on the Roland than on the other two instruments. The average dynamic ranges of H1 level are 18 dB on the Roland, 15 dB on the Yamaha, and 17 dB on the Bösendorfer. These values are similar to those for peak rms level, except on the Roland, where the range of H1 level is more restricted than that of peak rms level.

On the Bösendorfer, average H1 level could be examined as a function of nominal hammer velocity. This function was logarithmic, similar to that for peak rms level.

3. *H2 level*. It has been observed previously that, whereas the relative prominence of H1 in the piano tone spectrum increases with pitch, that of H2 decreases (Hall and Askenfelt, 1988). Therefore, H2 level was expected to decrease as a function of pitch at each MIDI velocity. Again, however, unsystematic variability was to be expected on the mechanical pianos. Figure 3 shows the H2 data.

On the Roland, random variability across pitches was fairly small, and the predicted decrease in H2 level with increasing pitch was present. However, this decrease was much more pronounced at lower MIDI velocities, leading to a fanning out of the functions in the figure. It is well known that higher hammer velocities increase the relative prominence of higher harmonics in the spectrum and thereby brighten the timbre of the piano sound (Benade, 1990). At lower pitches, this effect is probably conveyed primarily by harmonics above H2, whereas in higher pitches, which have fewer harmonics, it shows up increasingly in H2. At least, this is how it was simulated in the Roland synthesis algorithm.

On the Yamaha, the overall decline in H2 level was more pronounced but marred by the expected unsystematic variability. (That there were dips at A3, A4, and A5 may well be a coincidence.)

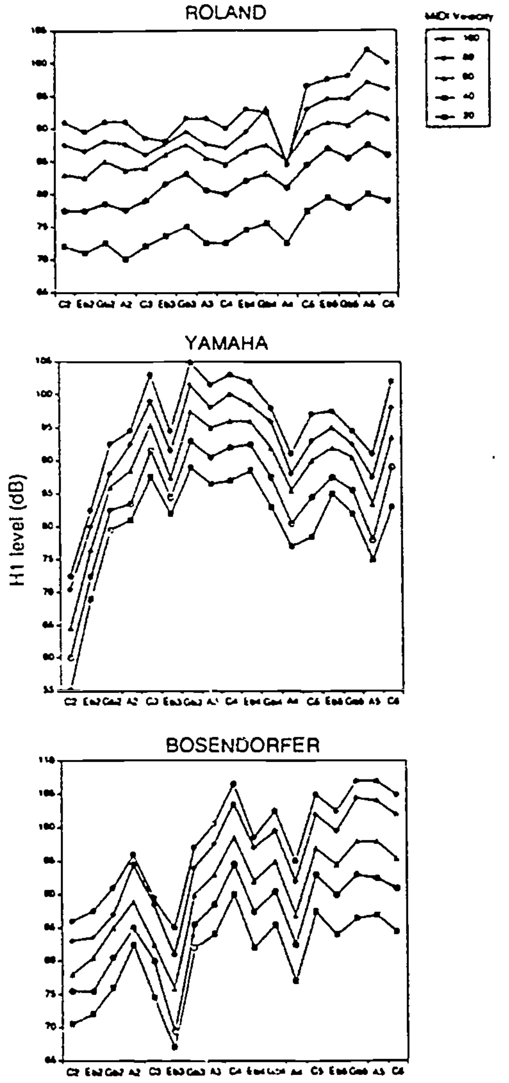


Figure 2. H1 level as a function of pitch and MIDI velocity on three instruments.

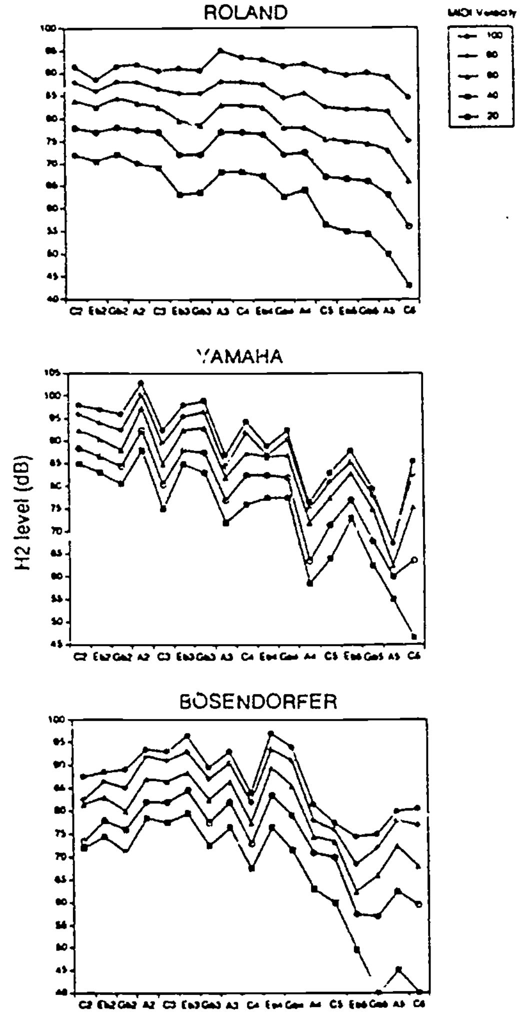


Figure 3. H2 level as a function of pitch and MIDI velocity on three instruments.

The decline was present at all MIDI velocities, and there was no significant fanning out of the functions, except at the highest pitch (C6). On the Bösendorfer, there was little overall decline in H2 level up to G♭4; a precipitous decline between G♭4 and E♭5 was followed by another plateau. There was a pronounced fanning out of the functions in the highest octave, mainly due to H2 level becoming very weak at the lowest MIDI velocity.

The functional relationship of H2 level to MIDI velocity, like that of H1 level to MIDI velocity, was nonlinear and negatively accelerated (quadratic) on all tnree instruments. Because of the fanning out observed in Figure 3, this function became steeper at higher pitches. On the Bösendorfer, H2 level could also be examined as a function of nominal hammer velocity. That relationship was approximately logarithmic and quite similar to that for H1, though the logarithmic fit was less good for H2.

B. Relationships among level measures. The principal aim of this study was to examine whether peak rms level could be estimated from H1 and/or H2 level (and vice versa) Ignoring at first the unsystematic variation in relative level across pitches, H1 and H2 level were examined as a function of average peak rms level at each individual pitch. In most instances, these functions were strikingly linear. Significant deviations from linearity (defined arbitrarily, but rigorously, as r-squared 0.99) were noted. Figure 4 plots the slopes of these linear functions as a function of pitch; deviations from linearity are indicated by asterisks.

For H1 level, the slopes on each instrument varied mainly between 0.8 and 1, with a tendency to increase with pitch. On the Roland, 6 of the 17 pitches examined showed a deviation from linearity in the direction of negative acceleration and hence had a reduced linear slope; the remaining 11 pitches yielded linear functions. On the Yamaha and Bösendorfer, all functions were linear. The lowest pitch (C2) on the Yamaha, for which H1 level was unusually low, showed a steeper slope than other pitches; otherwise, the results were comparable for the two mechanical instruments.

The functions relating H2 level to peak rms level within pitches also were predominantly linear. On the Roland, the slope of the linear function increased steadily with pitch, from below 1 (similar to H1) to almost 2. The slopes were abnormally steep for pitches E♭3 and G♭3, but there was no deviation from linearity. On the Yamaha, the H2 slopes were more variable and did not increase systematically with pitch until about A4. The slope was abnormally low for A5 and abnormally high for C6. Significant nonlinearities were observed at two pitches, E♭4 and C6. On the Bösendorfer, the slopes increased more gradually with pitch but showed significant nonlinearities in the highest octave. Thus, the function relating H2 level and peak rms level, unlike the corresponding function for H1, was not always linear on the mechanical instruments and also more variable in slope. On the Roland, on the other hand, H1 was a less consistently linear function of peak rms level than was H2.

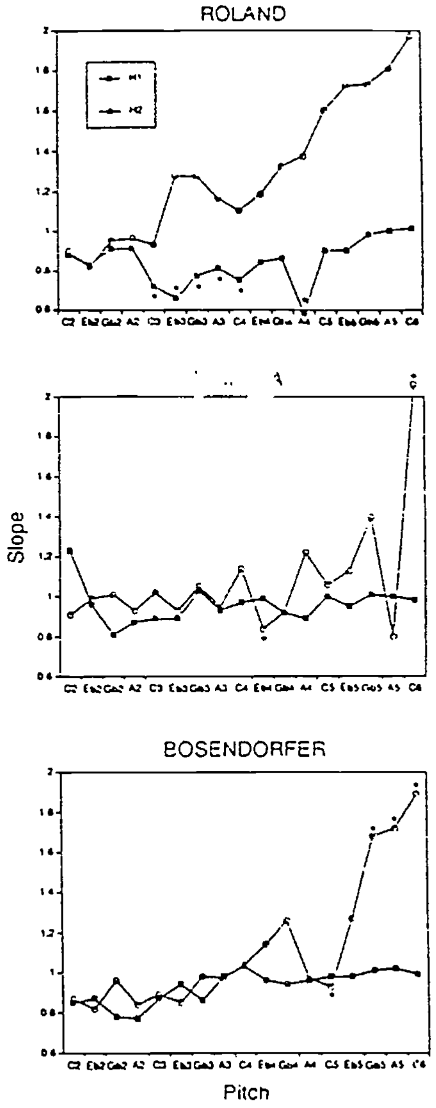


Figure 4. Slopes of the linear functions relating H1 or H2 level to peak rms level as a function of pitch on three instruments. Asterisks indicate deviations from linearity (r-squared < 0.99).

The linear relationship between H1 level and peak rms level within pitches, and the relative stability of that relationship across pitches on the mechanical instruments, are encouraging for the purpose of estimating one from the other. However, a crucial question remains. It was observed that both peak rms level and H1 level varied substantially and unsystematically as a function of pitch (Figures 1 and 2). If these variations were highly correlated, so that H1 level went up and down with quasi-random variations in peak rms level, then the variability would not prevent estimation of one from the other. However, if these variations were only weakly correlated, the predictability of peak rms level from H1 level (or vice versa) would be seriously impaired.

This issue is addressed partially in Figure 5, which plots the relationship between H1 level and peak rms level for a constant representative MIDI velocity, arbitrarily selected to be 60. The data points in these scatter plots represent the different pitches. On the Roland, there was a very narrow range of peak rms levels (corresponding to the variability across pitches in Figure 1) but a much wider range of H1 levels: The former varied by less than 4 dB, while the latter varied over a 10 dB range. There was only a weak correlation between the two; *for the same peak rms level, H1 level could vary by as much as 8 dB*. Also, note that the slope of a linear function relating the two variables would be much steeper than the slope of the function relating these variables when pitch was fixed and MIDI velocity varied (Figure 4).

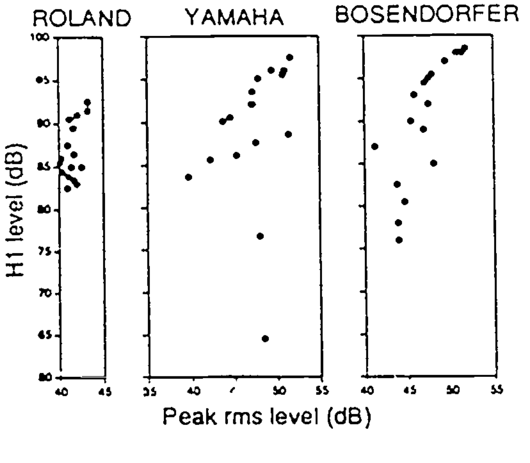


Figure 5. **H1 level as a function of peak rms level at a constant MIDI velocity of 60, on three instruments.**

On the Yamaha, the relationship was even poorer, despite a much wider range of peak rms levels. This was mainly due to the exceptionally low H1 levels at the two lowest pitches (cf. Figure 2). Without these two data points, the correlation would clearly be higher, and the slope of the linear regression function would be close to 1. However, as on the Roland, H1 level would still vary by as much as 8 dB for the same peak rms level.

A correlation between H1 level and peak rms level was also observed on the Bösendorfer, but here, as on the Roland, the slope of a linear regression function would be about twice as steep as that of the function for fixed pitch (Figure 4). H1 level varied by 10 dB or more for tones of similar peak rms level.

When attempting to infer peak rms level from H1 level measured in acoustic recordings of an unknown instrument, within-pitch and between-pitch variability cannot be distinguished. Thus, what is most relevant in the present data is the overall correlation between the two acoustic variables, across all pitches and MIDI velocities. These correlations are not as high as one would like them to be: 0.90 on the Roland, but only 0.61 on the Yamaha (where the two lowest pitches provided highly deviant data), and 0.84 on the Bösendorfer. Judging from scatter plots of all the data points, the error in estimating peak level from H1 level can exceed +/-5 dB.

Even larger variability was observed in H2 level for tones with identical or highly similar peak rms values. The correlation between H2 level and peak rms level across all pitches and MIDI velocities was 0.83 on the Roland, and the range of peak rms values for a given H2 level could be as wide as 15 dB at the lower levels. A similar degree of variability was observed on the Yamaha, where the correlation was 0.80. Finally, on the Bösendorfer the correlation was dismal, only 0.44, and estimation errors in excess of +/-10 dB would seem possible.

A final possibility to be explored was that H1 and H2 levels combined might show a tighter relationship to peak rms level than either of these variables alone, because of a tradeoff between them. One method was to simply average the H1 and H2 levels. On the Roland and the Yamaha, this led indeed to an improvement in the correlation with peak rms level over the correlations for H1 and H2 individually. On the Roland, the correlation was 0.96, and errors in estimating peak rms level would generally be within +/-2 dB. On the Yamaha, the correlation was 0.90, and estimation errors could be as large as +/-5 dB. However, on the Bösendorfer the correlation was only 0.82, not higher than that for H1 alone, and estimation errors in excess of +/-5 dB seemed possible. Alternatively, when the relation between H1 and H2 levels and rms peak level was examined by means of multiple regression, the multiple correlations were 0.97, 0.91, and 0.91, respectively, on the three instruments. The Bösendorfer data did look somewhat better from that perspective, but the estimation error to be reckoned with was still substantial.

III. SUMMARY AND CONCLUSIONS

This empirical study was conducted to investigate how accurately the peak rms level of recorded piano tones could be estimated from the levels of their first two harmonics, measured near onset. Within pitches, where the level variation was caused almost entirely by variations in hammer velocity, the relationship between peak rms level and H1 level was linear on both mechanical instruments investigated, with H1 level increasing at or slightly below the rate of peak level increase. Between C2 and C4, the relationship between H2 level and peak rms level was similar to that between H1 level and peak rms level; above C4, however, H2 level tended to increase faster than peak level, and large variations in the slope of the linear function, as well as some significant nonlinearities, were found on both mechanical instruments. These results suggest that H1 level is a more reliable predictor of peak rms level than is H2 level, although a combined measure of H1 and H2 may be superior to H1 alone.

Although the relationship between H1 level and peak rms level was quite regular within pitches, there were two unexpectedly large sources of variability across pitches. One of them concerned peak rms level, the other the relationship between H1 and H2 levels and peak level. The first kind of variation was observed primarily on the two mechanical pianos; it was rather small on the electronic instrument. Its most likely cause lies in the resonance patterns of the piano sound board, though irregularities of the mechanical action from key to key and room acoustics may also play a role. The variation was unsystematic (though not totally random, it seems) and showed quite different patterns on the Yamaha and Bösendorfer pianos. The stability of the pattern for any given instrument across variations in microphone placement and room acoustics was not investigated here; the study merely established the existence and magnitude of such variability in one essentially arbitrary recording situation. Although the variability may pattern differently if the recording situation were changed, its magnitude would probably be similar.

The other kind of unsystematic variability, that of H1 and H2 levels across tones of different pitch having comparable peak rms levels, was observed not only on the mechanical pianos but on the electronic piano as well; this suggests that its origin lies neither in the hammer-string interaction, nor in the soundboard response, nor in room acoustics, which creates a puzzle. However, it is possible that the Roland's proprietary synthesis algorithm models piano sounds exhibiting natural spectral variation due to some of these sources, while disregarding the more gross differences in overall sound level. The cause of this spectral variation is not well understood at present, nor is its stability known for any given instrument.[10] Again, the present study merely demonstrates the variability; an explanation would require a more systematic inquiry.

The first type of variation, that in peak level with pitch, is worrisome, but perhaps more to the pianist or listener than to the scientist analyzing the recorded sound. Pianists may have to adjust to it and compensate for it to some extent in order to realize the fine dynamic gradations that musical expression demands.[11] However, it is impossible in any case to infer the original hammer velocity from the recorded sound because of additional distortions introduced by microphone placement, room acoustics, and sound engineering. The performance analyst must take the recorded sound at face value, for what it is worth. This first kind of variation, therefore, is not really an obstacle to estimating peak rms level from H1 (and H2) level.

The second kind of variation, however, significantly reduces the accuracy of such estimates. As Figure 5 demonstrated, the error may be considerable. The present data are not sufficient to derive a precise estimate of the average error, but errors as large as +/-5 dB seem possible. The average error is going to be smaller, of course, and the possibility of obtaining sufficiently systematic and interpretable results in the analysis of recorded piano performances is by no means ruled out.

REFERENCES

Benade, A. H. (1990). *Fundamentals of musical acoustics* (Dover Publications, New York). [Reprint of 1976 edition by Oxford University Press.]

Fletcher, N. H., & Rossing, T. D. (1991). *The physics of musical instruments* (Springer-Verlag, New York).

Hall, D. E., & Askenfelt, A. (1988). Piano string excitation V: Spectra for real hammers and strings. *Journal of the Acoustical Society of America, 83*, 1627-1638.

Palmer, C., & Brown, J. C. (1991). Investigations in the amplitude of sounded piano tones. *Journal of the Acoustical Society of America*, 90, 60-66.

Repp, B. H. (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. *Journal of the Acoustical Society of America, 88*, 622-641.

Repp, B. H. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei." *Journal of the Acoustical Society of America*, 92, 2546-2568.

Savage, W. R., Kottick, E. L., Hendrickson, T. J., & Marshall, K. D. (1991). Air and structural modes of a harpsichord. *Journal of the Acoustical Society of America*, 91, 2180-2189.

Suzuki, H. (1986). Vibration and sound radiation of a piano soundboard. *Journal of the Acoustical Society of America*, 80, 1573-1582.

Whalen, D. H., Wiley, E. R., Rubin, P. E., & Cooper, F. S. (1990). The Haskins Laboratories' pulse code modulation (PCM) system. *Behavior Research Methods, Instruments & Computers*, 22, 550-559.

FOOTNOTES

[1] This had been done by matching the peak sound levels (measured with SOUND DESIGNER software) of Bösendorfer tones as closely as possible to those of MIDI-controlled "Piano 1" tones produced at the same pitch on a Roland RD1000 digital piano. If the sound synthesis algorithms of the Roland RD250S and RD1000 digital pianos are identical (which the author believes to be the case), then the MIDI scale of the Bösendorfer should correspond to that of the Roland RD250S. The present measurements indeed showed them to be quite similar, though not identical. The relationship between MIDI velocity (MV) and nominal hammer velocity (HV) in m/s on the Bösendorfer was approximately logarithmic: MV = 51.7 + 45.7ln(HV).

[2] The exact durations, which varied across instruments, are irrelevant because the analyses concerned only the initial portions (less than 100 ms) of the tones.

[3] No attempts were made to vary the recording environment or assess the effects of room acoustics. The relative arbitrariness of each recording situation corresponds to the equally arbitrary (or at least unknown) circumstances of commercial recordings.

[4] This method differs from that of Palmer and Brown (1991), who determined the maximum amplitude on a linear scale from a waveform display. The peak rms level measure used here is approximately a logarithmic transform of their measure. Note that absolute values have no meaning in either case, since they depend on recording and playback levels.

[5] This window was not the same as that from which the peak rms level was derived, although for the higher-pitched tones it overlapped or included it. There were two practical reasons for using different windows. One was that, in his ongoing performance analysis project, the author did not have the resources to compute multiple spectra for each of thousands of tones, which would have been required to determine peak H1 and H2 levels; measurements of H1 levels from spectra at tone onsets constitute the data still awaiting analysis. The second reason was that the ILS software used to compute the Fourier spectrum did not yield an estimate of rms level, whereas the in-house program used to determine rms level did not permit the computation of a spectrum with a predetermined time window. With some further programming effort, it would of course have been possible to compute the rms level over the initial 51.2 ms of each tone and/or the H1 and H2 levels in the 12.8 ms window from which the peak rms level was computed. These additional measures, however, would only have provided intermediate steps in the concrete problem of making sense of the H1 level data obtained near tone onset.

[6] Another relevant piece of evidence comes from the Yamaha recordings, which originally included each tone played with a MIDI velocity of 120. These sounds, however, proved to be nearly identical to those with a MIDI velocity of 100, suggesting that the range beyond 100 was not functional on this instrument. From C2 to G♭3, the peak rms levels of tones with MIDI velocities of 100 and 120 differed by no more than +/-0.2 dB. From A3 to C6, the tones with the nominally higher MIDI velocities were from 0.1 to 0.9 dB more intense, suggesting some marginal effectiveness of velocity specifications beyond 100. The random variability of peak rms level due to inherent unreliability (token variation) and measurement error thus was surely well below 0.5 dB.

[7] The author has not seen Lieber's original study. According to Rossing (personal communication), the spacing of ticks on the (unlabeled) y-axis in the reproduced figure is 10 dB, which makes the variation depicted similar in magnitude to that observed here. See also Savage et al. (1991, Figure 6) for some harpsichord data showing again comparable variation.

[8] That the Roland had a larger dynamic range than the real pianos was unexpected, but perhaps this was an artifact of the electronic recording procedure, which avoided any loss of energy due to sound transmission and absorption. Although less than two thirds of the total MIDI velocity range (0-127) was used, there was evidence that the Yamaha was not differentially responsive at the upper end of the MIDI scale (see Footnote 6); a similar observation was made informally at the lower end of the scale. Palmer and Brown (1991) found that their Bösendorfer (a different specimen from the one recorded here) did not respond differentially to low hammer velocities. Thus the dynamic ranges found here may not be far from the maximum capacity of the mechanical instruments, at least under MIDI control.

[9] Their Figure 3, however, suggests a poor fit for some pitches. A general problem with their figures is that they do not identify the data points for individual pitches, which makes it difficult to gauge the consistency of their data. Palmer and Brown also found a deviant (faster or slower, but still linear) growth in maximum amplitude with hammer velocity for some pitches. In the present data, however, the variability of the function relating peak rms level to MIDI or hammer velocity across individual pitches seemed relatively small. On the other hand, the large variability in relative peak level across pitches observed here was less evident in the Palmer and Brown study. Apart from the fact that they used a smaller number of pitches, the reasons for these differences in findings are not clear at present.

[10] Pilot data had been collected both on the Roland and the Yamaha for the pitches C3, C4, and C5. On both instruments, H1 level for C5 turned out to be at least 5 dB above the H1 levels for C3 and C4. This corresponds to the present data for the Roland (cf. Figure 2), but not at all to those for the Yamaha, which show just the opposite. The pilot recordings from the Yamaha had been obtained with the microphone positioned inside the piano. The pattern of the observed variation thus may depend (at least) on microphone placement.

[11] Note, however, Benade's (1990) comment that "...the ordinary sound pressure recipes even for adjacent piano notes look unrecognizably different from one another when measured, though they may sound very well-matched to our ears" (p. 437). The level measures examined here are acoustic, not perceptual primitives, and it is quite possible that listeners do not perceive the measured variability as such.

Probing the Cognitive Representation of Musical Time: Structural Constraints on the Perception of Timing Perturbations*

Bruno H. Repp

To determine whether structural factors interact with the perception of musical time, musically literate listeners were presented repeatedly with eight-bar musical excerpts, realized with physically regular timing on an electronic piano. On each trial, one or two randomly chosen time intervals were lengthened by a small amount, and the listeners had to detect these "hesitations" and mark their positions in the score. The resulting detection accuracy profile across all positions in each musical excerpt showed pronounced dips in places where lengthening would typically occur in an expressive (temporally modulated) performance. False alarm percentages indicated that certain tones seemed longer a priori, and these were among the ones whose actual lengthening was easiest to detect. The detection accuracy and false alarm profiles were significantly correlated with each other and with the temporal microstructure of expert performances, as measured from sound recordings by famous artists. Thus the detection task apparently tapped into listeners' musical thought and revealed their expectations about the temporal microstructure of music performance. These expectations, like the timing patterns of actual performances, derive from the cognitive representation of musical structure, as cued by a variety of systemic factors (grouping, meter, harmonic progression) and their acoustic correlates. No simple psycho-acoustic explanation of the detection accuracy profiles was evident. The results suggest that the perception of musical time is not veridical but "warped" by the structural representation. This warping may provide a natural basis for performance evaluation: expected timing patterns sound more or less regular, unexpected ones irregular. Parallels to language performance and perception are noted.

INTRODUCTION

In every society, people who listen to music have certain tacit expectations about how it ought to be performed. Professional musicians on the whole aim to satisfy these expectations in their performances, while also exploring imaginative deviations from the norm. This applies especially to the tonal art music of Western culture, with which this article is concerned. The primary cause of the mutual attunement of performer and listener presumably lies in the musical structure apprehended by both. It is the musical structure that calls for certain basic properties of a performance, without which it would be considered crude and deficient. Stylistic variation and individual preferences appear as quantitative modulations of these basic properties.

The present research is restricted to the temporal aspect of music performance, which is arguably the most important of the many physical dimensions along which performances vary. It is common knowledge that musical notes are not meant to be executed with the exact relative durations notated by the composer; rather, performers are expected to vary the intervals

This research was supported by NIH grant RR-05596 to Haskins Laboratories. Some of the results were presented at the 121st Meeting of the Acoustical Society of America in Baltimore, MD, and at the "Resonant Intervals" Interdisciplinary Music Conference in Calgary, Alberta, Canada, both in May 1991. For helpful comments on an earlier version of the manuscript, I am grateful to Carol Fowler, Carol Krumhansl, Mari Riess Jones, and two anonymous reviewers. Additional insights were provided by Patrick Shove in many discussions.

between tone onsets according to the expressive requirements of the musical structure. Music that is executed with mechanical precision generally sounds dull and lifeless, and this is particularly true of the highly individual and expressive music written during the nineteenth century. Objective measurements of the "timing microstructure" of expert performances, usually by pianists, have amply documented that deviations from exact timing are ubiquitous and often quite large (e.g., Clarke, 1985a; Gabrielsson, 1987; Hartmann, 1932; Henderson, 1936; Palmer, 1989; Povel, 1977; Repp, 1990b; Shaffer, 1981).[1]

These deviations are by no means random or unintended. Individual performers can replicate their own timing microstructure for a given piece of music with high precision (see, for example, Henderson, 1936; Repp, 1990b). Only a small proportion of the variance is not under the artist's control. There are also significant commonalities among the performance timing patterns of different artists playing the same music, despite considerable individual differences (Repp, 1990b, 1992). Certainly, musical interpretation is far from arbitrary. To a large extent, artists follow certain implicit rules in translating musical structure into timing variations. Musical listeners, in turn, presumably expect to hear variations that follow those rules, within broad limits. If these expectations could be measured, they might provide a window onto the listener's cognitive representation of musical structure, just as the actual performance microstructure informs us about the artist's structural conception (cf. Palmer, 1989; Todd, 1985).

The objective study of the principles that underlie systematic timing variations in serious music performance has barely begun. Some facts are already well established, however. One is that most timing deviations are *lengthenings* rather than shortenings relative to some hypothetical underlying regular beat, and lengthenings also are larger in size than shortenings.[2] The functions of lengthening are manifold (Clarke, 1985b), but perhaps the most important function is the demarcation of structural boundaries. Lengthening commonly occurs in performances not only at the ends of major sections, where composers may have prescribed a *ritardando* in the score, but also at the ends of subsections and individual phrases. The amount of lengthening tends to be proportional to the structural significance of the boundary; this regularity has been captured in Todd's (1985) formal model of timing at the phrase level. That model generates a timing microstructure for bar-size units by additively combining prototypical timing patterns nested within the levels of a hierarchical phrase structure; this leads to greater lengthening where boundaries at several levels coincide. The resulting timing pattern essentially reflects the *grouping structure* of the music (cf. Lerdahl & Jackendoff, 1983) and, with adjustment of some free parameters, can approximate actual performance timing profiles quite well (Todd, 1985).

Another important function of lengthening is to give emphasis to tones that coincide with moments of harmonic tension or that receive metric accent. Thus, to some extent lengthening is also determined by the *harmonic and metric structure* of a piece. A lengthened tone also delays the onset of a following tone, which may then be perceived as relatively more accented, especially on instruments such as the piano, where the intensity of tones decays over time. While this last function may be specific to keyboard instruments, the phenomena of (phrase-) final lengthening and emphatic lengthening are also well documented in speech (e.g., Carlson, Friberg, Frydén, Granström, & Sundberg, 1989; Lindblom, 1978) and appear to be very general prosodic devices. There is also a parallel with breathing and pausing patterns in speech production, which are likewise governed by prosodic structure (Grosjean & Collins, 1979; Grosjean, Grosjean, & Lane, 1979).

The focus of the present study, however, is not on the performer but on the listener. Moreover, it is not on the more obvious perceptual function of timing microstructure, which is to convey or reinforce various structural aspects of the music (see, for example, Palmer, 1989; Sloboda, 1985). Rather, the aim of this research was to demonstrate that, just as lengthening is implemented more or less consistently by professional performers (as part of the "art of phrasing"), so listeners with musical inclinations expect it to occur in certain places. There may be individual differences among listeners, just as there are among performers, with regard to the preferred extent of the expected lengthenings, but there may well be substantial agreement as to their location (unless the music is structurally ambiguous). If musical listeners did not have expectations about the microstructure of music performance, their ability to express preferences for certain performances over others could not be explained.[3] What is not known, however, is whether these expectations interact with ongoing music perception. That is, do listeners perceive the timing microstructure of music veridically and then compare it against some internal standard (either remembered literally from previous exposures to performances of the music or generated on-line according to some implicit cognitive rules), or do the expectations elicited by the structure of ongoing music interact with and "warp" listeners' perception of musical time?

To address this question, a simple technique was devised to probe listeners' temporal expectations. Several researchers concerned with music performance synthesis have observed informally that musically appropriate timing variations sound regular (as long as they are relatively small), whereas musically inappropriate variations are perceived as distortions (e.g., Clynes, 1983, p. 135; Sundberg, 1988, p. 62). Accordingly, the present experiments used a task which required listeners to detect a temporal perturbation (a small lengthening) in an otherwise isochronous performance. The hypothesis was that *listeners would find it more difficult to detect lengthening in places where they expect it to occur,* particularly at the ends of structural units, in strong metric positions, and at points of harmonic tension. The null hypothesis was that detectability of lengthening would not vary systematically across the musical excerpt - or if it did, that the variation could be explained by the influence of primitive psychoacoustic factors. (Such possible factors will be considered in the General Discussion.) If the null hypothesis were rejected and the pattern of detection performance reflected structural properties of the music, this would confirm that time perception in music is contingent on the perception of musical events. It would further reveal the level of detail in listeners' expectations of timing microstructure, and thereby the listeners' cognitive grasp of the musical structure.

Two similar experiments were conducted, each using different musical materials. The materials were chosen from the piano literature, because it was considered important to present listeners with music sufficiently complex and engaging to guarantee its processing as a meaningful structure rather than as a mere sequence of tones, even after many repetitions. The standard psychophysical procedure of extensive training and prolonged testing of individual subjects was deliberately avoided to retain a semblance of natural music listening. The group of listeners was treated as a single super-subject, representative of the average (musical) listener. Their detection performance, expressed in terms of an *accuracy profile* across each musical excerpt, was compared to the timing microstructure profiles of real performances by great artists. It was hypothesized that a significant correlation would exist between the two profiles, with lengthening in performance corresponding to dips in the accuracy profile, if listeners' expectations derive from a cognitive representation of musical structure similar to that in an artist's mind.

EXPERIMENT 1

Musical material

The musical excerpt represented the first eight bars of the third movement of Beethoven's Piano Sonata No. 18 in E-fiat major, Op. 31, No. 3. Its choice was influenced by the fact that it had been the subject of a detailed performance analysis (Repp, 1990b). The score is reproduced in Figure 1.

The piece is a minuet in 3/4 meter, and the tempo indication (omitted in Figure 1) is *allegretto e grazioso* (i.e., moderately fast and graceful). As can be seen, the predominant note value is the eighth-note, which served as the temporal unit for the purpose of this experiment. Bars and eighth-notes are numbered above the score in Figure 1. (Disregard the small two-digit numbers for the time being.) Altogether, there are 47 eighth-note intervals in the excerpt (2 in bar 0, 6 in each of bars 1-7, and 3 in bar 8). Their onsets are always marked by at least one note, except for interval 6 in bar O, which contains only a sixteenth-note.

Horizontally, the music can be divided into three strands: a melody in the upper voice, which exhibits a complex rhythm and a variety of note values; a bass in the lower voice, which serves as a counterpoint to the melody and marks the harmonic progression; and a middle voice consisting of a steady eighth-note pulse that completes the harmonic structure and serves mainly to mark time.

Vertically, the music can be divided into a four-level binary hierarchy (a *grouping structure)* of nested units. The schematic diagram above the music in Figure 1 is one of several possible ways of representing this structure. Vertical lines represent the onsets of units, whereas continuous horizontal lines show the length of a unit from the onset of the first note to the onset of the last note.[4] The smallest units are conceived here as coherent melodic fragments or *melodic gestures*. Horizontal dashed lines represent time between these gestures. At the lowest level, the music thus comprises eight units, which straddle the bar lines.

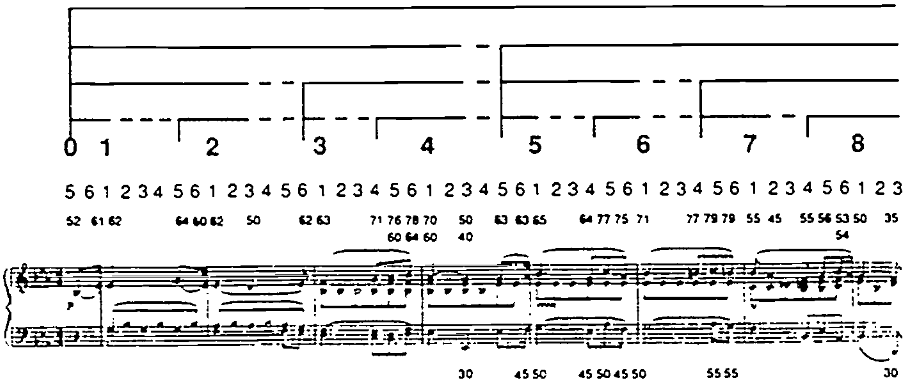


Figure 1. **Score of the musical excerpt used in Experiment 1. The computer-generated score follows the Breitkopf & Härtel Urtext edition, but bar 0 is from the first repeat, while bar 8 is from the second repeat. Above the score are the numbers of bars and eighth-notes, and a schematic representation of the hierarchical grouping structure. The small two-digit numbers above and below the score represent MIDI velocities for melody tones.**

These melodic gestures are of varying length, which bespeaks Beethoven's rhythmic ingenuity. At the second level, pairs of these short units can be grouped into longer units (phrases). The third level groups these phrases into two sections, and the highest level comprises the whole excerpt (which, in the original music, is followed by another eight-bar section that completes the structure of the minuet).[5]

According to this structural analysis, then, there is a hierarchy of unit boundaries at which final lengthening might be expected to occur in a performance. The "deepest" boundary (where units end at all four levels) is in bar 8 and coincides with the end of the excerpt; the next-deepest boundary is in bar 4; somewhat shallower boundaries are in bars 2 and 6; and the shallowest boundaries are in bars 1, 3, 5, and 7. Accordingly, the most pronounced lengthening (in fact, a *ritardando)* is expected in bar 8, a lesser lengthening in bar 4, and so on. As to the precise location of the expected lengthenings, however, especially at the shallower boundaries where only a single time interval may be affected, the *metric structure* must be taken into account. Final lengthening is generally expected on the last accented tone of a melodic gesture; if a gesture ends with an unaccented tone, it will usually be the preceding accented tone that is lengthened. The gestures ending in bars 2, 4, 7, and 8 have such "weak" endings, in which the penultimate accented tone acts as a harmonic suspense or *appoggiatura.* The last accented tone in each gesture always coincides with the downbeat (first tone) in a bar, so lengthening is generally expected on these metrically strong (and harmonically salient) tones. The metric structure of a piece forms a hierarchy similar to that of the grouping structure (Lerdahl & Jackendoff, 1983). Even though the two structures can be distinguished on theoretical grounds, for our present purpose they lead to the same predictions: final lengthening and emphatic lengthening mostly coincide.

The harmonic structure of the music will not be considered further, as it was not expected to have a major independent influence on the timing microstructure. One special feature of the piece should be noted, however, and that is the prescribed sudden *piano* in bar 7, which follows a *crescendo* through bars 5 and 6. Performers slow down substantially before this dynamic stepdown (Repp, 1990b), presumably to enhance its expressive effect.

An expert performance

Figure 2 presents the timing profile of an expert performance, obtained from a commercial recording by a famous pianist, Murray Perahia (CBS MT 42319). This performance, one of the finest in the set of 19 examined by Repp (1990b), was carefully remeasured for the present study.[6]

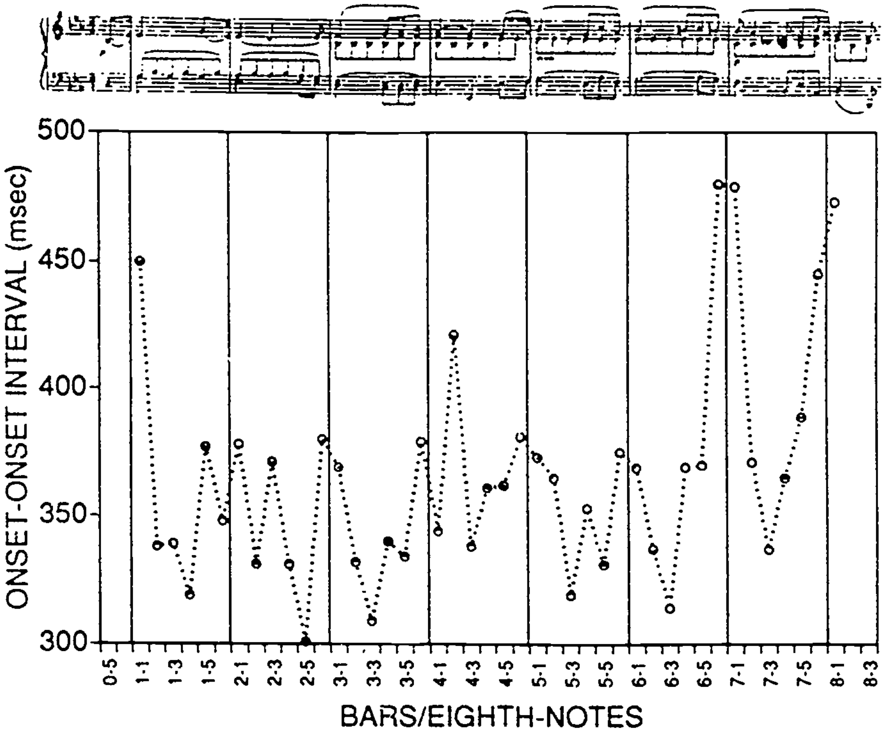


Figure 2. Timing pattern of an expert performance of the musical excerpt used in Experiment 1.

Tone onsets were determined in visual displays of the digitized acoustic waveform, and the durations of successive eighth-note onset-onset intervals (OOIs) were calculated.[7] These measurements were averaged over four repetitions of the same music (except for bar 1, which occurs in only two literal repetitions in the complete performance). The timing patterns were highly similar across the four repeats, and measurement error is believed to be less than 1%. Data for the initial upbeat (not an eighth-note) and for the final notes (which are not really final in the original music) are omitted in Figure 2.

Each data point in the timing profile represents the interval between the onset of the current eighth-note tone and the onset of the following tone. The pattern can roughly be characterized as a series of ups and downs, with troughs within bars and peaks near bar lines. These peaks are the predicted lengthenings. They generally fall on the first eighth-note interval in each bar, with the preceding interval often lengthened as well. An exception occurs in bar 4, where the second rather than the first eighth-note interval is lengthened; note, however, that both accompany the phrase-final accented tone, which is a quarter-note. Pronounced lengthening (in fact, a gradual *ritardando* spanning four eighth-notes) is found at the end of the excerpt, reflecting the deepest structural boundary there, and the smaller but still prominent peak in bar 4 is associated with a boundary at the next level of depth. The boundaries at the lower levels in the hierarchy are marked with smaller peaks, with two exceptions: there is substantial lengthening at the beginning of the piece (bar 1) and especially in connection with the sudden dynamic change from bar 6 into bar 7. Finally, it should be noted that the tempo of this performance invariably increased (i.e., OOIs decreased) between melodic gestures (marked by dashes in the diagram in Figure 1), as if to gain momentum for the next gesture.

These observations, even though they are derived from only a single performance, generally Confirm the predictions made about the timing microstructure. Below they will be compared to the perceptual results.

Methods

Stimuli. The musical stimuli were generated on a Roland RD-250S digital piano under the control of a microcomputer running a MIDI sequencing program (FORTE). The score was entered manually, such that each eighth-note tone occupied 60 "ticks," the internal time unit of the program. All tones were given their exact notated values (i.e., 30 ticks for a sixteenth-note, 120 ticks for a quarter-note, etc.), except for tones that were immediately repeated, which needed to be separated from the following tone by a brief silence; this silence arbitrarily replaced the last 5 ticks (in eighth-notes) or 2 ticks (in sixteenth-notes) of the nominal tone duration. There were no other temporal modifications in this isochronous performance; thus, for example, the slurs in the score were ignored. Except for repeated tones, therefore, the performance was entirely *legato,* without use of the sustain pedal, and the onsets of simultaneous tones were virtually simultaneous, within the-accuracy of the MIDI system. The tempo of the performance was determined by setting the "metronome" in the FORTE program to 88 quarternotes per minute. Accordingly, one tick corresponded to 5.68 ms, and the duration of a full eighth-note interval was 341 ms. The total performance lasted about 16.4 s.

To increase the musical appeal of the excerpt, the melody tones were assigned relative intensities (coded in the FORTE program as MIDI velocities) derived from a performance on the Roland keyboard by the author, an amateur pianist. These MIDI velocities are shown as the small two-digit numbers above or below the corresponding notes in Figure 1. All other tones (including all in the middle voice) were assigned a constant velocity of 40.[8]

The perfectly isochronous performance was presented to the subjects only as an initial example. In the experimental stimuli, a single eighth-note interval was lengthened by a small amount. The lengthening was achieved by extending all tones occupying that interval by the same number of ticks and by consequently delaying the onset of the following tone(s) and of all subsequent tones. (That is, the lengthening was *not* compensated by shortening the following tone.) Since there were 47 eighth-note intervals in the excerpt, there were 47 possible stimuli with a single lengthened interval. When an interval was bisected by the onset of a sixteenth-note tone, each sixteenth-note interval was lengthened by half the amount.

Four amounts of lengthening were used, based on pilot observations: 10 ticks (56.8 ms or 16.7%) for three initial examples; 8 ticks (45.4 ms or 13.3%) in the first block of trials; 6 ticks (34.1 ms or 10%) in the second block; and 4 ticks (22.7 ms or 6.7%) in the third block. The music was reproduced on the Roland with (synthetic, but fairly realistic) "Piano 1" sound and at an intermediate "brilliance" setting. Recordings were made electronically onto cassette tape. The experimental tape contained three repetitions of the isochronous performance, followed by the three examples of easily detectable lengthening (16.7%). Three blocks of 47 stimuli each followed, in order of increasing difficulty. The order of stimuli within each block was random. The interstimulus interval was 5 s, with longer intervals after each group of 10 and between blocks.

Subjects. Twenty musically literate subjects were paid to participate in the experiment. Most of them had responded to an advertisement in the Yale campus newspaper. They were 10 men and 10 women ranging in age from 15 to 52. A wide variety of musical backgrounds was represented, ranging from very limited musical instruction to professional competence. All subjects had played some musical instrument(s) at some time in their lives (piano, violin, guitar, and others), but only half of them still played regularly.[9]

Procedure. Subjects were mailed a cassette tape with instructions and answer sheets, and they listened at home on their own audio system.[10] They first viewed a sheet with the musical score on it, with bars and eighth-notes numbered. Then they listened to the examples; on the sheet, the lengthened intervals were indicated numerically as 3-3, 6-2, and 1-5, meaning the third eighth-note interval in bar 3, the second in bar 6, and the fifth in bar 1. After making sure that they heard these lengthenings as slight hesitations in the computer performance (they were allowed to rewind the tape and listen again to the examples, if necessary), the subjects turned to the answer sheets, each of which displayed the score on top. Subjects gave their responses in the numerical notation just explained. They were encouraged to follow along in the score with their pencil tip and to hold it where they perceived a hesitation, but to continue listening until the end of the music before writing down their response. They were asked to write down a question mark if they did not hear any hesitation; wild guesses were discouraged, though a response was welcome if subjects "had a hunch" of where the lengthening might have been. Subjects were asked to take the whole test in a single session, but to take short rests between blocks. Rewinding the tape was strictly forbidden during the test proper, and there was no indication that this instruction was not followed. At the end of the test, subjects completed a questionnaire about their musical experience before returning all materials to the author by mail.

Results and discussion

Overall accuracy. With 47 possible positions of the lengthened time interval, chance performance in this task was about 2%. It was immediately evident that the subjects performed well above chance at all levels of difficulty. However, they were frequently off by one, sometimes two, eighth-notes. Since these near-misses, in view of the low guessing probability, must have reflected positive detection of the lengthened interval in nearly all cases, all responses within two positions of the correct interval were considered correct. (Their distribution will be analyzed below.) By that criterion, chance performance was about 10% correct.

Overall performance was 57% correct. Not surprisingly, performance declined across the three blocks of trials as the amount of lengthening decreased, despite the possible benefits of practice; the scores were 70%, 61%, and 41% correct. This decline was of little interest in itself, and the data were combined across the three blocks in all following analyses. It should be noted that an average performance level of about 50% was optimal for observing variations in detection accuracy as a function of position.

The detection accuracy profile. Average detection accuracy as a function of position is shown in Figure 3. Each data point is based on 60 responses: 3 (blocks) from each of 20 subjects. It can be seen that there were enormous differences in the detectability of lengthening across positions, with scores ranging from 0 to 90% correct.

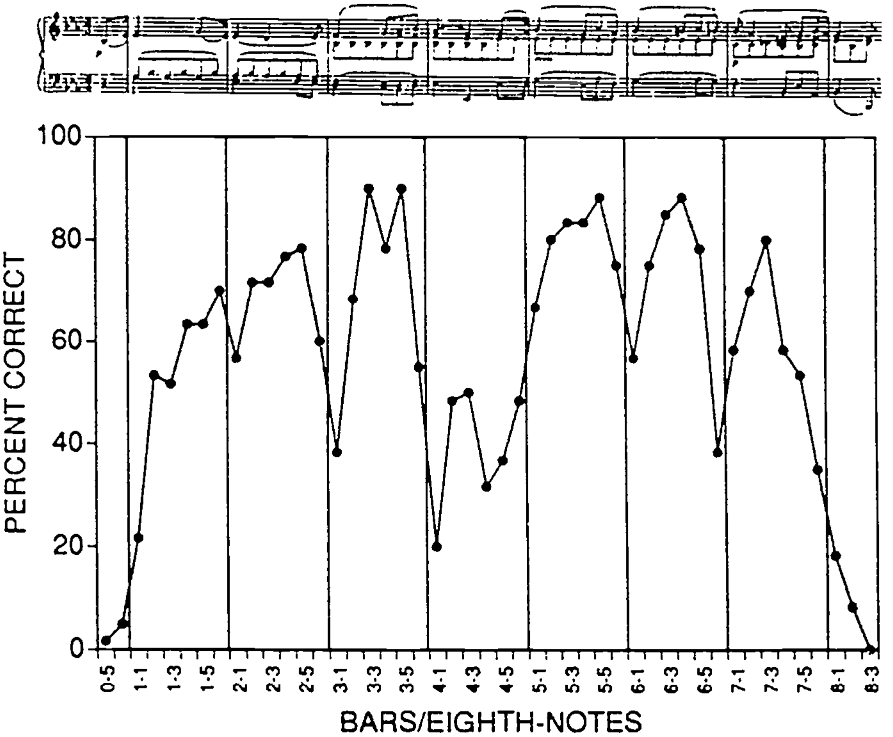


Figure 3. Percent correct detection as a function of the position of the lengthened eighth-note interval in Experiment 1.

Lengthening was never detected in the first two intervals and in the last interval. In the case of the initial intervals, this may be explained by the absence of an established beat at the beginning of the piece, and the final interval evidently did not have a clearly demarcated end. The poor score for the penultimate interval (position 8-2) and the steeply declining performance for the preceding intervals cannot be explained on these trivial grounds, however; they appear to be due to subjects' expectation of substantial lengthening (*ritardando*) toward the end of the excerpt.

The remainder of the accuracy profile is characterized by peaks and valleys, the peaks generally in the middle of bars and the valleys near bar lines. This pattern roughly the inverse of that observed in Murray Perahia's performance (Figure 2). The correlation is -0.59 ($p < .001$).[11] In general where Perahia slowed down, lengthening was more difficult to detect, and where he speeded up, detection performance improved. The correlation is not perfect, nor should it be expected to be, since both data sets presumably contain other sources of variability and are related only via the structural properties of the music.

In fact, the perceptual data seem to reflect the musical structure more closely than does the performance timing profile. The valleys in the accuracy profile generally coincide with the first note in each bar, which bears the metric as well as gestural accent. This coincidence is evident in bars 2, 3, 4, and 6. In bars 1 and 8, the drop to chance performance for the initial and final intervals, respectively, obscures any local valley on the downbeat, and a similar argument may be made for bar 5, where performance is recovering from an especially large dip in bar 4. The early peak in the last interval of bar 6 is due to the sudden dynamic change on the following note, which normally requires a preparatory lengthening or "micropause" in performance (see Figure 2). Thus the following valley is obscured here, too. Only the pattern in bar 4 remains unexplained: after a momentary rise in performance, there is a second dip in the fourth interval, which is not part of any melodic gesture. This interval, however, marks the end of the first half of the musical excerpt; at this point one could reasonably stop the performance. Thus the dip at position 4-4 may reflect a more abstract structural boundary of the kind represented in the conventional grouping hierarchy of Lerdahl and Jackendoff (1983). Bar 4 also contains a major disagreement with Perahia's performance: Perahia lengthens the second interval in this bar, whereas the dip in the accuracy profile occurs on the first interval. The detection data seem more internally consistent in that respect than Perahia's performance. On the whole, therefore, the perceptual results confirm the predictions based on metric and grouping structure. A final observation is that detection scores invariably increase between melodic gestures, where Perahia's performance just as consistently gained speed.

False alarms. The 43% incorrect responses consisted of question marks (26%) and false alarms (17%, $n = 479$). We now direct our attention to these latter responses, which were clearly unrelated to the intervals actually lengthened. That is, subjects did not hear the lengthened interval but nevertheless believed they had a hunch of where it might have been. Given the low probability of a correct guess, these trials thus served as catch trials, which may reveal the listeners' expectations.

Indeed, the false alarms were far from evenly distributed across positions. Their frequency distribution is shown in Figure 4. It is evident that it followed a pattern not unlike that of the "hits" shown in Figure 3. Here, too, there are peaks within bars and troughs near bar lines, although the peaks are narrower than those for the hits. The correlation between the hit and false-alarm profiles is 0.64 ($p < .001$); the correlation of the false-alarm profile with Perahia's performance timing profile is -0.40 ($p < .01$). This uneven distribution of false alarms signifies that there were certain intervals that *a priori* sounded longer than others, and these were ones in which actual lengthening also was easy to detect. These intervals were *not* likely to be lengthened in performance: typically, they either immediately preceded the onsets of melodic gestures or were gesture-initial but unaccented. Thus, under conditions of isochrony, intervals that were expected to be short sounded somewhat long, and intervals that were expected to be long sounded somewhat short.

The variation in hit rates greatly exceeded that in false-alarm probabilities; note that the false-alarm percentages have been magnified ten times in Figure 4 relative to the hit percentages in Figure 3, so as to make their pattern more visible. The comparison is not quite fair, perhaps, because the scoring criterion for correct responses included responses that were off by one or two positions ("near-misses"). However, even when only the percentages of "true hits" are considered (see Figure 5 below), their variation is nearly ten times as large as that of the false alarms.

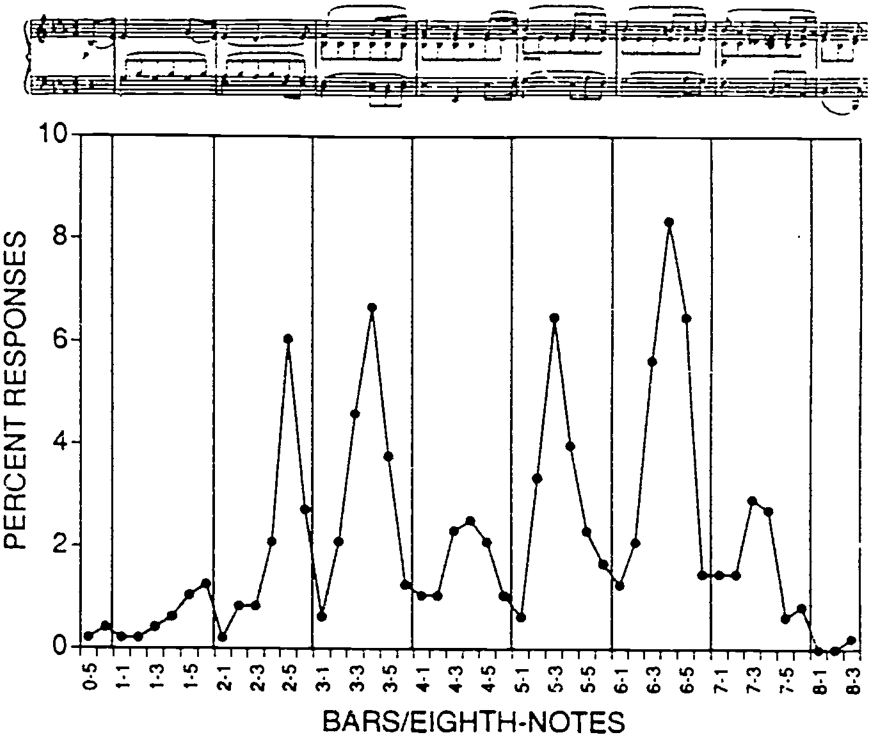


Figure 4. Distribution of false alarms across eighth-note positions in Experiment 1.

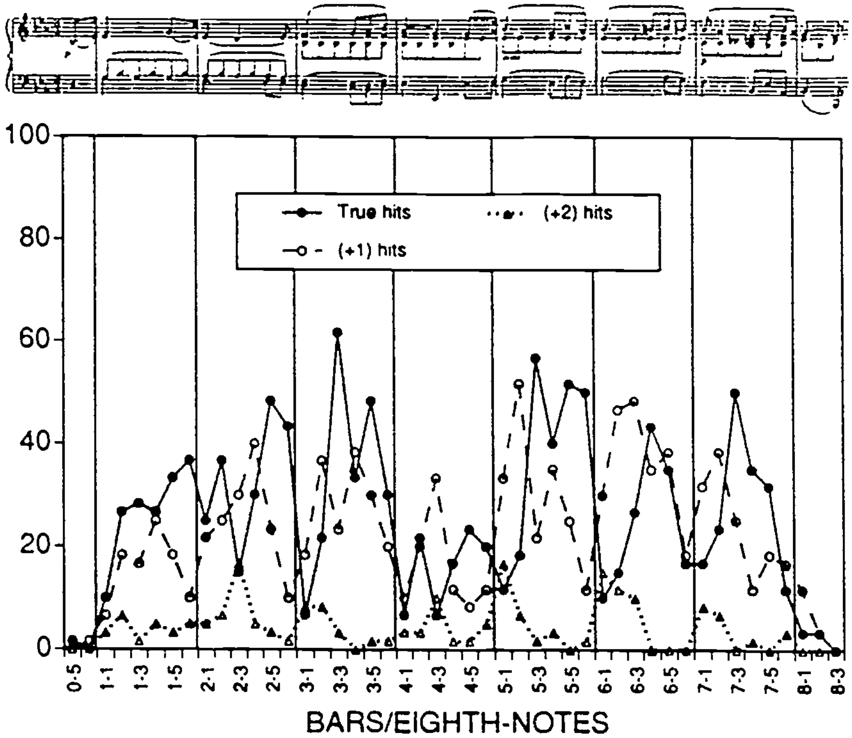


Figure 5. Percentage profiles of true hits (i.e., on-target correct responses), and of (+1) and (+2) hits (i.e., near-misses) in Experiment 1.

Still, hits and false alarms are probably independent manifestations of the same underlying cause, namely, listeners' structurally determined expectations about timing microstructure. These expectations affect the *criterion* for perceiving a particular interval as lengthened.

Near-misses. Averaged across all 47 positions, correct responses (57% overall) were distributed as follows with respect to the correct position: 1% (-2 positions), 3.6% (-1), 25.7% (on target), 22.6% (+1), and 4.2% (+2). Thus there was a pronounced tendency to locate the lengthened interval in the immediately following position. Other types of near-misses were much less frequent, though (+2) and (-1) responses were about twice as frequent as expected by chance (2%). One possible explanation for this postponement tendency is that subjects necessarily heard the following interval at the time that they realized that an interval had been lengthened; thus, they probably arrested their pencil at that point in the score and forgot to backtrack when writing down their response. Another possibility is that subjects neglected the instructions and attributed the "hesitation" to the tone whose onset was delayed, rather than to the preceding lengthened interval. The delayed tone may also have sounded slightly more intense because of the additional decay of the preceding tone during the lengthened interval (cf. also Clarke, 1988, p. 19, who notes that a delay "heightens the impact" of the delayed note).

If subjects simply had forgotten to backtrack, the profile of (+1) responses should have paralleled that of the true hits; in other words, the relative frequencies of true hits and (+1) responses should have been constant across positions. This was not the case, however. The percentage profiles of true hits, (+1) hits, and (+2) hits are shown in Figure 5.[12] It is evident that each type of response had a distribution with peaks and troughs, but that the peaks in the (+ 1) and (+2) profiles were generally one and two positions, respectively, to the left of the peaks in the true hit profile (which generally coincide with the peaks in the false-alarm profile—cf. Figure 4). In other words, (+1) responses were most frequent when the lengthened interval occurred immediately before an interval that was *a priori* more likely to be perceived as lengthened, and (+2) responses were most frequent when another interval intervened between these two. Thus, the positional distribution of correct responses was modulated by the same perceptual biases that governed the overall accuracy and false alarm profiles.

These observations receive statistical support: the correlation between the true hit and (+1) profiles aligned as in Figure 5 is 0.24 (n.s.), but it becomes 0.63 ($p < .001$) when the (+1) profile is shifted to the right by one position. Similarly, the correlation between the original true hit and (+2) profiles is -0.26 (n.s.), whereas it is 0.55 ($p < .001$) when the (+2) profile is shifted to the right by two positions.[13] The shifted (+1) and (+2) profiles are also significantly correlated with the false-alarm percentages shown in Figure 4 (0.76 and 0.48, respectively; both $p < .001$), as is the true hit profile (0.60, $p < .001$). Thus, (+1) and (+2) responses, just like true hits, reflect the fact that structurally weak intervals seem *a priori* longer than others.

Yet another way of demonstrating the similarity of the hit and false-alarm profiles is to plot the *frequency distribution* of all correct responses (i.e., the relative frequency with which each position was given as a correct response) and to compare this distribution with that of the false alarms (cf. Figure 4). These two distributions are superimposed in Figure 6. Their similarity is striking ($r = 0.77$, $p < .001$). Virtually all the peaks coincide, and both functions consistently show minima immediately after bar lines. The main difference is that the correct response distribution is flatter, presumably because of the positional constraint imposed by lengthenings that were actually detected, whereas the false-alarm distribution reflects pure perceptual bias. The two distributions evidently reflect the same underlying tendencies (i.e., listeners' expectations).

Individual differences and musical experience. Individual subjects' overall accuracy in the detection task varied considerably, from 26% to 83% correct. One measure of interest was the extent to which each individual subject's accuracy profile (based on only three responses per position, alas) resembled the grand average profile shown in Figure 3. These individual "typicality" correlations ranged from 0.37 to 0.84. The correlation between these correlations and overall percent correct was 0.69 ($p < .001$), indicating that the more accurate subjects also showed the more typical profiles.

Correlations were also computed between the individual accuracy and typicality measures on one hand, and sex, age, and several indices of musical experience derived from the questionnaire responses on the other hand. No striking relationships emerged. There was no sex difference.

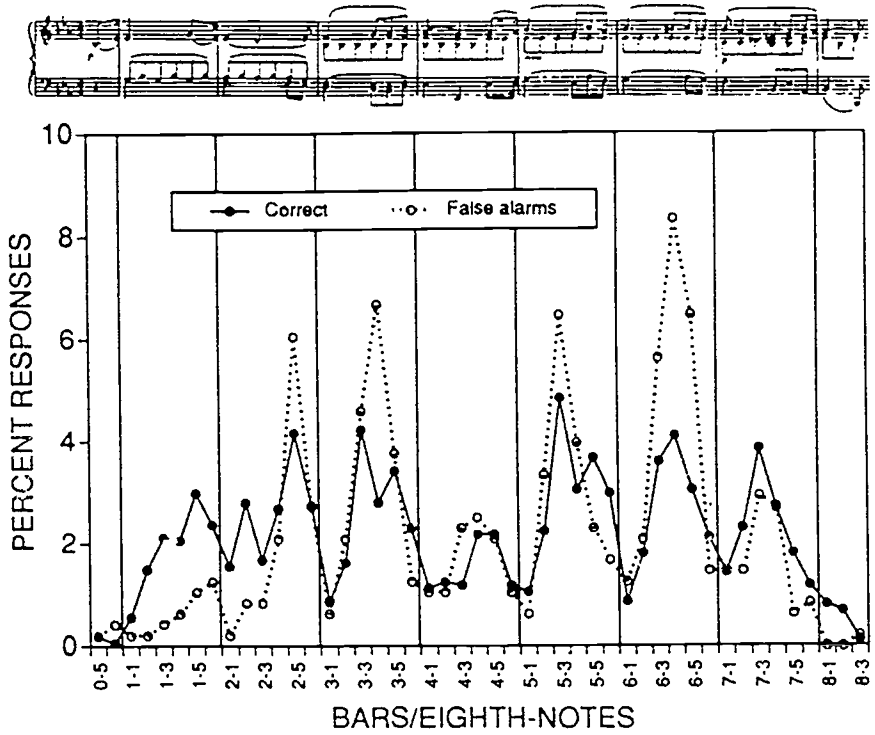


Figure 6. Frequency distribution of all correct responses combined (according to the response given, not according to the actual location of the lengthened interval) in Experiment 1, compared with the false-alarm distribution (from Figure 4).

Correlations with age were negative but nonsignificant. Correlations with years of musical instruction (ranging from 1 to 34, added up across all instruments studied) were positive but nonsignificant. With regard to hours per week of active music making (ranging from 0 to 30), the correlation with the typicality measure, 0.42, reached the $p < .05$ level of significance. In view of the multiple correlations computed, however, this could be a chance finding. Correlations with hours per week spent listening to music were negligible. Finally, pre-experimental familiarity with the music seemed to play no role. (Twelve subjects indicated that they were totally unfamiliar with the music, 4 subjects were somewhat familiar, and 4 subjects professed greater familiarity.) In the absence of significant correlations, it seems highly unlikely that subjects' timing expectations derived from remembered previous performances of the Beethoven minuet; rather, these expectations must have been induced by the musical structure during the experiment, despite the absence of systematic timing microstructure.

EXPERIMENT 2

Although the results of Experiment 1 demonstrate convincingly the influence of musical structure on listeners' judgments, they were based on only one musical excerpt. Moreover, measurements from a single expert performance were available for comparison. Experiment 2 attempted to replicate the findings using a different musical excerpt, for which detailed timing measurements from 28 different expert performances were available (Repp, 1992). This provided a unique opportunity to relate listeners' performance expectations (as reflected in the detection accuracy profile) to a representative measure of actual performance microstructure.

Musical material

The music used in this experiment constituted the initial eight bars of "Träumerei" ("Rêverie"), a well-known piano piece from the cycle "Kinderszenen," Op. 15, by Robert Schumann. The key is F major, and there is no verbal tempo indication. The score, without slurs, is reproduced in Figure 7.[14]

Again, the predominant note value is the eighth-note, which served as the temporal unit in the experiment. The time signature is common time (4/4), so that there are 8 eighth-notes per bar. The excerpt contains 62 eighth-note intervals, not counting the final chord.[15] However, not all of these intervals are marked by tone onsets; 12 of them are unmarked, as indicated by the dashes in the numbering above the score in Figure 7. Thus there were only 50 OOIs to be measured or manipulated, some of which were multiples of eighth-note intervals.

The music is more complex than that of Experiment 1 in several other respects. Horizontally, four voices (soprano, alto, tenor, bass) can be distinguished, most clearly in bars 3-4 and 7-8. Unlike the Beethoven minuet, where the bass voice merely supported the melody and followed essentially the same grouping structure, the voices in "Träumerei" are more independent of each other and constitute different layers of sometimes staggered melodic gestures.

The vertical structure is represented in Figure 7 by the diagram above the score. At the higher levels, it is similar to that of the Beethoven minuet: one large 8-bar section can be divided into two 4-bar sections, which in turn can be divided into four 2-bar phrases. Even at the lowest level, bars 1-2 and 5-6 resemble those in the minuet, each being composed of two melodic gestures, the second longer than the first. Where the Schumann piece differs from the Beethoven minuet, and becomes considerably more complex, is in bars 3-4 and 7-8.

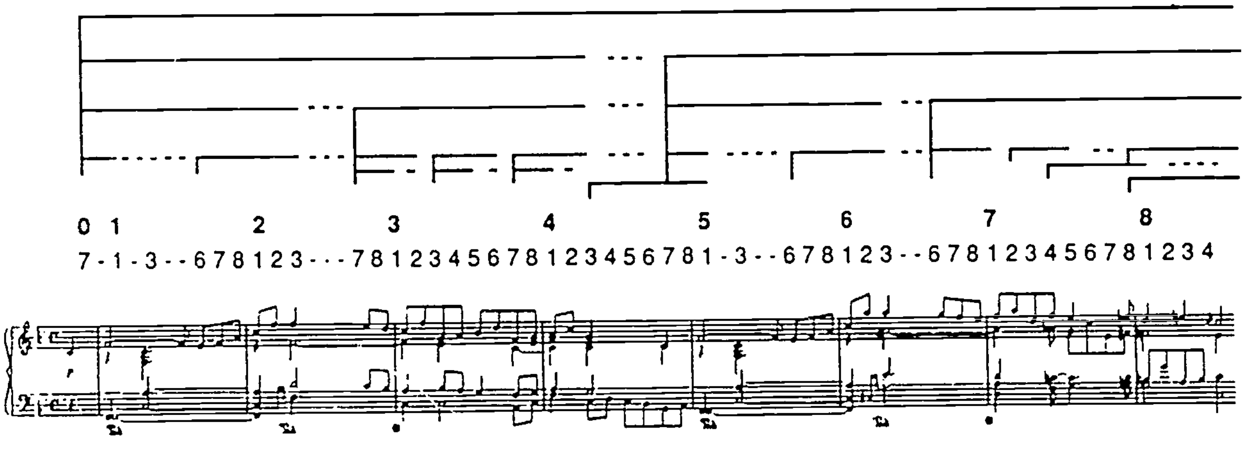


Figure 7. Score of the musical excerpt used in Experiment 2. The computer-generated score follows the Clara Schumann (Kalmus) edition, but all slurs and expressive markings have been removed. Above the score are the numbers of bars and eighth-notes, and a schematic representation of the hierarchical grouping structure.

The phrase spanning bars 3-4 is constituted of four shorter gestures. The first two, in the soprano voice, each comprise 4 eighth-notes and are accompanied by even shorter gestures in the tenor voice. The third gesture in the soprano voice, joined now by tenor and bass, has an additional final note which overlaps the first note of the fourth gesture, now in the bass. That gesture essentially comprises 7 notes, ending on the low F in bar 5, and thus overlaps with the first gesture of the second major section. In fact, the offsets of these two gestures coincide; thus the last gesture of the first major section provides a link with the second major section. Bars 7-8 are even more complex, and quite different from bars 3-4. Instead of being divided between soprano and bass, the melodic gestures descend from soprano to alto to tenor, overlapping each other in the process. The first gesture begins with an extra note and comprises either 8 eighth-notes or two groups of 4. Anticipating performance data to be discussed below, we assume that it consists of two 4-note gestures. The end of the second gesture overlaps the beginning of the third gesture in the alto (6 eighth-notes), which in turn overlaps the 6-note gesture in the tenor (possibly even extends through it). Although the overall melodic contour seems similar to bars 3-4, the notes are grouped differently. Alternative phrasings are conceivable, though gesture onsets are made unambiguous by the entrances of the different voices.

The metric structure of the piece is irregular. The melodic gestures in bars 1-2 and 5-6 negate the accent on the downbeat and instead postpone it to the second quarter-note beat in bars 2 and 6, respectively. This contradicts the placement of the bar lines in the score; perhaps, if Schumann had lived 100 years later, he would have changed the meter to 5/4 in bars 1 and 5 and to 3/4 in bars 2 and 6. A similar but weaker asymmetry exists in bars 3-4 and 7-8, where there is a strong accent on the second beat of bars 4 and 8, respectively. These metric irregularities are not visible in the score (except for the occurrence of rich chords on the second beats in bars 2, 4, and 6); their existence is derived from musical intuition and knowledge of standard performance practice. Apart from a knowledge of where accents are likely to fall, however, the metric structure as such contributes little to predicting lengthening beyond what can be derived from the grouping structure shown in Figure 7. The same can be said about the harmonic structure, though the salient chromatic transitions in positions 7-4 and 7-8 should be noted; they are likely to receive some emphasis in performance, even though they are metrically weak. Another special feature is the occurrence of grace notes in intervals 2-2 and 6-2, which are likely to result in extra lengthening of the intervals accommodating them.[16]

Performance microstructure

Salient aspects of performance timing are illustrated in Figure 8. This figure shows a timing microstructure extracted from a sample of 28 famous pianists' performances by means of principal components analysis (Repp, 1992). The values represent the factor scores of the first principal component, rescaled into the millisecond domain. Thus, this pattern, while not necessarily corresponding to an outstanding performance, captures what is common to many different performances and contains very little unsystematic variation. The OOIs are plotted in terms of eighth-note intervals; longer OOIs are represented as "plateaus" of multiple eighth-note intervals. The grace notes in bars 2, 6, and 8 are not represented.

Again, we can observe a pattern of peaks and valleys, with the peaks reflecting lengthening, mostly at the ends of melodic gestures. Thus, there is a peak at the end of the excerpt, and another at the end of bar 4, which marks the end of the first major section (even though the second section has already begun). Furthermore, the long intervals (in bars 1, 2, 5, and 6), all of which are gesture-final, are relatively extended. The lengthened first upbeat reflects a common start-up effect. The second eighth-note intervals in bars 2 and 6, which are penultimate in their respective gestures, are extra long because they must accommodate the two grace notes (a written-out *arpeggio)* in the bass voice. Finally, a series of small peaks can be seen in bars 3-4 and 7-8. Those in bars 3-4 reflect final lengthening for the eighth-note gestures, while those in bars 7-8 reflect the brief ge..tures in the bass voice, which are coupled with salient harmonic progressions.

This generic performance pattern may serve as the basis for predictions about the detectability of actual lengthening of individual intervals in an otherwise isochronous performance. Lengthening should be most difficult to detect where there are peaks in the performance timing profile, and easiest where there are valleys.

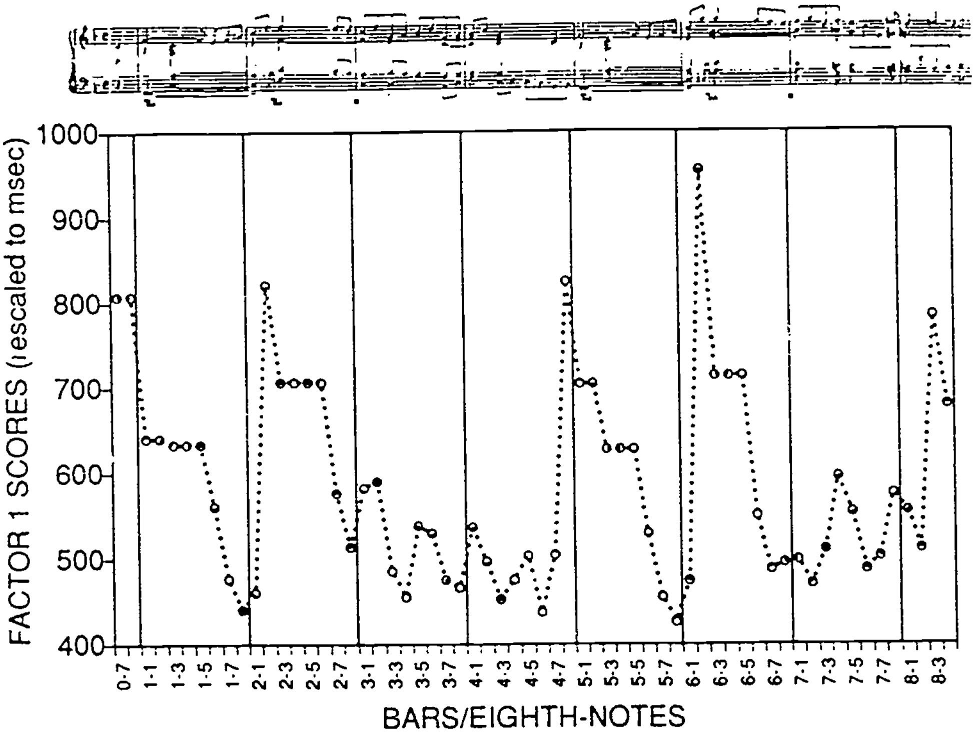


Figure 8. Generic performance timing microstructure for the musical excerpt of Experiment 2, derived by principal components analysis from the timing patterns of 28 performances by famous pianists (Repp, 1992).

Methods

Stimuli. The score was entered manually into the FORTE program, with 100 ticks per eighth-note interval. The tempo was set at the equivalent of 60 quarter-notes per minute, so that the duration of an eighth-note interval was 500 ms, and the temporal resolution (1 tick) was 5 ms. The total duration was about 33 s. All tones lasted for their nominal duration, except: (1) those that were immediately repeated, whose last 25 ms were replaced with silence, (2) the left-hand chords tied over into bars 2 and 6, which were shortened by 500 ms, so that they ended with the first eighth-note interval in these bars; and (3) the quarter-note interval preceding the grace note in bar 8, which was shortened by 250 ms, making it effectively a dotted eighth-note interval followed by a sixteenth-note interval (the grace note) lasting 250 ms. The grace notes in bars 2 and 6 started 165 and 330 ms, respectively, into the eighth-note interval; the first of them ended with the onset of the second. Except for repeated tones, then, the performance was strictly *legato*. Sustain pedal was added as indicated, with pedal onset 25 ms after, and pedal offset 25 ms before, the relevant tone onsets.[17]

A fixed intensity microstructure was imposed on the tones (see Table 1). The relative intensities were modeled after acoustic measurements of a recorded performance by a distinguished pianist, Alfred Brendel (Philips 9500 964), and were adapted to the dynamic range of the Roland digital piano. The precise methods of this transfer will not be described and defended here; suffice it to say that the intensity pattern seemed musically appropriate and pleasing to the ear.

Table 1. *Intensity microstructure (in MIDI velocities) of the musical excerpt used in Experiment 2. As much as possible, the tones have been assigned to four voices (* = grace notes).*

Bar /eighth-note		Bass		Tenor		Alto		Soprano	
0-7								72	
1-1		42						73	
1-3			26	25	55	61			
1-6								62	
1-7								74	
1-8								70	
2-1								76	
2-2								80	
2-3	48*	40*		59		62		79	
2-7				59				62	
2-8				55				86	
3-1		36		76				73	
3-2								78	
3-3		45		60		67		71	
3-4				76				78	
3-5				59				73	
3-6								88	
3-7		23		41		80		84	
3-8		53		51			74	76	
4-1		46		56			76	76	
4-2								81	
4-3		48		52		60	69	69	
4-4		61							
4-5		68							
4-6		63							
4-7		49						89	
4-8		49							
5-1		25						76	
5-3			30	43	58	69			
5-6								66	
5-7								75	
5-8								74	
6-1								82	
6-2								98	
6-3	55*	65*		57	55	63		91	
6-6								69	
6-7								76	
6-8								75	
7-1		32		65				80	
7-2								90	
7-3								90	
7-4		33		23		36		84	
7-5		48				57		79	
7-6						76			
7-7						77			
7-8		42		57		68		89	
8-1		37		59		62		80	
8-2				58					
8-3				51				55	
8-4								60	54*
8-5		15		61		32		51	

With the final chord and the grace notes excluded, there were 50 possible intervals to be lengthened, most of them corresponding to an eighth-note, but some longer. All intervals were lengthened in proportion to their nominal duration; thus, for example, the half-note interval in bar 2 was extended by four times the amount applied to an eighth-note interval. The grace notes in bars 2 and 6 were never lengthened; when the eighth-note interval that contained them was lengthened, the onset of the first grace note (and the corresponding pedal onset) was delayed by that amount. The grace note in bar 8, on the other hand, was lengthened by half the amount of the lengthening applied to the simultaneous eighth-note interval.

For reasons of economy, it was decided to lengthen *two* intervals in each presentation of the excerpt: one in the first half and one in the second half. The second half was defined as beginning with the upbeat on the fourth beat in bar 4; that way, there were 25 possible intervals to be lengthened in each half. The two intervals lengthened in each stimulus were randomly chosen, with the restriction that they never occupied structurally analogous positions in the two halves of the excerpt.

An experimental tape *(Test A)* was recorded, containing three blocks of stimuli preceded by six examples. The first three examples were entirely isochronous; in each of the next three, two intervals were lengthened by 16%. Each of the three successive blocks contained 25 stimuli, arranged in random sequence with ISIs of 5 s, 10 s after each group of 10, and longer intervals between blocks. The amounts of lengthening employed in the three blocks were 14% (70 ms), 12% (60 ms), and 10% (50 ms), in that order.

The degrees of lengthening were chosen on the basis of the author's impressions during stimulus construction; however, as might have been anticipated from Experiment 1, they proved too easy to detect in eighth-note intervals for most listeners, so that there was a ceiling effect in some regions of the accuracy profile. A supplementary test tape *(Test B)* was therefore constructed. It contained three initial examples with 10% lengthening and two additional blocks of stimuli with 8% (40 ms) and 6% (30 ms) lengthening, respectively. Only the eighth-note data from this test were considered.[18]

Subjects. The subjects were musically literate Yale undergraduates who attended a course on the psychology of music. Participation was voluntary, and subjects were paid for their services. Sixteen subjects completed Test A, and 7 of them listened to Test B some weeks later. One additional subject, who had not received Test A, completed a special test comprising two blocks of stimuli with 10% and 8% lengthening, respectively; his 10% data were included in the Test A totals, and his 8% data in the Test B totals.

Procedure. The procedure was the same as in Experiment 1, with two exceptions: first, as already mentioned, subjects gave two responses per presentation rather than just one. Second, rather than giving a numerical response, subjects circled the lengthened note(s) in a copy of the score. (It was sufficient to circle any of several simultaneous notes.) The answer sheets provided a separate miniscore for each presentation in the sequence. The possible responses were first illustrated on a sheet showing also the locations of the lengthened intervals in the initial examples.

Results and discussion

Overall accuracy. With 25 possible choices in each half of the excerpt, the chance level of performance was 4% correct-twice as high as in Experiment 1. Also, subjects seemed to be on target, relatively more often than in Experiment 1. Therefore, the criterion for accepting responses as correct was tightened, and only responses that were within one position of the correct interval were considered correct. (Long intervals were counted as a single position.) That way, chance level was 12% correct.

Overall, subjects scored 74% correct on Test A. Average performance declined across the three blocks from 78% to 75% to 70% correct. This surprisingly small effect of increasing the difficulty of the task may have been due in part to a counteracting effect of practice; in part, the reason was that the task was a little too easy, leading to a ceiling effect for most eighth-note intervals. (The scores for eighth-note intervals only were 82%, 79%, and 74%, respectively, in the three blocks.) However, even in Test B, where only eighth-notes were scored, performance was still quite good: 59% and 50% correct, respectively. The amount of lengthening in the last block was 6%, or 30 ms, which evidently was still quite detectable.

Again, since absolute performance levels were not of particular interest, the data were combined across all levels of difficulty in each test.

The accuracy profile. Figure 9 shows percent correct detection as a function of position in the music. The solid line shows the results of Test A, where each data point is based on 49 responses.

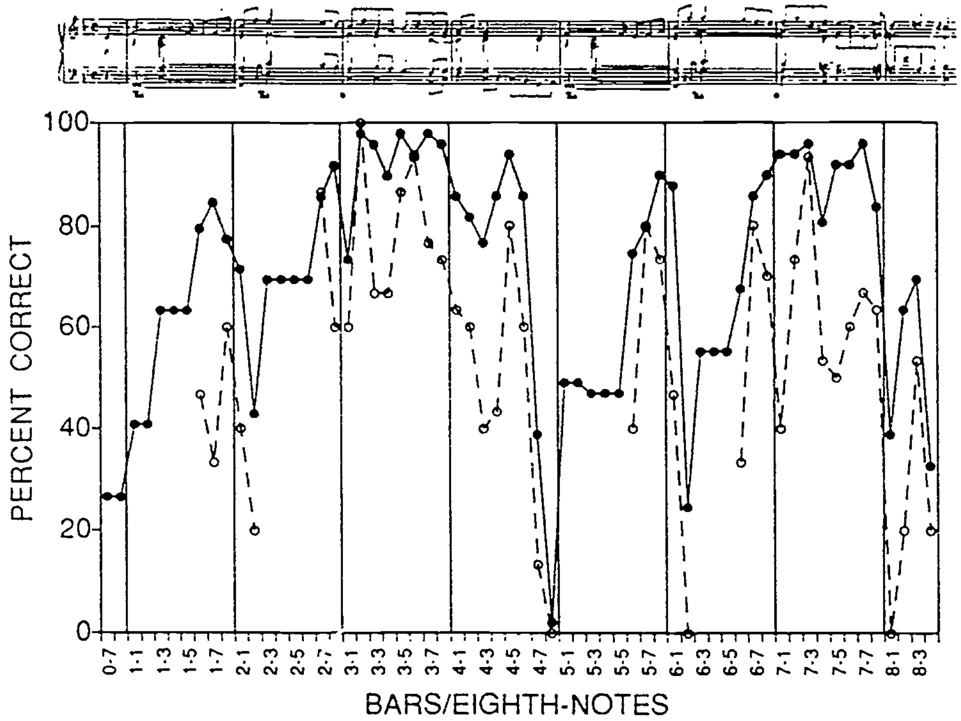


Figure 9. **Detection accuracy profiles obtained in Experiment 2. The solid line represents Test A results, the dashed line Test B results (eighth-notes only).**

The dashed line represents the results of Test B (eighth-notes only), where each data point derives from 15 responses. As in the generic performance timing profile (Figure 8), the results for long notes are shown as plateaus; they really represent only a single data point.

Consider the Test A results first. It can be seen that performance varied dramatically across positions, from chance to nearly perfect detection. Lengthening was moderately difficult to detect in all long intervals. It was easy to detect in most eighth-note intervals, but with significant exceptions. Performance was at chance for the last eighth-note interval of bar 4, and rather poor for the preceding interval. These tones represent the end of the melodic gesture in the bass voice which marks the end of the first major section of the excerpt and coincides with the upbeat to the second major section. Lengthening was also difficult to detect in the eighth-note interval preceding the final chord, in the first eighth-note interval of bar 8, and in the second eighth-note intervals in bars 2 and 6, which coincide with grace notes. Most of these intervals also tend to be lengthened in performance; the accuracy profile is very nearly the inverse of the performance timing profile shown in Figure 8. The correlation between these two profiles is -0.75 ($p < .001$).[19] Thus the results are again consistent with the hypothesis that lengthening is more difficult to detect where it is expected.

Clearly, the gross accuracy profile mirrors the major lengthenings encountered in performance. It is less clear whether there is also any correspondence at the more molecular level of the eighth-note intervals in bars 3-4 and 7-8, because of the paucity of errors in these regions in Test A. Here the Test B results are useful. As can be seen in Figure 9, the Test B results generally confirm but magnify the tendencies observed in the Test A profile. The correlation between the Test A and Test B eighth-note profiles is 0.84 ($p < .001$). The correlations between each of these two profiles and the performance timing profile (eighthnotes only) are -0.70 and -0.49, respectively (both $p < .001$). The second correlation is a good deal lower, suggesting that the match at this more detailed level is not so close, and that the significant correlations may be largely due to the several eighth-notes that yielded extremely poor scores.

Let us compare, therefore, the exact locations of peaks and valleys in the performance timing and accuracy profiles. The ascending eighth-note gestures straddling bars 1-2 and 5-6 exhibit an acceleration of tempo followed by a dramatic slowing on the penultimate note (Figure 8). This pattern is mirrored rather nicely in the perceptual data (Figure 9). In bars 3 and 4, however, the correspondence breaks down. In the performance profile, there are small peaks in positions 3-1 and 3-2, 3-5 and 3-6, 4-1 and 4-5. These locations do not correspond to valleys in the accuracy profile; on the contrary, there are accuracy peaks in some of these positions. In fact, at this local level (from position 2-6 to position 4-6) there is a *positive* correlation between the performance and Test B accuracy profiles: $r(16) = 0.57(p\ (.02)$. In bar 7, there are small performance timing peaks in positions 4-5 and 4-8. The first two locations do have a corresponding valley in the Test B accuracy profile, but the last one does not. Conversely, in the accuracy profile there are pronounced dips in positions 6-6, 7-1, and 8-1 that do not correspond to lengthenings in performance. Nor is there a dip relating to the pronounced *ritardando* at the end of the excerpt. Thus, although there are clear peaks and troughs in the accuracy profile, they do not mirror the performance profile at this detailed level. We will examine later what factors they might reflect (see General Discussion).

False alarms. The false-alarm responses were pooled across Tests A and B. There were 300 such responses altogether, or 9.4% of all responses. As in Experiment 1, they were not evenly distributed. Figure 10 shows their frequency histogram.

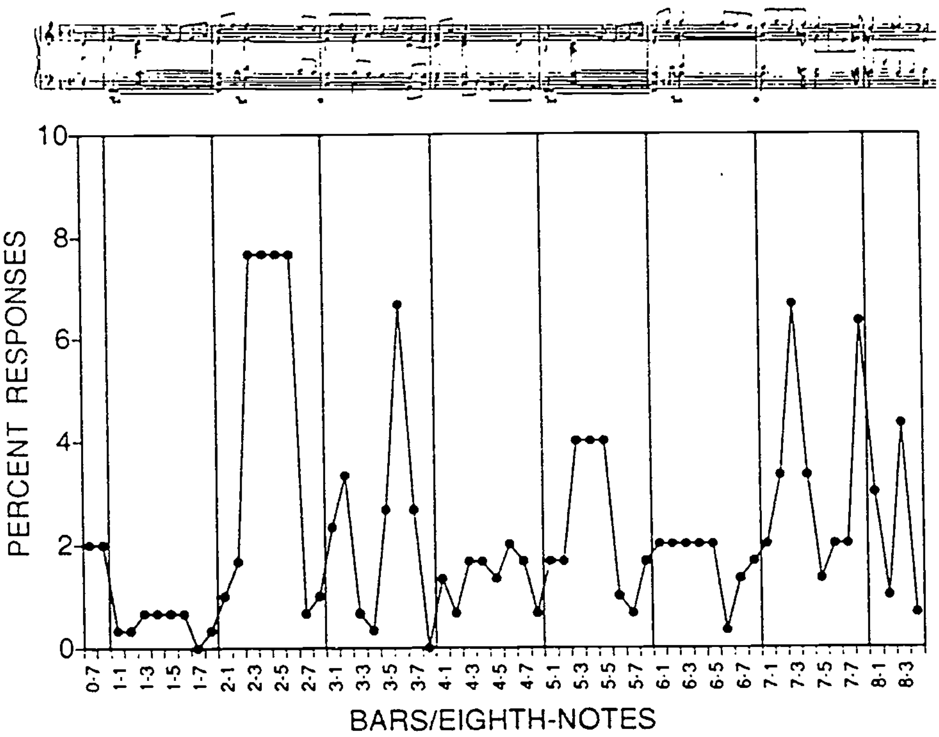


Figure 10. **False-alarm distribution in Experiment 2.**

There are some seven intervals that tended to attract false alarms. Two of them are long (in bars 2 and 5); the others are eighth-note intervals. The eighth-note peaks in the false-alarm profile do coincide with peaks in the accuracy profile, especially that of Test B. Otherwise, however, the correspondence is not close. The overall correlations are 0.12 (Test A, all intervals, n.s.) and 0.31 (Test B, eighth-note intervals only, $p < .05$). Thus, again, there were certain intervals that *a priori* seemed longer than others. As in Experiment 1, some of these intervals were located immediately preceding the onset of melodic gestures. The long notes, of course, may have attracted false alarms simply because of their length. The false alarm peaks in positions 7-8 and 8-3, and the corresponding peaks in the accuracy profile, have no obvious explanation at present.

Near-misses. As in Experiment 1, there was a tendency to postpone responses by one position, but it was less pronounced, perhaps due to the slower tempo of the music. Mislocations to the preceding interval were rather infrequent. The percentages of these two types of near-misses are shown, together with the "true hit" accuracy profile, in Figure 11. The results of Tests A and B have been combined in this figure; thus the scores for long intervals (which derive from Test A only) appear somewhat elevated. Unlike Experiment 1, there was no convincing tendency for the peaks in the (+ 1) profile to be shifted one position to the left with respect to the peaks in the true hit profile. The correlation between these two profiles is 0.46 (*p*(.001); it is only slightly higher (0.54) when the (+l) profile is shifted one position to the right. Thus, the influence of perceptual bias on correct responses was less evident than in Experiment 1.

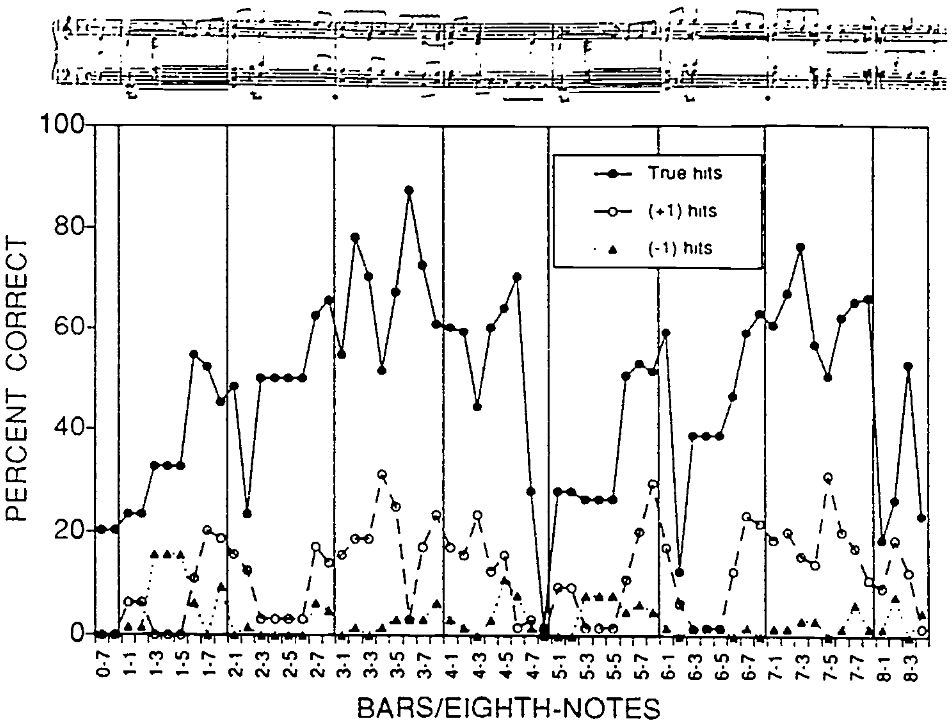


Figure 11. Accuracy profiles for true hits, and for (+1) and (-1) hits (near-misses) in Experiment 2.

Individual differences and musical experience. Individual overall accuracy scores ranged from 52% to 95% correct in Test A, and from 42% to 80% correct in Test B. None of the (generally positive) correlations with measures of musical experience reached significance. Even though these subjects had, on the whole, received more musical training than the subjects of Experiment 1, only two subjects indicated more than a passing familiarity with the music.[20] The subject with the highest scores in both tests had received only limited musical instruction and neither played nor listened to classical music. The next highest score was achieved by an accomplished pianist, but the third-highest score belonged to one of the two subjects in this group who had no musical training at all. It is possible that structurally guided expectations inhibit rather than facilitate the detection of lengthening, which would counteract any advantage musicians may have in general rhythmic sensitivity.

GENERAL DISCUSSION

On the whole, the two experiments reported here support the hypothesis that lengthening of musical intervals is more difficult to detect where lengthening is expected. Expectation was defined here mainly with reference to the hierarchical grouping structure and its reflection in the timing microstructure of expert performances. In both experiments, there was a highly significant negative correlation between a representative performance timing profile and listeners' detection accuracy profile: Where performers typically slow down, listeners had trouble detecting lengthened intervals as hesitations. The reason for this is presumably that lengthened intervals in these positions sound regular, while lengthened intervals in other positions are perceived as hesitations that disrupt the rhythm.

It is noteworthy that this result was obtained in the context of a psychophysical detection task. It would be less surprising, for example, if listeners had been asked to provide aesthetic judgments of noticeably lengthened intervals and had been found to be in agreement with expert performance practices. The present listeners, however, were not making aesthetic judgments but listened very carefully for a perturbation in an isochronous sequence of sonic events. Some listeners were remarkably accurate in this task. Nevertheless, they consistently missed certain intervals—usually the ones that were most likely to be lengthened in music performance.

One interpretation of these findings is that, despite the simple task, listeners processed (and learned) the musical structure, which in turn elicited expectations of timing microstructure via some internal representation of performance rules. These top-down expectations *interacted* with subjects' perception of the temporal intervals; they made certain intervals seem longer than others, on which actual lengthening then was especially easy to detect, whereas other intervals seemed relatively short and were difficult to perceive as lengthened. Microstructural expectations thus "warped" the musical time scale in accordance with the perceived grouping structure. This may be a manifestation of the "dynamic attending" discussed by Jones and Boltz (1989).

If this interpretation is correct, several corollaries follow. First, most, if not all, listeners must have grasped the musical grouping structure, at least at the higher levels of the hierarchy, even though a number of them had little music training beyond an ability to read notation. Second, the processing of musical structure and the activation of performance expectations seem to be obligatory. Even though listeners were exposed to many repetitions of mechanically regular performances, there was no indication that their expectations weakened (i.e., that the accuracy profile became flatter) in the course of the experimental session. Third, the origin of these expectations must be in subjects' past experiences with music performance in general, or perhaps in even broader principles of prosody and of the dynamic structure of events. They cannot derive from exposure to performances of the specific music excerpts, for many subjects (especially in Experiment 1) were in fact unfamiliar with these excerpts and never were presented with an expressively timed performance in the course of the experiment.

These conclusions are intriguing and potentially important for our understanding of musical perception and judgment. There is an alternative interpretation of the data, however, which we need to examine closely now: the observed variations in detection accuracy may be direct reflections of the complex *acoustic properties* of the musical stimuli, rather than of expectations elicited by the musical structure as such. The apparently obligatory nature of the phenomenon, its consistency across listeners of widely varying musical backgrounds, and its independence from specific experience with the music would all be consistent with a "bottom-up" account.

It is difficult, of course, to dissociate acoustic from musical structure: music *is* an acoustic structure, and acoustic variations along a number of dimensions in fact define the musical structure for the listener.[21] Although acoustic structure, therefore, is indispensable in music, it may be asked whether there are basic psycho-acoustic principles that make it possible to explain the local variation in detection accuracy without explicit reference to musical structure. If that were the case, it would seem that composers have taken advantage of certain principles of auditory perception in building their musical structures, so as to make recovery of the grouping structure easy and natural. A good example of this kind of argument may be found in David Huron's work (Huron, 1989, 1991; Huron & Fantini, 1989), which demonstrates that J. S. Bach in his polyphonic compositions unwittingly implemented principles of auditory scene analysis (Bregman, 1990).

In another relevant study, Krumhansl and Jusczyk (1990) demonstrated that 4-6-month-old infants are sensitive to phrase boundaries in music: infants preferred listening to Mozart minuets that were interrupted by 1-s pauses at phrase boundaries, rather than to minuets that were interrupted in the middle of phrases. Since it seemed unlikely that infants that young would have had sufficient exposure to music to acquire performance expectations or even a representation of the tonal system (cf. Lynch, Eilers, Oller, & Urbano, 1990), Krumhansl and Jusczyk searched for and identified (in their materials) several acoustic correlates of phrase endings: drops in melody pitch, longer melody notes, and presence of octave intervals. If these factors are operative for infants, they should provide phrase boundary cues for adults as well.

Psycho-acoustic studies with simple sequences of tones have demonstrated that both infants and adults find it more difficult to detect an increment in the duration of a between-group interval than of a within-group interval (Fitzgibbons, Pollatsek, & Thomas, 1974; Thorpe & Trehub, 1989; Thorpe, Trehub, Morrongiello, & Bull, 1988). In these studies, two groups of identical tones were segregated by a change in frequency or timbre. Musical grouping structures are considerably more complex, of course, and a number of sometimes conflicting factors may be relevant at the same time (cf. Lerdahl & Jackendoff, 1983; Narmour, 1989).

Five such factors were considered in the following analysis. The second and third variables identified by Krumhansl and Jusczyk (1990) were not included because they seem to pertain only to major phrase boundaries, and more to the effects of interruption than of lengthening, which typically is observed before the end of a long tone (if shorter tones accompany it).

Possible psycho-acoustic factors

Pitch change. It is well known from many psychophysical studies that the difficulty of detecting or accurately judging a temporal interval between two tones increases with the pitch distance between the tones (see Bregman, 1990, for a review and many references; also, Hirsh, Monahan, Grant, & Singh, 1990). It could be that, in the present experiments, lengthening was more difficult to detect when the lengthened tone was followed by a large pitch skip. This simple hypothesis is actually difficult to evaluate because, in the musical excerpts used, there were often several simultaneous tones and pitch progressions in several voices. The additional assumption was made, therefore, that listeners paid attention to the principal melody tones, and to other tones only if they did not coincide with a melody tone. This is quite plausible because the melody tones were also louder than the others. This assumption reduces the Beethoven minuet of Experiment 1, for example, to the melody in the upper voice, interspersed with the single notes of the middle voice accompaniment whenever no new melody note occurs (cf. Figure 1). The melody and the middle-voice accompaniment each move in small pitch steps, usually less than 4 semitones, but at points where there is a switch from one voice to the other, larger intervals occur. Large descending pitch skips from the upper to the middle voice occur after the first beat in each bar, which always coincides with a long melody note. This is indeed where dips in the accuracy profile tended to occur in Experiment 1 (Figure 3). Conversely, however, the ascending skips from the middle voice back to the upper voice are associated with *peaks* in detection performance, as well as in the false-alarm profile (Figure 4). Since there is no obvious psycho-acoustic reason why ascending skips should *facilitate* detection performance, the absolute pitch distance hypothesis fails to explain the pattern in the data. Directional pitch change, however, may be a relevant variable. It was expressed in semitones and included in the regression analysis reported below.

Intensity. A second possible hypothesis is that the relative intensities of the tones influenced the detectability of lengthening. The musical excerpts in both experiments had an intensity microstructure derived from real performances. It could be that louder tones sound longer than softer tones, so that their lengthening is easier to detect. The opposite could also be true, however, if listeners compensate perceptually for a positive correlation between lengthening and intensity in performance. That is, listeners may expect a loud (accented) tone to be longer but, since it has the same duration as the others, it may sound relatively short. To test these hypotheses, the tone intensities (expressed as MIDI velocities) were included in the regression analysis. Whenever several tones coincided, the intensity of the loudest tone was taken.

Intensity change. Another reasonable hypothesis is that the change in intensity from one tone to the next may have played a role. A loud tone may mask the onset of a following soft tone, so that the loud tone sounds longer and the soft tone shorter. This hypothesis was evaluated by including the intensity difference between each tone and the next (taking the loudest tone in a chord) in the regression analysis.

Tone density. Some eighth-note intervals were marked by the onset of a single tone, others by several simultaneous tones. It might be hypothesized that clusters of tones sound inherently longer (or perhaps shorter?) than single tones, so that lengthening then is correspondingly easier (or more difficult) to detect in chords. This hypothesis, which is perhaps the least plausible, was tested by including the number of simultaneous tone onsets as a variable.

Change in tone density. The final variable to be considered was the change in tone density from one eighth-note interval to the next. The relevant hypothesis is similar to that for intensity change.

Regression analysis. Two stepwise linear multiple regression analyses were conducted on the data of each experiment, with either percent correct or false-alarm frequencies as the dependent variable, and the five independent variables listed above. In Experiment 1, the first two and the last two eighth-note intervals were omitted because their extremely low scores were probably due to other causes and would have dominated the correlations. In Experiment 2, only the eighth-note data were analyzed (from Test B only in the case of percent correct), and the three eighth-note intervals that contained grace notes were omitted. That way, the Experiment 1 analyses included 43 intervals, and the Experiment 2 analyses included 40. The intercorrelations among the variables are of greater interest than the regression equations themselves which, with one exception, included only a single independent variable. The intercorrelations are shown in Table 2.

Table 2. *Intercorrelations among the variables in Experiments 1 and 2.*

(a) Experiment 1 (*n* = 43)

	Percent correct	False alarms	PC	I	IC	D
Pitch change (PC)	0.38	0.40				
Intensity (I)	-0.07	0.08	-0.38			
Intensity change (IC)	0.51	0.53	0.85	-0.54		
Density (D)	-0.29	-0.07	-0.47	0.72	-0.53	
Density change (DC)	0.49	0.44	0.85	-0.38	0.79	-0.61

(If $r > 0.47$, $p < .001$; $r > 0.38$, $p < .01$; $r > 0.29$, $p < .05$)

(b) Experiment 2 (*n* = 40)

	Percent correct (Test B)	False alarms (A+ B)	PC	I	IC	D
Pitch change (PC)	-0.10	-0.23				
Intensity (I)	0.32	0.45	-0.39			
Intensity change (IC)	0.12	-0.21	0.68	-0.60		
Density(D)	-0.10	0.11	-0.24	0.21	-0.27	
Density change (DC)	0.4	0.21	0.15	0.16	0.03	-0.67

(If $r > 0.49$, $p < .001$; $r > 0.39$, $p < .01$; $r > 0.30$, $p < .05$)

In Experiment 1, three variables (pitch change, intensity change, and density change) correlated significantly with the perceptual results. However, they were also highly intercorrelated, so that only one of them contributed to the regression equation. That was the variable with the highest correlation, intensity change, although density change clearly was an equivalent predictor, at least for percent correct. In other words, in the Beethoven minuet, ascending pitch changes went together with increases in amplitude and with increases in tone density, all of which occurred at the onsets of melodic gestures and facilitated detection of lengthening preceding the change. Conversely, decreases in these three variables went along with poor detection scores. It thus appears that intensity and density decreases, and to a lesser extent pitch decreases, served as natural boundary markers. Of course, this was in large part due to the switching between melody and accompaniment. In the false-alarm analysis, the intensity variable made an additional contribution to the regression equation beyond the effect of intensity change; residual false-alarm rates were positively correlated with intensity. However, the regression equations explain only 26% of the variation in the accuracy profile and 48% of the variation in the false-alarm distribution.

The results of Experiment 2 were different. Although pitch change and intensity change were correlated with each other, neither of them correlated with either density change or the dependent variables. Density change, however, was a significant, though weak, predictor of detection performance, accounting for 17% of the variance. Intensity, too, showed a weak correlation with percent correct, but did not make a significant additional contribution to the regression equation. Intensity was a stronger correlate of the false-alarm percentages, but again the variance accounted for was small (20%). No additional variables made a significant contribution to the regression equation.

In summary, then, it appears that tone density change was most consistently related to detection performance: when a chord was followed by a thinner texture or a single (accompanying) tone, listeners tended to have difficulty perceiving lengthening of the intervening interval. In other words, these intervals sounded relatively short to the subjects.[22] This cannot have been due to masking of following tone onsets, which predicts the opposite. Thus it is not clear whether the effect of density change represents an independent psycho-acoustic effect, or whether it is simply a conventional marker of metric strength and/or grouping boundaries that elicited the performance expectations of musical listeners. Experiments employing this variable in nonmusical contexts or with musically inexperienced listeners remain to be done. Also, it must be kept in mind that most of the variation in the accuracy profiles remains unexplained in terms of local acoustic variables. A larger proportion of the variance, at least at a global level, is accounted for by the correlations with the expressive timing profile of expert performances. Whatever the factors are that influence perception, they seem to constrain production as well. These factors are perhaps better captured in an abstract description of the hierarchical metric and grouping structure, which is served by a variety of acoustic variables. A purely reductionistic explanation may not be the most parsimonious one when real music is the subject. Nevertheless, analytical research remains to be done to sort out the many interacting factors that may constrain perception of lengthening.

Parallels with language

The activities of producing and perceiving melodic gestures and phrases are analogous in many ways to those of producing and perceiving prosodic constituents in spoken language (see, for example, Hayes, 1989). The language situation most comparable to performing composed music would be reading text aloud or giving a memorized speech (rather than conversational speech, which is more analogous to improvisation in music). Experienced speakers, just like musicians, introduce prosodic variations that make the speech expressive and group words into phrasal units. The closest parallels may be found in timing. The phenomenon of *final lengthening* in constituents of varying sizes is well documented in speech production (e.g., Edwards, Beckman, & Fletcher, 1991; Klatt, 1975; Lehiste, 1973), and its parallelism with final lengthening in music has been pointed out by Lindblom (1978) and Carlson et al. (1989), among others. It has also been shown that pauses in speech, when they occur, tend to occur at phrase boundaries (e.g., Hawkins, 1971; Grosjean & Deschamps, 1975), and their duration tends to be proportional to the structural depth of the boundary (Gee & Grosjean, 1983; Grosjean, Grosjean, & Lane, 1979). Listeners, in turn, use perceived lengthening and pausing to determine the boundaries between prosodic units, as has been shown in experiments using syntactically ambiguous or semantically empty speech (e.g., Scott, 1982; Streeter, 1978). The same cues are likely to aid infants in discovering meaningful units in spoken language (Bernstein Ratner, 1986; Hirsh-Pasek et al., 1987; Kemler Nelson, Hirsh-Pasek, Jusczyk, & Wright-Cassidy, 1989); in fact, the infant study by Krumhansl and Jusczyk (1990), cited above, replicated for musical phrases what Hirsh-Pasek et al. (1987) had found for clauses in speech: infants prefer to hear pauses at structural boundaries signalled by prosodic variables.

Final lengthening (sometimes called pre-pausal lengthening) and pausing seem to serve the same function in speech, with pauses taking over wherever lengthening reaches limits of acceptability. In the present research, it was not necessary to distinguish between lengthening and pausing because the *legato* character of the music made silent pauses inappropriate. Because of the equivalent function of lengthening and pausing as structural markers, it seems quite appropriate to subsume them under a single temporal variable (onset-onset time). The contrast between lengthening and pausing in music pertains to the performance dimension of *articulation* (or connectedness, i.e., *legato* versus *nonlegato),* which did not vary in the pieces considered here. This dimension does not have a clear analogue in speech, where "articulation" refers to something quite different.

There does not seem to be a language study in the literature quite analogous to the present experiments, in which all points of potential lengthening (or pausing) were probed perceptually in a sizeable stretch of speech. However, there is little doubt that results parallel to the present findings would be obtained. Klatt and Cooper (1975) showed that phonetic segment lengthening is more difficult to detect in (absolute) phrase-final position, an effect also obtained in simpler sequences of syllables or nonspeech sounds (e.g., Benguerel & D'Arcy, 1986). Boomer and Dittman (1962) found that pauses are more difficult to detect at a terminal juncture (clause boundary) than within a clause, and Butcher (1980) reported that the pause detection threshold increases as a function of the depth of the structural boundary at which the pause occurred. Duez (1985), who investigated the perception of pauses in continuous speech, concluded that pauses that occurred at unexpected. places were *less* often detected than pauses in boundary locations. This finding may have been due to a response bias: Martin and Strange (1968) found that listeners tended to mislocate pauses towards structural boundaries. Still, Duez's findings are at variance with the present results, which showed no bias to mislocate lengthening at phrase boundaries- quite the opposite. The explanation may be that, in speech, lengthening may occur at structural boundaries even when no pause follows. Pauses tend to be heard following lengthened syllables even if no silence is present (Martin, 1970), which confirms the functional equivalence of lengthening and pausing.

It must be kept in mind that the speech in these pause detection tasks was naturally produced, whereas the music excerpts in the present study were mechanically regular. While it would be straightforward to conduct a pause detection experiment with music that contains expressive timing microstructure, it is difficult to envision an experiment in which lengthened segments or syllables would have to be detected in mechanically regular speech. Even the poorest synthetic speech includes many forms of temporal variation that are inherent to consonant and vowel articulation. Nevertheless, even with all this variation present, listeners are quite sensitive to changes in the durations of individual segments, particularly vowels (e.g., Huggins, 1972; Nooteboom, 1973; Nooteboom & Doodeman, 1980; however, see also Klatt & Cooper, 1975). In music, on the other hand, listeners are not particularly accurate in detecting local temporal changes in a temporally modulated performance (Clarke, 1989).

Nooteboom (1973, p. 25) concluded that "The internal representation of how words should sound appears to be governed by rather strict temporal patterns...." The norms of a particular language community seem to dictate a rather narrow range of possible temporal realizations for words at a given speaking rate. Nevertheless, these norms do permit modulations of timing and intensity (together with pitch) when words occur in context, not only to demarcate syntactic boundaries, but also to convey focus, emphasis, and emotional attitude. Listeners also seem to have expectations of these modulations (Eelting, 1991), though probably within relatively wide margins. It is these contextual modulations that have close parallels in the expressive microstructure of music, where phrase structure, emphasis, and emotions are conveyed by timing and intensity (as well as notated pitch) variations. Listeners also have expectations about these variations, and although it is difficult to gauge at present how precise these expectations are, they may be quite analogous to those for speech.

Specific expectations due to musical training and sophistication are probably superimposed on a substrate of very general principles applying to speech, music, and perhaps auditory events in general. These general principles include group-final lengthening as well as lengthening and raising of intensity to convey emphasis. Louder and longer events attract attention, so speakers and performers use this device to draw the listener's attention to certain points in the acoustic structure. Final lengthening may be a general reflection of relaxation and of the replenishing of energy (e.g., through breathing) to start a new cycle of activity. Thus the prosodic and microprosodic structures of speech and music may rest on the same general foundation, which is derived from our experience with real-world events and activities.

This common foundation may not only arise from a general desire to imbue music and speech with "living qualities" (Clynes & Nettheim, 1982), but it may be a direct consequence of the hierarchical event structure that music and speech have in common, and of the processing requirements that production and perception of such structures entail. This argument has been pursued by such authors as Cooper and Paccia-Cooper (1980), Gee and Grosjean (1983), and Wijk (1987) for speech production, and by Todd (1985) for music performance. The characteristic slowing down or pausing at structural boundaries enables speakers and performers to plan the next unit, and it in turn enables listeners to group what they hear into meaningful parts by closing off one unit and "laying the foundation" (Gernsbacher, 1990) for the next one. The presence of an appropriate performance (micro)structure may not be absolutely necessary for basic comprehension, but it certainly facilitates "structure building" (Gernsbacher, 1990) for the listener. What the present results seem to demonstrate is that, for music at least, listeners build their mental structures anyway, based on whatever clues are available in the music played. However, the absence of timing microstructure may be perceived as a lack of humanity and consideration on the part of the performer (a computer, in this instance). A poor music performance, like a poor public speech, may be offensive to a sensitive listener, whereas expert performers are loved by the audience.

Implications for musical aesthetics

The evaluation of musical performances is a topic that has barely been touched by psychologists, though it is of vital interest to music critics, performers, and teachers. Perceptual results of the kind reported here, in conjunction with analyses of musical microstructure in representative samples of performances (Repp, 1990b, 1992), begin to provide an objective foundation for judgments of performance quality, particularly in educational contexts. If it is indeed true, as the present findings suggest, that certain performance variations sound more "natural" than others at a perceptual level, then it is presumably part of the performer's task to provide these variations ("to discharge faithfully their aesthetic responsibilities," as Narmour, 1989, p. 318, expresses it). Once a good model of these natural variations is available (which is still a task for the future; but see Todd, 1985, 1989, and Gee & Grosjean, 1983, for speech), performances may be judged with respect to the degree to which they meet the model's predictions, and these .judgments may well be supported by the evaluations of expert listeners and critics.[23] Furthermore, such a normative model then could provide a basis for characterizing the nature of differences among performances that are perceived as equally adequate with respect to the model. Such differences may include the expression of different underlying structures for the same music, as well as different scalings of the magnitude of the microstructural variations. Musical performance (questions of technical accuracy aside) could thus be decomposed into two aspects: (1) expression of structure, which is open to objective evaluation; and (2) choice of scale, which is a matter of individual preference, within broad limits. Similarly, listeners' expectations could be partitioned into a shared structural component (leaving aside instances of structural ambiguity) and individual scale preferences, if any. This is surely an oversimplification of such complex activities as musical performance and judgment, but it is a modest beginning of trying to analyze these processes, which traditionally have been shrouded in mystery.

REFERENCES

Benguerel, A.-P., & D'Arcy, J. (1986). Time-warping and the perception of rhythm in speech. *Journal of Phonetics, 14,* 231-246.

Bernstein Ratner, N. (1986). Durational cues which mark clause boundaries in mother-child speech. *Journal of Phonetics, 14,* 303-309.

Boomer, D. S., & Dittmann. A. T. (1962). Hesitation pauses and juncture pauses in speech. *Language and Speech, 5,* 215-220.

Bregman, A. S. (1990). *Auditory scene analysis.* Cambridge, MA: MIT Press.

Butcher, A. (1980). Pause and syntactic structure. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler* (pp. 85 90). The Hague: Mouton.

Carlson, R., Friberg, A., Frydén, L., Granström, B., & Sundberg, J. (1989). Speech and music performance: Parallels and contrasts. *Contemporary Music Review, 4*, 391-404.

Clarke, E. F. (1985a). Some aspects of rhythm and expression in performances of Erik Satie's "Gnossienne No. 5." *Music Perception, 2*, 299-328.

Clarke, E. F. (1985b). Structure and expression in rhythmic performance. In P. Howell, I. Cross, & R. West (Eds.), *Musical structure and cognition* (pp. 209-236). London: Academic Press.

Clarke, E. F. (1988). Generative principles in music performance. In J. Sloboda (Ed.), *Generative processes in music* (pp. 1-26). Oxford: Clarendon Press.

Clarke, E. F . (1989). The perception of expressive timing in music. *Psychological Research, 51*, 2 9.

Clynes, M. (1983). Expressive microstructure in music, linked to living qualities. In J. Sundberg (Ed.), *Studies of music performance* (pp. 76-181). Stockholm: Royal Swedish Academy of Music.

Clynes, M., & Nettheim, N. (1982). The living quality of music: Neurobiologic patterns of communicating feeling. In M., Clynes (Ed.), *Music, mind, and brain* (pp. 47-82). New York: Plenum Press.

Cooper, W. E., & Paccia-Cooper, J. (1980). *Syntax and speech*. Cambridge, MA: Harvard University Press.

Duez, D. (1985). Perception of silent pauses in continuous speech. *Language and Speech, 28*, 377-389.

Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *Journal of the Acoustical Society, 89*, 369 382.

Eefting, W. (1991). The effect of "information value" and "accentuation" on the duration of Dutch words, syllables, and segments. *Journal of the Acoustical Society of America, 89*, 412-424.

Fitzgibbons, P. J., Pollatsek, A., & Thomas, I. B. (1974). Detection of temporal gaps within and between tonal groups. *Perception & Psychophysics, 16*, 522-528.

Gabrielsson, A. (1987). Once again: the theme from Mozart's Piano Sonata in A major (K. 331). A comparison of five performances. In A. Gabrielsson (Ed.), *Action and perception in rhythm and music* (pp. 81-103). Stockholm: Publications issued by the Royal Swedish Academy of Music, No. 55.

Gee, J. P., & Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology, 15*. 411 458.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, N J: Erlbaum.

Grosjean, F., & Collins, M. (1979). Breathing, pausing and reading. *Phonetica, 36*, 98-114.

Grosjean, F., & Deschamps. A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français: Vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica, 31, 144-184*.

Grosjean, F., Grosjean, L., & Lane, H. (1979). The patterns of silence: Performance structures in sentence production. *Cognitive Psychology, 11*, 58-81.

Hartmann, A. (1932). Untersuchungen über metrisches Verhalten in musikalischen Interpretationsvarianten. *Archiv für die gesamte Psychologie. 84*, 103-192.

Hawkins, P. R. (1971). The syntactic location of hesitation pauses. *Language and Speech, 14*, 277 288.

Hayes, B. (1989). The prosodic hierarchy in meter. In P. Kiparsky & G. Youmans (Eds.), *Phonetics and phonology, Vol. 1: Rhythm and meter* (pp. 201-260). New York: Academic Press.

Henderson. M. T. (1936). Rhythmic organization in artistic piano performance. In C. E. Seashore (Ed.), *Objective analysis of music performance* (pp. 281-305). Iowa City, IA: The University Press (University of Iowa Studies in the Psychology of Music, Vol. IV).

Hirsh, I. J., Monahan, C. B., Grant, K. W., & Singh, P. G. (1990). Studies in auditory timing: I. Simple patterns. *Perception & Psychophysics, 47*, 215-226.

Hirsh-Pasek, K., Kemler Nelson, D. G., Jusczyk, P. W., Wright-Cassidy, K., Druss. B., & Kennedy. L. (1987). Clauses are perceptual units for young infants. *Cognition. 26*, 269-286.

Huggins, A. W. F. (1972). Just noticeable differences for segment duration in natural speech. *Journal of the Acoustical Society of America, 51*, 1270-1278.

Huron, D. (1989). *Voice segregation in selected polyphonic keyboard works by Johann Sebastian Bach*. Unpublished doctoral dissertation, University of Nottingham, UK.

Huron, D. (1991). The avoidance of part-crossing in polyphonic music: Perceptual evidence and musical practice. *Music Perception, 9*, 93-104.

Huron, D., & Fantini, D. (1989). The avoidance of inner-voice entries: Perceptual evidence and musical practice. *Music Perception. 9*, 43-48.

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review, 96*, 459-491.

Kemler Nelson, D. G., Hirsh-Pasek, K., Jusczyk, P. W., & Wright-Cassidy, K. (1989). How the prosodic cues in motherese might assist language learning. *Journal of Child Language, 16*, 53-68.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics, 3*, 129-140.

Klatt, D. H., & Cooper, W. E. (1975). Perception of segment duration in sentence contexts. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 69 89). New York: Springer-Verlag.

Krumhansl, C. L., & Jusczyk, P. W. (1990). Infants' perception of phrase structure in music. *Psychological Science, 1*, 70-73.

Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception. *Journal of the Acoustical Society of America, 54*, 1228-1234.

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Lindblom, B. (1978). Final lengthening in speech and music. In E. Gårding, G. Bruce, & R. Bannert (Eds.), *Nordic prosody* (pp. 85-101). Lund University: Department of Linguistics.

Lynch, M. P., Eilers, R. E., Oller, D. K., & Urbano, R. C. (1990). Innateness, experience, and music perception. *Psychological Science, 1*.272 276.

Martin, J. G. (1970). On judging pauses in spontaneous speech. *Journal of Verbal Learning and Verbal Behavior, 9*, 75-78.

Martin, J. G., & Strange, W. (1968). The perception of hesitation in spontaneous speech. *Perception & Psychophysics, 3*. 427-438.

Narmour, E. (1989). On the relationship of analytical theory to performance and interpretation. In E. Narmour & R. Solie (Eds.), *Explorations in music, the arts, and ideas: Essays in honor of Leonard B. Meyer* (pp. 317-340). New York: Pendragon.

Nooteboom, S. G. (1973). The perceptual reality of some prosodic durations. *Journal of Phonetics, 1*. 25-45.

Nooteboom, S. G., & Doodeman, G. J. N. (1980). Production and perception of vowel length in spoken sentences. *Journal of the Acoustical Society of America, 67*. 276-287.

Palmer, C. (1989). Mapping musical thought to musical performance. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 331-346.

Povel, D.-J. (1977). Temporal structure of performed music: some preliminary observations. *Acta Psychologica, 41*, 309-320.

Repp, B. H. (1990a). Further perceptual evaluations of pulse microstructure in computer performances of classical piano music. *Music Perception, 8*, 1-33.

Repp, B. H. (1990b). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. *Journal of the Acoustical Society of America, 88*, 622-641.

Repp, B. H. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei." *Journal of the Acoustical Society of America, 92,* 2546-2568.

Scott, D. R. (1982). Duration as a cue to the perception of a phrase boundary. *Journal of' the Acoustical Society of America, 71, 996-1007.*

Shaffer, L. H. (1981). Performances of Chopin, Bach, and Bartók: Studies in motor programming. *Cognitive Psychology, 13,* 326-376.

Sloboda, J. A. (1985). Expressive skill in two pianists: Metrical communication in real and simulated performances. *Canadian Journal of Psychology, 39,* 273-293.

Streeter, L. A. (1978). Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America, 64.* 1582 1592.

Sundberg, J. (1988). Computer synthesis of music performance. In J. A. Sloboda (Ed.), *Generative processes in music* (pp. 52 69). Oxford: Clarendon Press.

Thorpe, L. A., & Trehub, S. E. (1989). Duration illusion and auditory grouping in infancy. *Developmental Psychology, 25,* 122-127.

Thorpe, L. A., Trehub, S. E., Morrongiello, B. A., & Bull, D. (1988). Perceptual grouping by infants and preschool children. *Developmental Psychology, 24.* 484-491.

Todd, N. P. (1985). A model of expressive timing in tonal music. *Music Perception, 3,* 33-58.

Todd, N. P. (1989). A computational model of rubato. *Contemporary Music Review, 3,* 69-88.

Wijk, C. van (1987). The PSY behind PHI: A psycholinguistic model for performance structures. *Journal of Psycholinguistic Research, 16,* 185-199.

FOOTNOTES

[*] *Cognition, 44,* 241-281 (1992).

[1] I am not concerned here with the variation in tone duration as such (i.e., in onset-offset intervals), which signal differences in articulation (i.e., *legato* vs. *staccato*), an important dimension in its own right. "Timing microstructure" here refers exclusively to variation in tone onset-onset intervals, regardless of whether these intervals are filled or partially empty.

[2] This assertion is based on the observation that the frequency distribution of the actual durations of nominally equal tone inter-onset intervals in performed music is strongly skewed towards long values (Repp, 1992). Research remains to be done to show what point in that distribution corresponds to the perceived underlying beat, but that point is likely to be near the peak of the distribution.

[3] Although performances usually differ along many dimensions, Repp (1990a) has provided some evidence that listeners can express consistent preferences among performances distinguished by timing microstructure alone.

[4] The slurs in the score (which presumably are Beethoven's own) do not indicate the melodic grouping structure; rather, they mark *legato* connections within bars and detached articulations across bar lines. This "articulation structure," which is out of phase with the grouping structure, is ignored here. That the slurs do not represent the grouping structure can be easily proven by the armchair "pause test": it would be much more natural to stop a performance of the music at the ends of melodic gestures than at the ends of slurs.

[5] This performance-oriented representation of the grouping structure differs from the more score-oriented formalism employed by Lerdahl and Jackendoff (1983) in that it does not exhaustively apportion the music to units at the lowest level. For example, the first melodic gesture is assumed to end with the first melody note in bar 1, or more precisely with the onset of the first following note: the following three notes are part of a middle-voice background and hence do not belong to the melodic gesture. In fact, they belong to a different (secondary) grouping structure, that of the middle voice. Lerdahl and Jackendoff (1983) did not deal with the case of several simultaneous grouping structures.

[6] The earlier measurements unfortunately were not at the level of detail required for the present purposes.

[7] Sixteenth-note tones were skipped. Asynchronies among the onsets of nominally simultaneous tones could not be resolved; they were generally small (cf. Palmer, 1989). Chords were thus treated as if they were single tones. A terminological effort is made throughout this paper to distinguish *notes* (printed symbols) from *tones* (single-pitched instrument sounds) and (onset-onset) *intervals* (which include all tones that sound simultaneously). Mixed terms such as "eighth-note interval" and ambiguous terms such as "chord" cannot be avoided, however.

[8] The reason for not varying the intensity of the accompanying tones was that the author's technical skills were not deemed reliable at that very subtle level. The possible influence of the "intensity microstructure" on the perceptual responses will be investigated in the General discussion. Suffice it to note here that the metrically accented downbeats were not usually more intense than neighboring melody tones.

[9] The author routinely served as a pilot subject, and since his data were quite typical, they were included to replace the data of one subject who performed at chance level in all conditions.

[10] The systematicity of the data proved this "take-home" method to be quite successful for this kind of experiment, in which precise control over volume and sound quality was not essential. Subjects listened to the music in familiar surroundings at a time of their own choosing and thus largely avoided the tense atmosphere that hovers over laboratory experiments.

[11] The correlation does not include intervals 0-5, 0-6, 8-2, and 8-3.

[12] For clarity, (-1) responses, which were less systematic, are not included in the figure, nor are (-2) responses, which occurred only by chance.

[13] The analogous operation for (-1) responses did not yield an increase in the correlation, though it reached significance (0.40, $p < .01$), and (-2) responses did not yield any significant correlation at all (0.14).

[14] Most editions of the music contain tempo suggestions in the form of metronome values that go back to the composer or his wife Clara, but which are generally considered too fast for contemporary tastes. Slurs were omitted in this experiment in order to avoid visually biasing subjects' auditory perception of the grouping structure. In the Beethoven minuet of Experiment 1, slurs generally ended at bar lines and thus cut across the melodic gestures. The slurs in the Schumann piece (at least in the several editions that were consulted) coincide with the grouping structure, except for a single long slur across bars 6-8. It seemed prudent to eliminate this visual source of grouping information from the answer sheets.

[15] The final chord appears in this form only in the present excerpt, to provide an appropriate conclusion. In the original music, the eighth-note movement continues in the bass voice.

[16] These intervals were measured between the onsets of the two F's in the soprano voice; the other tones in the chord in positions 2-3 and 6-3 were sometimes strongly asynchronous (see Repp, 1992). Similarly, the interval 4-8 included a grace note, in addition to being subject to terminal lengthening.

[17] These were. for pedal onsets, the first tones in bars 1 and 5; for pedal offsets, the first tones in bars 3 and 7; and for both offsets and onsets, the first grace notes in bars 2 and 7.

[18] Due to an oversight, the longer intervals had been lengthened by the same small amount as eighth-note intervals, which was

virtually impossible to detect. These stimuli thus served essentially as catch trials.

[19] Each of the long notes was a single data point in any correlation computed. Their representation as multiple data points in the figures is for graphic purposes only.

[20] All subjects in this study had heard the excerpt several times in a lecture given by the author several weeks to several months preceding the experiment. The presentation included an expert performance—the recording by Alfred Brendel.

[21] Even though musical structure is commonly discussed with reference to the printed score, it must be based on the analyst's sonic image to have any value; purely visual structure in music notation is irrelevant, though it often parallels the auditory/cognitive structure.

[22] This statement is based on the fact that false alarms were rare on these chords. The argument throughout this paper has been that intervals that are expected to be long sound short to listeners under conditions of isochrony. This is in contrast to the "duration illusion" discussed by Thorpe and Trehub (1989) and by earlier authors cited there, according to which a between-group interval sounds longer than a within-group interval. One possible reason for this discrepancy is that the duration illusion concerns silent intervals. If group-final sounds are expected to be lengthened and therefore are perceived as relatively short, a following silence may be perceived as relatively long.

[23] I am not trying to suggest here that the most "natural" performances are necessarily the most preferred. They should be perceived as pleasing and, at their best, as beautiful. Truly noteworthy performances will often be deviant in some respect, however, just as fascinating faces, landscapes, or personalities have nonideal properties.

Haskins Laboratories Status Report on Speech Research
1992, SR-111/112, 321-328

A Review of Yoh'ichi Tohkura, Eric Vatikiotis-Bateson, and Yoshinori Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure**

Bruno H. Repp

This attractive volume is the result of a workshop held at ATR (Advanced Telecommunications Research) Laboratories, Kyoto, in November 1990, immediately preceding the International Conference on Spoken Language Processing in Kobe. The contributions are arranged into two major sections (Speech Perception; Speech Production and Linguistic Structure), each of which has three subsections. Each of these contains a number of articles followed (with one exception) by several shorter commentaries; the exact distribution is: 5+2, 2+2, 5+4; 2, 3+3, 5+2. Thus there are 22 articles and 13 commentaries in all. Nearly all authors are well-established researchers, roughly one third of them Japanese, the majority from the United States, plus a few European representatives.

The first subsection deals with the rather specific topic of "Contextual Effects in Vowel Perception," with papers by Sumi Shigeno, Robert Allen Fox, Caroline B. Huang, and Masato Akagi, followed by comments from Dominic W. Massaro and Sieb G. Nooteboom. Context effects are not only interesting from a general psychophysical perspective but constitute one of the central problems faced by automatic speech recognition. The contributions by Huang and Akagi represent this latter perspective, while Shigeno and Fox are primarily concerned with modelling human perception. **Shigeno**'s study (a slightly condensed version of Shigeno, 1991) represents a continuation of her earlier careful work on contextual effects in the perception of isolated vowels as a function of category membership, spectral distance, and temporal proximity. The most salient result is that successive vowels show contrastive effects as long as they belong to different categories, but assimilative effects when they belong to the same category. In agreement with dual-process models of categorical perception, Shigeno suggests that assimilation occurs in auditory memory, whereas contrast arises from categorical representations. **Fox** reports four experiments aimed at demonstrating that phonotactic regularities of the language constrain vowel perception in syllabic contexts. He compares the identification of vowels in open syllables and in the context of a following /r/, which neutralizes the tense-lax contrast in English. The first experiment, using a synthetic /ɪ–ɛ–æ/ continuum in the two contexts, shows a weakening of the perceptual /ɛ–æ/ distinction preceding /r/. The second experiment, a multidimensional scaling analysis of similarity judgments on a larger set of naturally produced syllables, supposedly also shows a reduction of the perceptual distance between tense and lax vowels in the /r/ context, although I have difficulty seeing this; the /r/ context just seems to rotate the

I am grateful for the support of Haskins Laboratories during the preparation of this review. Thanks are also due to Yoh'ichi Tohkura for helpful comments on an earlier draft.

perceptual space, but not to affect the distances. Experiment 3, however, confirms a deleterious effect of following /r/ on the identification of natural vowels, and Experiment 4 shows that a following /l/ (which permits tense-lax contrasts) does not have a similar effect. Fox argues that the observed phonotactic effect is perceptual rather than a response bias, though the basis for that claim remains to be elucidated. For one thing, perceptual effects can take the form of a bias (e.g., Repp, Frost, & Zsiga, 1992).

Huang is concerned with the information conveyed by the dynamic formant trajectories within (monophthongal) vowels in CVC contexts. She first shows that an automatic classification algorithm gains little if the trajectories are used to infer single formant target values; however, a substantial improvement results when three points from each vowel's trajectories are presented to the algorithm. Further improvements result when information about duration and consonantal context is added. These data are then compared to those of human listeners who identified the excised vowel nuclei or the full CVC syllables. Though it is not quite clear how degree of agreement between machine and human performance was determined, it seems to be highest for the three-point condition, which confirms the perceptual importance of formant trajectories. Of course, this is hardly a new insight (see, e.g., Nearey & Assmann, 1986), but it is good to see it put to use in the context of automatic speech recognition. **Akagi**, in the subsequent paper, presents a detailed and elegant study of contextual effects among successive vowel-like stimuli, based on a dual-process model of interaction, both at the level of individual spectral peaks and at the level of phoneme boundaries. The model seems similar to Shigeno's, though Akagi makes no reference to it or to any other related work. The results, obtained with vowels preceded by either single-formant stimuli or full vowels at various temporal intervals, provide a detailed map of assimilation and contrast effects. The results for full vowel contexts resemble those of Shigeno, showing assimilation when spectral distance is small and contrast when it is large, with little effect of temporal separation. For single-formant stimuli, however, temporal factors seem to be most important, with assimilation at separations of less than 70 ms, and contrast beyond. Akagi's study is very clever and the data are informative, but it must be said that they were obtained by presenting two female subjects with more than 100,000 (!) stimuli, resulting from a large factorial design, each stimulus to be identified as either /u/ or /a/. The necessity and indeed the humanity of such an excessive design must be seriously questioned.

The two commentaries on the preceding four articles are brief. **Massaro** predictably takes this opportunity to recite the canons of his "fuzzy logical model" of speech perception. Moreover, he admonishes the authors to "keep the big picture in mind in their day-to-day struggle with the wonders of speech perception" and to design their experiments according to the tenets of his own "paradigm for speech perception research," even though this factorial paradigm has rarely yielded data of the richness of any of the four studies commented on. **Nooteboom**'s ideas are more constructive and to the point. He argues that categorical perception explains the assimilation/contrast findings of Shigeno and Akagi, and that temporal proximity is likely to enhance these effects by pulling speech stimuli into a single stream of sounds. Within-stream contrast serves to enhance phoneme boundaries.

The second subsection is devoted to "Perceptual Normalization of Talker Differences" and offers articles by Tatsuya Hirahara and Hiroaki Kato, Howard Nusbaum and Todd Morin, and Kazuhiko Kakehi, followed by a commentary by David Pisoni. The topic, like that of the preceding subsection, is one of signal importance to both human and machine speech recognition, and the work reported is very interesting. **Hirahara and Kato** provide another example of the meticulous and wide-ranging parametric studies that Japanese researchers seem to excel in. By presenting an array of 200 isolated vowels varying in formant frequencies and, independently, in fundamental frequency (F0) for identification to 24 listeners, they mapped out the perceptual space and demonstrated shifts in category boundaries in F1-F2 space caused by F0. These shifts were much reduced, however, when F0 (in Bark) was subtracted from each coordinate of the vowel space, which provided an effective normalizing procedure. In other words, vowel quality remained constant as long as F0 was shifted along with the other formants on a Bark scale. The authors argue that low harmonics of a relatively high F0 may act as a "perceptual formant" because they are not subject to auditory integration; this would account for changes in perceived vowel quality when F0 alone is changed. An intriguing difference in the Japanese /a/-/o/ boundary between male and female listeners is also reported.

Nusbaum and Morin, in the subsequent paper, describe five experiments using a common paradigm: the comparison between single (blocked) and multiple (mixed) talker conditions. For several types of stimuli (isolated vowels, consonants and vowels in CV syllables, monosyllabic words) they find slightly reduced identification accuracy and particularly longer reaction times in the multiple-talker condition, which they take as evidence of a normalization process. In their most interesting manipulation, they added a secondary task (either one or three numbers to hold in memory) and showed that memory load further slows reaction times in the mixed, but not in the blocked condition. This indicates that the normalization process requires cognitive resources that the memory task competes for. Nusbaum and Morin further show that eliminating F0 by using whispered stimuli reduces accuracy in the mixed condition only, and that mixing similar talkers (of the same sex) has little detrimental effect (though it is not clear whether the talkers could actually be discriminated). The authors distinguish two kinds of normalization processes: "contextual tuning" and "structural estimation"; the first seems to operate in the blocked, and the second in the mixed condition. This is an interesting set of experiments, although the tasks are somewhat artificial, some methodological information is missing, and at least one important earlier study (Summerfield & Haggard, 1975) is not mentioned.

Kakehi's paper is rather brief and evidently a summary of a study published previously in Japan. Using a large set of Japanese syllables, he examined the time it takes to adapt to a single speaker (about four trials), as well as the time to lose that adaptation (about 7 trials, though there is a significant irregularity in the data that the author glosses over). There were also talker-specific variations. **Pisoni** says little about the preceding papers but rather reports some relevant research from his own laboratory, showing effects of talker variation on reaction time and memory performance. His important conclusion is that talker-specific information is not "stripped off" but is retained in memory along with the phonetic structure of an utterance. This observation raises the question (in my mind, at least) whether the normalization process referred to by Nusbaum and Morin and many others is really a process at all, in the sense that it changes mental representations of speech, or whether it is simply a reflection of cognitive uncertainty caused by stimulus variability.

The third and most substantial subsection deals with the "Perception and Learning of Non-Native Language." It includes articles by Reiko A. Yamada and Yoh'ichi Tohkura, by Scott E. Lively, David B. Pisoni, and John S. Logan, by Winifred Strange, by Jacques Mehler and Anne Christophe, and by Patricia K. Kuhl, followed by commentaries by Howard C. Nusbaum and Lisa Lee, by Anne Cutler, by Shigeru Kiritani, Fumi Katoh, Akiko Hayashi, and Toshisada Deguchi, and by Morio Kohno. The first two papers, and the third in part, deal with the narrow problem of Japanese speakers' difficulties in distinguishing the American English phonemes /r/ and /l/, which has attracted an unusual amount of research effort, much of it using synthetic /r/-/l/ continua of the kind developed at Haskins Laboratories. **Yamada and Tohkura**, in their presentation (which overlaps substantially with Yamada and Tohkura, 1992), report data showing that Japanese listeners perceive /w/ in the boundary region of /r/-/l/ continua, that their identification accuracy for synthetic and natural stimuli is correlated, that they are sensitive to stimulus range, and that (unlike native speakers of English) they seem to use second-formant transition information to distinguish English /r/ and /l/. The most interesting and extensive part of their study compared 120 Japanese subjects who had never lived abroad with 122 Japanese who had lived in the U.S. for some time. Not only did the second group outperform the first in /r/-/l/ identification accuracy, but there was a clear relationship between the age at which subjects had moved to the U.S. and their accuracy: Virtually all the high performers had moved before the age of 11. Yamada and Tohkura's work, by the way, shows something that American studies tend to downplay: There are quite a few speakers of Japanese who can discriminate English /r/ and /l/ perfectly well.

Lively, Pisoni, and Logan summarize various attempts to train Japanese speakers in the laboratory to improve their discrimination of /r/ and /l/. After criticizing earlier approaches that used synthetic speech and discrimination tasks, they report the results of their own studies using natural speech, a variety of utterances, and multiple speakers in a categorization task (partially reported also in Logan, Lively, & Pisoni, 1991). Over 15 sessions, the subjects' performance improved significantly but not impressively (between 6 and 9%). In view of considerable variation in accuracy for different speakers' tokens in the training phase, the results of post-training generalization tests, in which utterances from a single novel speaker were presented., seem uninterpretable. The test should have included multiple novel talkers, or at least the single novel talker should have been rotated with the talkers used in training in a counterbalanced design. Still, the training methods of Lively et al. seem intuitively reasonable, and their data provide useful information about contextual variation and speaker differences in /r/-/l/ discriminability. They also relate their approach to work on exemplar-based storage models of memory (e.g., Hintzman, 1986). **Strange**, in the following article, reviews more broadly the methodological variables relevant to training studies and points to a crucial factor (surprisingly neglected in the two preceding papers), viz., the relation of the non-native categories to be discriminated to the phonetic categories of the native language. After reporting some (previously unpublished) evidence that the nature of the training task makes little difference in Japanese listeners' discrimination of the /r/-/l/ distinction (though all training was done with synthetic speech), she surveys several experiments that involved not only /r/ and /l/, but a number of other phonetic distinctions, including some that are difficult to discriminate by native speakers of English (a welcome relief from the heavy focus on what "others" cannot do). These data provide some interesting glimpses of effects of phonetic context and of individual differences among subjects with regard to trainability--a dimension almost totally ignored in this type of research, but highly relevant to the natural task of second-language acquisition. It is fair to conclude, however, that there is so far remarkably little success in training subjects in the laboratory to discriminate non-native categories, a skill that seems to be very difficult for most adult subjects to acquire. It would be interesting to conduct such training studies with children who should be more malleable in that regard, though not necessarily more responsive to boring laboratory training procedures. Perhaps, if the training procedures were placed in a motivating social context, they would meet with greater success.

The remaining two articles in this section focus not on the perception of non-native categories (except tangentially) but on the acquisition of the native phonology and phonetics; therefore, the subsection should really have been given the heading "Acquisition and Perception of Native and Non-Native Language." **Mehler and Christophe** focus on the question of the units of speech perception. They first review recent research by Mehler (in collaboration with Cutler, Norris, and Segui) who used a syllable monitoring task to demonstrate that speakers of French employ a syllabic segmentation strategy, whereas speakers of English do not. A study in which the two languages were interchanged and another study with true bilinguals suggested that speakers have only one processing strategy that they apply to native (or dominant) and non-native (or non-dominant) languages alike, though apparently a strategy can be "inhibited" as well. Recent work is cited which extends the paradigm to Spanish and Catalan, with intermediate results: These speakers seem to have a syllabification strategy at their disposal, but seem to employ it only in certain contexts or under certain conditions. The authors conclude that "speech processing procedures depend on the maternal language of the speaker," and I find it surprising that they had seriously considered an alternative hypothesis. In the second part of their paper, Mehler and Christophe discuss the important problem of speech unit acquisition in young infants, on which Mehler and his group have conducted pioneering research with neonates. They cite several studies in progress, which suggest that "infants do not perceive speech as a string of phonemes, but in terms of some higher-order unit(s)." One recent study seems to suggest that infants are sensitive to subtle word boundary cues in fluent speech. Interesting and important as this research is, I cannot avoid the feeling that Mehler and his colleagues are conceptually enslaved by the idea of (phonological) units. A more fluid model of perceptual differentiation might be appropriate, in which the units are not preordained by phonological theories but emerge autonomously from patterns of repetition and statistical frequency in the input. Such a view, incidentally, also entails that "processing" strategies will be language-specific, to the degree that languages are different from each other.

Kuhl, in the final article in this subsection, summarizes in somewhat schoolmasterly fashion her research on vowel category "prototypes" in human adults, 6-months old infants, and macaque monkeys (see also Kuhl, 1991). Both variants of the human species are shown to have a prototype or best exemplar for the /i/ category, whereas monkeys do not. The evidence comes from a discrimination task which shows that, if a prototype exists, discrimination is more difficult in its vicinity than at some distance from it in the acoustic vowel space. This is interesting research, but it is still rather limited in its restriction to a single category of isolated vowels. More recent cross-linguistic work on English and Swedish infants is, unfortunately, only mentioned very briefly; unlike Mehler and Christophe, Kuhl seems unwilling to share preliminary results of work in progress. The most interesting part of the results is that for the infants who, by the age of 6 months, seem to have acquired a notion of what a good /i/ sounds like. Kuhl seriously considers the possibility that some prototypes might be innate (a hypothesis that seems absurd to me) but ultimateiy favors an explanation in terms of exposure to speech. Monkeys, of course, are at a serious disadvantage because they have not been exposed to human speech in the same way, nor do they presumably care much about human vowels. They may well have a prototype for some conspecific vocalization that has communicative significance for them.

Of the four commentaries in this subsection, only two turn out to be relevant to the topic. **Nusbaum and Lee** go on for too long about too little, ending up with the suggestion that perceptual learning can be understood as a reshaping of the distribution of attention over the speech signal, which seems little more than a restatement of the problem. **Cutler's** very brief commentary reiterates the unsurprising conclusion (cf. Mehler and Christophe) that native phonology constrains native as well as non-native language processing, and then mentions a recent paradigm-bound finding of longer phoneme monitoring times for vowels than for consonants. The remaining two "commentaries" are actually short reports of research that seem to have found shelter in the wrong place. The paper by **Kiritani et al.** deals with perceptual normalization of vowels in children and infants and clearly belongs in the preceding subsection. The paper by **Kohno**, while quite intriguing, does not really fit anywhere in this volume. Following up an important but rarely cited study by Hibi (1983), he presents further evidence of a qualitative change in the processing of rhythmic sequences at a rate of about 3 per second, which presumably reflects a shift from a non-integrative to an integrative processing of successive auditory events.

Part II of the book begins with two articles by John J. Ohala and Hiroya Fujisaki, respectively, that are assigned to a separate subsection entitled "Introduction," though neither serves a truly introductory function with respect to the following papers. **Ohala**, in characteristically enlightening fashion, discusses the difficulty of distinguishing phonetic variation due to active causes (style of speech, coarticulation, etc.) from similar variation that, though it may originally have been caused by the same factors, has become phonologized and part of the language norm. He summarizes the methodology of historical linguistics and points out its limitations, which include the inability to establish the causal basis of sound change. He goes on to cite some empirical research that begins to address this question. To cite just one example, Solé (1992)--originally in collaboration with Ohala--has presented rather strong evidence that anticipatory nasalization of vowels is part of the language norm (i.e., a sound change) in English, but not in Spanish and several other languages, where it is merely a passive coarticulation effect. **Fujisaki** presents the latest version of his model for generating Japanese F_0 contours, which has been influential for over two decades, even though it was originally described (like almost all of Fujisaki's brilliant work) only in technical reports and conference proceedings. The model has two basic components: phrase commands that generate impulse-like changes in F_0 and accent commands that generate stepwise changes. A model of the underlying physiological mechanism is also briefly discussed, which postulates the strain of the vocal cord as the principal variable being controlled.

The following subsection, entitled "Articulatory Studies," contains articles by Kevin G. Munhall, J. Randall Flanagan, and David J. Ostry, by Eric Vatikiotis-Bateson and Janet Fletcher, and by Mary E. Beckman and Jan Edwards, followed by three brief commentaries, respectively by Osamu Fujimura, René Collier, and Kiyoshi Honda. To anticipate, two of these commentaries are in fact on Fujisaki's presentation: **Collier** argues in favor of modelling the perceptually significant aspects of intonation, rather than the raw F_0 contour of the acoustic signal. However, to the extent that a model actually succeeds in closely approximating the F_0 contour (as Fujisaki's model seems to do), Collier's proposal seems superfluous. **Honda** adds some interesting physiological observations on the active control of F_0 lowering. **Fujimura**'s comments, unfortunately, provide only an obscure promise of some global model to come. Also, his contribution evidently was not edited or updated since the 1990 workshop, as his introductory paragraph does not jibe with the contents of the present volume.

The three articulatory studies largely bypassed by the commentators are actually quite interesting, though they suffer from a problem endemic to speech production research: insufficient numbers of subjects and large inter-subject variability. **Munhall et al.** largely escape that criticism by restricting themselves to mere examples of data in the context of theoretical observations. They sketch their approach to articulatory modelling, which is based on the notion of a kinematic space whose coordinates are joint angles or displacements, rather than effector movements in Cartesian space. Preliminary observations suggest simple linear relationships among the coordinate variables, which represent built-in constraints on articulatory trajectory formation. This elegant and promising approach contrasts with the somewhat undisciplined presentation by **Vatikiotis-Bateson and Fletcher**, who review a profusion of extremely complex and variable data that are difficult to make head or tail of. Their claim is very interesting and important: that local changes in a phrase may affect articulatory patterns throughout the whole utterance. However, it is not clear whether they have solid evidence to support their claim, or whether they are dealing with variability pure and simple. **Beckman and Edwards** tell a more coherent story, although they seem guilty of rather casual and selective reporting of data, possibly focusing on those most favorable to their conclusions. Their theoretical framework is the task-dynamic model developed at Haskins Laboratories, and they report results (in part already presented by Edwards, Beckman, & Fletcher, 1991) that illustrate how the dynamic control variables of stiffness, amplitude, and phasing are differentially employed to produce different types of stress and lengthening phenomena in (painfully stilted) laboratory-style utterances. It is to be hoped that the extremely promising theoretical ideas of these authors and those of the two preceding papers (each group including a former associate of Haskins Laboratories) will soon be supported by sufficiently extensive data.

The last subsection, "Acoustic Studies," includes papers by Nobuyoshi Kaiki and Yoshinori Sagisaka, by Nick Campbell, by Anne Cutler, by Jacques Terken and René Collier, and by Sieb G. Nooteboom and Wieke Eefting, with commentaries by Yoshinori Sagisaka and by Mary E. Beckman. **Kaiki and Sagisaka** report a statistical analysis of segment durations in a sizeable corpus of Japanese speech, produced by a single speaker (a professional announcer). Somewhat confusingly, the technique is described as "categorical factor analysis," though in fact it seems to be a version of linear regression analysis. Many independent variables were considered, including position in utterance groups of various sizes, length of utterance group, part of speech, etc. Still, the full regression model derived from one half of the data did not predict the segment durations in the other half as accurately as one would like. One of the more surprising detailed results is a substantial shortening of vowels in sentence-final position. **Campbell's** subsequent discussion, based on the same data base, reveals this effect as being due entirely to the prevalence of the Japanese past-tense particle /-ta/, which always occurs in sentence-final position. Apart from arguing convincingly for the necessity of detailed linguistic analysis of speech corpora, Campbell mainly demonstrates the utility of z-scores as a way to eliminate differences in intrinsic segmental duration from cross-segment comparisons. **Cutler**, in the following paper, briefly summarizes some of her research on word boundary cues, referring the reader to papers published elsewhere for details. Her work suggests that speakers of English expect words to start with strong (unreduced) syllables, in agreement with the statistical predominance of such words in the language. When induced to speak very clearly, speakers may introduce subtle durational cues to word boundaries, mainly pre-boundary lengthening or pausing. One limitation of this research (which it shares with Mehler's) is its adherence to binary theoretical concepts such as strong/weak, where in fact syllables probably vary in "strength" depending on segmental composition and context. Also, her (admittedly abbreviated) discussion does not convey the richness of possible lexical entries (ranging from individual letter names to words, compounds, familiar phrases, and memorized texts) and the somewhat arbitrary distinctions imposed by the orthography between what counts as two words and what counts as one. The "word boundary problem" may in fact be that of prosodic grouping in general.

Terken and Collier examine the influence of syntactic structure on prosody in a Dutch corpus obtained (like the Japanese corpus mentioned above) from a single professional speaker. They find that a major syntactic boundary (NP-VP) is marked by pitch inflection, pausing, and lengthening, whereas lesser syntactic boundaries are marked by pitch inflections only, if they are marked at all. The material is limited in structural variety, however, and the authors appropriately characterize their work as a pilot study within a general effort to improve the naturalness of a text-to-speech system. In the final article, **Nooteboom and Eefting** examine how a speaker adjusts prosody to meet a listener's needs. They report a study that demonstrates that the lengthening of "new" relative to "given" lexical items is a concomitant of the presence versus absence of pitch accent (in Dutch). They also show (in a study reported in more detail in Eefting, 1992) that the judged naturalness of speech suffers when the durational characteristics are not in agreement with accentuation. **Sagisaka**, in the first of the two commentaries, surprisingly does not respond to Campbell's critical examination of his own results but instead comments on long range control of timing and on differences in prosodic organization between Japanese and English. **Beckman** concludes with some perceptive comments on rhythm in different languages, but surprisingly refers to the "three papers" in this section and to a paper by Fant et al., which is not contained in this volume. Apparently, this commentary was not updated to reflect the final contents of the book.

In summary, despite some peculiarities of organization, this is a generally well-edited and consistently interesting collection of articles, some of which present original findings not available elsewhere. The "commentaries" are less successful, on the whole, and could have been omitted without much damage. The book would be a worthwhile addition to the library of anyone interested in the contemporary speech scene, but many potential buyers will find the price tag prohibitive.

REFERENCES

Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America*, 89, 369-382.

Eefting, W. (1992). The effect of accentuation and word duration on the naturalness of speech. *Journal of the Acoustical Society of America, 91*, 411-420.

Hibi, S. (1983). Rhythm perception in repetitive sound sequence. *Journal of the Acoustical Society of Japan, 4*, 83-95

Hintzman, D. L. (1986). "Schema abstraction" in a multiple trace memory model. *Psychological Review, 93*, 411-428.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics. 50*, 93-107.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify /r/ and /l/. *Journal of the Acoustical Society of America, 89*, 874-886.

Nearey, T., & Assmann, P. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America, 80*, 1297-1308.

Repp, B. H., Frost, R., & Zsiga, E. (1992). Lexical mediation between sight and sound in speechreading. *Quarterly Journal of Expenmental Psychology, 45A,* 1-20.

Shigeno, S. (1991). Assimilation and contrast in the phonetic perception of vowels. *Journal of the Acoustical Society of Amenca, 90,* 103-111.

Solé, M.-J. (1992). Phonetic and phonological processes: The case of nasalization. Language *and Speech, 35,* 29-43.

Summerfield, A. Q., & Haggard, M. P. (1975). Vocal tract normalization as demonstrated by reaction times. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 115-142). London: Academic Press.

Yamada, R. A., & Tohkura, Y. (1992). The effects of expenmental variables on the perception of American English /r,l/ by Japanese listeners. *Perception & Psychophysics, 52,* 376-392.

FOOTNOTE

*Tokyo, Ohmsha, and Amsterdam: IOS Press, 1992. 463 pp. $115. This review will appear in *Language and Speech, 36(1)* (1993).

Appendix

SR #	Report Date	NTIS #	ERIC #
SR-21/22	January-June 1970	AD 719382	ED 044-679
SR-23	July-September 1970	AD 723586	ED 052-654
SR-24	October-December 1970	AD 727616	ED 052-653
SR-25/26	January-June 1971	AD 730013	ED 056-560
SR-27	July-September 1971	AD 749339	ED 071-533
SR-28	October-December 1971	AD 742140	ED 061-837
SR-29/30	January-June 1972	AD 750001	ED 071-484
SR-31/32	July-December 1972	AD 757954	ED 077-285
SR-33	January-March 1973	AD 762373	ED 081-263
SR-34	April-June 1973	AD 766178	ED 081-295
SR-35/36	July-December 1973	AD 774799	ED 094-444
SR-37/38	January-June 1974	AD 783548	ED 094-445
SR-39/40	July-December 1974	AD A007342	ED 102-633
SR-41	January-March 1975	AD A013325	ED 109-722
SR-42/43	April-September 1975	AD A018369	ED 117-770
SR-44	October-December 1975	AD A023059	ED 119-273
SR-45/46	January-June 1976	AD A026196	ED 123-678
SR-47	July-September 1976	AD A031789	ED 128-870
SR-48	October-December 1976	AD A036735	ED 135-028
SR-49	January-March 1977	AD A041460	ED 141-864
SR-50	April-June 1977	AD A044820	ED 144-138
SR-51/52	July-December 1977	AD A049215	ED 147-892
SR-53	January-March 1978	AD A055853	ED 155-760
SR-54	April-June 1978	AD A067070	ED 161-096
SR-55/56	July-December 1978	AD A065575	ED 166-757
SR-57	January-March 1979	AD A083179	ED 170-823
SR-58	April-June 1979	AD A077663	ED 178-967
SR-59/60	July-December 1979	AD A082034	ED 181-525
SR-61	January-March 1980	AD A085320	ED 185-636
SR-62	April-June 1980	AD A095062	ED 196-099
SR-63/64	July-December 1980	AD A095860	ED 197-416
SR-65	January-March 1981	AD A099958	ED 201-022
SR-66	April-June 1981	AD A105090	ED 206-038
SR-67/68	July-December 1981	AD A111385	ED 212-010
SR-69	January-March 1982	AD A120819	ED 214-226
SR-70	April-June 1982	AD A119426	ED 219-834
SR-71/72	July-December 1982	AD A124596	ED 225-212
SR-73	January-March 1983	AD A129713	ED 229-816
SR-74/75	April-September 1983	AD A136416	ED 236-753
SR-76	October-December 1983	AD A140176	ED 241-973
SR-77/78	January-June 1984	AD A145585	ED 247-626
SR-79/80	July-December 1984	AD A151035	ED 252-907
SR-81	January-March 1985	AD A156294	ED 257-159
SR-82/83	April-September 1985	AD A165084	ED 266-508
SR-84	October-December 1985	AD A168819	ED 270-831
SR-85	January-March 1986	AD A173677	ED 274-022
SR-86/87	April-September 1986	AD A176816	ED 278-066
SR-88	October-December 1986	PB 88-244256	ED 282-278

SR-89/90	January-June 1987	PB 88-244314	ED 285-228
SR-91	July-September 1987	AD A192081	**
SR-92	October-December 1987	PB 88-246798	**
SR-93/94	January-June 1988	PB 89-108765	**
SR-95/96	July-December 1988	PB 89-155329	**
SR-97/98	January-July 1989	PB 90-121161	ED32-1317
SR-99/100	July-December 1989	PB 90-226143	ED32-1318
SR-101/102	January-June 1990	PB 91-138479	ED325-897
SR-103/104	July-December 1990	PB 91-172924	ED331-100
SR-105/106	January-June 1991	PB92-105204	ED340-053
SR-107/108	July-December 1991	PB92-160522	ED344-259
SR-109/110	January-June 1992	PB93-142099	ED352594
SR-111/112	July-December 1992		

AD numbers may be ordered from:

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22151

ED numbers may be ordered from:

ERIC Document Reproduction Service
Computer Microfilm Corporation (CMC)
3900 Wheeler Avenue
Alexandria, VA 22304-5110

In addition, Haskins Laboratories Status Report on Speech Research is abstracted in *Language and Language Behavior Abstracts*, P.O. Box 22206, San Diego, CA 92122

**Accession number not yet assigned

Contents—Continued

Haskins Laboratories Status Report on Speech Research

SR-111/112
JULY-DECEMBER 1992

Contents

—Continued Inside—