DOCUMENT RESUME

ED 331 100 CS 507 425

AUTHOR Studdert-Kennedy, Michael, Ed.
TITLE Status Report on Speech Research, July-December 1990.
INSTITUTION Haskins Labs., New Haven, Conn.
SPONS AGENCY National Inst. of Child Health and Human Development (NIH), Bethesda, MD.; National Inst. of Health (DHHS), Bethesda, MD. Biomedical Research Support Grant Program.; National Inst. on Deafness and Other Communications Disorders, Bethesda, MD.; National Science Foundation, Washington, D.C.
REPORT NO SR-103/104
PUB DATE Jul 90
CONTRACT NO1-HD-5-2910
NOTE 216p.; For previous report, see ED 325 897.
PUB TYPE Collected Works - General (020) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC09 Plus Postage.
DESCRIPTORS *Articulation (Speech); Communication Research; Cross Cultural Studies; *Language Processing; Language Research; *Recall (Psychology); Speech Habits; *Vowels
IDENTIFIERS Speech Perception; Speech Research; *Vocalization

ABSTRACT

One of a series of semiannual reports, this publication contains 13 articles which report the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. Articles and their authors are as follows: "The Role of Contrast in Limiting Vowel-to-Vowel Coarticulation in Different Languages" (Sharon Y. Manuel); "Anticipatory Velar Lowering: A Coproduction Account" (Fredericka Bell-Berti and Rena Arens Krakow); "Converging Sources of Evidence for Dissecting Articulatory Movements into Core Gestures" (Suzanne E. Boyce and others); "Rotation and Translation of the Jaw During Speech" (Jan Edwards and Katherine S. Harris); "Linguistic Structure and Articulatory Dynamics: A Cross Language Study" (Eric Vatikiotis-Bateson and J. A. Scott Kelso); "Gestural Specification Using Dynamically Defined Articulatory Structures" (Catherine P. Browman and Louis Goldstein); "Stimulus Order Effects in Vowel Discrimination" (Bruno H. Repp and Robert G. Crowder); "The Haskins Laboratories' Pulse Code Modulation (PCM) System" (D. H. Whalen and others); "Factors Contributing to Performance on Phoneme Awareness Tasks in School-Aged Children" (Anne E. Fowler); "Short-Term Serial Recall Performance by Good and Poor Readers of Chinese" (Nianqi Ren and Ignatius G. Mattingly); "Recall of Order Information by Deaf Signers: Phonetic Coding in Temporal Order Recall" (Vicki L. Hanson); "The Processing of Inflected Words" (Leonard Katz and others); and "Steady-State and Perturbed Rhythmical Movements: A Dynamical Analysis" (Bruce A. Kay and others). An appendix lists DTIC and ERIC numbers for publications in this series since 1970. (SR)

Haskins Laboratories Status Report on

Speech Research

SR-103/104
JULY-DECEMBER 1990

NEW HAVEN, CONNECTICUT

Haskins Laboratories 270 Crown Street New Haven, CT 06511-6695
Telephone (203) 865-6163 FAX (203) 865-8963

April 18, 1991

Ms Maureen Roberts
ERIC Processing and Reference Facility
2440 Research Boulevard
Rockville, MD 20850

Dear Ms Roberts:

Enclosed please find a copy of our Status Report on Speech Research (SR-103/104 July-December 1990) for processing in the ERIC system. Please send me the accession number when it has been processed.

Thanking you for your assistance, I remain

Sincerely,

Alice Dadourian

Alice Dadourian

Copies of SR-103/104: 212 content pages per copy

- Contents: Recycled white stock
  - 104 printed 2 sides
  - 2 printed 1 side—v, unnumbered vii
  - (pages vi and viii are unnumbered and blank—blank pages have been inserted)
- Covers and Spine: Unvarnished, white, heavy stock

Distribution Statement

Editor
Michael Studdert-Kennedy

Production Staff
Yvonne Manning
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
Contract NO1-HD-5-2910

National Institute of Health

Biomedical Research Support Grant RR-05596

National Science Foundation

Grant ; NS-8820099

National Institute on Deafness and Other Communication Disorders

Grant DC 00121
Grant DC 00183
Grant DC 00403
Grant DC 00016
Grant DC 00594

Investigators

Arthur Abramson*
Peter J. Alfonso*
Thomas Baer*
Eric Bateson*
Fredericka Bell-Berti*
Catherine T. Best*
Susan Brady*
Catherine P. Browman
Franklin S. Cooper*
Stephen Crain*
Robert Crowder*
Lois G. Dreyer*
Alice Faber†
Laurie B. Feldman*
Janet Fodor*
Carol A. Fowler*
Louis Goldstein*
Carol Gracco†
Vincent Gracco
Vicki L. Hanson*
Katherine S. Harris*
Leonard Katz*
Rena Arens Krakow*
Andrea G. Levitt*
Alvin M. Liberman*
Diane Lillo-Martin*
Leigh Lisker*
Anders Löfqvist*
Virginia H. Mann*
Ignatius G. Mattingly*
Nancy S. McGarr*
Richard S. McGowan
Patrick W. Nye
Lawrence J. Raphael*
Bruno H. Repp
Philip E. Rubin
Elliot Saltzman
Donald Shankweiler*
Michael Studdert-Kennedy*
Michael T. Turvey*
Douglas Whalen

*Part-time
†NRSA Training Fellow

Technical/Administrative Staff

Philip Chagnon
Alice Dadourian
Michael D'Angelo
Betty J. DeLise
Lisa Fresa
Vincent Gulisano
Donald Hailey
Raymond C. Huey*
Marion MacEachron*
Yvonne Manning
Joan Martinez
Maura Murphy
William P. Scully
Richard S. Sharkany
Zefang Wang
Edward R. Wiley

Students*

Melanie Campbell
Sandra Chiang
Margaret Hall Dunn
Terri Erwin
Elizabeth Goodell
Joseph Kalinowski
Laura Koenig
Betty Kollia
Simon Levy
Salvatore Miranda
Maria Mody
Weijia Ni
Mira Peter
Nian-qi Ren
Christine Romano
Joaquim Romero
Arlyne Russo
Jeffrey Shaw
Caroline Smith
Mark Tiede
Qi Wang
Yi Xu
Elizabeth Zsiga

Contents

Haskins Laboratories Status Report on

Speech Research

Haskins Laboratories Status Report on Speech Research
1990, SR-103/104, 1-20

The Role of Contrast in Limiting Vowel-to-vowel Coarticulation in Different Languages*

Sharon Y. Manuel†

Languages differ in their inventories of distinctive sounds and in their systems of contrast. Here we propose that this observation may have predictive value with respect to how extensively various phones are coarticulated in particular languages. This hypothesis is based on three assumptions: (1) there are "output constraints" on just how a given phone can be articulated; (2) output constraints are, at least in part, affected by language-particular systems of phonetic contrast; and (3) coarticulation is limited in a way that respects those output constraints. Together, these assumptions lead to the expectation that, in general, languages will tend to tolerate less coarticulation just where extensive coarticulation would lead to confusion of contrastive phones. This prediction was tested by comparing acoustic measures of anticipatory vowel-to-vowel coarticulation in languages which differ in how they divide up the vowel space into contrastive units. The acoustic measures were the first and second formant frequencies, measured in the middle and at the end of the target vowels /a/ and /e/, followed by /pV/, where /V/ was /i,e,a,o,u/. Two languages (Ndebele and Shona) with the phonemic vowels /i,e,a,o,u/ were found to have greater anticipatory coarticulation for the target vowel /a/ than does a language (Sotho) that has a more crowded mid- and low-vowel space, with the phonemic vowels /i,e,ɛ,a,ɔ,o,u/. The data were based on recordings from three speakers of each of the languages.

INTRODUCTION

Languages differ in their inventories of distinctive sounds and in their systems of contrast. The premise of this paper is that this observation may have predictive value with respect to how extensively certain phones are coarticulated in particular languages. We provide an empirical test of our theoretical perspective by comparing vowel-to-vowel coarticulation in languages that differ in how they divide the vowel space into contrastive units.

I would like to thank the faculty and students of the University of Zimbabwe for their generous help in the data collection. I am grateful to Suzanne Boyce, Harvey Gilbert, Marie Huffman, Rena Krakow, Harriet Magen, and Kenneth Stevens and three anonymous reviewers for valuable comments on an earlier draft of this paper, and to Louis Goldstein, Alvin Liberman, and Ignatius Mattingly for advice on my doctoral dissertation, on which this paper is based. I would also like to thank Tracy Sheppard, Linnea Bankey, and Yvonne Manning-Jones for help with manuscript preparation. This research was supported principally by NIH Grant HD-01994 to Haskins Laboratories and more recently by a NIH Postdoctoral Training Grant (DC-00005) to MIT.

Words are distinguished from one another by their phone composition, and since articulatory gestures (which ones, when they occur) and their acoustic consequences distinguish phones, we might expect speakers to exercise some effort to ensure that the acoustic consequences of articulatory gestures remain distinct (see, for example, Lindblom, 1983; Lindblom & Engstrand, 1989; Martinet, 1952, 1957; Stevens, 1989). That is, there may be *output constraints* (tolerances) on phones, definable in terms of articulatory or acoustic dimensions, that limit how much their articulatory gestures are allowed to stray from an ideally distinctive pattern. For example, a speaker who intends to say the word *tick* must make a tight tongue tip-alveolar closure—too weak a constriction will produce something that might be heard as *sick* or *thick*.

In speech, the articulatory requirements of one phone are often anticipated during the production of a preceding phone. This phenomenon, known as anticipatory coarticulation, results in contextually induced variability in the portion of the signal

that we conventionally associate with a given phone. An example is the difference in how /t/ is produced in *tea* and *tree*. In the word *tree* the tongue may make a relatively more posterior contact with the roof of the mouth and may itself be retroflexed, in anticipation of the following retroflex consonant /r/. Similarly, in a vowel-nasal sequence such as in *pan*, the velum typically begins (and may complete) its lowering movement, associated with the nasal /n/, while the vocal tract is still open for the vowel /æ/ and well before the oral occlusion for the /n/ is achieved.

Since coarticulation affects the very primitives of contrast between phones, i.e., articulatory gestures and their acoustic consequences, extreme coarticulation would possibly put speakers at risk of blurring or even obliterating phonetic contrasts. We might expect, then, that speakers generally limit coarticulation such that it does not destroy the distinctive attributes of gestures. That is, coarticulation might be limited so that the output constraints on distinctive gestures (i.e., gestures that have consequence for distinctiveness—analogous to distinctive features) are not violated (see also Engstrand, 1988; Manuel & Krakow, 1984; Manuel, 1987a, 1987b; Martinet, 1952, 1957; Schouten & Pols, 1979; Tatham, 1984; Keating, 1990).

But what counts as a distinctive gesture? Clearly, *what counts* varies from phone to phone and from language to language. For example, if the oral tract is completely closed, it matters quite a bit whether or not the velar port is appreciably open, since this is the major articulatory difference (and is responsible for the major acoustic difference) between voiced oral and nasal stops (e.g., /d/ vs. /n/). On the other hand, vowels in English are not distinguished by virtue of the fact that they are nasalized or not (though of course they normally are relatively nonnasal, the exceptions being when they occur in a sequence with nasal consonants[1]). It is therefore not surprising that English tolerates the presence of a substantial velar lowering gesture during vowels in nasal contexts.

With respect to language-particular aspects of distinctiveness, we note that different languages essentially take a universally available articulatory/acoustic space and divide it up differently. For example, Swedish distinguishes [+round] from [-round] front vowels, whereas English does not. Similarly, vowels in French are distinctively [+nasal] or [-nasal], but as noted above, English vowels are nasalized only in the context of a nasal consonant.

In some languages, very similar articulatory gestures may result in linguistically distinct phones. In order to maintain distinctiveness between those gestures, each must have a fairly narrow constraint on its production. On the other hand, if a language makes only a single phone in a particular region of some phonetic dimension, we might expect a fair amount of variability in production to be tolerated (see Keating and Huffman, 1984, for related discussion). For example, Malayalam makes two distinctive voiceless coronal stops—alveolar and dental, whereas English only makes one distinctive voiceless stop in that area of the vocal tract. English speakers have been shown to exhibit more variability in production of English voiceless coronal stops than do Malayalam speakers for Malayalam alveolar or dental stops (Jongman, Blumstein, & Lahiri, 1985). This observation suggests that the output constraints on gestures associated with particular phones, and consequently the degree to which those phones are susceptible to coarticulatory influence of neighboring phones, can be predicted to a large extent by examining the way a language uses a particular phonetic space.[2]

The results of several previous studies can be indirectly or directly related to the role that contrast plays in constraining coarticulatory behavior in different languages (e.g., Chasaide, 1979; Lubker & Gay, 1982; Magen, 1984; Manuel & Krakow, 1984; Öhman, 1966). Öhman's early study of vowel-to-vowel coarticulation showed that for English and Swedish V_1CV_2 utterances, the articulators begin assuming configurations needed for V_2 before the occlusion is reached for the medial consonant. However, similar effects were generally not found for Russian. As Öhman points out, Russian contrasts palatalized and unpalatalized consonants, and it is the position of the tongue body that distinguishes these two sets. Presumably in Russian the tongue body is not free to begin assuming the configuration for V_2 during the V_1 to C transition because the consonant itself makes (possibly conflicting) requirements on the tongue body.

Manuel and Krakow (1984) compared vowel-to-vowel coarticulation in English, a language with a relatively crowded vowel space (13 to 15 vowels, depending on dialect), and two Bantu languages (Shona and Swahili) for which the vowels are well spread-out (the phonemic vowels of Shona and Swahili are /i,e,a,o,u/). The prediction in Manuel and Krakow was that since the vowels of English are closer together in the articulatory/acoustic space, the range of productions for each English

vowel should be relatively small, while in contrast, since the vowels of Shona and Swahili are more spread apart, they could tolerate larger ranges of productions without danger of neutralizing phonemic differences. The main result of the experiment was as predicted. Vowel-to-vowel coarticulation (as reflected by values of the first and second formant frequencies measured in the middle of the vowels) was in fact stronger in Shona and Swahili than in English.

The results of the Manuel and Krakow study are consistent with the idea that proximity of contrastive phonetic units affects coarticulation. However, because the study was based on a single speaker of each language, we cannot conclude with certainty that Manuel and Krakow's results reflected language differences rather than simply speaker-to-speaker differences, given that speakers of a language may vary somewhat in amount and type of coarticulatory patterns (Lubker & Gay, 1982; Nolan, 1985) In addition, English vowels are often diphthongized, whereas Shona and Swahili vowels tend to be monophthongal. It is possible that the restricted degree of coarticulation in English was not due to constraints on the quality of vowel *nuclei*, but was instead due to demands of the diphthongal second part of the English vowels. Despite these confounding problems, the major finding was predicted by the assumption that distribution of vowels affects their output constraints, which in turn affect the amount of vowel-to-vowel coarticulation. We are unaware of any other hypothesis that has such predictive power for these results.

The current experiment is an expansion of the basic paradigm used by Manuel and Krakow. The three languages analyzed are all in the same family (Southern Bantu), and they are phonologically, morphologically, and syntactically similar to each other. Several speakers of each language were recorded, to allow comparison of coarticulation both between speakers of the same language, and across languages.

We included two languages (Shona and Ndebele) with the phonemic vowels /i,e,a,o,u/ and one language (Sotho) that has a more crowded vowel inventory /i,e,ɛ,a,ɔ,o,u/. It should be kept in mind that what is at issue here is not the number of vowels *per se*, but their distribution. As can be seen in Figure 1, the Sotho vowel space is more crowded in the low- and mid-vowel region than are the Shona and Ndebele vowel spaces. As a mnemonic for which languages have relatively less crowded (LC) and more crowded (MC) vowel spacing, we will use the shorthand Ndebele(LC), Shona(LC) and Sotho(MC).

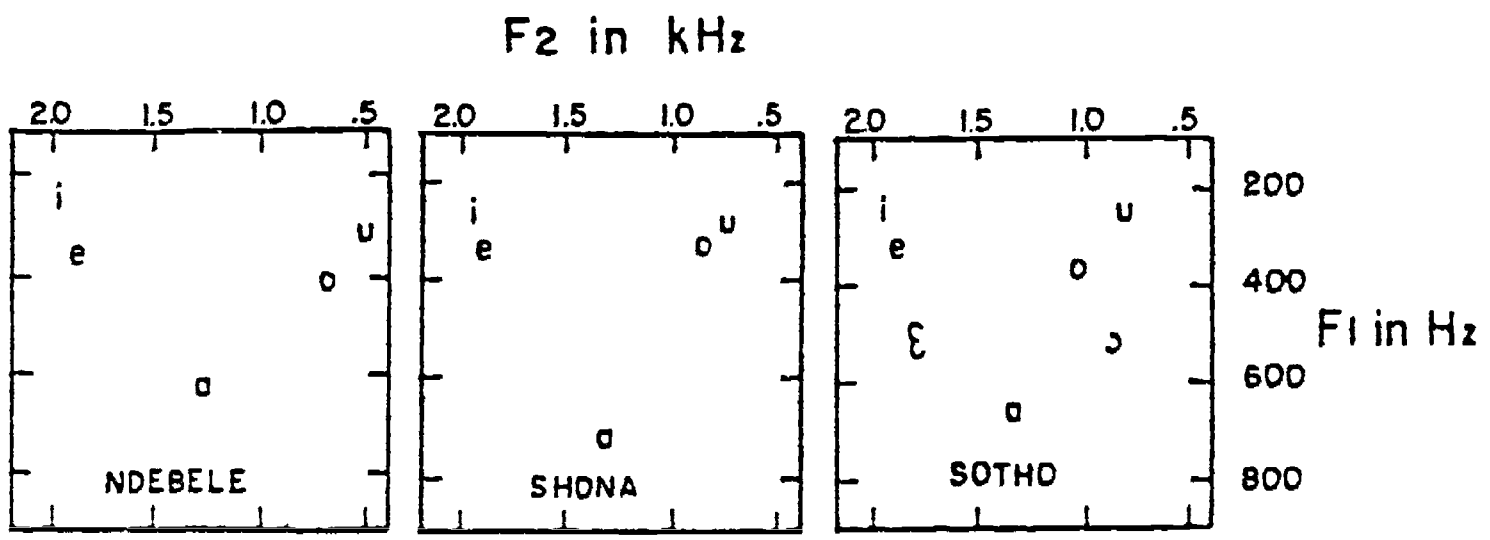


Figure 1. Examples of phonemic vowels of Ndebele, Shona and Sotho. Data are from one speaker of each of the languages.

In Sotho(MC), the vowels /ɛ,ɔ/ intervene between phonemic /a/ and phonemic /e,o/. If this relative crowding affects output constraints and limits coarticulation in Sotho(MC), we would expect to find smaller vowel-to-vowel coarticulatory effects on /a/, /e/, and /o/ in Sotho(MC) than in Shona(LC) and Ndebele(LC). Specifically, we expect Sotho(MC) /a/ to show less movement into the /ɛ,ɔ/ space. For the vowel /e/, we expect less movement into the /ɛ/ space, and for /o/, less movement toward the /ɔ/ space. Here we examine anticipatory coarticulation for the target vowels /a/ and /e/. The acoustic measures of coarticulation were the first (F1) and second (F2) formant frequencies.

Method

A. Further specifics of languages and subjects

The languages we studied are characterized by open syllables, and the vowels are monophthongal (see Cole, 1955; Doke, 1954). Word stress is not distinctive, but rather it is fixed on the penultimate syllable. Tone is phonemic in all three of the languages. Phonetically, the vowels /i/ and /u/ are quite high in these languages. Similarly, the mid vowels /e/ and /o/ are also phonetically rather high, with /e/ approaching the quality of [ɪ]. The low vowel /a/ is more fronted than its English counterpart. The Sotho(MC) vowel /ɛ/ is phonetically slightly higher than the vowel in English ***bet*** and Sotho(MC) /ɔ/ is similar to the vowel in English ***caught*** (for dialects which distinguish the vowels in ***caught*** and ***cot***). Of possible relevance to the present study is the claim that Sotho(MC) has a phonological rule which raises its mid vowels /e,ɛ,o,ɔ/ when they are followed in the next syllable by a higher vowel or a syllabic nasal, as well as by some palatal consonants. This process has the effect of raising the target vowel by a half step. For example, raised /e/ is higher than nonraised /e/ but not as high as /i/. Similarly, raised /ɛ/ is higher than nonraised /ɛ/, but lower than nonraised /e/.

The data reported here are from three adult male native speakers of each of the languages studied. All subjects were fluent speakers of English as a second language. The subjects were paid for their participation.

B. Materials

The /a/ and /e/ vowels that we analyzed were a subset of a larger data base. This larger database was comprised of nonsense trisyllables of the form /pV$_1$pV$_2$pV$_3$/, where the vowels could be any one of the five vowels /a,e,i,o,u/. The consonant /p/ was chosen because its articulation is relatively independent of articulations involving the tongue. These nonsense trisyllables were spoken in carrier phrases which were selected so that the last phoneme of the word preceding the target trisyllable was an /a/, and the first syllable of the word following the trisyllable was /pa/. The carrier phrases and their glosses are shown below[3]:

Shona(LC): Taura pepapa *pachena*
"Say 'pepapa' clearly."

Ndebele(LC): Ngiak .uluma *phephapha* pakati
"I spea : 'pepapa' in the middle."

Sotho(MC): Ke ua *phephapha* phakisa
"Sa / 'pepapa' quickly."

For each combination of target vowel and following vowel context (e.g., target /e/ followed by the context vowel /i/), we selected for analysis three of the 10 possible types of trisyllables which contained that particular combination. If the target vowel was in V$_1$, then V$_2$ was the context vowel, and V$_3$ was either /a/ (e.g., /pepipa/), or was identical to V$_2$ (e.g., /pepipi/). When the context vowel was /a/, these two types were not distinct. If the target vowel was itself V$_2$, then V$_3$ was the context vowel, and V$_1$ was always /a/ (e.g., /papepi/). We will pool data from these three types of trisyllables, as no relevant effects are found when the types are considered separately (see Manuel, 1987).

C. Recordings

Subjects were told that the purpose of the experiment was to find out how speakers of different languages produce speech sounds. They were instructed to read five randomized lists of the 125 trisyllables, inserting each trisyllable into the appropriate carrier phrase. Subjects were asked to produce all syllables on a low tone, and to speak at a normal rate. Despite this instruction, some subjects spoke very rapidly, particularly the Sotho(MC) speakers (this may have been due to the semantic content of the carrier phrase or because, according to some of the subjects, in their culture there is a popular game in which high value is put on speaking rapidly).

The recordings were made in a sound-treated recording studio at the University of Zimbabwe. The speech was recorded on audio tape at 3 3/4 ips on a NAGRA tape recorder. At the end of each list, subjects were asked to repeat trisyllables that had obviously been misread.

We had hoped to record those Sotho(MC) vowels which were phonetically closest to the Shona(LC) and Ndebele(LC) mid vowels, /e/ and /o/. Generally, in Sotho(MC) all of the allophones of

both /e/ and /ɛ/ are represented orthographically by *e*, and the allophones of /o/ and /ɔ/ as *o*; sometimes diacritics are employed to distinguish the phonemic vowels. Before reading the test materials, the Sotho(MC) speakers were presented with the five orthographic vowels *a, e, i, o, u,* and asked how they pronounced those vowels in isolation. For every speaker the quality of the isolated e was judged to be [e], and that of *o* to be [o]. We also monitored the speakers as they read the lists of trisyllables, and generally judged orthographic *e* and *o* to be produced as [e] and [o], respectively. However, subsequent listening to the audio recordings revealed that for two of the Sotho(MC) speakers there was a very large variability in the production of orthographic *e*, ranging from [ɪ] to [e], and one Sotho(MC) speaker produced a lower mid vowel, more like English /ɛ/. This observation indicates that, for the vowel /e/, we cannot be certain whether subjects were consistently producing a single phonemic vowel, and if not, which one in a given case. Though this results in some noise in the data, it does not crucially affect the ability to test the hypotheses under consideration. First, as shown in Figure 1, both Sotho(MC) /e/ and /ɛ/ have closer neighboring vowels than does the mid vowel /e/ of Shona(LC) and Ndebele(LC). Second, as will be discussed below, when we look at the context effects on target vowel /ɑ/, we concentrate on the effects of the /i/ and /u/, with less attention to the mid-vowel contexts.

D. Acoustic measurements

The speech was low-pass filtered with a cutoff frequency of 5 kHz and digitized at a 10 kHz sampling rate. The first (F1) and second (F2) formant frequencies were measured using linear predictive (LPC) analysis in the ILS package. A 20 ms analysis window was moved in 5 ms increments over the trisyllable. The formant peaks for each analysis frame were calculated using the root-solving procedure.

Here we will report on measures of F1 and F2 that were made at two points in the target vowels.[4] The first measurement point was made in the middle of the vowel. This point was selected by examining the waveform and formant tracks and choosing a point, in the middle region of the vowel, which seemed to be most clearly associated with maximal F1 values. It was expected that this point approximated the time of maximal jaw opening. We expected the anticipatory effects of the following vowel to be seen more clearly at the end than in the middle of the target vowel. Therefore the second measurement point, the end point, was made as close to the /p/ closure (as observed in the waveform) as possible but with plausible F1 and F2 values.[5] This procedure yielded four values for each token [two formant values (F1, F2) by two measurement points (*middle, end*)].

Occasionally this procedure failed to yield F1 or F2 at one of the selected measurement points; in these cases a point immediately preceding or following the desired point was selected. For some vowels the entire region of interest failed to yield formant values (this was particularly problematic for F1) and an attempt was made to determine the values by examining smoothed spectra in that region. Finally, for each speaker, the mean and standard deviation were calculated separately for each of the four measures for both target vowels. Values which were more than two standard deviations above or below the mean were omitted from further analysis. This last procedure resulted in a loss of from 0 to 3% of the values from each measure for any one subject. Thus, while five tokens of each type were recorded, not all (occasionally as few as two) were ultimately usable; some trisyllables were misread, and for others it was impossible to determine the formant frequencies. Since we have pooled data from three types of utterances, for each speaker, and for each combination of target vowel and context vowel, there are from 9 to 15 (with a mean of 13.4) tokens. The exception is when the context vowel was /ɑ/, for which there are only two types of utterances, and from 7 to 10 (with a mean of 9.1) tokens per target for each speaker.

Results

A. Non-Context effects: Distribution of /ɑ/ and /e/

Each speaker's average F1 and F2 values in the middle of target vowels /ɑ/ and /e/ are plotted in the F1/F2 space in Figure 2. For comparison, values for these speakers' /i,o,u/ are also shown; no further analysis was done on these reference vowels. For all subjects except Sotho(MC) speaker 1, /e/ has a relatively high F2 and low F1 value. The general trend of a high F2 and low F1 for most of the speakers' /e/ vowels is not surprising, given the auditory impression that this usually was a fairly high vowel. As noted earlier, some Sotho(MC) orthographies are ambiguous as to whether orthographic *e* represents the phoneme /e/ or /ɛ/, and among those orthographic traditions which do distinguish the two, the conventions on use of diacritics are different. It seems not unlikely that Sotho(MC) speaker 1 produced the lower phoneme, while the other two speakers produced the higher one.

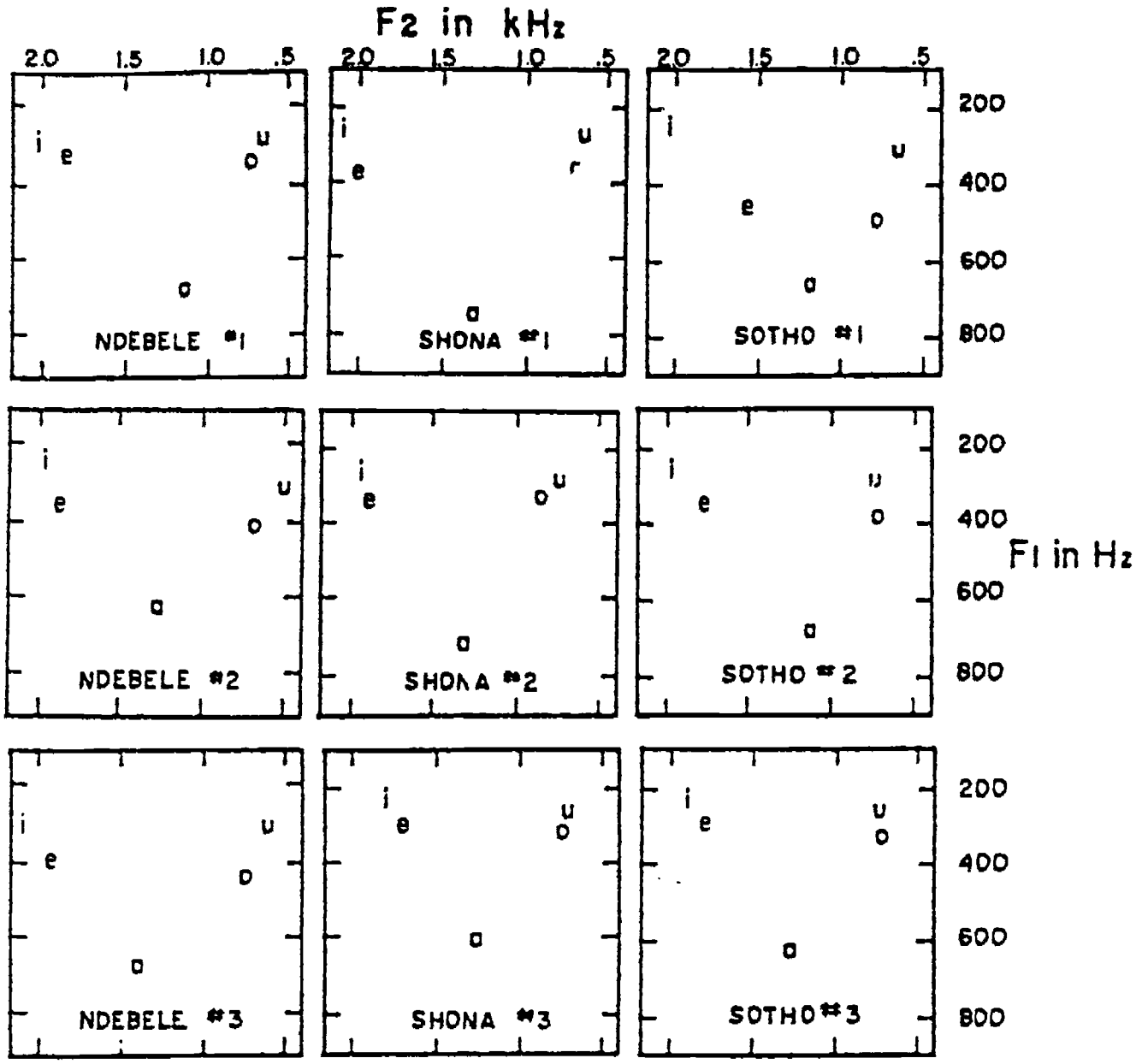


Figure 2. Average F1 and F2 values for the middle of target vowels /ɑ/ and /e/ for the nine subjects. Values for the vowels /i/, /o/, and /u/ are from the medial vowel in /pipipi/, /popopo/, and /pupupu/, respectively, and are based on from two to five tokens each. Values for /ɑ/ and /e/ are based on more contexts, as indicated in the text.

B. Expected effects on F1 and F2 of fronting/backing and raising/lowering contexts

In Figures 3a-c we show the expected acoustic effects of variation in production of the vowel /ɑ/. In these figures, the subscript assigned to the vowel /ɑ/ identifies the following vowel context. The locations of the points indicate schematically how the formant frequencies for the vowel /ɑ/ are expected to be influenced by the different following vowel contexts. Fronting and backing of /ɑ/ should cause changes in F2, as shown in Figure 3b. Raising of /ɑ/ should primarily affect F1, with little effect on F2, as shown in Figure 3a. A combination of raising and front/back movement should be reflected in movement of both F1 and F2, as shown in Figure 3c. For target vowel /ɑ/, we are interested in how much it moves toward the phonetic /e/ and /ɔ/ spaces. All of the vowel contexts we used here (other than /ɑ/ itself) are contexts which would potentially move target /ɑ/ into the phonetic spaces of interest.

The potential context effects for the target vowel /e/ are shown in Figures 4a-c. We looked at target /e/ as a function of two context vowels (/i,u/) which might raise the target /e/, three vowels which might back /e/ (/ɑ,u,o/), one fronting context (/i/), and one lowering context /ɑ/. As shown in Figure 4a, for a front vowel like /e/, variation along the front/back dimension should be reflected acoustically in F2: at a given height, the more front the vowel is, the higher its F2. Changes in

the height of /e/ have effects *both* in F1 and F2, since in general the higher a front vowel is, the lower its F1 and the higher its F2 (Fant, 1960). Thus if /e/ is sensitive to only the height of a following vowel, we might expect to see its contextually induced variants as shown in Figure 4b. Note that in this case a following /u/ actually results in a variety of /e/ which has a higher F2 than if the target /e/ is followed by /ɑ/. Finally, if /e/ is affected by both the front/back and height character of the following vowel, we might expect a pattern somewhat like that of Figure 4c.

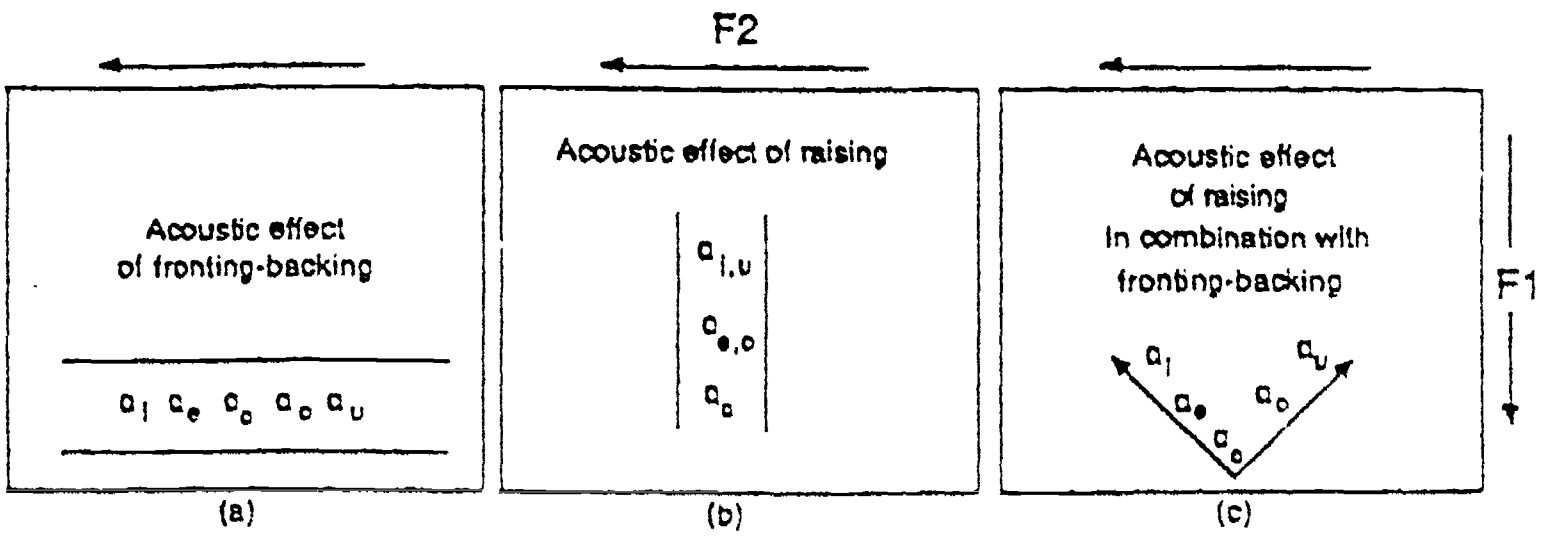


Figure 3. Expected acoustic correlates of context-induced variation in the target vowel /ɑ/: *a)* fronting-backing; *b)* raising; *c)* combination of raising and fronting-backing. The subscripts indicate the quality of the following vowel; for example, $ɑ_i$ indicates the vowel /ɑ/ when it is followed by the vowel /i/.

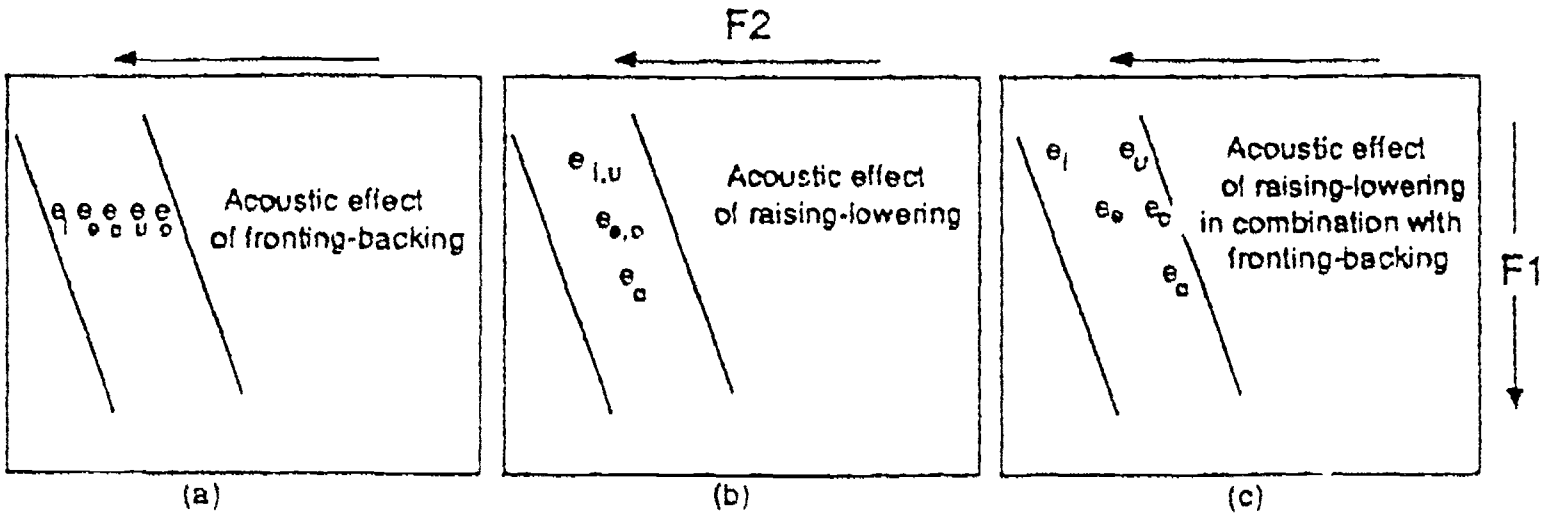


Figure 4. Expected acoustic correlates of context-induced variation in the target vowel /e/: *a)* fronting-backing; *b)* raising-lowering; *c)* combination of raising and fronting-backing. Subscripts indicate the quality of the following context vowel (e.g., e_i indicates the vowel /e/ when it is followed by the vowel /i/).

C. Obtained effects of vowel context: General

To get an overall impression of the anticipatory coarticulation effects, we averaged the data across the three speakers of each language. These averaged data are plotted in the F1/F2 space as a function of context and measurement point in Figures 5a-c (individual subject data are given in Table 1). In Figures 5a-c, for each language, there are two scatters of points, one scatter for the vowel /ɑ/ (in the lower portion of the graphs), and one for the vowel /e/ (in the upper left portion of the graph). Each of these scatters is composed of two smaller sets of data, one for measurements made in the middle of the vowel (points enclosed by inner loops), and one for measurements made at the end of the vowel (points enclosed by outer loops). Each symbol represents the context in which the target vowel occurred (e.g., filled squares indicate values for the target vowels followed by the context vowel /i/).

D. Context effects on the vowel /ɑ/

Beginning with the Ndebele(LC) data shown in Figure 5a, we see that all F1 values for target /ɑ/ are lower at the end than in the middle of the vowel, presumably due to the labial closure. The consonant closure generally had a lowering effect on F2 as well, as can be seen by comparing the midpoint and endpoint F2 values for target vowel /ɑ/ followed by context vowel /ɑ/ (open circles).

In general, for Ndebele(LC), the following vowel context had a large effect on the target vowel /ɑ/. Relative to the /ɑ/ context, high vowels /i,u/ tend to lower F1 of target vowel /ɑ/. The mid-vowel contexts /e,o/ also lower F1 of target vowel /ɑ/, though to a lesser extent. Front vowels /i,e/ raise F2, and the back rounded vowels /o,u/ lower F2. Even in the middle of /ɑ/ there is an effect of the following vowel. The effects of context on F2 increase as the measurement point moves closer to the end of the target vowel, and therefore to the contextual vowel itself. The vowel context effect on F1 does not appear to be much larger at the end than in the middle of the target vowel.

The vowel-context effects for Shona(LC) and Sotho(MC) are shown in Figures 5b and 5c, respectively. The influence of the context vowel is smaller in these languages, particularly in Sotho(MC), where there is very little effect of following vowel context at the steady-state point. In Sotho(MC) there is essentially no difference in F2 for target /ɑ/ produced in the contexts of a following /ɑ/, /o/, and /u/, whereas in both Ndebele(LC) and Shona(LC) there are some differences between these contexts, although they are small in the steady state. At later points in the target vowel, a front/back effect emerges for Sotho(MC), but it is much smaller than that seen in Ndebele(LC) or Shona(LC). Height effects in Sotho(MC) also appear to be relatively small. In general, Sotho(MC) /ɑ/ appears to spread less into the F1/F2 space than does /ɑ/ in Ndebele(LC) or in Shona(LC). In fact, most of the movement of Sotho(MC) /ɑ/ is straight up the middle of the vowel space, and has the form expected if this movement were due mostly to the intervocalic /p/. In what follows, we will examine the F1 and F2 effects separately for the target vowel /ɑ/.

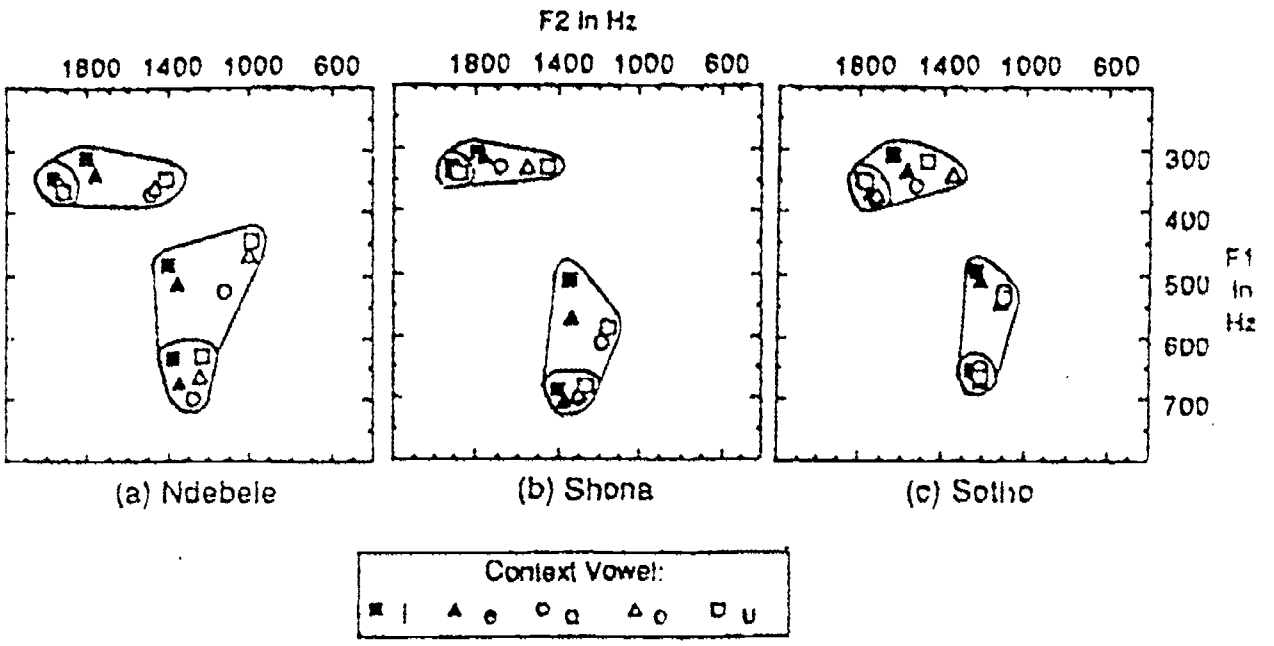


Figure 5. Acoustic effects of following context vowel on the target vowels /ɑ/ and /e/ in the three languages. Inner loops enclose data from the middle of the target vowels, outer loops enclose measurements made at the end of the target vowels. Data are averaged over the three speakers of each language.

Table 1. *Individual subject d. a.*

Context Vowel	i		e		a		o		u	
Measurement Point	MID	END	MID	END	MID	END	MID	END	MID	END
					F2 of Target Vowel /a/					
Speaker										
Sotho(MC) #1	1256	1266	1212	1199	1248	1136	1227	1130	1218	1131
Sotho(MC) #2	1169	1151	1160	1180	1118	1017	1108	1021	1142	1063
Sotho(MC) #3	1348	1281	1339	1272	1275	1169	1296	1174	1282	111.
Shona(LC) #1	1397	1384	1380	1353	1346	1227	1314	1183	1276	1157
Shona(LC) #2	1408	1348	1384	1342	1361	1230	1310	1192	1301	1196
Shona(LC) #3	1386	1313	1344	1298	1208	1128	1260	1141	1224	1124
Ndebele(LC)#1	1235	1295	1209	1234	1200	1070	1145	901	1129	892
Ndebele(LC)#2	1413	1458	1369	1385	1301	1134	1244	1029	1221	997
Ndebele(LC)#3	1476	1458	1470	1444	1332	1176	1350	1082	1353	1098
					F1 of Target Vowel /a/					
Sotho(MC) #1	672	540	659	575	661	564	670	597	667	556
Sotho(MC) #2	660	552	686	563	674	602	692	607	701	625
Sotho(MC) #3	621	385	625	382	608	414	653	423	622	423
Shona(LC) #1	755	673	748	695	772	725	750	687	741	695
Shona(LC) #2	703	418	736	534	753	630	713	561	690	561
Shona(LC) #3	602	447	630	494	587	479	621	504	609	504
Ndebele(LC)#1	688	483	694	566	705	589	681	445	674	402
Ndebele(LC)#2	612	450	620	478	692	499	625	456	574	445
Ndebele(LC)#3	605	513	710	491	699	489	682	496	645	488
					F2 of Target Vowel /e/					
Sotho(MC) #1	1681	1546	1597	1461	1624	1463	1581	1397	1620	1419
Sotho(MC) #2	1847	1744	1813	1619	1746	1486	1774	1312	1871	1573
Sotho(MC) #3	1848	1634	1850	1649	1828	1628	1804	1358	1815	1435
Shona(LC) #1	2054	1965	2041	1921	2002	1864	2053	1717	2054	1480
Shona(LC) #2	1934	1790	1969	1763	1931	1612	1905	1459	1879	1419
Shona(LC) #3	1756	1635	1723	1596	1715	1560	1734	1468	1706	1463
Ndebele(LC)#1	1918	1754	1914	1726	1919	1580	1904	1518	1883	1468
Ndebele(LC)#2	1985	1885	1932	1804	1866	1394	1907	1377	1922	1245
Ndebele(LC)#3	1991	1777	1931	1741	1996	1497	1934	1501	1964	1501
					F1 of Target Vowel /e/					
Sotho(MC) #1	437	356	475	398	477	431	477	413	445	379
Sotho(MC) #2	326	311	349	331	373	365	358	343	318	298
Sotho(MC) #3	280	260	287	277	291	283	284	269	294	284
Shona(LC) #1	371	319	377	329	373	351	370	352	375	353
Shona(LC) #2	328	327	344	337	342	351	339	346	337	327
Shona(LC) #3	290	280	295	282	308	290	303	293	309	309
Ndebele(LC)#1	317	283	319	303	335	319	329	318	335	311
Ndebele(LC)#2	329	309	343	325	377	378	365	368	359	342
Ndebele(LC)#3	383	342	399	379	385	381	400	391	399	378

D.1 Front/back effects (F2) on target /a/. To simplify the analysis of front/back coarticulation, we concentrated on the two contexts that tended to yield the highest and lowest F2 values, that is /i/ and /u/, respectively. The F2 data for /a/ followed by /i/ and by /u/, averaged across speakers within each language, are shown in Figures 6a-c. In these figures, which can be thought of as highly schematic formant trajectory plots, F2 is shown on the vertical axis, and the two measurement points

are shown on the horizontal axis. Individual subject data are shown in a similar fashion in Figures 6d-l.

Again, it is clear that on average, Ndebele(LC) speakers displayed a large amount of front/back coarticulation. The difference in the /i - u/ contexts was 141 Hz in the middle of the vowel, and 408 Hz at the end of the target vowel. These Ndebele(LC) data were submitted to an analysis of variance with repeated measure factors of the /i/ vs. /u/ contexts and the two measurement points. While the individual speaker data show that Ndebele(LC) speaker 2 contributes more to the average /i - u/ difference than do the other two Ndebele(LC) speakers, the front/back contrast is significant [$F(1,2) = 107.6$, $p < 0.01$]. The front/back contrast is much larger at the end of the vowel, and this was reflected in an interaction of the front/back effect with the measurement point [$F(1,2) = 238.5$, $p < 0.01$]. Simple main effects tests confirm that the front/back contrast is significant at the 0.05 level in the middle of the vowel, and at the 0.01 level at the end of the vowel.

For Shona(LC), the /i - u/ difference is about the same (130 Hz) as Ndebele(LC) in the middle of the vowel, but considerably smaller (189 Hz) than Ndebele(LC) at the end of the vowel. The Shona(LC) speakers all showed patterns similar to each other, and the front/back contrast was in fact significant [$F(1,2) = 112$, $p < .01$]. Although each Shona(LC) speaker had larger /i - u/ differences at the end of the vowel than in the middle, the interaction between measurement point and the /i - u/ difference was not statistically significant [$F(1,2) = 6.16$, $p > 0.05$]

On average, the Sotho(MC) speakers showed much less of an /i - u/ difference (44 Hz) in the middle of the target vowel than did speakers of either Shona(LC) or Ndebele(LC). While each speaker did show at least some tendency for the /i/ context to give higher F2 values than the /u/ context, no Sotho(MC) speaker showed as large a difference as even the *smallest* difference found among the Shona(LC) and Ndebele(LC) speakers. Furthermore, while there was a main effect of context [$F(1,2) = 24.9$, $p < 0.05$], simple main effects indicated that in the middle of the vowel, the /i - u/ contrast was not significant for Sotho(MC) speakers [$F(1,2) = 14.1$, $p > 0.05$]. At the end of the vowel, Sotho(MC) speakers did have a substantial /i - u/ difference (130 Hz), and this was statistically significant [$F(1,2) = 30.4$, $p < 0.05$]. Even so, this difference at the end of the vowel was only as large as the /i - u/ difference in the *middle* of the Shona(LC) /ɑ/ vowel.

D.2 Context effects on F1 of target /ɑ/. The averaged data shown in F1/F2 plots in Figures 5a-c indicate that there is some effect of context on the F1 value of target vowel /ɑ/. For example, in the middle of target vowel /ɑ/ for the Ndebele(LC) speakers, the high-vowel contexts (/i/ and /u/) lower F1 by about 65 Hz, relative to target /ɑ/ followed by the low vowel /ɑ/. The mid vowels /e/ and /o/ lower F1 by a smaller amount, about 30 Hz. A similar pattern can be seen at the end of the vowel, with additional lowering of F1 for the /o/ and /u/ contexts, compared to the /e/ and /i/ contexts. In the other languages, there is also a general tendency for F1 to lower when the target vowel is followed by high vowels. However, the details of the patterns are somewhat different. For example, for Shona(LC) speakers, at the end of target vowel /ɑ/, the /u/ context has a higher F1 than does the /i/ context, whereas the opposite pattern was found in Ndebele(LC). In Sotho(MC), a following /ɑ/ context actually gives a lower F1 value than does a following /u/. As can be seen in Table 1, there was a fair amount of subject-to-subject variability in the F1 data. This variability may have been due in part to general problems of obtaining F1 data, as discussed above and in Note 5.

As we did for F2, we simplified the F1 analysis by focusing on only some of the possible context contrasts. In this case we compared the *average* of the two high-vowel contexts (/i/ and /u/) to the data for the low-vowel context /ɑ/. The results, averaged across the speakers within each language, are shown in Figures 7a-c.

For Ndebele(LC), the average difference between the high-vowel and low-vowel contexts was 66 Hz in the middle of the vowel, and 62 Hz at the end of the vowel. The average Shona(LC) data also show differences between the high-vowel and low-vowel contexts: 21 Hz in the middle of the vowel and 61 Hz at the end of the vowel. Despite the overall tendency for F1 to be lower in the high-vowel contexts, neither Ndebele(LC) nor Shona(LC) show a statistically significant F1 effect for high- vs. low-vowel contexts. However, as there are only three speakers for each language, the statistical tests are particularly susceptible to speaker-to-speaker variability, which was rather large for the F1 measure in Ndebele(LC) and Shona(LC). When we combine the data from all six speakers of these two languages, the high vs. low effect does reach significance [$F(1,5) = 10.04$, $p < 0.05$].

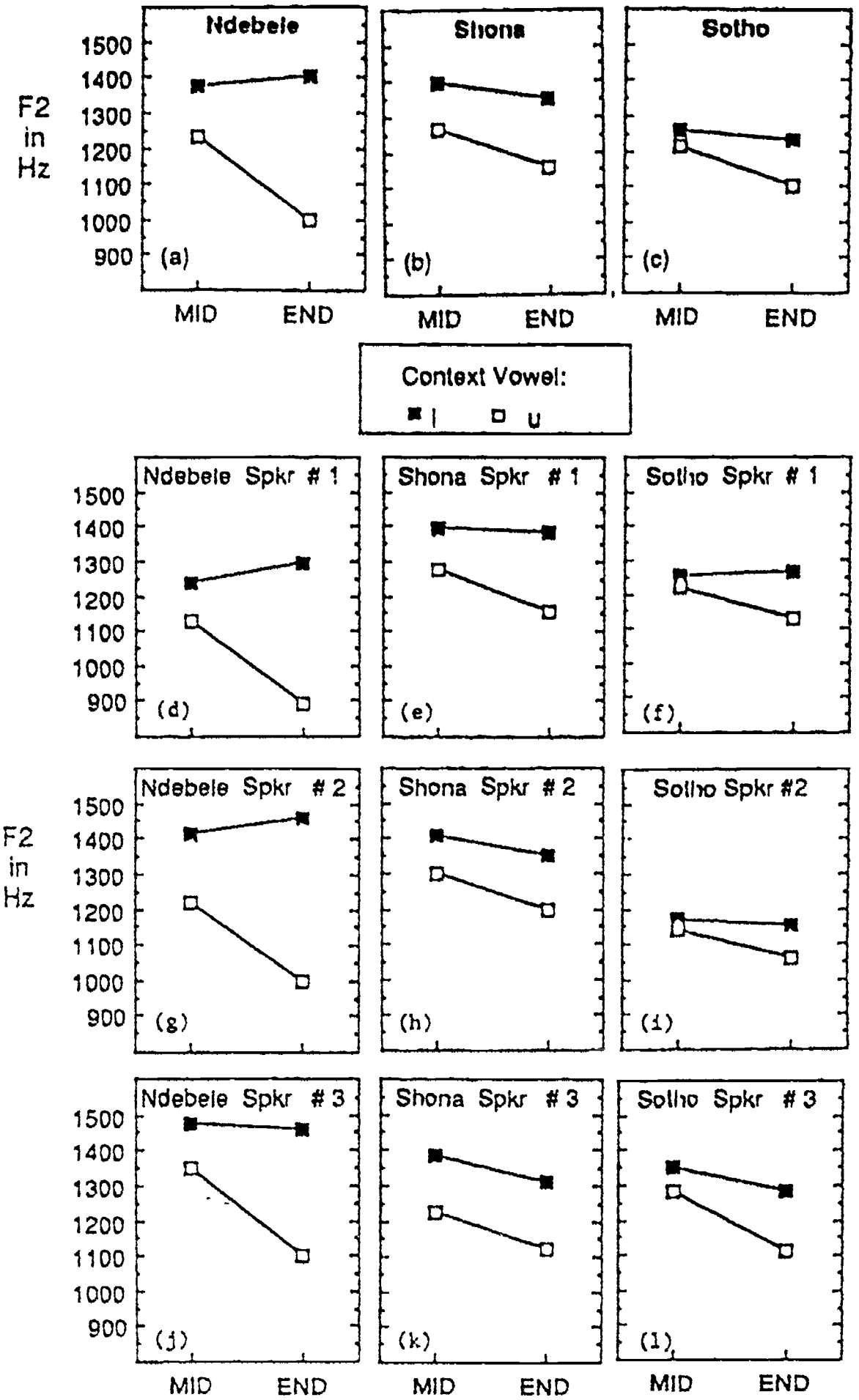


Figure 6. F2 of the target vowel /ɑ/ followed by the front vowel /i/ (filled squares) and by the back vowel /u/ (open squares). Data are shown for measurements made in the middle and end of the target vowel. Plots a-c are averaged over the three speakers of each of the languages. Plots d-l are individual subject data.

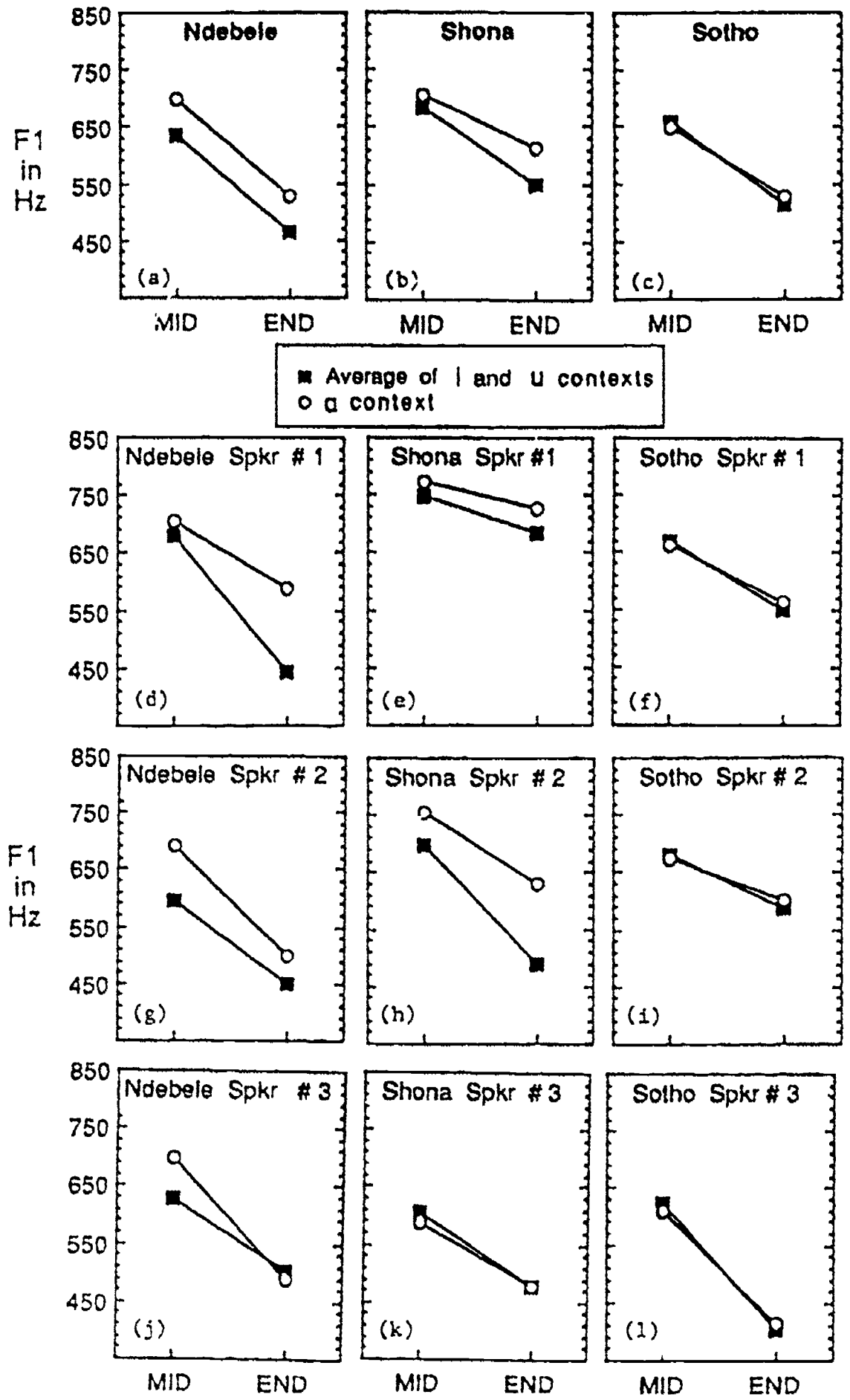


Figure 7. Open circles are F1 values for target vowel /ɑ/ in the context of a following low vowel (/ɑ/), and filled squares show average of F1 value from the /i/ and /u/ contexts for target vowel /ɑ/. Plots a-c are averaged over the three speakers of each of the languages. Plots d-l are individual subject data.

For the Sotho(MC) speakers, there was very little difference in the high and low contexts, either in the middle of the vowel (which was 7 Hz in the *wrong* direction) or at the end of the vowel, where the high-vowel contexts decreased F1 by only 13 Hz. In this case, the average data represent very well the individual subject data, as every Sotho(MC) speaker showed the same pattern (Figures 7f, i, l). Thus, the very small effects, which reversed in direction from the middle to the end of the vowel, gave an interaction of the context and measurement point factors [$F(1,2) = 226.7, p < 0.01$], and significant effects of the high/low context at both the middle of the vowel [$F(1,2) = 23, p < 0.05$] and end of the vowel [$F(1,2) = 56.3, p < 0.05$].

E. Context effects on target vowel /e/

Referring back to Figures 5a-c, we see that compared to target vowel /ɑ/, target vowel /e/ shows much less F1 fall as a function of movement toward the oral tract closure for /p/ than does target vowel /ɑ/. This relatively small amount of F1 fall is presumably because /e/ is already a high vowel, with a lower F1, than target vowel /ɑ/.

Perhaps the most striking aspect of the /e/ data, and of crucial relevance here, is that in none of the languages do the /e/ values encroach very much on the phonetic /ɛ/ space. The vowel /e/ has a relatively high F1 value even when followed by /ɑ/ in Ndebele(LC) and Shona(LC). The production of flanking /p/ consonants may have encouraged the subjects to maintain a relatively high jaw position during the target vowel, thus limiting or canceling the lowering effects of a following low vowel.

To show more clearly the context effects that do exist, in Figures 8a-c we have expanded the upper left portion of the F1/F2 plots that were shown in Figures 5a-c. The points enclosed by loops are from measures made in the middle of target vowel /e/, and the other points are from the measure at the end of the vowel. Again, beginning with Ndebele(LC), in the middle of target /e/ there is very little effect of context, though /i/ has a tendency to both lower F1 and raise F2. A similar but smaller effect can be seen in the middle of Shona(LC) /e/. In the middle of Sotho(MC) target /e/, both a following /i/ and /u/ give lower F1 and higher F2 values than a following /e/ context does. This pattern was found for Sotho(MC) speakers 1 and 2, and is consistent with what we would expect from a phonological rule that raises /e/ when it is followed by a high vowel (see Figure 4b).

If we look at the measurements made at the end of target /e/, a front/back effect emerges in all languages, such that for a given height of the context vowel, the back vowel contexts result in lower F2 values. This effect is strongest in Ndebele(LC). At the end of the vowel, on average each of the languages shows a height effect in that high front contexts yield lower F1 values than mid front vowels, and high back vowels yield lower F1 values than do mid or low back vowels. However, at a given height specification, the back vowels give lower F1 values. This may be due to rounding for the /o/ and for /u/ contexts.

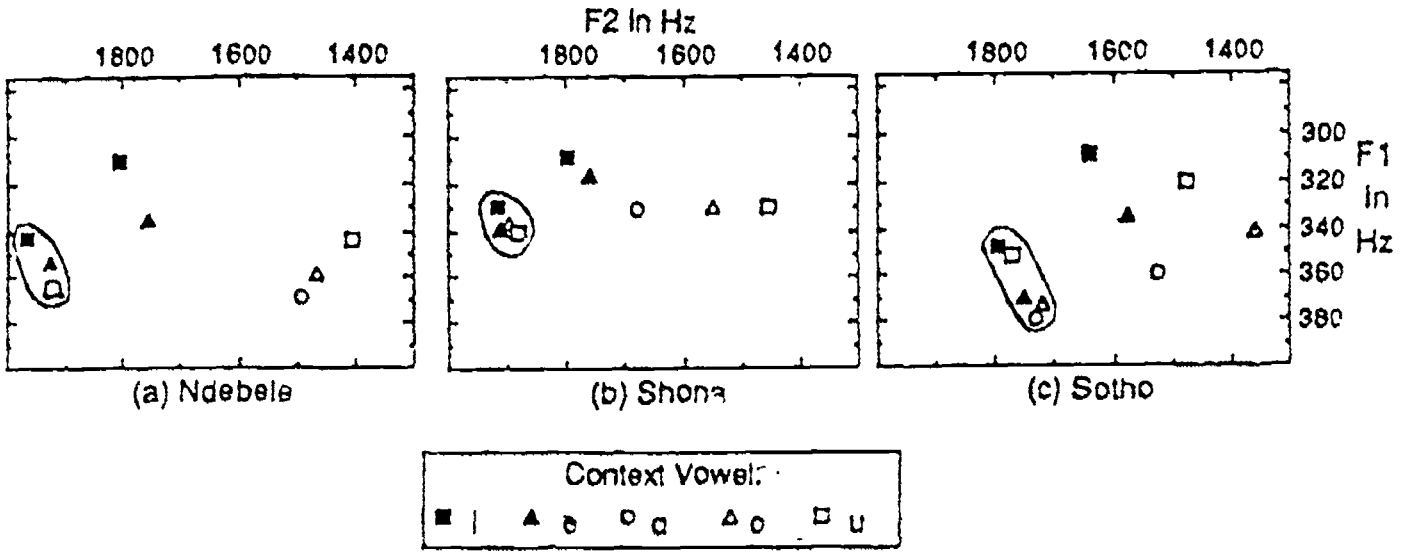


Figure 8. Expanded view of Figure 5, showing the acoustic effects of following-vowel context on the /e/. Loops enclose data from the middle of the target vowels; remaining points show measurements made at the end of the target vowels. Data are averaged over the three speakers of each language.

We are primarily interested in seeing if the Ndebele(LC) and Shona(LC) /e/ show more movement into the /ɛ/ phonetic space, a space that is used phonemically in Sotho(MC), but not in the other languages. The only vowel context which would be expected to cause both lowering and backing is /ɑ/. Consequently, we focused our attention on the F1 and F2 values of /e/ followed by /ɑ/ vs. /e/ followed by /e/.

The difference in the /e - ɑ/ contexts for F1 was small, both in the middle and end of the target vowel, and did not differ substantially from language to language, especially when individual speaker differences were taken into account. In general, there was a high degree of speaker-to-speaker variability with respect to anticipatory coarticulation effects for target vowel /e/. The only language for which there was a significant effect of context on F1 was Shona(LC) [$F(1,2) = 41.3, p < 0.05$], but the /e - ɑ/ context difference was actually smallest, and quite negligible, in this language (on average only 3 Hz in the middle of the vowel and 15 Hz at the end of the vowel for the /e/ vs. /ɑ/ contexts). For F2, none of the languages had a statistically significant /e - ɑ/ context effect, even though all Ndebele(LC) subjects had at least a 140-Hz higher F2 value for the /e/ than for the /ɑ/ context, measured at the end of the vowel. On average, Shona(LC) had an 81-Hz, and Sotho(MC) a 50-Hz /e - ɑ/ context difference at the end of the vowel. Possibly, with more speakers of each language, the context effects on F2, and even the small context effects on F1, would prove to be significant. When we pooled the data for the nine speakers together, there was a significant contrast of the /e/ and /ɑ/ contexts at the end of the vowel for both F1 [$F(1,8) = 14.5, p < 0.01$] and F2 [$F(1,8) = 9.4, p < 0.05$].

Discussion

A. Inventories of contrast as predictors of output constraints

Our prediction was that Sotho(MC) speakers would exhibit smaller anticipatory coarticulation effects for /ɑ/ and /e/ than would Shona(LC) or Ndebele(LC) speakers, because too much anticipation of an upcoming sound in Sotho(MC) would move the articulators (and therefore acoustics and perception) into competing distinctive vowel spaces. For example, we assumed that /ɑ/ would have a tighter output constraint in Sotho(MC) than in Shona(LC) or Ndebele(LC), because of the proximity of Sotho(MC) /ɑ/ to other phonemic Sotho(MC) vowels /ɛ, ɔ/.

When the target vowel was /ɑ/, our prediction was borne out. Even when we take into account speaker-to-speaker variability, Shona(LC) and Ndebele(LC) show larger anticipatory coarticulation effects on /ɑ/ than does Sotho(MC). Most of the Sotho(MC) /ɑ/ movement is straight up the middle of the vowel space, and is mostly due to the oral tract closure associated with the intervocalic /p/. By limiting *lateral* movement, speakers of Sotho(MC) avoid excessive movement into the phonemic /ɛ/ and /ɔ/ spaces of this language.

We had expected to see similar effects for the target vowel /e/, with Shona(LC) and Ndebele(LC) exhibiting more movement of /e/ toward the phonetic /ɛ/ space. However, none of the languages studied showed much movement into the /ɛ/ space, and what effects there were did not seem to be less robust in Sotho(MC) than in the other languages. One reason for this result might be that the target vowel is flanked on both sides by the consonant /p/. This /p/ may itself impose constraints on the height of the jaw, and may limit the anticipatory lowering movement of the following /ɑ/ during the target vowel /e/.

Thus far we have focused on the direct comparison of Sotho(MC), Shona(LC) and Ndebele(LC), which after all, do not differ very much in terms of their respective spacing of phonemic vowels. Actually, all of these languages might be expected to have relatively large coarticulatory effects as compared to a language like English (Very Crowded), which has many more vowels crowded into the vowel space. In the languages in the present study, the effects of V_2 on V_1 are clearly observable in the V_1C transitions, but they also extend into the middle portions of V_1, and in Shona(LC) and Ndebele(LC) these effects are remarkably large. Examples of vowel-to-vowel effects in Ndebele(LC), Sotho(MC), and English(VC) can be seen Figure 9, which compares the medial /ɑ/ in /pɑpɑpɑ/ and /pɑpɑpi/ for a single token for one speaker each of the three languages (see also Manuel and Krakow, 1984). The effects in English(VC) are very small, as we would expect given the proximity of English(VC) /æ/ and /ɑ/. Somewhat more coarticulation is seen in Sotho(MC), in which the nearest front vowel to /ɑ/ is /ɛ/. Finally, it is not surprising that Ndebele(LC) shows very extensive effects, since the next closest front vowel to /ɑ/ is /e/ in Ndebele(LC).

Given these observations, we fully expect that anticipatory coarticulation in Shona(LC) and Ndebele(LC) may extend further back into the vowel, and perhaps is present even at the

beginning of V_1. Interestingly, English schwa, a phone that it is known to behave as if it had very lax output constraints, allows quite extensive anticipatory coarticulation. As Magen (1989) has recently shown for English trisyllables like /bəbʌbi/ and /bəbʌbə/, the effects of the third vowel are seen throughout a medial schwa and can extend into the first vowel.

Strictly speaking, since we have not tested a number of seven-vowel languages against a number of five-vowel languages, our data alone do not allow us to generalize to other five- vs. seven-vowel comparisons. It is possible, then, that our results have nothing to do with output constraints or phonemic contrasts, but are due to some other (perhaps arbitrary) fact about the languages studied. Further testing of other languages is necessary to determine whether or not the present data reflect a general tendency, or are due to a fortuitous sampling accident.

Having noted this qualification, we observe that similar results have been found in other studies that reflect on the effect of contrast in constraining coarticulation (e.g., Lubker & Gay, 1982; Magen, 1984; Manuel & Krakow, 1984; and Öhman, 1966). Note, however, Nartey's (1979) cross-linguistic analysis of fricatives found that variation in fricative production was not correlated with the number or distribution of contrastive fricatives in a language. It may be that the precision needed to make fricatives, or the more categorical nature of consonant perception or production, or perhaps simply arbitrary amounts of pickiness in some of the languages studied, is such that any added constraint from the system of contrasts is negligible.

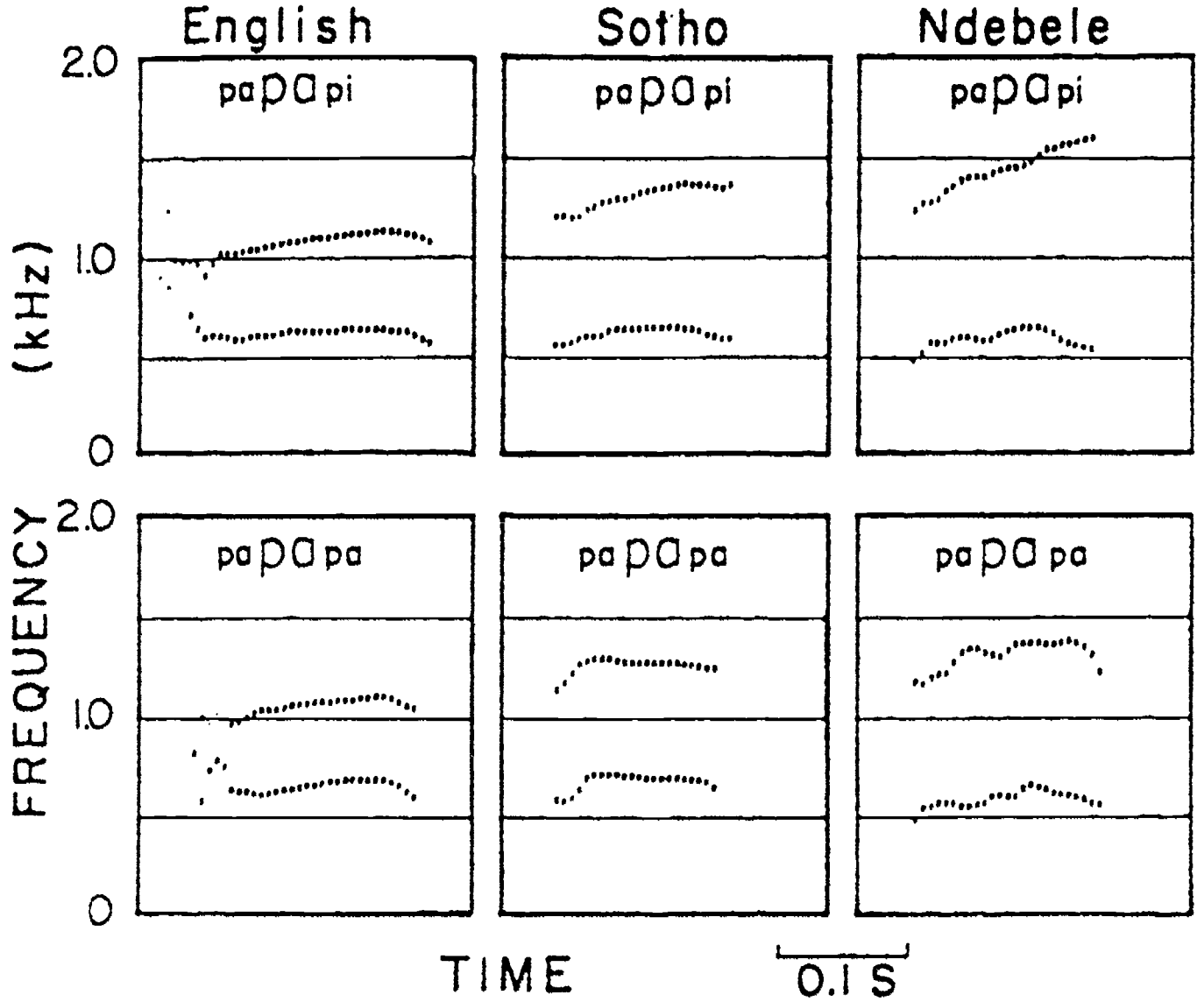


Figure 9. First and second formants for the middle vowel /ɑ/ in a single token of /pɑpɑpɑ/ and /pɑpɑpi/ from a speaker of each of the languages indicated. The rising F2, in anticipation of following /i/, is particularly striking for the Ndebele speaker. The Ndebele and Sotho tokens shown here are part of the present general analysis. The English token was recorded specifically for this illustration.

B. Other factors affecting output constraints

Are output constraints determined solely by the distribution of contrastive phones in a language? It would seem not. In the present study there appear to be differences between the 2 five-vowel languages in amount of coarticulation. Ndebele(LC) seems to have more front/back coarticulation on /a/ than does Shona(LC), and this is not obviously predicted by anything else we know about the languages. One way of thinking about the various contributions to output constraints is that minimally phonemes must have some audible, distinctive output, and that languages are free to restrict the output of particular phonemes even further. Ladefoged (1983) has pointed out that languages may choose to be more or less particular about how they make their phonemes. Our claim here, however, is that there are likely to be general constraints on the *lack* of fastidiousness.

There may be a number of other factors that contribute to acceptability, and the role of contrast may be to set maximal laxity for output constraints. One such factor is probably formal vs. informal styles of speech. In formal situations, people are generally more careful in their speech. But what is "careful" if not a more precise, narrowed range of production for particular phonemes? In situations where communication is difficult (e.g., noisy conditions, talking to children, nonnative speakers, or the hearing-impaired) people tend to "overarticulate" (Lindblom & Engstrand, 1989; Lindblom & MacNeilage, 1986). Overarticulation can be partially modeled as a slowing of articulation, but in all probability also involves a compression of the range of acceptable productions of phonemes, in terms of their spatial characteristics (Picheny, Durlach, & Braida, 1985; 1986). The very fact that people can vary their range of productions at will, and in formal or careful speech conditions tend to narrow those ranges toward clear exemplars of the phonemes, implies that at some level speakers have an awareness of the notion "best production" and the range of acceptable productions.

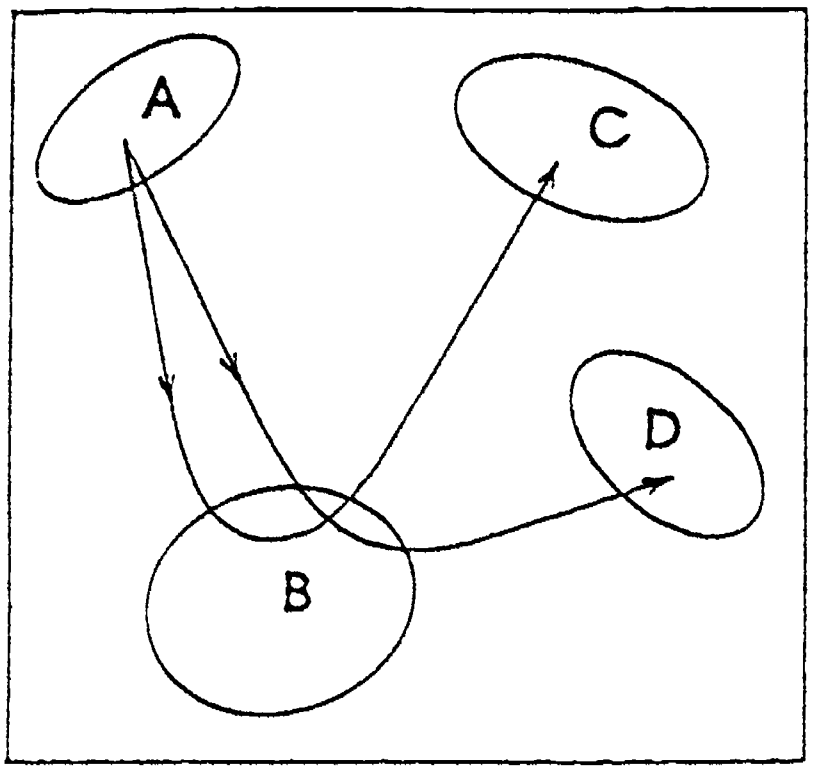
Are output constraints ever violated? Obviously they are, since phonological processes such as nasal place assimilation (e.g., ***input*** > [ɪmpʊt]) are common and neutralize the underlying difference between contrastive phonemes. However common these neutralizations are, though, it is equally obvious that they are not so rampant in any one language so as to lead to widespread or wholesale loss of phonemic contrast. Of course, another way of looking at these phenomena is to say that the output constraints themselves have been relaxed, not that the output constraints have been violated. It is not clear what, if any, data would distinguish between these two hypotheses. In any case, as pointed out by Martinet (1952, pp. 129), there are a variety of forces at play in speech, and sometimes the forces which push or pull one phoneme into the territory of another "may simply be more powerful than the functional factors working toward conservation."

C. Incorporating output constraints into models of coarticulation

The concept of output constraints, and the role of contrast in setting output constraints, can potentially be of value in trying to understand particular instances of coarticulatory behavior, and can be used to try to restrict models of coarticulation. In this section we briefly explore how this concept might be incorporated into more formal models of coarticulation. Note that while the following discussion is limited to only a few types of models, the role of output constraints is (or should be) a concern for any coarticulation model.

We have suggested elsewhere (Manuel, 1987a; 1987b) that output constraints could be thought of as target spaces in some appropriate dimension(s) and that speakers generally try to move smoothly from one space to the next, crucially always trying to maximize efficiency (pick the easiest overall route, according to some measure of "ease" of articulation). The idea was that the movement from target to target is affected by the narrowness of the target spaces themselves: extremely narrow targets don't allow for much variability in the movement from one target to the next, whereas large targets allow various trajectories through a given space. This concept is shown schematically in Figure 10, which demonstrates the effect of the narrowness of the medial target on a non-speech "connect the circles" drawing task. The task is to start in Circle A, move through Circle B, and end up in Circle C or D. When Circle B is quite small (analogous to a very restrictive output constraint), the trajectory from A to B is rather insensitive to whether the following target is C or D. In contrast, when Circle B is large, the trajectory from A to B is highly affected by the location of the following target, that is, whether it is C or D. A similar proposal, suggesting interpolation between and through targets of various sizes, has been made by Keating (1990). Working within a connectionist

framework, Jordan (1990) has recently developed more explicit models for comparing the effect of loose output constraints (*don't care* conditions on outputs) with strict output constraints (*strongly care* conditions on outputs). Jordan's model learns a path through several target specifications, and Jordan demonstrates that varying these constraints affects just which path his model learns.



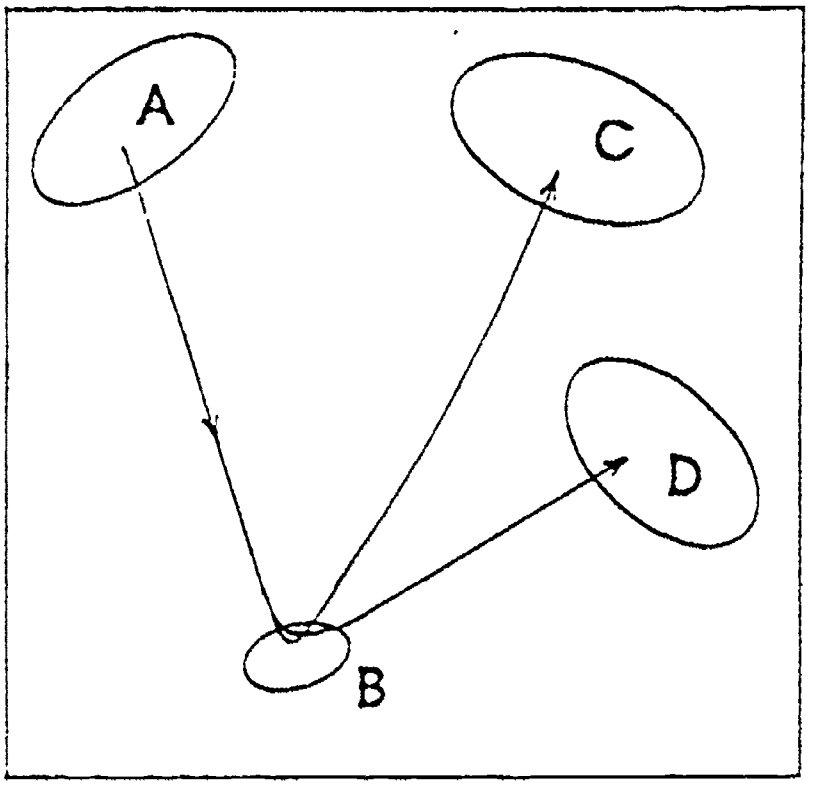


Figure 10. When Target B is large (upper panel), the trajectory from A to B is affected by the location of the next Target (C vs. D). In contrast, when Target B is small (lower panel), the trajectory from A to B is more restricted, and is minimally influenced by the location of the next target.

The above accounts assume, at least implicitly, that given a starting point A and two following goals B and C, the actor/speaker somehow calculates an overall path through the two upcoming goals. But it may be that, for motor behaviors in general, the actual surface path is the result of combining an underlying invariant command to move from the starting point A to Goal B, and an underlying invariant command to move from Goal B to Goal C.

The idea that surface paths may reflect simultaneous input from distinct, invariant commands to move to different targets has been applied to speech, most notably in the *coproduction* models of, for example, Bell-Berti and Harris (1982), Browman and Goldstein (1990), and Fowler (1981; 1986). In coproduction models, coarticulation is achieved by allowing a particular articulator (or articulatory system) to respond simultaneously to invariant commands associated with adjacent or neighboring phones. For example, the anterior /k/ closure in /aki/ (vs. the more posterior closure for /aka/) can be described by assuming that while the tongue dorsum is still responding to a command (associated with the /k/) to make a velar closure, it also begins responding to commands, associated with the goal of making the following /i/, to move anteriorly. The actual movement of the tongue dorsum reflects both of these inputs. While the basic concept of simultaneous response to two phones is not new, it has recently been explicitly modeled (e.g., Saltzman, Rubin, Goldstein, & Browman, 1987).

In coproduction models, varying degrees of coarticulatory effects, such as those seen in different languages, can be achieved in two ways. First, the amount of temporal overlap of the commands can be increased. For example, given a sequence of phones XY, Y will have a greater effect on the production of X the more the commands for X and Y overlap. Second, the particular weights used to combine the various commands can be varied. Given simultaneous input from X and Y, the inputs could average, they could sum, or they could combine in some other linear or nonlinear fashion such that Y has a greater or smaller affect on the production of X (see Boyce, 1988: Munhall & Löfqvist, 1987; Saltzman, Goldstein, Browman, & Rubin, 1988; Saltzman & Munhall, 1989). Whatever the combinatorial algorithm is and however it is determined, once it *is* determined, coarticulation presumably proceeds in a mechanical way.

An important issue for development of such coproduction models is whether or not there are any *general* principles which determine the combinatorial algorithms for overlapping gesture-commands. In this paper we have argued that the goals, or output constraints, of one phone essentially limit interference from another phone. To the extent that this is true, the algorithms combining commands for a sequence of phones would not be determined in an *ad hoc* random fashion. Rather, we expect that the combinatorial algorithms would be such that, in a sequence of phones XY, the commands for X have the effect of *suppressing* the commands for Y. (For a discussion of suppression, see Saltzman et al., 1988; Saltzman & Munhall, 1989.) In some cases the output constraints on X will be so weak that commands for Y will not be suppressed very much, but in other cases the output constraints on X will lead to extreme suppression of the commands for Y. For example, in /aki/ there is a relatively lax output constraint on exactly where contact is made on the roof of the mouth for the /k/, and this constraint does not heavily suppress the forward movement for the following /i/ vowel. At the same time, since /k/ is a stop consonant, there is a strict constraint on the degree of oral tract constriction, and this constraint suppresses any contrary gestures from the following /i/ which would preclude making a complete oral closure.

The present results for Ndebele(LC), Shona(LC), and Sotho(MC) show that coarticulation can be both temporally and spatially very extensive. From the point of view of coproduction models, if anticipatory coarticulatory effects are found early in the first vowel of a V_1CV_2 utterance, we can conclude that the commands for the second vowel begin (at least) that early. In addition, since the first vowel clearly dominates production at least until the consonant is approached, those commands associated with the second vowel must have a *lesser* effect on the surface trajectory in that time period than do the commands associated with the first vowel itself. The fact that V_2 has a greater or lesser effect on V_1 in different languages could, in a mechanical sense, be modeled as being due to different tolerances or output constraints for particular phonetic gestures. These tolerances, which are predictable (at least in part) from certain general principles, lead to language-dependent amounts of suppression of the V_2 gestures in different languages. That is, the determination of the combinatorial algorithm for neighboring phones is affected by requirements for the maintenance of intelligible speech, i.e., maintaining distinctions between phones. Those requirements vary from one

speaking style or condition to the next, and from language to language, partially because languages have different systems of contrast. Contrast affects output constraints, and output constraints determine how gestures associated with neighboring phonetic units are combined.

SUMMARY

The data presented here show that the vowel /ɑ/ is more susceptible to anticipatory coarticulation with a following transconsonantal vowel in Shona(LC) and Ndebele(LC), which have no near phonemic neighbors to /ɑ/, than in Sotho(MC), which does have relatively near neighboring and contrasting phonemic vowels. This result is consistent with the idea that coarticulation is limited by output constraints on phones, and that these output constraints are determined, in part, by the need to maintain phonological distinctions in a language.

REFERENCES

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

Boyce, S. E. (1988). *The influence of phonological structure on articulatory organization in Turkish and in English: Vowel harmony and coarticulation.* Doctoral dissertation, Yale University, New Haven, CT.

Browman, C. P., & Goldstein, L. (1990). 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 Associates

Chasaide, A. (1979). Laterals of Gaoth-Dobhair Irish and Hiberno-English. *Occasional Papers in Linguistics and Language Learning, 6*, pp. 55-78. The New University of Ulster.

Cole, D. T. (1955). *An introduction to Tswana grammar*. London: Longmans, Green and Co.

Doke, C. M. (1954). *The Southern Bantu languages*. London: Oxford University Press.

Engstrand, 0. (1988). Articulatory correlates of stress and speaking rate in Swedish VCV utterances. *Journal of the Acoustical Society of America, 83*, 1863-1875.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Flege, J. E. (1989). Differences in inventory size affect the location but not the precision of tongue positioning in vowel production. *Language and Speech, 32*,123-147.

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research, 46*, 127-139.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*, 3-28.

Jongman, A., Blumstein, S. E., & Lahiri, A. (1985). Acoustic properties for dental and alveolar stops. *Journal of Phonetics, 13*, 235-251.

Jordan, M. I. (1990). Serial order: A parallel distributed processing approach. In J. L. Elman & D. E. Rumelhart (Eds.), *Advances in connectionist theory: Speech*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Keating, P. A. (1990). The window model of coarticulation: articulatory evidence. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonetics I: Between the grammar and the physics of speech* (pp. 451-470). Cambridge: Cambridge U.P.

Keating, P. A., & Huffman, M. (1984). Vowel variation in Japanese. *Phonetica, 41*, 191-207.

Ladefoged, P. (1983). The limits of biological explanation in phonetics. *UCLA Working Papers in Phonetics, 57*, 1-10.

Lindblom, B. (1983). Economy of speech gestures. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 217-245). New York: Springer-Verlag.

Lindblom, B., & Engstrand, 0. (1989). In what sense is speech quantal? *Journal of Phonetics, 17*, 107-121.

Lindblom, B., & MacNeilage, P. (1986). Action theory: Problems and alternative approaches. *Journal of Phonetics, 14*, 117-132.

Lubker, J. F., & Gay, T. (1982). Anticipatory labial coarticulation: Experimental, biological and linguistic variables. *Journal of the Acoustical Society of America, 71*, 437-448.

Magen, H. (1984). Vowel-to-vowel coarticulation in English and Japanese. *Journal of the Acoustical Society of America*, Suppl. 1, *75*, S41.

Magen, H. S. (1989). *An acoustic study of vowel-to-vowel coarticulation in English*. Unpublished doctoral dissertation, Yale University, New Haven, CT.

Manuel, S. Y. (1987a). *Acoustic and perceptual consequences of vowel-to-vowel coarticulation in three Bantu languages*. Unpublished doctoral dissertation, Yale University, New Haven, CT.

Manuel, S. Y. (1987b). Output constraints and cross-language differences in coarticulation. *Journal of the Acoustical Society of America, Suppl 1*, 82, S115.

Manuel, S. Y., & Krakow, R. A., (1984). Universal and language particular aspects of vowel-to-vowel coarticulation. *Haskins Laboratories Status Report on Speech Research, SR-77/78*, 69-78.

Martinet, A (1952). Function, structure, and sound change. *Word, 8*, 1-32. Reprinted in P. Baldi & R Werth (Eds.), *Readings in historical phonology: Chapters in the theory of sound change* (pp. 121-159). University Park, PA: Pennsylvania State University Press, 1978.

Martinet, A. (1957). Phonetics and linguistic evolution. In B. Malberg (Ed.), *Manual of phonetics* (pp. 252-272). North Holland, Amsterdam.

Munhall, K. G., & Löfqvist, A. (1987). Gestural aggregation in speech. *PAW Review, 2*, 13-15. (University of Connecticut Press).

Nartey, J. N. A. (1979). A study of phonemic universals, especially concerning fricatives and stops. *UCLA Working Papers in Phonetics, 46*.

Nolan, F. (1985). Idiosyncrasy in coarticulatory strategies. *Cambridge Papers in Phonetics and Experimental Linguistics 4*, 1-9.

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

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985) Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research, 28*, 96-103.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research, 29*, 434-446.

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

Saltzman, E. L., Goldstein, L., Browman, C. P., & Rubin, P. (1988). Dynamics of gestural blending during speech production. *Neural Networks, 1*, 316.

Saltzman, E. L, Rubin, P., Goldstein, L., & Browman, C. P. (1987). Task-dynamic modelling of interarticulator coordination. *Journal of the Acoustical Society of America, Suppl 1*, 83, S15.

Schouten, M. E. H., & Pols, L. C. W. (1979). Vowel segments in consonantal contexts: A spectral study of coarticulation—Part 1. *Journal of Phonetics*, 7, 1-23.

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

Tatham, M. A. A. (1984). Towards a cognitive phonetics. *Journal of Phonetics*, 12, 37-47.

FOOTNOTES

Journal of the Acoustical Society of America, 88(3), 1286-1298 (1990).

†Now at Wayne State University, Department of Communication Disorders and Sciences, Detroit, MI.

[1]As is well known, in a word like *can't* there may be no nasal murmur in the signal. That is, there may be no period of time where there is simultaneously (1) a wide velopharyngeal opening, (2) oral closure, and (3) substantial vocal fold vibration. In the absence of a nasal murmur, the nasal quality of the vowel itself can distinguish *cat* from *can't*.

[2]It may well be that the range of normal token-to-token variability (vs. coarticulation induced variability) does not put sufficient pressure on output constraints to test the hypothesis that output constraints vary as a function of phoneme distribution. For vowels spoken in a single context, Flege (1989) found that speakers of Spanish (a language with vowels which are well-spread apart in the acoustic and articulatory space) do not show more token-to-token variability for vowel height than do speakers of English, a language with a more crowded vowel space.

[3]Ndebele and Sotho contrast voiceless ejective [p'] (orthographic *p*) with voiceless aspirated [p^h] (orthographic *ph*). Consistent with the orthographic traditions of the three languages, the stimulus lists for these two languages used *ph*, while those for Shona used *p*, which also represents an aspirated [p^h].

[4]In Manuel (1987a) we report data from an additional measurement point, made at 20 ms before the end of the target vowel. While there are differences in the values obtained at this point and the measurement point made at the end of the vowel, the conclusions reached are not affected by the omission of those data.

[5]At points very close to the end of the vowel, where there is a rapid change in frequency, and amplitude is dropping, the LPC technique sometimes picks spurious values for the formants. This was particularly true for F1, which tended to be "lost" earlier than F2. Occasionally the end point was actually made as much as 20 ms, though more often it was 5 or 10 ms, from closure. In addition, for some speakers F1 was hard to determine even in the middle of the vowel, particularly for the vowel /ɑ/, in which F1 had a wide bandwidth, perhaps due to tracheal coupling. Note that the /p/ in these utterances was often heavily aspirated.

Haskins Laboratories Status Report on Speech Research
1990, SR-103/104, 21-38

Anticipatory Velar Lowering: A Coproduction Account*

Fredericka Bell-Berti† and Rena Arens Krakow††

Feature Spreading and Coproduction Models make fundamentally different assumptions about the nature and organization of speech motor control, and yet each model is supported by some, but not all, of the existing empirical data. This has led some researchers to conclude that speakers probably use alternative strategies at different times. This study suggests that the identification of coarticulatory influences requires the concurrent identification of intrinsic articulatory characteristics of the segment. Moreover, the evidence for feature spreading or variable coarticulation strategies derives from t'.e misidentification of such intrinsic characteristics as context effects. This velar coarticulation study used a controlled comparison between CV_nN and CV_nC minimal pairs. Vocalic string duration was manipulated by varying the number of segments and speech rate, allowing us to alter the time between the onsets of vocalic and subsequent consonantal gestures. Velar lowering occurred in CV_n sequences, whether or not a nasal consonant followed, and similar vocalic gestures were observed across minimally contrastive environments with and without the nasal consonant. Moreover, velar lowering for the nasal consonant began in close temporal proximity to the nasal murmur. These results strongly support the coproduction model and provide insight into previously conflicting reports.

INTRODUCTION

In early studies of speech production, coarticulation was viewed as problematic because researchers' intuitions about the speech stream (the notion of "segments") did not match the acoustic or articulatory output that the new technology revealed (Harris, 1970; Kent & Minifie, 1977; Liberman, Cooper, Harris, & MacNeilage, 1962). To continue to maintain the notion that speech is, at some level, segmentally structured, required an account of speech production that takes a string of discrete segments as its input and outputs a stream of overlapping and asynchronous gestures. Over the years, two such general types of explanations of observed coarticulatory phenomena have been offered: "feature spreading" models (e.g., Henke, 1966; Joos, 1948; Kozhevnikov & Chistovich, 1965; Moll & Daniloff, 1971) and "coproduction" models (e.g., Bell-Berti & Harris, 1981; Browman & Goldstein, 1986; Fowler, 1980; Öhman, 1966).

Studies of anticipatory nasal coarticulation (including data on velar lowering movements, nasal airflow, velopharyngeal port opening, or levator palatini relaxation) alor[illegible] with studies of labial coarticulation (including data on upper lip protrusion, lower lip protrusion, or orbicularis oris contraction) hav[illegible] been widely cited in studies proposing the different coarticulation models (e.g., Bell-Berti & Harris, 1982; Benguerel & Cowan, 1974; Daniloff & Moll, 1968; Kent, Carney, & Severeid, 1974; Moll & Daniloff, 1971). The present study re-examines the theoretical and empirical bases for the two models with a focus on nasal coarticulation (We refer the reader to similar discussions of lip rounding in Boyce, 1988; Boyce, Krakow, Bell-Berti, & Gelfer, 1990; Gelfer, Bell-Berti & Harris, 1989). We begin by reviewing the models themselves.

In "feature spreading" models, an articulatory planning (or "look-ahead") component determines what movements are required for upcoming segments and initiates them as soon as their onset would not interfere with the more immediate

This work was supported by NIH grant DC-00121 to the Haskins Laboratories. We wish to thank Katherine Harris, Ignatius Mattingly, Leigh Lisker, Mary Boyle, and Carol Fowler for their helpful comments on an earlier version of this manuscript.

articulatory requirements (Figure 1a). In these theories, the temporal extent of such anticipatory adjustments is limited only by characteristics of other segments (e.g., Henke, 1966), or of the syllabic (e.g., Kozhevnikov & Chistovich, 1965) or syntactic structure of the utterance (e.g., McClean, 1973). In the absence of such context-conditioned limitations, however, these theories claim that coarticulation may be unlimited in extent. Thus, for example, velar lowering for a nasal consonant is predicted to spread back to the first vowel in a CV_nN sequence,[1] but not to go beyond the first vowel because of the conflicting high velum specification of the initial oral consonant (Figure 1a).

In contrast, "coproduction" models claim that coarticulation is a function of the overlap of gestures whose onsets bear a stable relation to other aspects of the articulation of a given segment (Figure 1b). For example, Bell-Berti (1980) has proposed that the onset of velar lowering for a nasal consonant (or velar raising for an obstruent consonant) bears a stable temporal relation to the achievement of the oral tract constriction that is characterized in units of time, not numbers of segments. Furthermore, empirical evidence, from studies of the velum and from the lips, in support of these models suggests that gestural onsets are of limited temporal extent (e.g., Bell-Berti, 1980; Bell-Berti & Harris, 1982; Gelfer et al., 1989). In this type of model, there is no "look-ahead" component that alters the plan for a given segment as a function of context. That is, in coproduction models coarticulation does not affect the nature of the individual gestures; rather, it is manifested in the way in which co-occurring and successive stable gestures combine[2] (Bell-Berti & Harris, 1981; Munhall & Löfqvist, submitted; Saltzman & Munhall, 1989).

Clearly, the two types of models, coproduction and feature spreading, make fundamentally different assumptions about the nature and organization of speech motor control. However, each model appears to account for some, but not all, of the data. Thus, we are confronted with a new problem of coarticulation. That is, we are left either to accept the unparsimonious explanation that production strategies may vary according to articulator, speaker, and token, or to consider the possibility that some of the data have been interpreted inappropriately. We will, in fact, argue that one of the major failures within studies of coarticulation, on the part of both feature spreading and coproduction theorists, has been the assumption that one can identify coarticulatory influences of a segment's context without first identifying intrinsic articulatory characteristics of the segment itself. We now turn to our explanation of this notion with respect to nasal coarticulation.

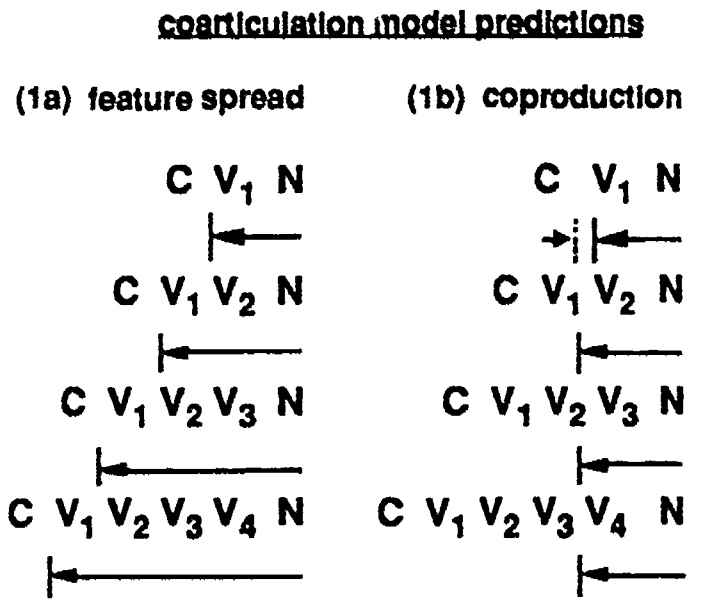


Figure 1. **Predictions of feature-spread and coproduction models.** ***(1a)*** **Feature-spread models predict that velar lowering in anticipation of a nasal consonant extends to the beginning of the vocalic sequence preceding the nasal consonant, regardless of vocalic string duration or number of segments.** ***(1b)*** **Coproduction models predict that velar lowering during a vocalic sequence preceding a nasal consonant begins at a stable time before the nasal consonant, regardless of vocalic string duration.**

In spite of some evidence to the contrary, a general assumption in many studies of anticipatory velar lowering has been that the articulatory plan calls for a uniformly low velar position for nasal consonants, a uniformly high velar position for oral consonants, and a uniformly neutral position for vowels. According to this view, for anticipatory coarticulation to occur, the neutral specification of the vowel must be replaced, through feature copying, with the next specified velar position. Data from a number of studies showing that the velum begins to lower at the consonant release in CV_nN strings have been taken to support feature spreading models (e.g., Kent et al., 1974; Moll & Daniloff, 1971; cf. Ohala, 1971).

In contrast, Bell-Berti reported data explicitly calling into question the uniformity, with respect to velar height, of each of the three categories of sounds that have been described, nasal consonants, oral consonants, and vowels. In one study she showed variable velar positions for different vowels as a function of vowel height, as well as effects of vowel height on velar position for adjacent oral or nasal consonants (Bell-Berti,

Baer, Harris, & Niimi, 1979). In a subsequent study, Bell-Berti (1980) showed that, within a given vowel environment, velar raising continued through long oral consonant sequences (of 400 ms or so), achieving the highest positions for the longest sequences. These data suggest an additive effect from the overlapping contributions of intrinsic velar positions for adjacent oral consonants. The results also showed that the effect of the following vowel on velar position during the consonant sequence extended back only about 250 ms; that is, the vowel affected velar position during the pre-consonantal vowel when the consonant sequence was very short, but had no effect on the pre-consonantal vowel or the early part of the consonant sequence when the consonant sequence was very long. This suggests a limited temporal window for coarticulatory effects of velar position, and finds support in the work of Ushijima and Hirose (1974) on Japanese and Benguerel, Hirose, Sawashima, and Ushijima (1977) on French. Furthermore, whereas Bell-Berti's results are incompatible with feature-spreading models, they are predicted by coproduction accounts of speech production.

In an attempt to sort out the two general types of models, Bladon and Al-Bamerni (1982) concluded that both means of coarticulation are employed by individual speakers. Thus, they reported that, in some instances for each speaker, velar lowering occurred at variable times before a nasal consonant correlating with the duration of the vocalic string. In other instances, velar lowering gestures showed a "two-stage" movement, with the first stage occurring well in advance of the nasal consonant, in accord with feature spreading models, and the second stage occurring as though time-locked to the onset of the closure for the nasal consonant, in accord with coproduction models. Based on these results, Bladon and Al-Bamerni suggest a model of speech production that allows variable production strategies. Similar results in studies of lip protrusion have led Perkell to the same conclusion (Perkell, 1986; see also Perkell & Chiang, 1986). Neither Perkell nor Bladon and Al-Bamerni proposed a specific source for a speaker's choice among variable patterns.

In the present study, we will suggest that much, if not all, of the evidence for feature spreading (whether in variable strategy accounts or in straight feature spreading accounts), may be due to the mistaken identification of an intrinsic articulatory component of a given segment for a coarticulatory effect of an adjacent or near-adjacent segment (Bell-Berti, 1980). In part, the mistake is due to the adoption of phonologically-based characterizations in some models of speech motor control. For example, with respect to velar activity and nasalization, it has seemed appropriate to assume that, for English, phonological description needs to be concerned only with the specification [+] or [-] NASAL for consonants, but that it need not be concerned with such a specification for vowels, since vowel nasalization is not distinctive. In the articulatory domain, however, we must also be concerned with documented differences in velar position for different vowel phonemes in different languages (e.g., Bell-Berti, 1980; Bell-Berti et al., 1979; Fritzell, 1969; Henderson, 1984; Moll, 1962; Passavant, 1863; Ushijima & Sawashima, 1972).[3] Moreover, velar height for oral vowels has been shown to be lower than that for oral consonants. Thus, when the coarticulation accounts describe the earliest onset of velar lowering in a CV_nN sequence as the onset of coarticulation, they may, in fact, only have identified an expected transitional movement of the CV sequence, a movement that is unrelated to the nasal consonant. To argue otherwise requires clear evidence that such movements occur only in the context of the upcoming nasal.

This study, in contrast to previous studies of velar coarticulation, included six minimally contrastive CV_n sequences followed by either an oral or a nasal consonant (i.e., CV_nC vs. CV_nN). The sequences were designed to allow us to systematically distinguish coarticulatory effects of the nasal consonant from intrinsic articulatory gestures of the adjacent segments. In addition, we adopted the strategy used in most earlier studies of lengthening the vocalic sequence by increasing the number of segments (up to three) to examine the extent of anticipatory nasal coarticulation after factoring out the intrinsic effects. We also used changes in speaking rate as an alternative way of manipulating sequence duration, also allowing us to explore the effects of suprasegmental manipulation on the extent of coarticulation. We argue here that the coproduction account receives overwhelming support when intrinsic segmental effects are distinguished from coarticulatory effects.

Methods

Subjects

The subjects were three native speakers of dialects of American English that are spoken in the Metropolitan New York area. None of the

subjects reported a history of speech or hearing disorder. Subject 1 is a co-author of the paper.

Speech samples

Table 1 lists the target sequences for this study: half of the utterances, the "nasal utterances," contained a post-vocalic nasal consonant; the other half, the "oral utterances," did not. The target words varied in the number of vocalic segments (from one to three) preceding the consonant. Each target was embedded in a carrier phrase, "It's _____ again." According to the feature spreading models, the longer the vocalic string, the earlier we can expect to see the onset of velar lowering for the nasal consonant. Effects of sequence length were examined by comparing sequences with different numbers of vocalic segments and also by comparing sequences produced at different speech rates. Given the matched sequences without nasal consonants, it was possible to determine at what point in the vocalic string velar lowering for the nasal consonant was evident as distinct from velar lowering for the vocalic string itself. Note that we are taking /l/ to be vocalic in nature in terms of velar height because of evidence that it is produced with a velar position more like that of vowels than oral consonants in English (see, for example, Kuehn, 1976; Moll & Daniloff, 1971; Ohala, 1971; Schourup, 1973).

The test words were presented to the subjects on individual index cards. Subjects were asked to produce six repetitions of each utterance in each of two test orders, for a total of 12 tokens of each of the 12 utterance types per subject. The subjects were asked to produce the items at a self-selected "conversational" speaking rate. To allow us to examine the relation between speech rate and gestural organization, Subject 3 was asked to produce an additional 12 repetitions of the list at a rapid rate. Because Subject 2's "conversational" rate was rapid, we had data from two subjects (Subjects 1 and 3) at a slower rate, and from two subjects (Subjects 2 and 3) at faster rates (Figure 2). Subjects made a small number of errors and, in a few instances, produced additional repetitions, so that there were occasionally fewer and occasionally more than 12 tokens of each utterance type. Thus, altogether we recorded and analyzed 560 phrases: 271 oral and 289 nasal phrases.

Table 1. *Experimental utterance set.*

ORAL		NASAL	
/ɑsɑl/	(asal)	/ɑnsɑl/	(ansal)
/lɑsɑl/	(lasal)	/lɑnsɑl/	(lansal)
/ʌ ɑsɑl/	(a asal)*	/ʌ ɑnsɑl/	(a ansal)*
/ʌ lɑsɑl/	(a lasal)	/ʌ lɑnsɑl/	(a lansal)
/seɪ ɑsɑl/	(say asal)	/seɪ ɑnsɑl/	(say ansal)
/seɪ lɑsɑl/	(say lasal)	/seɪ lɑnsɑl/	(say lansal)

*Although these are not standard English sequences, subjects were able to produce these sequences as requested, without using "an" as the indefinite article.

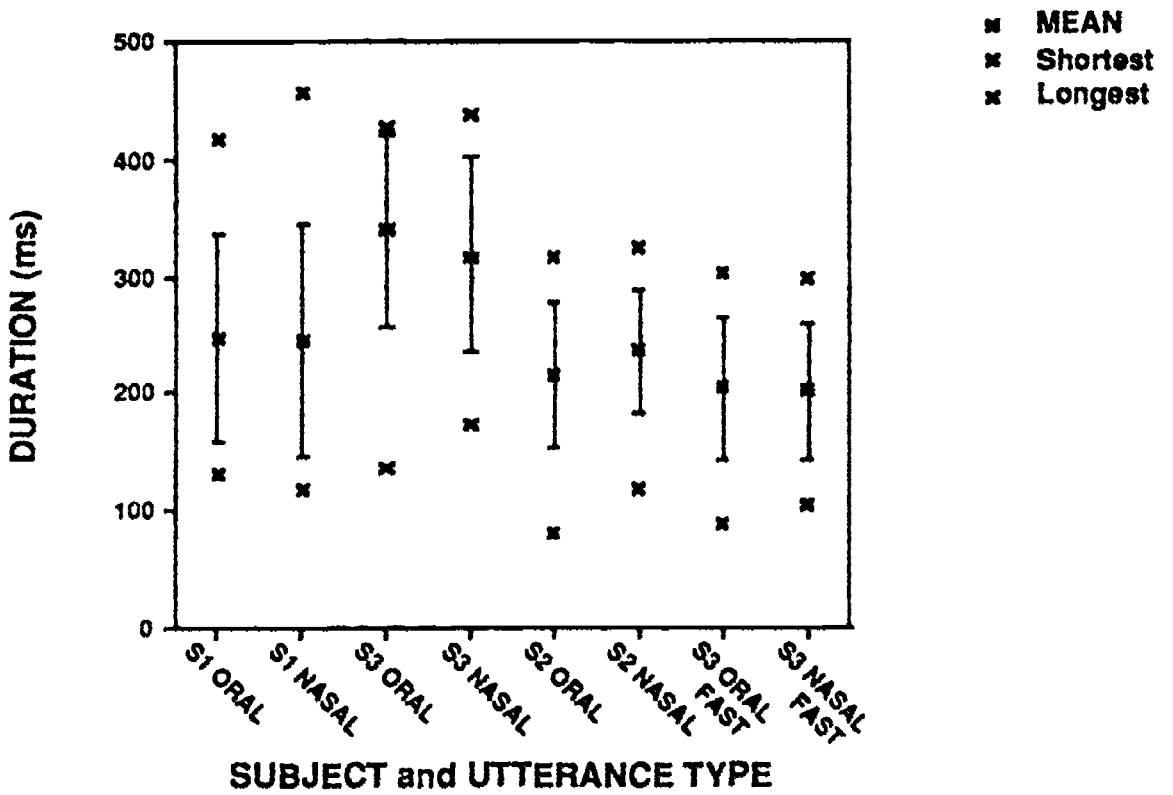


Figure 2. Mean vocalic sequence durations, standard deviations, and ranges (in ms) of oral and nasal utterances at the self-selected normal rate for all three subjects and at the fast rate for Subject 3.

Instrumentation

The Velotrace, a mechanical device developed by Horiguchi and Bell-Berti (1987) for the purpose of tracking the time-varying position of the velum, was used to monitor velar kinematics. Figure 3 provides a schematic representation of the Velotrace. It consists of three major parts: a curved internal lever whose tip rests on the nasal surface of the velum, an external lever that remains in full view outside of the nose, and a push rod (carried on a support rod) that connects the internal and external levers. Because of this connection, movements of the velum that result in changes in the angle of the internal lever with respect to its fulcrum are reflected in corresponding angular movements of the external lever. The levers are connected so that when the internal lever is raised, the external lever moves toward the subject. The Velotrace has been shown to track even very rapid movements of the velum accurately (Horiguchi & Bell-Berti, 1987).

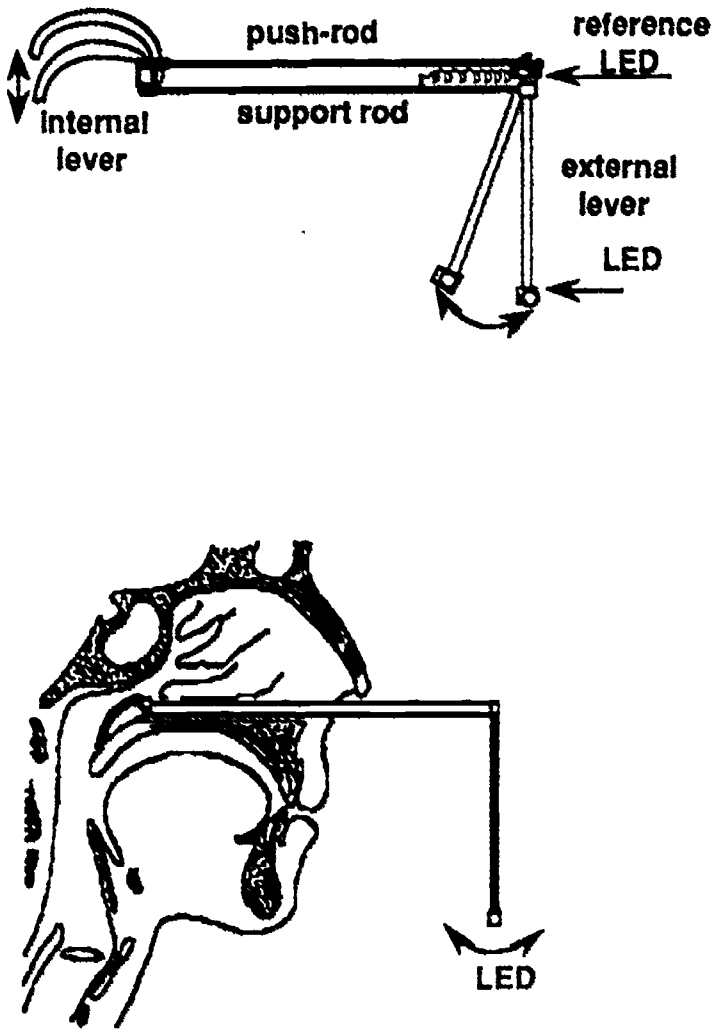


Figure 3. Schematics of the Velotrace, above, and of the device positioned in the nasal cavity.

An optoelectronic system (Kay, Munhall, V. Bateson, & Kelso, 1985) was used to track the movements of infrared diodes (LEDs) attached to the end of the external lever, and to the fulcrum of the Velotrace (for reference). The positions of the LEDs in the sagittal plane were tracked by a position-sensitive detector. The output was then converted into pairs of x and y coordinates for each LED. The acoustic speech signal was simultaneously recorded with the x and y coordinates onto a multi-channel instrumentation recorder.

Calibration of the displacement of the external lever was accomplished by moving one diode a known distance (two centimeters) in the focal plane of the optoelectronic position sensor using a precision calibration device, and then recording its output. After digital sampling, the observed change in sampled values corresponding to this two-centimeter movement was used to calibrate the recorded channels of articulator and reference signals. Note that the calibrated values of the Velotrace reflect the magnitude of movement of the external lever, rather than that of the velum itself. The external lever is about two times the length of the internal lever and so the obtained displacements are likewise larger than the actual displacements. Furthermore, it is not possible to compare the absolute magnitudes of velar gestures across subjects because, depending upon the precise positioning of the internal lever on the nasal velar surface, the magnitude of the movements of the Velotrace lever may differ.

Procedure

Each subject was seated in a dental chair with a headrest adjusted to support the head in a stable and upright position. The Velotrace was positioned after the application of a topical anesthetic and decongestants to the nasal mucosa. The fulcrum of the internal lever was positioned in the nasal cavity above the end of the hard palate, with the internal lever resting on the nasal surface of the velum and the support rod on the floor of the nasal cavity. A special headband kept the Velotrace stable. A "shot-gun" microphone was positioned in front of the subject for the purpose of obtaining a high-quality audio recording of the session, and a videotape recorder was set up to provide an audio-visual record as well. Before, during, and after the experiment, the subject was asked to produce a number of speech sounds and non-speech postures for the purpose of checking the Velotrace signal. These included: (1) sustained /m/; (2) sustained /s/; and (3) nasal breathing. The signals obtained during these test maneuvers

were monitored on-line and recorded on FM tape for further analysis. The sustained /s/ is associated with an extreme raised position of the velum, while the sustained /m/ and nasal breathing positions are associated with extreme lowered positions. These positions provided a means for determining whether the Velotrace was tracking the full excursions of the velum in the test utterances of the experiment. We found that for one of our subjects (S3), it was not possible to identify the end of the lowering gesture for some tokens because the Velotrace signal apparently "leveled off" at the nasal breathing position. Since velar lowering offset was not one of our crucial measures and only a limited number of tokens were affected, this problem did not affect the results of the study.

Analysis

In the present study, we examined the vertical movements of the velum, the traditional indicator of nasal coarticulation, because they reflect changes in velar port size beyond the point at which the port is closed. The Velotrace and reference signals were digitized at 200 Hz. Once sampled, the Velotrace reference signal was smoothed (using a 25 ms smoothing window) and then subtracted from the raw Velotrace signal in order to correct for head movement. The resulting Velotrace signal (minus the reference) was smoothed with the 25 ms window as well and its velocity was obtained by taking the first derivative of the smoothed movement signal. The velocity signal was then smoothed with the standard 25 ms window. We also examined the acoustic speech signal, which was sampled at 10,000 Hz.

We identified a specific number of events in the acoustic waveform and in the movement signal (Figure 4). For all utterances, we marked the acoustic offset of the /s/ as "s1" in "It's _____" or "It's say_____." We marked the acoustic onset of the /n/ as "n" in the target sequences with nasal segments and the acoustic onset of the medial /s/ as "s2" in the matched target sequences containing only oral segments. We determined the duration of the vocalic sequence (that is, "n-s1" for the nasal utterances and "s2-s1" for the oral utterances). We identified kinematic events using the displacement and velocity functions: Examining the displacement and velocity functions for the velum, we determined the earliest onset of velar lowering in each of the test utterances. In addition, we identified the onset of all subsequent lowering gestures (e.g., Figure 4b), since many tokens showed a pattern of multi-stage lowering (cf. Bladon & Al-Bamerni, 1982). Kinematic measures were made on both "oral" and "nasal" tokens using a procedure for determining movement onsets and offsets that included a velocity-noise criterion around the zero-crossings (see Krakow, 1989).

Results

1. Temporal extent of velar lowering and nasal coarticulation

Traditionally, the first evidence of velar lowering before a CV_nN sequence has been identified as the beginning of the gesture for the nasal consonant (e.g., McClean, 1973; Moll & Daniloff, 1971), a gesture that has been shown to occur earlier in the vocalic sequence as the duration of the vocalic sequence increases. However, before linking the beginning of downward velar movements with the production of an upcoming nasal consonant, one must show that such movements do not occur in the absence of a nasal consonant. The approach we have taken here, supported by the reports of velar-position variations across oral utterances and differences among oral consonants and among vowels, was to examine minimally contrastive utterances for the presence of velar lowering for vowels preceding both nasal and oral consonants (Figure 4).

We began our analysis by examining the nasal sequences and comparing our results with those of previous studies which had provided support for the feature spreading model. We used the procedure of those earlier studies; that is, identifying the earliest instance of velar lowering in relation to the occlusion for the nasal consonant. In agreement with those studies (e.g., McClean, 1973; Moll & Daniloff, 1971) we observed that velar lowering began earlier in advance of a nasal consonant when the preceding vocalic sequence was lengthened by the addition of segments. And, as observed in the other studies, we found a strong positive correlation (Figure 5a) between the duration of the vocalic sequence and the duration of the interval between the onset of velar lowering and the acoustic onset of the nasal consonant (i.e., the duration of the "anticipation").

We then turned to the oral utterances, to determine if there is velar lowering in vocalic sequences occurring in the absence of an upcoming nasal consonant, and if so, whether the temporal characteristics of the lowering gesture are similar to those of velar lowering gestures before nasal consonants. We took the equivalent measure that we used for our nasal utterances: we compared the time of the earliest velar lowering for vocalic sequences preceding the post-vocalic obstruent /s/

(in the oral sequences) with the durations of those vocalic sequences. This comparison also revealed a strong positive correlation between these measures, one that was not different from the correlation for the nasal sequences (Figures 5a and 5b). The individual-subject data (Table 2) also reflect the essential similarity of this measure for nasal and oral utterances.

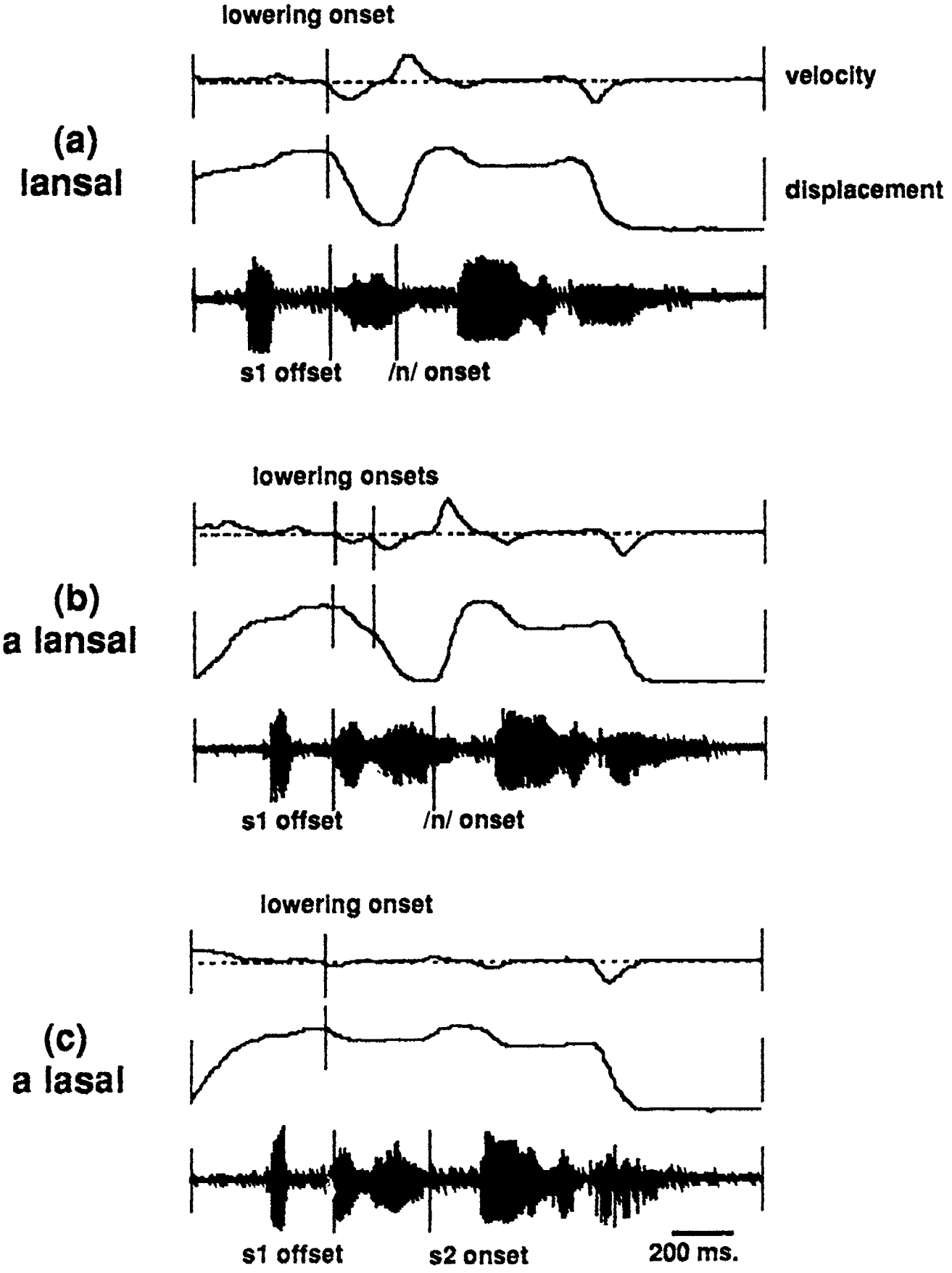


Figure 4. Representative velocity, displacement, and acoustic signals for three tokens (two nasal and one oral), with displacement and ac stic landmarks labelled.

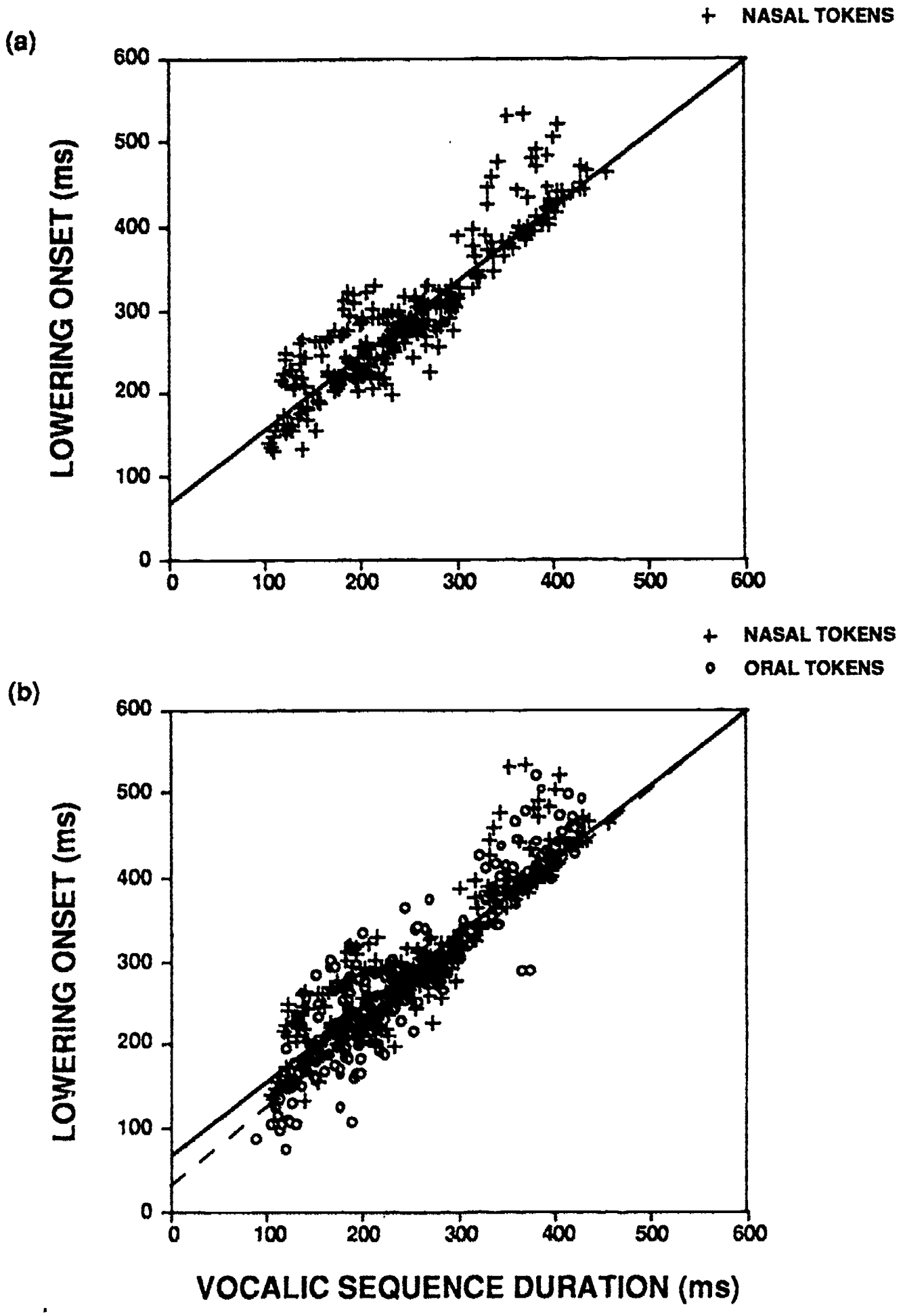


Figure 5. (a) Scatterplots of the onset of velar lowering before /n/ vs. the duration of the vocalic string, above, with data pooled across subjects (r=.923). *(b)* Scatterplots of the onset of velar lowering before /s/ vs. the duration of the vocalic string, pooled across subjects (r=.913), superimposed on the data shown in (a).

Table 2. *Correlations between onset of velar lowering and vocalic sequence duration.*

	Nasal Utterances	Oral Utterances	t=
Subject 1	.932	.916	0.6347†
Subject 2	.857	.897	0.9929†
Subject 3 (normal rate)	.991	.947	4.7313*
Subject 3 (fast rate)	.974	.979†	0.3446†

†p >.05
*p <.01

This result reveals the error of identifying the earliest velar lowering onset in a CV_nN string as anticipatory nasal coarticulation. Instead, these data suggest that at least some portion of the velar lowering movements previously attributed to feature spreading from nasal segments is, in fact, related to the articulation of the vowel string itself, since there is velar lowering in both contexts, but a nasal consonant in only one. Indeed, these results are not surprising given the cross-language evidence that the velum lowers in the transition from an oral consonant to a following vowel even in oral sequences (e.g., Bell-Berti, 1980; Bell-Berti et al., 1979; Henderson, 1984; Ushijima & Sawashima, 1972).[4]

However, we also noted something about which most earlier studies have made no comment: a consistent difference in the velar lowering patterns of our sequences as a function of their length. That is, as segments were added, increasing the duration of the vocalic sequences, multi-stage lowering movements became increasingly evident (Figure 6).

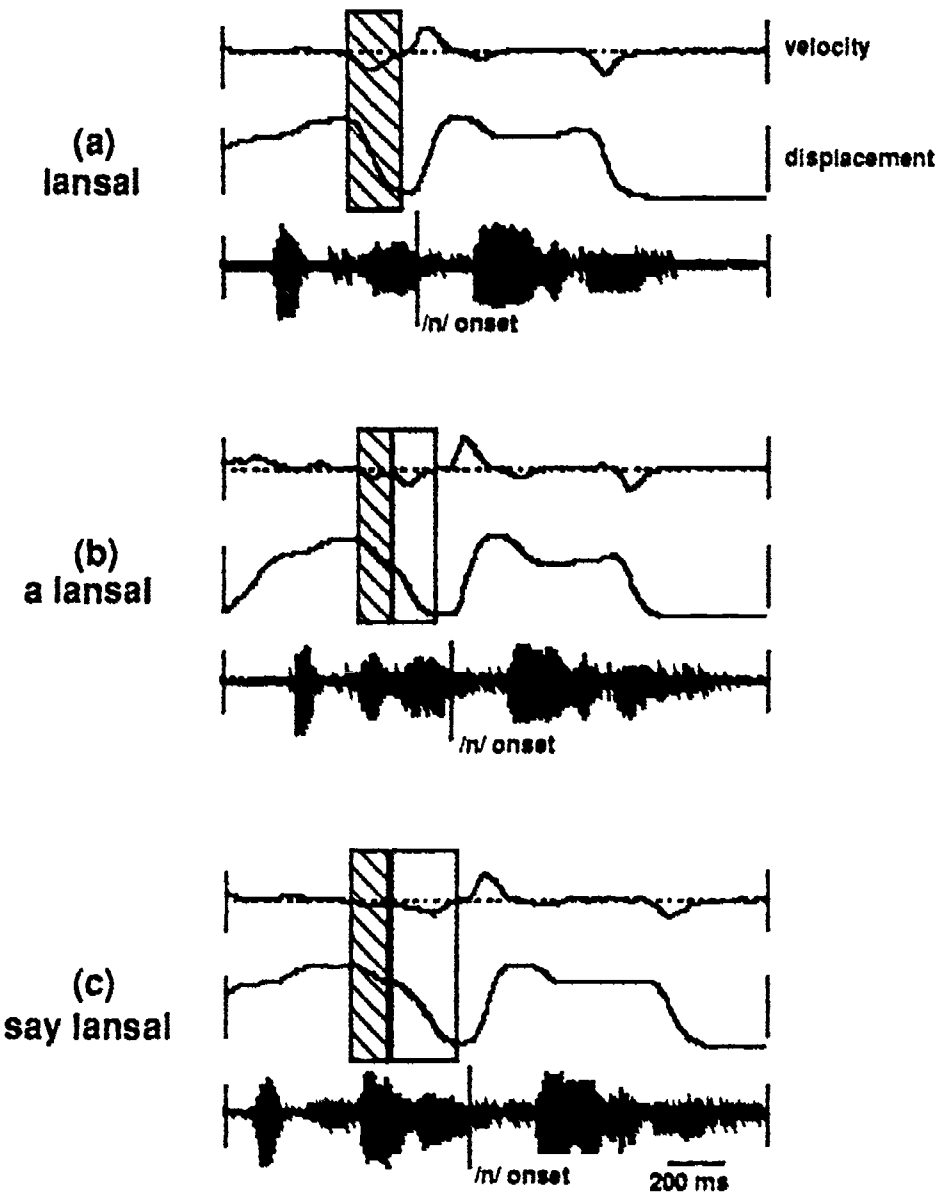


Figure 6. Representative velocity, displacement, and acoustic signals for tokens of three nasal utterances with successively longer vocalic sequence durations achieved by the addition of segments: /lansal, ʌ lansal, seɪ lansal/.

In the longer strings, we typically saw a large velar lowering movement in close temporal proximity to the nasal consonant and a shallower lowering movement earlier in the vowel string. Note that the feature spreading approach predicts a single smooth lowering gesture; it does not predict the appearance of discrete vocalic and consonantal velar gestures in sequences consisting of an oral consonant, some number of vowels, and a nasal consonant. Such patterns are, however, fundamental to the coproduction approach and the particular pattern seen here is precisely what this approach predicts. That is, separately identifiable vocalic and consonantal velar gestures are expected to be evident when they have enough time in which to appear, as in sequences of longer duration.[5] In contrast, in the shorter sequences the vocalic and consonantal gestures would be predicted to overlap in such a manner that only a single lowering movement would be evident.

2. Isolating the vocalic and consonantal gestures

In order to apportion the velar lowering movement between the vocalic and consonantal articulations, we first examined the "normal" speech rate utterances of each of our subjects in which the vocalic sequences were systematically lengthened by adding segments; we then examined the "fast" utterances of our third subject to examine the effects of changing speech rate.

2.1. Segmental manipulation. Recall that in Figure 6 we showed that velar lowering movements are affected by sequence lengthening due to segment addition: with increasing vocalic sequence duration, separate (vocalic and consonantal) gestures become increasingly evident as distinct movement components. Considering relative durations across the utterances of our speakers in the normal rate condition, we found that shorter utterances had only a single, smooth, lowering movement while longer utterances had two-stage lowering movements. The velocity functions reinforce this point. Note that in the longest of the three utterances shown (Figure 6c), the velocity function returns to zero between the two components, whereas in the mid-length utterance (Figure 6b), the velocity only approaches zero between the two components, and in the shortest utterance, it reflects only a single smooth movement.

The pattern of increased separation between vocalic and consonantal gestures with increased vocalic sequence duration was evident in the data of subjects 1 and 3; for subject 2, on the other hand, we observed two-stage patterns in only a few of the longest utterances. An explanation for our cross-subject differences is that there were differences in our subjects' speech rates, with Subjects 1 and 3 using a slower "normal" rate than Subject 2. As shown in Figure 2, the longest vocalic sequence durations in Subject 2's utterances were substantially shorter than the corresponding utterances of Subject 1 and Subject 3's "normal" speech rate. In fact, the duration range of Subject 2's utterances was most like that of Subject 3's "fast" speech rate utterances. We argue that our results show both within- and between-subject effects of sequence duration, with Subject 3 showing considerably fewer multi-stage velar lowering movements at the faster speech rate. The upper panels in Figure 7 provide examples of tokens of short nasal utterances produced by each of the three subjects: Each token reveals a single large velar lowering movement. The lower panels in Figure 7 show the movements for tokens with long vocalic sequences (lengthened by adding segments). Here, the more complex, multi-stage patterns are clearly evident, as are the between-subject differences. That is, two separate lowering components are readily observable in the longer tokens of Subjects 1 and 3, whereas for Subject 2, the component (vocalic and consonantal) gestures are not completely separate, although there is a complex lowering pattern for the longer sequence, a pattern that is quite distinct from that obtained from this subject's productions of shorter sequences. The nature of the complexity suggests to us the presence of at least two underlying components, one (or more) vocalic, and another, consonantal.

While we have established the greater likelihood of multi-stage or complex gestures occurring with sequences of longer duration, we wish to strengthen our claim that the vocalic sequence is the source of the earlier stage, and that it only becomes evident when given an adequate timeframe. To do this, we compared the velar displacement functions for oral-nasal minimal pairs of sufficient duration for both the vocalic and consonantal gestures to become evident. Figure 8 offers two examples of such data for longer sequences, in which we see two discrete movements in the nasal utterance, an early gesture of small magnitude followed by a second gesture of much greater magnitude. In the corresponding oral utterances, we would expect to see a gesture that matches that of the first stage of the complex gesture; and, indeed, we find an early gesture (of small magnitude) that is highly similar to the early movement in the longer nasal utterances.[6]

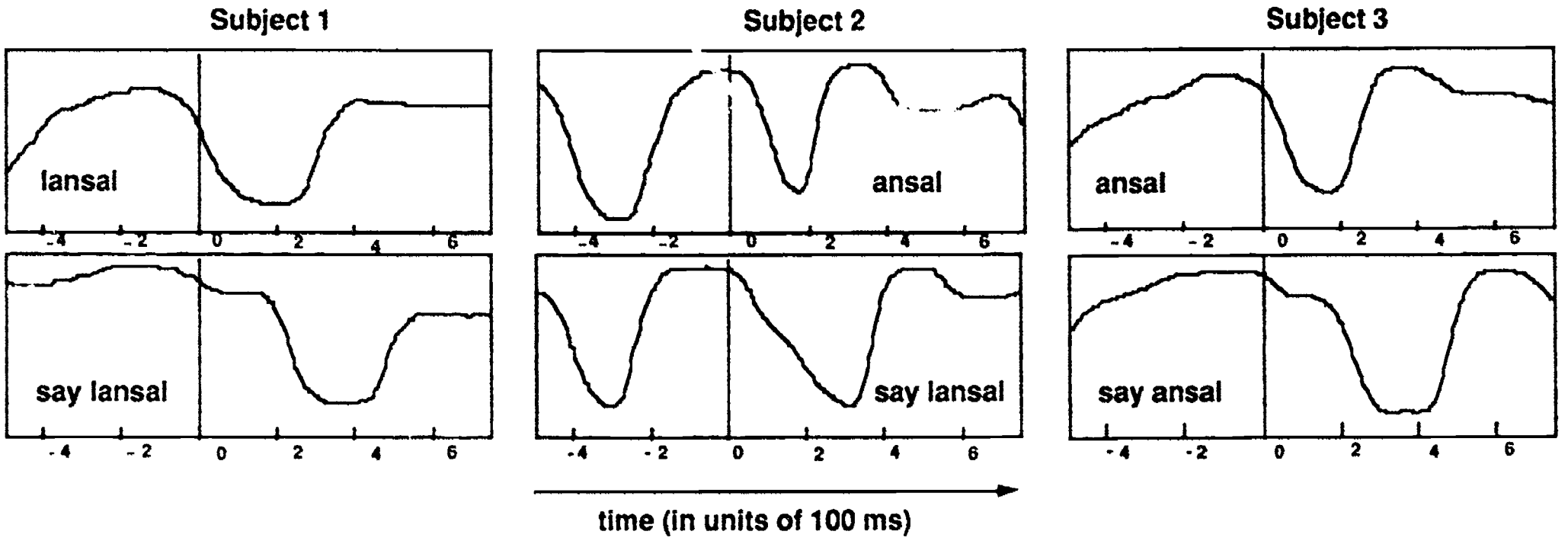


Figure 7. Displacement functions of representative short and long nasal utterances for all three subjects, showing simple and complex movements, respectively. Displacement is represented on the ordinate, with velar lowering indicated by a downward movement. Time is represented along the abscissa, with "0" marking the end of the /s/ of the carrier phrase. Because the data are displayed on the same time scale, we see the end velar lowering for the preceding utterance early in Subject 2's data, reflecting this subject's faster speaking rate.

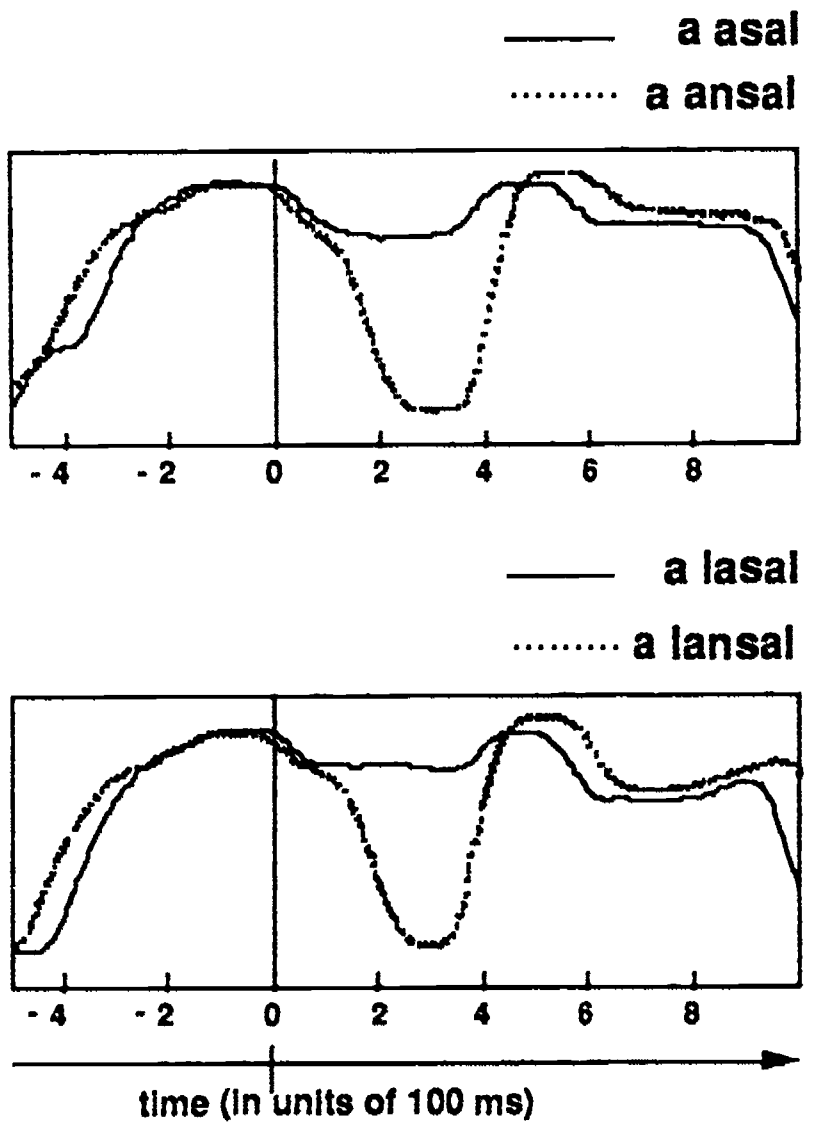


Figure 8. **Displacement functions of two representative minimally contrastive oral and nasal utterance pairs, demonstrating similar lowering onsets within each pair for Subject 3. Displacement is represented on the ordinate, with velar lowering indicated by a downward movement. Time is represented along the abscissa, with "0" marking the end of the /s/ of the carrier phrase.**

2.2. Rate manipulation. If the emergence of underlying gestures as separate on the surface is a function of reducing overlap by increasing segment duration, then the effects of speaking at a slower rate should resemble the effects of adding vocalic segments. This is not to say that the articulatory processes and strategies involved in each are the same—certainly they are not, but the two manipulations share one characteristic—that is, an increase in the duration of the vocalic string. If the predictions of the coproduction model are correct, then the result of this increase (whether due to adding vocalic segments or to speaking at a slower rate) should be a reduction in the overlap between velar lowering for the vocalic sequence and lowering for the nasal consonant. We have had an indication (by comparing the data of Subject 1 and Subject 3's self-selected "normal" rate with those of Subject 2) that speech-rate differences can affect the amount of overlap between successive gestures. To allow us to systematically examine the relation between speech rate and gestural organization for a given subject, Subject 3 was asked to produce the test utterances at a rapid speech rate in addition to the normal rate. When we explored the within-subject effects of speech rate by comparing her productions at the two different rates, we found that the slower utterances were more likely to show the multi-stage movements (Figure 9). Note that both the normal and fast rate oral utterances show a vocalic velar lowering movement, but that, in these tokens, only the normal rate nasal utterance reveals separate lowering gestures for the vowels and the nasal consonant. The faster nasal utterance is simply not long enough to allow the separate components to emerge as independent. However, even in the fast rate

utterances, once the vocalic portion reached some critical duration (roughly 250 ms), we did observe discrete movements for the oral and nasal portions of the utterance (Figure 10). Thus, we see the same effect of increased vowel-sequence duration, whether the increase was achieved by adding segments or by speaking at a slower rate: both manipulations resulted in an increased incidence of multi-stage lowering gestures. Furthermore, and as we have seen in Figure 8 for normal speaking rate utterances, the early lowering gestures in both the normal and fast nasal utterances are paralleled in the normal and fast oral utterances (cf. Figures 9 and 10).

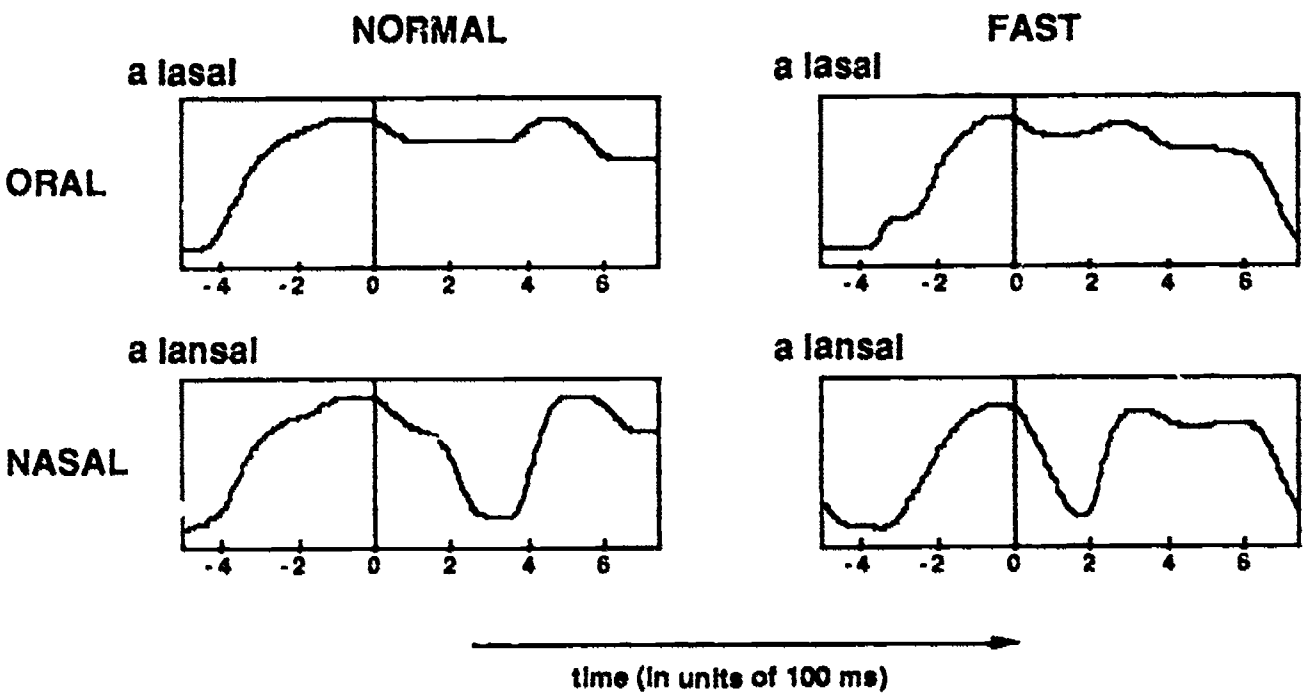


Figure 9. Displacement functions of representative tokens of a minimally contrastive utterance pair at two speaking rates. Displacement is represented on the ordinate, with velar lowering indicated by a downward movement. Time is represented along the abscissa, with "0" marking the end of the /s/ of the carrier phrase.

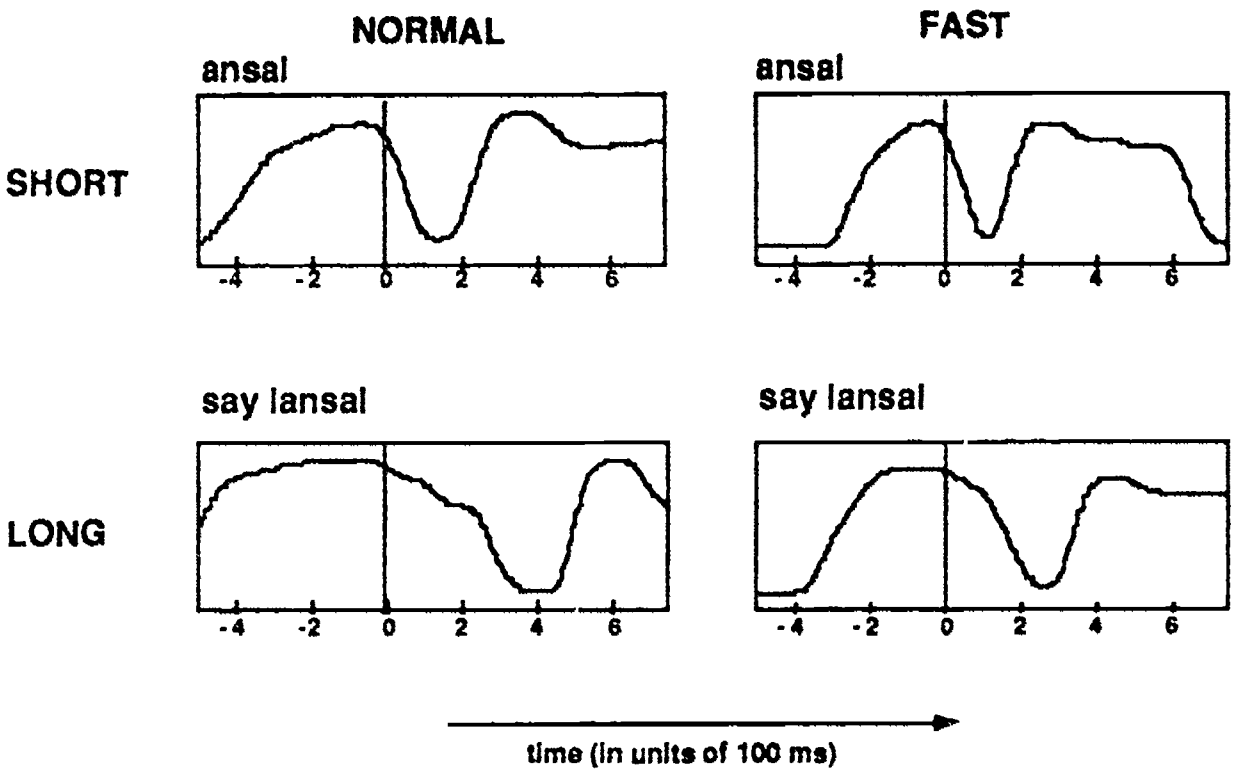


Figure 10. Displacement functions of representative tokens of short and long utterances at two speaking rates for Subject 3. Displacement is represented on the ordinate, with velar lowering indicated by a downward movement. Time is represented along the abscissa, with "0" marking the end of the /s/ of the carrier phrase.

Figure 2 shows that the same point can be made by looking at between-subject differences in rate. To show how similar these two independent variables (rate and segment number) were with respect to this outcome, Figure 11 shows the number of tokens with multi-stage lowering for Subjects 1 and 3 as a function of utterance type (i.e., number of vocalic segments) and rate (only Subject 3). With regard to Subject 2, recall that the durations of her vocalic sequences were considerably shorter than those of Subjects 1 and 3; not surprisingly, Subject 2 provided scanty evidence of multi-stage gestures, and for this reason her data are not included in Figure 11.[7] The incidence of multi-stage lowering gestures clearly increased along with the duration of the vocalic string. Least likely to show multi-stage movements were the faster (rate), shorter (number of segments) sequences and most likely, were the slower, longer sequences.

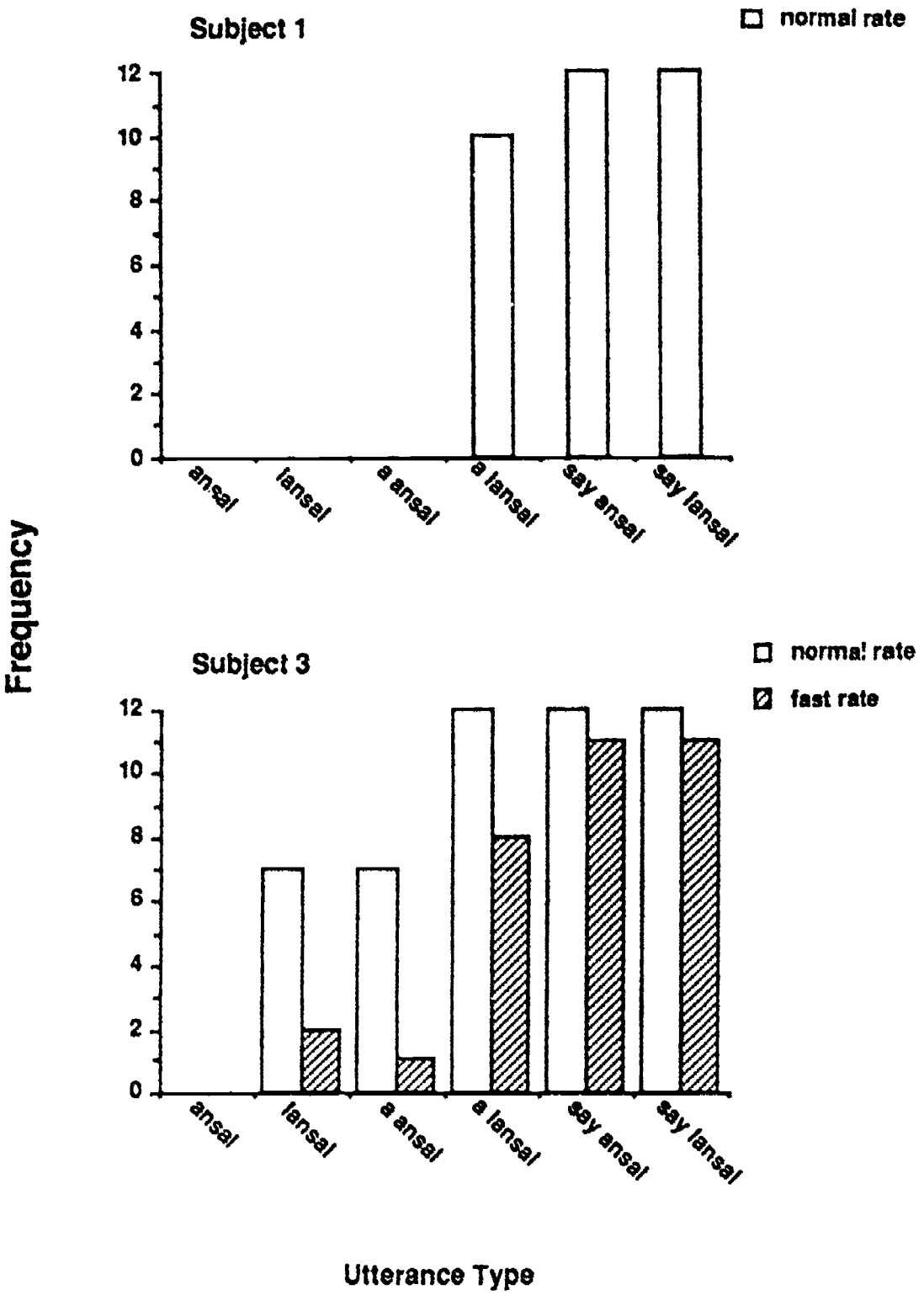


Figure 11. Histogram of frequency of multistage gestures in nasal utterances of Subject 1 and Subject 3.

To quantify the effect of vocalic sequence duration on the number of stages observed in the lowering gestures, we pooled the data for our subjects and divided the nasal utterances into two groups, those having a single stage and those having more than one stage. We also divided the utterances into two groups on the basis of vocalic sequence duration, a group of "short" utterances (those falling below the median duration of 241 ms) and a group of "long" utterances (those falling above the median duration). The resulting frequency distribution is given in Table 3.

Table 3. *Observed frequency of single and multiple gestures as a function of vocalic sequence duration (short or long).*

	short	long	totals
one	118 *(74.68%)*	40 *(25.32%)*	158
multi	27 *(20.61%)*	104 *(79.39%)*	131
totals	145	144	289

DISCUSSION

In feature spreading models, anticipatory velar lowering is predicted to extend back to the first vocalic segment in a string preceding a nasal consonant. That is, the motor plan for a nasal consonant is presumed to change with changes in the context in which it is embedded. Moreover, the effect of the nasal consonant is theoretically unbounded. In coproduction models, anticipatory velar lowering is simply viewed as an effect of the temporal overlap of the intrinsic velar gestures for adjacent vowels and consonants. That is, each segment, including the vowels, is specified as having an associated velar gesture, and the influence of a segment on its neighbors is presumed to be limited in duration.

This experiment is designed to compare the predictions of feature spreading and coproduction models systematically for sequences consisting of an oral consonant followed by some number of vowels and a nasal consonant. Using minimally contrastive sequences of the form CV_nC and CV_nN has allowed us to clarify the role of articulatory movements intrinsic to a particular segment and the coarticulatory influence of phonetic context in producing observed velar displacement patterns, and has provided strong support for the coproduction model of Bell-Berti and Harris (1981).

The results indicate that much of what has been attributed to the coarticulatory influence of a nasal consonant is actually velar lowering that would be observed in any CV_n sequence, independent of the presence of an upcoming nasal consonant. It should have come as no surprise to researchers that the velum lowers in the transition between an oral consonant and an oral vowel, since it has long been known that vowels are produced with their own intrinsic velar positions and that these positions are lower than those observed for oral consonants. With a short vocalic sequence between an oral consonant and a nasal consonant, it is difficult to separate the vocalic and nasal consonantal influences, because they appear as one continuous lowering movement. However, when a vocalic string is lengthened, either by segment addition or by speaking at a slower rate, separate vocalic and nasal consonantal movements emerge. What is more, the relatively shallow early part of the lowering movement observed in the CV_nN sequences matches that observed in minimally contrastive CV_nC sequences that end with an oral, rather than a nasal, consonant. Those ending with a nasal consonant, of course, also contain a sharp, extensive velar lowering gesture; however, this latter movement occurs at a relatively stable time in close temporal proximity to the acoustic onset of the nasal murmur and, by inference, the achievement of oral closure.

Previous research, carried out without minimally contrastive oral sequences, has led to the earliest onset of velar lowering being identified as the onset of nasal coarticulation. This approach has its origin in the phonological notion of segment underspecification and the specific hypothesis that lacking an oral/nasal contrast, English vowels are unspecified for velar position and, thus, strongly affected by their consonantal context. Underspecification is a crucial notion in phonology and also, it has been suggested, in some phonetic domains (see Keating, 1988, for a recent discussion of this issue). However, our results indicate that it is an inappropriate and misleading notion when applied to the organization of articulatory gestures, including velar movement (see Boyce, Krakow, & Bell-Berti, in press, for a discussion of the same issue concerning lip rounding). Moreover, even though phonological descriptions need only specify a binary distinction between "oral" and "nasal," the articulatory

organization requires n-ary values of velar height that also clearly have phonological implications. For example, the relation between vowel height and velar height is relevant to an understanding of why, in the languages of the world, distinctive vowel nasalization is more commonly found in low than in high or mid vowels (see Beddor, 1983).

Returning for a moment to the frequent observation that velar height for vowels is lower than that for oral consonants, we believe that this relation has explanatory power for reconciling the data of Bladon and Al-Bamerni (1982) with our own. They claimed to have found a combination of feature spreading and coproduction strategies that, we believe, is compatible with the predictions of the coproduction model alone if intrinsic velar positions for vowels and the effects of variable overlap of vocalic and consonantal gestures are included in the analysis. Their primary argument for concluding that speakers combine both feature spreading and coproduction strategies was their observation of movements in which there was 2-stage lowering: a first stage that began near the release of the initial (oral) consonant in a CV_nN string, and a second stage that was closely timed to the onset of the nasal consonant occlusion. Since the first stage appeared earlier in advance of the nasal consonant occlusion as the vowel string increased in duration, they considered this as evidence for feature spreading. The second stage was considered as evidence for time-locking (i.e., coproduction), because of its constant and relatively close temporal relation to the nasal consonant occlusion. However, this is precisely the pattern one would expect if the early, Stage 1, movement reflects lowering for the vowel and is unrelated to the upcoming nasal consonant, and the Stage 2 movement reflects lowering for the nasal consonant. This is the pattern that we observed when we lengthened the vocalic string sufficiently for the observation of discrete vocalic and nasal consonantal velar gestures.

Bladon and Al-Bamerni were also puzzled by the alternation between such 2-stage movements and single-stage lowering that appeared smooth and continuous and whose onset occurred earlier in relation to the nasal consonant as the vocalic string was lengthened. We also found this alternation; in our data, its source can be clearly seen to reside in changes in the duration of the vocalic string that led to more or less overlap between vocalic and nasal consonant gestures and, thus, more or less observed separation of the two lowering components. Unfortunately, it is not possible to reconcile our interpretation of this pattern with that of Bladon and Al-Bamerni because their report is not sufficiently detailed to make it possible for us to determine the durations of their CV_nN strings. We would obviously expect the patterns to be correlated with durational differences.

In our discussions, we have assumed that separate vocalic and nasal consonant gestures are present in the shorter or faster utterances, but that they are coincident, or, at least, overlap substantially, and are, therefore, not independently observable (Figure 10). We want to make it clear that we also believe that the vocalic lowering we have observed includes separate components for the individual vowels in the sequence, although the present experiment was not designed to separate those from one another. Nonetheless some of our slower and longer utterances showed multiple shallow gestures in the vocalic sequence (see for example, Figure 9—long normal rate utterance). The velum has often been referred to as a "slow" articulator, and Bell-Berti (1980) reported data suggesting that, all else being equal, movements for a segment begin about 250 ms before the oral articulation. Our present data support the view that there is a limited and stable timing relation between the onsets of velar movements and the corresponding oral gestures. It is, perhaps, not surprising, then, that one does not find evidence of each of the vocalic gestures in the displacement pattern, since few of them achieved individual durations of such length. Furthermore, to make such effects evident one must use segments of clearly differing intrinsic velar positions; this, in turn, requires further study, to determine, for example, the intrinsic positions for the liquid and glide segments that were used in creating our long vocalic sequences. Since we found that the number of gestures in a displacement trajectory varied with speaking rate and the length of the sequence (in segments), we suggest that the timing of velar gestures is stable across speaking rates. We therefore suppose that if very slow speech were studied, one would find more separate gestures in the displacement pattern.

CONCLUSION

To reconcile the di. ,reements about the adequacy of the two types of coarticulation models, we conducted a carefully controlled study of velar movements in CV_nC and CV_nN sequences. The results indicate that the interpretation of the earliest onset of velar lowering in CV_nN strings as coarticulation of the nasal consonant is unfounded. That is, similar lowering occurs in

strictly oral sequences. Instead, these data show that there are intrinsic velar positions for the vocalic sequence and for the upcoming nasal consonant, each of temporally limited extent. Observed patterns of coarticulation, then, are simply the result of the temporal overlap of the nasal consonant gesture with the gestures for the vocalic sequence. The portion of the vocalic sequence that is overlapped by the nasal consonant gesture increases as the duration of the vocalic sequence decreases. This study strongly supports the coproduction model, and shows how certain misinterpretations of data have led to the conflicting conclusions.

REFERENCES

Beddor, P. S. (1983). Phonological and phonetic effects if nasalization on vowel height. Doctoral dissertation, University of Minnesota, Minneapolis. (Reproduced by the Indiana University Linguistics Club.)

Bell-Berti, F. (1980). A spatial-temporal model of velopharyngeal function. In N. J. Lass (Ed), *Speech and language: Advances in basic research practice* (Vol. IV). New York: Academic Press.

Bell-Berti, F., Baer, T., Harris, K. S., Niimi, S. (1979). Coarticulatory effects of vowel quality on velar functio.. *Phonetica, 36,* 187-193.

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: Lip rounding. *Journal of the Acoustical Society of America, 71,* 449-454.

Benguerel, A.-P., & Cowan, H. A. (1974). Coarticulation of upper lip protrusion in French. *Phonetica, 30,* 41-55.

Benguerel, A.-P., Hirose, H., Sawashima, M., & Ushijima, T. (1977). Velar coarticulation in French: A fiberoptic study. *Journal of Phonetics, 5,* 149-158.

Bladon, R. A. W., & Al-Bamerni, A. (1982). One-stage and two-stage temporal patterns of velar coarticulation. *Journal of the Acoustical Society of America, 72,* S104(A).

Boyce, S. E. (1988). The influence of phonological structure on articulatory organization in Turkish and English: Vowel harmony and coarticulation. Unpublished doctoral dissertation, Yale University, New Haven.

Boyce, S. E., Krakow, R. A., & Bell-Berti, F. (in press). Phonological underspecification and speech motor organization. *Phonology.*

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

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

Clumeck, H. (1976). Patterns of soft palate movements in six languages. *Journal of Phonetics, 4,* 337-351.

Daniloff, R., & Moll, K. (1968). Coarticulation of lip rounding. *Journal of Speech and Hearing Research, 11,* 707-721.

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

Fritzell, B. (1969). The velopharyngeal muscles in speech: An electromyographic and cineradiographic study. *Acta Otolaryngologica, Suppl. 250.*

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 (L).

Harris, K. S. (1970). Physiological aspects of articulatory behavior. *Haskins Laboratories Status Report on Speech Research, 23,* 49-67.

Henderson, J. B. (1984). *Velopharyngeal function in oral and nasal vowels: A cross-language study.* Unpublished doctoral dissertation, University of Connecticut, Storrs.

Henke, W. (1966). *Dynamic articulatory model of speech production using computer simulation.* Unpublished doctoral dissertation, M. I. T., Cambridge.

Horiguchi, S., & Bell-Berti, F. (1987). The Velotrace: A device for monitoring velar position. *Cleft Palate Journal, 24,* 104-111.

Joos, M. (1948). Acoustic phonetics. *Language, 24,* 1-136.

Kay, B. A., Munhall, K. G., V.-Bateson, E., & Kelso, J. A. S. (1985). Processing movement data at Haskins: Sampling, filtering, and differentiation. *Haskins Laboratories Status Report on Speech Research, 81,* 291-303.

Keating, P. A. (1988). Underspecification in phonetics. *Phonology, 5,* 3-29.

Kent, R. D., Carney, P. J., & Severeid, L. R. (1974). Velar movement and timing: Evaluation of a model of binary control. *Journal of Speech and Hearing Research, 17,* 470-488.

Kent, R. D., & Minifie, F. D. (1977). Coarticulation in recent speech production models. *Journal of Phonetics, 5,* 115-133.

Kozhevnikov, V., & Chistovich, L. A. (1965). *Speech: Articulation and perception.* Washington, DC. Joint Publications Research Service.

Krakow, R. A. (1989). *The articulatory organization of syllables: A kinematic analysis of labial and velar gestures.* Unpublished doctora' ~ tation, Yale University, New Haven.

Kuehn, D. P. ,...'6). A cineradiographic investigation of velar movement in two normals. *Cleft Palate Journal, 13,* 88-103.

Liberman, A. M., Cooper, F. S., Harris, K. S., & MacNeilage, P. F. (1962). A motor theory of speech perception. *Proceedings of the Speech Communication Seminar.* Stockholm: Royal Institute of Technology.

McClean, M. (1973). Forward coarticulation of velar movements at marked junctural boundaries. *Journal of Speech and Hearing Research, 16,* 286-296.

Moll, K. L. (1962). Velopharyngeal closure of vowels. *Journal of Speech and Hearing Research, 5,* 30-77.

Moll, K. L., & Daniloff, R. G. (1971). Investigation of the timing of velar movements during speech. *Journal of the Acoustical Society of America, 50,* 678-684.

Munhall, K. G., & Löfqvist, A. (submitted). Gestural aggregation in speech. *Journal of Phonetics.*

Ohala, J. J. (1971). Monitoring soft palate movements in speech. *Project on Linguistic Analysis Reports, Phonology Laboratory, Department of Linguistics, University of California Berkeley,* J01-J015.

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

Passavant, G. (1863). *Ueber die Verschliessung des Schlundes beim Sprechen.* Frankfurt a.M.: J. D. Sauerlander. (Cited in Fritzell, 1969)

Perkell, J. S. (1986). Coarticulation strategies: Preliminary implications of a detailed analysis of lower lip protrusion movements. *Speech Communication, 5,* 47-68.

Perkell, J. S., & Chiang, C.-M. (1986). Preliminary support for a "hybrid model" of anticipatory coarticulation. In *Proceedings of the 12th International Congress of Acoustics,* July 1986, A3-6.

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

Schourup, A. (1973). A cross-language study of vowel nasalization. *Ohio State University Working Papers in Linguistics, 15,* 190-221.

Ushijima, T., & Hirose, H. (1974). Electromyographic study of the velum during speech. *Journal of Phonetics*, 2,315-326.

Ushijima, T., & Sawashima, M. (1972). Fiberscopic examination of velar movements during speech. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, 25-38.

FOOTNOTES

*Versions of this paper were presented at the Convention of the American Speech-Language-Hearing Association, Boston, MA, November 1988, and at the meeting of the Linguistic Society of America, Washington, DC, December 1989.

†Also Department of Speech, Communication Sciences, and Theatre, St. John's University, Jamaica, NY.

††Also Department of Speech-Language-Hearing, Temple University, Philadelphia.

[1]C= an oral consonant, V_n= any number of vowels, and N= a nasal consonant.

[2]In this model, gestures are modified only by having their onsets delayed by the requirements of preceding segments (i.e., carryover coarticulation), exemplified in the first row of Figure 1b.

[3]The languages studied include English, Hindi, German, and Japanese.

[4]Such patterns are also seen in CV sequences containing high and mid vowels (Bell-Berti, 1980; Bell-Berti et al., 1979; Henderson, 1984).

[5]These results also reflect the report of Bladon and Al-Bamerni, although they do not report consistency of this pattern, nor do they comment on the relation between vocalic sequence duration and number of stages.

[6]We expect the slope of the early stage observed in longer nasal utterances to become more like that of the oral utterances as vocalic sequence duration increases, pushing the onset of the nasal consonant gesture further from the onset of the (first) vocalic gesture, which would reduce the gestural overlap earlier in the vocalic sequence.

[7]The few tokens in which multi-stage gestures are found are, however, tokens of "long" utterances.

Haskins Laboratories Status Report on Speech Research
1990, SR-103/104, 39-50

Converging Sources of Evidence for Dissecting Articulatory Movements into Core Gestures*

Suzanne E. Boyce,† Rena A. Krakow,†† Fredericka Bell-Berti,††† and Carole E. Gelfer††††

Explaining and modelling the process of coarticulation is one of the central tasks of speech research. At the same time, the literature on the temporal extent of anticipatory coarticulation has been characterized by conflicting findings. In this paper, we bring together data from a number of different articulatory studies to argue that many of these inconsistencies can be resolved by careful comparison between minimally contrastive contexts. In particular, we examine reports of "one-" and "two-stage" coarticulatory patterns, and suggest that the occurrence of these patterns can be accounted for by factoring in the effects of neighboring segments and suprasegmental variables such as speaking rate. Our conclusions support the coproduction view of coarticulation as the consequence of overlap between spatio-temporally stable, context-independent gestures.

Over the last several decades, the temporal extent of coarticulation—in particular, anticipatory coarticulation—has been an ongoing subject of research and debate. Resolution of the issue has remained elusive, however, largely because the results reported by various investigators have been inconsistent with each other and/or internally inconclusive. In previous work (Bell-Berti & Krakow, this volume; Boyce, Krakow, & Bell-Berti, in press; Gelfer, Bell-Berti, & Harris, 1989), we have argued that experimental controls in the form of minimally contrastive contexts are under-utilized in studies of coarticulation. In this paper, we will argue that many of the apparent inconsistencies in the literature can be resolved by careful examination of such controls and that the results lend strong support to the "coproduction" model of articulation.

Conceptually, coarticulation is defined as the influence of segmental context on the articulatory/acoustic realization of a target segment. It is assumed that, because of perceptual or articulatory constraints on target and surrounding segments, there are limits on the temporal extent of coarticulation (see, for example, Lindblom, 1983, Manuel & Krakow, 1984). Thus, the problem of modelling coarticulation is often cast in terms of the question "what are the constraints on the temporal spread of coarticulation?"

The debate on this issue has focused on two particular speech production frameworks, the "look-ahead" or "feature-migration" models exemplified by Henke (1966), Daniloff and Moll (1968), Benguerel and Cowan (1974) and, in a revised version, Keating (1988), and the "coproduction" models of Fowler (1980), Bell-Berti and Harris (1981), Browman and Goldstein (1986), and Saltzman and Munhall (1989). In the "look-ahead" models, an articulatory planning component determines which movements will be required for upcoming segments and initiates them as soon as possible, barring any competing articulatory requirements. Although they differ in the level at which requirements are specified, with Henke (1966) defining specification in purely articulatory terms and Keating (1988) advocating specification at the level of phonological contrast, these models have certain basic assumptions in common. First, segments are assumed to be concatenated in non-

This research was supported by NIH grants NS-13617 and BRS RR-055996 to Haskins Laboratories and by NIH grant NS0-7040-15 to M.I.T. We would like to acknowledge the helpful comments of Catherine Browman, Jan Edwards, Marie Huffman, Sharon Manuel, Joe Perkell, and Elliot Saltzman, which greatly improved the quality and clarity of our paper.

overlapping time slots. Thus, if movement for a target segment is observed in a context segment, that movement is assumed to have "spread" into the domain of the context segment. Second, because every different context poses a different set of conditions, the complex of neuromuscular commands (the motor "plan") associated with the target segment, and consequently the time at which coarticulation begins, will change in a predictable manner according to context.

The coproduction models, on the other hand, assume that each segment has an associated articulatory control structure for instantiation of a particular feature. These core structures, or gestures, are posited to be (a) consistently present, and (b) relatively stable with regard to segmental context. It is assumed that the ordinary time course of an articulatory gesture extends both before and after the time that its effects dominate the acoustic signal and that much, if not most, coarticulation can be explained as resulting from local interactions between overlapping gestures. Thus, the underlying motor control structure for a particular segment remains essentially the same regardless of the phonetic identity of surrounding phones. In contrast to the look-ahead view, changes in observed patterns of movement in different contexts stem from local interactions between context and target gestures rather than from any change in the motor plan for the target segment. It should be noted that early versions of the look-ahead model made a distinction between anticipatory coarticulation, as requiring advance computation, and carryover coarticulation, as proceeding strictly from physical factors such as inertia. In the coproduction approach, this distinction is blurred, since both intentional activity and physical forces are subsumed under the notion of a characteristic gesture.

Experimental efforts to resolve the debate between these two classes of theory (look-ahead and coproduction) have shown inconsistent results. Based on studies of anticipatory lip-rounding, Benguerel and Cowan (1974), Daniloff and Moll (1968), Sussman and Westbury (1981) and Lubker (1981) concluded that the onset of rounding may vary depending on the identity of the context segment(s); in contrast, Bell-Berti and Harris (1974, 1979, 1982), Engstrand (1981), McAllister (1978), and Boyce (1988) concluded that the time course of rounding is stable over different phonetic contexts. Studies of anticipatory nasalization have led to similarly inconsistent findings (cf. Bell-Berti & Krakow, this volume; Kent, Carney, & Severeid 1974; Moll & Daniloff, 1971; Ohala, 1971).

One explanation for the variable outcome of coarticulation studies was put forward by Bladon and Al-Bamerni (1982) who proposed that observed coarticulatory patterns might be a combination of anticipatory feature spread plus stable gestures and that the inconsistent results of previous studies might be due to differences in measurement technique that weighted the effects of one or the other more heavily. In addition, Bladon and Al-Bamerni proposed that speakers may vary randomly in how they implement anticipatory coarticulation. Observation of token-to-token variability in lip-rounding patterns led Perkell and Chiang (1986) (see also Perkell, 1986 and Chiang, 1987) to a similar conclusion. For these investigators, then, coarticulation for a target segment, while showing general similarities from token to token, is not predictable in detail from the segmental context.

A different explanation was put forward by Gelfer et al. (1989), who showed that much of what investigators have taken to be anticipatory movement for an upcoming target segment may be due to movement associated with surrounding phones. In this paper, we will expand the original argument of these authors by bringing together data from different studies, from the literature or our own research, that examine coarticulation by different measurement techniques. To allow direct comparison with the existing studies cited above, we have used measures of lip movement for rounding and associated electromyographic (EMG) data, as well as measures of velic height for nasality. Taken together, these data demonstrate that, when utterances containing minimal contrasts for the features of interest are examined, it is often possible to separate the effects of target and context segments. Moreover, many of the apparent inconsistencies in the literature can be resolved by this method. We will argue further that the notion of stable but overlapping gestures can account for much of the variability observed in coarticulatory behavior, given certain assumptions about the nature of overlap and the effect of suprasegmental factors such as speaking rate and stress.

One of the crucial assumptions made by investigators studying coarticulation is that segments may be articulatorily neutral (or "unspecified") for a particular feature. This assumption seems to be drawn from current phonological theories which maintain that segments are distinguished from

one another at the underlying level by the specification of a few contrastive features (usually translatable as articulatory configurations in the vocal tract), rather than by a maximally detailed description. In the application of this principle to theories of speech production, it is generally assumed that what distinguishes segments from one another phonologically also dictates the parameters by which they are allowed to vary articulatorily. Normally, investigators have taken this to mean that as long as the conditions represented by the specified features are met, the rest of the vocal tract, for which features are not specified, will be free to vary. Thus, if a segment is unspecified at the point in speech production when abstract phonological units are translated into a motor plan, it will contribute nothing of its own to a trajectory between segments that are specified for that feature (Keating, 1988).

This assumption has been particularly important for studies testing the look-ahead model. Given a segment with a particular feature preceded by one or more segments regarded as neutral for that feature, researchers have tended to interpret the first sign of movement or muscle activity typical of that feature as an instantiation of that segment. To take an example, in a sequence such as /is#tu/, where the feature of interest is lip-rounding, the first evidence of lip protrusion, or EMG activity in the orbicularis oris (OO) muscle, is assumed to be associated with the rounded vowel /u/. If lip protrusion or OO activity is observed during /i/, /s/, or /t/, then coarticulation for lip rounding is taken to have begun, and the time at which it occurs is taken to indicate the extent to which coarticulation can spread to earlier segments. This argument presumes, of course, that the consonants /s/ and /t/ are indeed neutral with respect to rounding. Similarly, in a sequence such as /sei #an/, where the feature of interest is nasality and the vowels /ei/ and /a/ are considered neutral for nasality, the first sign of velopharyngeal port opening, or velic lowering towards an open nasal port, is taken to indicate the temporal extent of anticipatory nasal coarticulation.

Practically speaking, however, when a single neutral segment precedes a segment which is specified for the feature of interest, it is hard to disentangle the look-ahead and coproduction models. This is because while the look-ahead models predict that articulatory onset will occur at the beginning of the neutral segment, the stable trajectory predicted by the coproduction model may begin at that point as well. To solve this problem, investigators have varied the number of supposedly neutral segments preceding a target segment that is positively specified for the feature of interest (e.g., Bell-Berti & Harris, 1982, Sussman & Westbury, 1981); for instance, in addition to examining velic lowering in a CVN sequence, they examine lowering in sequences such as CVVN or CVVVN, etc. If the onset time of the movement associated with the target segment is found to vary in proportion to the number and/or duration of the neutral segments, this is taken as support for the look-ahead model. If, on the other hand, a consistent time interval intervenes between movement onset and some other stable landmark (such as the onset of the nasal murmur or the beginning of oral contact for the nasal consonant), this is taken as support for the coproduction model.

Assumptions of articulatory neutrality have often been questioned, however. It is well known, for instance, that English consonants such as /s/, /l/, and /ɹ/, although not phonologically contrastive for rounding, may be produced with a rounded configuration in unrounded environments (Brown, 1981; Delattre & Freeman, 1968; Leidner, 1973). In addition, numerous investigators have reported evidence that the velum has characteristic positions for different vowels (Bell-Berti, 1980; Bell-Berti, Baer, Harris, & Niimi, 1981; Moll, 1962; Ushijima & Sawashima, 1972). In many studies (e.g., Bell-Berti & Harris, 1981; Benguerel & Cowan, 1974; Chiang, 1987; Daniloff & Moll, 1968; Lubker, 1981; Perkell, 1986; Perkell & Chiang, 1986), investigators attempted to characterize subjects for gross articulatory activity associated with the feature of interest during so-called neutral segments. Such activity, if found, would normally be ascribed to a subphonemic (i.e., non-contrastive) specification for that feature. But, as noted by Gelfer et al. (1989), investigators have not attempted a comprehensive analysis of the less obvious characteristics of segments included in their experimental corpora before drawing conclusions concerning coarticulation. That effects uncovered by such an analysis may have significant import can be seen in the following sets of data.

Figure 1, reprinted from Gelfer et al. (1989), shows EMG data in the form of ensemble-averaged orbicularis oris inferior (OOI) traces, as spoken by a native speaker of American English. This figure shows muscle activity associated with rounding in utterances of the structure V#CV, VC#CV, and VCC#CV, where the first vowel is unrounded, the second vowel, rounded, and the consonants various combinations of /s/ and /t/.

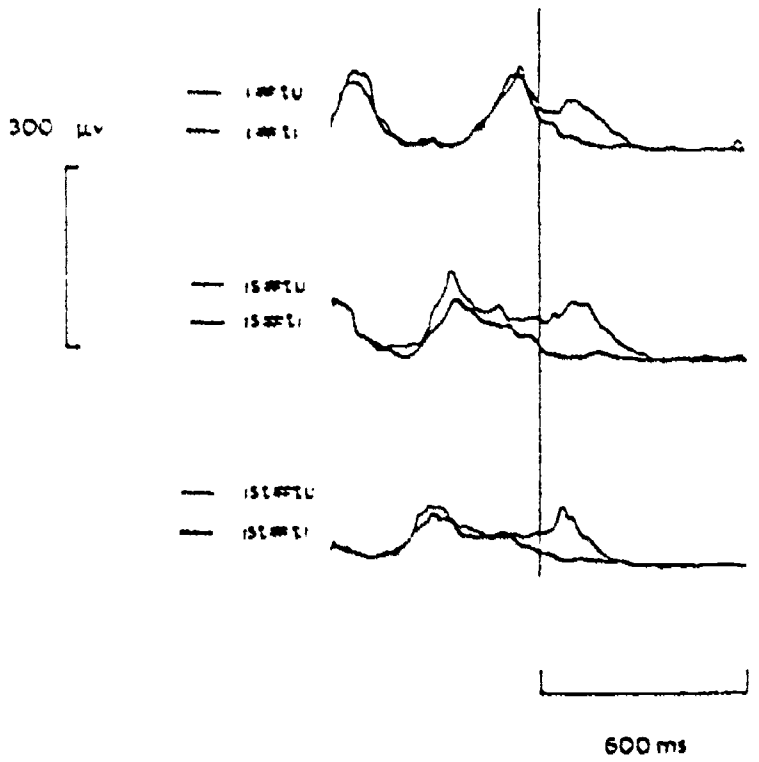


Figure 1. **Ensemble-averaged OOI EMG activity for 15-20 tokens of three pairs of utterances with increasing numbers of "neutral" consonants. The vertical line indicates the lineup point for averaging tokens, which was at the acoustic onset of the second vowel.**

As noted above, it is normally assumed that the first instance of lip movement or orbicularis oris muscle activity indicates the onset of motor execution for the rounded vowel (see also Bell-Berti & Harris, 1981; Lubker, 1981). Thus, the fact that, in /i#tu/, /is#tu/, and /ist#tu/, OOI activity seems to begin between 200 and 400 ms before the acoustic onset of the rounded vowel might suggest that rounding for /u/ actually begins at these times. (Association of these EMG activity peaks with protrusion movements was confirmed by reference to simultaneously recorded lip movement data.) At first glance, these results appear to support the look-ahead model. Although the look-ahead model would not predict that the OOI trace would split into an early and a late peak of activity (as it does in all three cases), the time interval between the first peak of OOI activity and the acoustic onset of the /u/ was longer for words with greater numbers of intervocalic consonants, such as /ist#tu/ and /ist#stu/, suggesting that anticipation begins at the first "neutral" consonant.

As Gelfer et al. point out, however, this early movement and EMG activity is echoed in the traces for words where /u/ has been replaced with /i/, as in /is#ti/. Since no rounded vowel is present, the early activity must be attributed to subphonemic lip movement for either or both of the intervening consonant(s), and the two peaks of activity seen in, for example, /is#tu/ must be assigned separately to the intervocalic consonants and to the rounded vowel. The fact that EMG activity is evident with the first consonant of the consonant string can no longer be viewed as unambiguous evidence of coarticulatory anticipation; rather, it reflects the characteristics of the intervocalic consonant(s). (Note that the second subject in this study showed similar patterns.)

A similar demonstration may be made in the case of what are known as "one-stage" and "two-stage" movement onsets for nasal and rounding coarticulation. Bladon and Al-Bamerni (1982) compared nasograph (Ohala, 1971) traces for various utterances incorporating one to three oral vowels followed by a nasal consonant. They found two patterns of velopharyngeal port opening, (a) a "one-stage" pattern beginning early in the vowel string and continuing smoothly to a maximum open position during the nasal consonant; and (b) a "two stage" pattern, consisting of slight port opening during the vowel string followed by an abrupt high-velocity opening stage beginning just before the nasal consonant. Figure 2, adapted from Bladon and Al-Bamerni (1982), shows these two patterns, which occurred in different repetitions of the Kurdish sequence [tagutsɛːɛmaːkhaː]. The authors pointed out that timing for the "one-stage" pattern conformed to the predictions of the look-ahead model of coarticulation (i.e., port opening onset occurred at the acoustic onset of the first neutral segment /ɛː/).

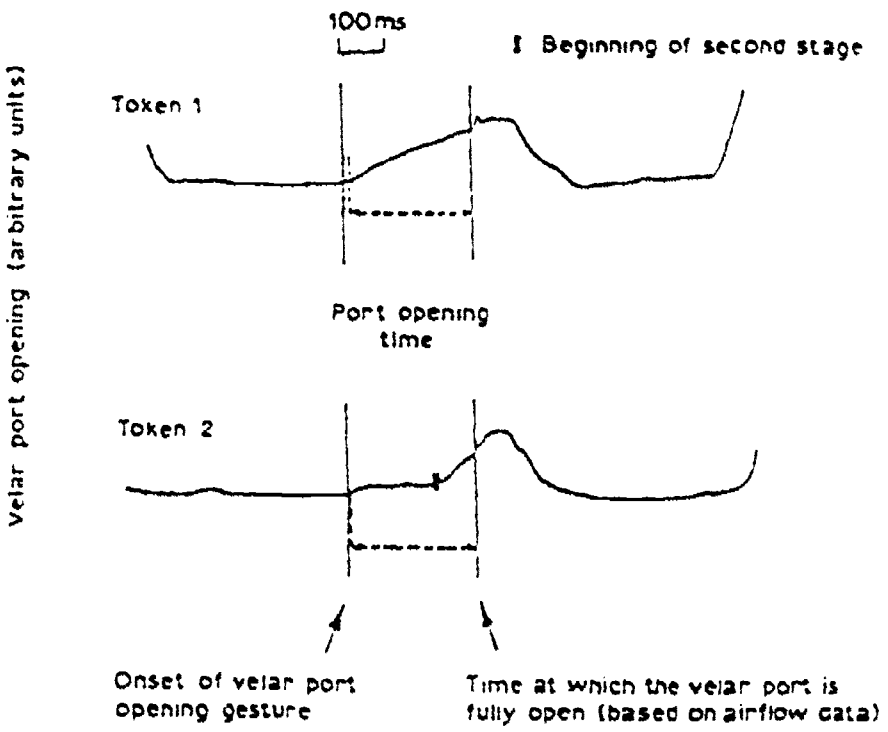


Figure 2. **Nasograph traces for two tokens of [tagutseɛemaɪkhaɪ] as spoken by a native speaker of Kurdish. Adapted from Bladon and Al-Bamerni (1982).**

However, with regard to the two-stage pattern, they argued that while the early stage conformed in timing to the predictions of the look-ahead model, the second stage conformed to the predictions of the coproduction model (i.e., the onset of the high-velocity stage occurred at a relatively fixed time in relation to the nasal murmur). Especially puzzling was the finding that, as Figure 2 shows, a given speaker might produce either pattern during different repetitions of the same utterance. Bladon and Al-Bamerni were unable to discover any factors which would predict when the different patterns occurred; accordingly, they suggested that the occurrence of one- and two-stage patterns might be random. Further, they proposed that findings in the literature supporting one model or the other were probably a function of measurement criteria emphasizing either the first or second stage onset of the two-stage pattern.

One possibility not explored in Bladon and Al-Bamerni's (1982) study was that early nasalization might be due to intrinsic, subphonemic velar position for these posited "neutral" segments. That the two-stage pattern owes its shape to a combination of intrinsic positions for the vowels as well as the nasal consonant has been argued by Bell-Berti and Krakow (1988). In what follows, we review some of their arguments. We will also suggest that at least some of the seemingly random alternations between single and multi-stage movement patterns are a function of suprasegmental variations leading to differences in the extent to which vocalic and consonantal gestures overlap in time.

Figure 3 shows velic movement data, as recorded with the Velotrace (Horiguchi & Bell-Berti, 1987), for characteristic tokens of two utterances, /ɪts#ʌ#lɑsəl#əgɛn/ ("It's a lasal again") and /ɪts#ʌ#lənsəl#əgɛn/ ("It's a lansal again"), spoken by a native speaker of American English. The tokens were aligned at the acoustic release of the /s/ in "It's" for both utterances. (Note that for Velotrace data in this paper, a low position on the vertical scale indicates a low position of the velum. This is in contrast to the figure from Bladon and Al-Bamerni's study, in which a high position in the Nasograph trace indicates an open velopharyngeal port, and thus a low velic position.) As can be seen, the utterance with a nasal consonant—"It's a lansal again"—appears to show two separate stages of movement leading up to the nasal consonant, a pattern similar to that reported by Bladon and Al-Bamerni (1982). What appears to be the "first stage" of the velic lowering movement begins roughly at the release of the /s/ in "It's"; the "second stage" begins during the sequence /ʌlə/.

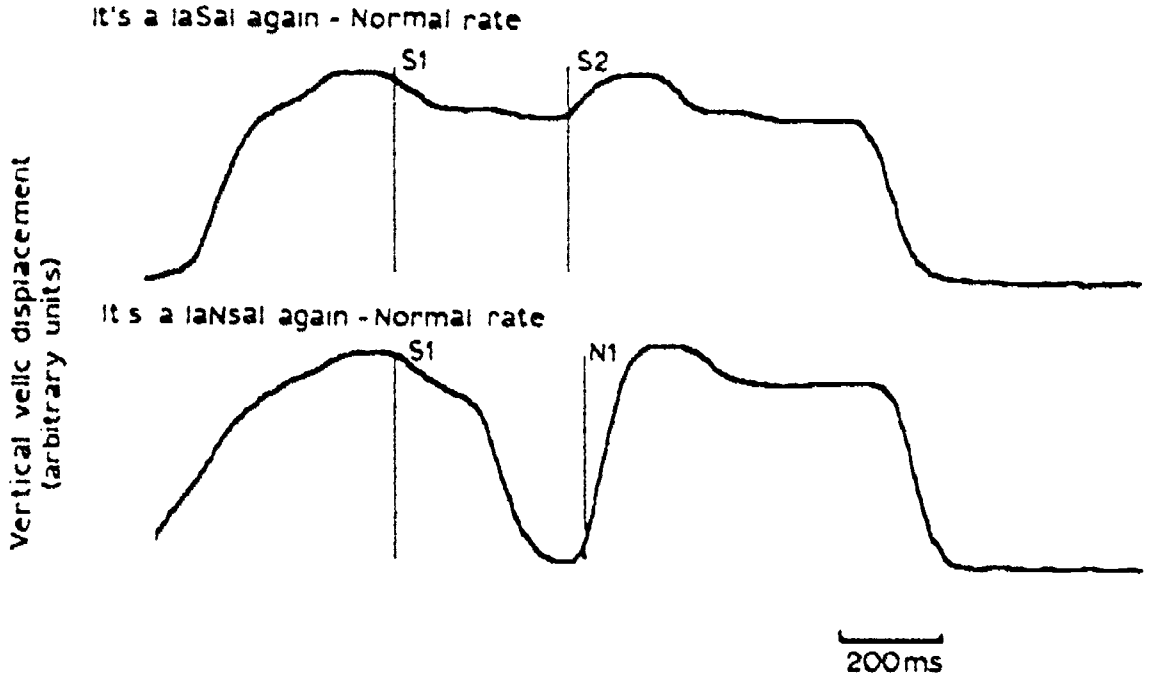


Figure 3. Velotrace movements for oral and nasal versions of "It's a lə(n)səl again," spoken at a normal conversational rate. Vertical lines labelled S1 indicate the acoustic offset of the /s/ frication in "It's," S2 indicates the onset of the /s/ frication in "lə(n)səl" and N1 indicates the onset of the nasal murmur for /n/. Tokens were aligned at S1.

Looking at the trace for the oral utterance "It's a ləsəl again," however, we see that it is nearly identical to that for "It's a lənsəl again" up to the point at which the so-called second stage begins. That is, "It's a ləsəl again" shows a slight velic lowering movement that corresponds nearly exactly in time and amplitude with the early portion of the "two-stage" movement in "It's a lənsəl again." Lowering of this type in the transition between consonants and vowels has been described previously for oral sequences (Kent et al. (1974). Thus, it seems highly likely that the first stage movement involved in "It's a lənsəl again" is (a) characteristic of the sequence /sʌlə/, at least for this speaker, and (b) not related to anticipatory velic lowering for the nasal consonant. We conclude from this that the two-stage movement results, not from anticipation of velic lowering during an unspecified segment, but from combination of segments with consistent individual articulatory specifications for velic height.

The notion of "two-stage" anticipatory coarticulation was also examined by Perkell and Chiang (1986) using upper lip protrusion data for sequences such as /li#kut/, /li#sut/, /lis#kut/ with up to four medial consonants, as spoken by four native speakers of English. They partitioned each movement (whether or not it had an obvious "two-stage" shape) into two components: (a) a first stage, stretching between the first sign of protrusion (the beginning of positive velocity) preceding the acoustic onset of the /u/ and the point of maximum acceleration in that interval, and (b) a second stage, from the acceleration peak to the protrusion peak (end of positive velocity). (Typically, in a trace with a visually evident two-stage pattern, peak acceleration tends to co-occur with the beginning of the second stage.) Figure 4, adapted from Perkell and Chiang (1986), illustrates upper lip protrusion traces from several tokens of the representative utterance /li#sut/, as spoken by a single speaker. As this figure shows, a wide range of variability was found in the onset of protrusion with regard to the acoustic onset of /u/. Timing between the onset of /u/ and the maximum acceleration of the lip protrusion movement was also variable. The authors concluded that, in spite of frequent similarities in the sizes and shapes of the second stage of the protrusion movement among the different tokens, there was still too much variability between the times of maximum acceleration and acoustic vowel onset, and between vowel onset and protrusion peak, to demonstrate the existence of a stable core gesture.

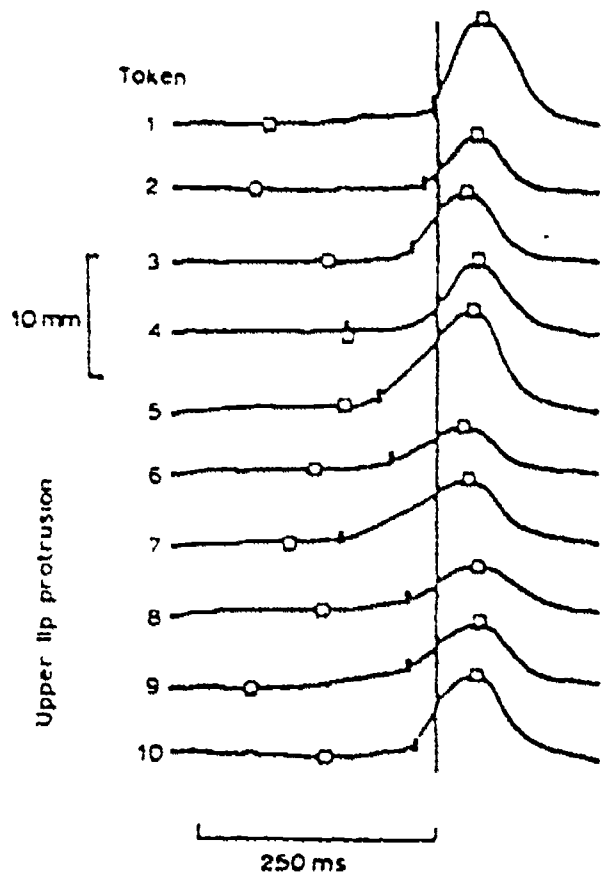


Figure 4. Upper lip protrusion data obtained by strain-gauge for ten tokens of /li#sut/. The vertical line indicates acoustic onset of the second vowel. Adapted from Perkell and Chiang (1986). □: beginning and end of protrusion; ■: point of maximum acceleration.

At the same time, variability in the timing of the "first stage" onset *vis-à-vis* acoustic offset of the /i/ was greater than predicted by look-ahead models. The best predictor of movement patterns was the identity of the intervocalic consonant sequence, leading Perkell and Chiang to assign a greater role to the characteristics of supposedly "neutral" consonants than the look-ahead models had originally proposed.

Perkell and Chiang concluded from these findings that neither the look-ahead model nor the coproduction model could be correct in their strong form. Instead, they proposed a "hybrid" model, in which the characteristics of movements are determined in part by their segmental goals, and in part by token-specific interactions among acoustic, aerodynamic, and biomechanical factors associated with neighboring segments. Implicit in this conclusion is the assumption that since such factors are minutely balanced from moment to moment during speech production, their effect on the timing and spatial characteristics of individual gestures cannot be precisely predicted. This conclusion was affirmed in a later study involving more subjects (Chiang 1987).

Interestingly, in some respects Perkell and Chiang's hybrid model makes predictions of variability that are little different from those of the coproduction model. In the hybrid model, the point at which a gesture appears to start will change as a function of the constraints imposed by its context. In the coproduction model, the notion that gestures overlap and combine means that the point at which one gesture appears to dominate the signal is also dependent on its context. This is illustrated schematically in Figure 5, which shows, in the left hand panel, two theoretical gestures associated with two overlapping segments, and, in the right hand panel, a smoothed trajectory representing their articulatory output as a "two-stage" pattern. The point at which the skirts of the two gestures overlap is, approximately, where we would expect to identify the beginning of the second stage.[1] If we think of the later gesture as protrusion for an /u/ vowel and of the earlier gesture as belonging to a preceding consonant, it is readily apparent that the point at which protrusion appears to begin will vary depending upon the size and shape of the consonant gesture, and its timing with respect to the vowel gesture. If the consonant gesture also involves protrusion, the apparent beginning of the "first stage" will depend on these characteristics of the individual consonant. Similarly, the point at which the "second stage" emerges from the "first stage" will depend on the characteristics of the overlap between the two gestures. Thus, Perkell and Chiang's conclusion that these aspects of the movement trajectory are constrained by context is also predicted by the coproduction model.[2]

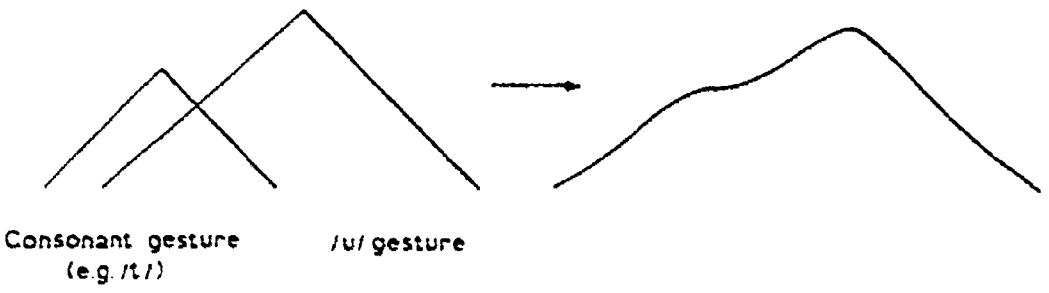


Figure 5. Schematized drawing of hypothesized consequences of gestural overlap.

To show how variability of this type can be teased out from movement patterns, some data from Boyce (1988) are shown below. Figure 6 shows upper lip protrusion movement for six tokens of the nonsense words /kiktluk/ and /kiktlik/, embedded in the carrier phrase "It's a _____ again." The data were recorded from a native American English speaker using a modified Selspot system (Kay, Munhall, Vatikiotis-Bateson, & Kelso, 1985). Tokens of /kiktluk/ and /kiktlik/, collected at different points during the experiment, are paired by similarity in shape. Both /kiktluk/ and /kiktlik/ tokens are lined up at the acoustic onset of the second vowel. To provide a parallel analysis with that of Perkell and Chiang (1986), the /kiktluk/ traces are marked for protrusion onset (onset of positive velocity), protrusion peak (onset of negative velocity), and for peak acceleration. There were typically two acceleration peaks of similar magnitude for /kiktluk/ tokens in the interval between the onset of protrusion and its peak, and both are shown.

In these data, the /kiktluk/ traces show two peaks of protrusion. The first peak occurs considerably before the acoustic offset of /i/ and may be related to the /s/ of "It's" in the carrier phrase. It is followed by a dip in the signal whose minimum occurs slightly prior to the acoustic offset of /i/ and generally coincides with the onset of positive velocity. The first maximum acceleration point occurs soon after. The main protrusion peak occurs approximately 100 ms after the acoustic onset of the /u/ vowel. Typically, the /u/ protrusion shows a second inflection between the protrusion onset and peak. The second maximum acceleration point occurs at this time.

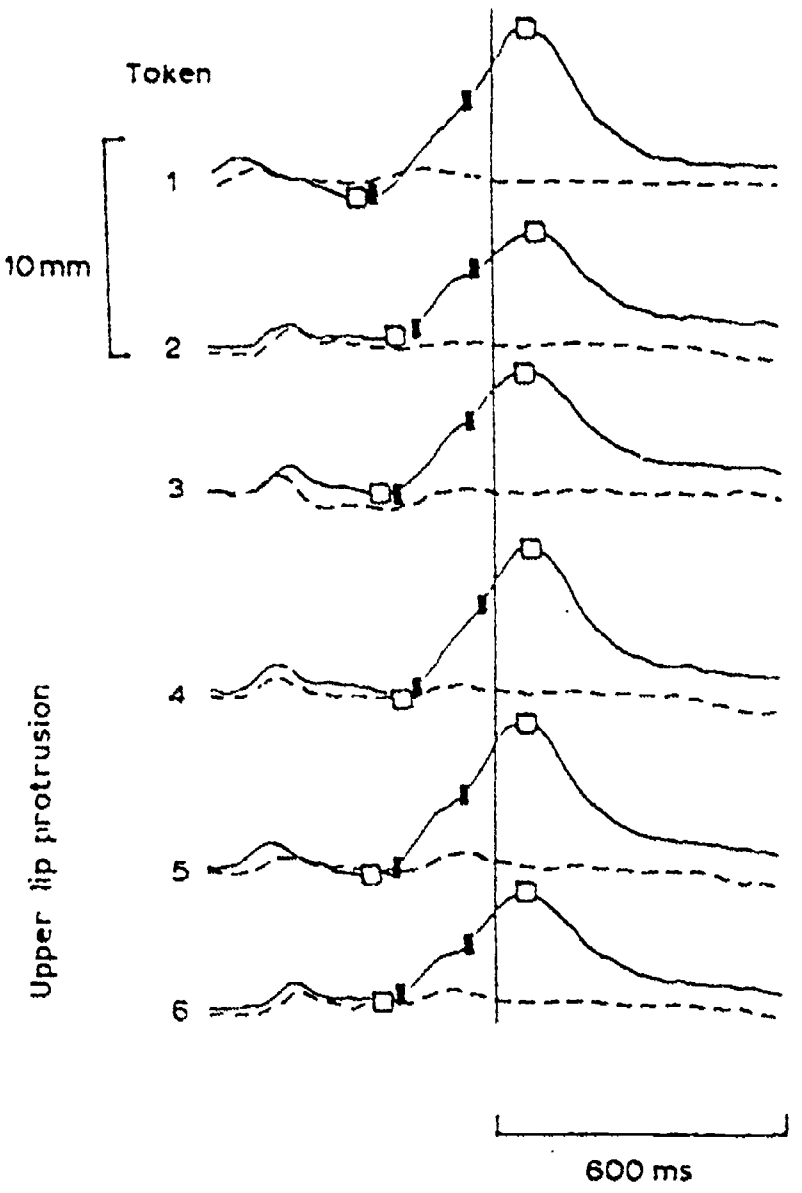


Figure 6. Upper lip protrusion traces, obtained by Selspot, for six pairs of tokens paired by similarities in shape and timing. The vertical line indicates acoustic onset of the second vowel. Baselines have been adjusted to facilitate comparison and do not necessarily reflect experimental baselines. ——: kiktluk; - - -: kiktlik; □: beginning and end of protrusion; ■: point of maximum acceleration.

Looking at the /kiktlik/ trace, however, we see that some portion of the protrusion movement in /kiktluk/ occurs in the absence of the rounded vowel. In particular, the inflection at the second maximum acceleration point, which identifies the beginning of the second stage, is often temporally aligned with a clear peak in the corresponding /kiktlik/ trace. In addition, protrusion onset for the /u/ in /kiktluk/, as determined by the calculation of positive velocity, appears to coincide with a parallel movement occurring in /kiktlik/. Probably, the minimum in the lip protrusion trace reflects a retraction movement for the /i/, whereas the following peak is associated with one or more of the intervening consonants. Thus, at least part of the first stage protrusion movement seen in /kiktluk/ tokens appears to be due to the consonant sequence found in both /kiktlik/ and /kiktluk/, rather than being fully attributable to the rounding gesture for the /u/. (Similar demonstrations using upper and/or lower lip data can be made for the seven other subjects in Boyce's study.

It is clear from these data that the time at which protrusion appears to begin cannot be assumed, *a priori*, to reflect the time course of the underlying protrusion gestures associated with /u/. Rather, it is necessary to identify the articulatory movements that are characteristic of each segment in a sequence before drawing any conclusion about the beginning of coarticulation. As in the velic movement data, the beginning of the second stage may reflect the boundary of interaction between gestures for adjacent segments.

One question that remains is how to account for the seemingly random incidence of one- vs. two-stage movements for different repetitions of the same utterance. In Bell-Berti and Krakow (1991), it was proposed that changes in speech rate might lead to alternations between simple and multi-stage movements. In what follows, we review their arguments and later we will suggest that, in addition to rate, stress may influence the shape of articulatory movements such that they appear as "one"- or "two-stage" (cf. Krakow, 1987).

Figure 3 (above) showed movement traces from Bell-Berti and Krakow (1991) for tokens of /ɪts#ʌ# lɑsɑl#ʌgɛn/ ("It's a lasal again") and /ɪts#ʌ# lɑnsɑl#ʌgɛn/ ("It's a lansal again"), produced at a normal conversational rate. Figure 7 shows corresponding productions of the same utterances produced by the same subject at a self-selected "rapid rate." As can be seen, the rapidly produced "It's a lansal again" (seen in Figure 7) contains only a single velic lowering movement in the vicinity of the nasal consonant; in contrast, the normally produced token (seen in Figure 3) contains two component lowering movements. One way to account for this is to say that in the faster utterance, as in the slower utterance, there are independent (underlying) gestures for the vowel and the nasal consonant. They appear as a single movement in the faster rendition because there isn't ample time for the two movements to occur in temporally separate space.

This interpretation is strengthened by the fact that the rapidly produced oral utterance, "It's a lasal again" (seen in Figure 7) shows a small lowering movement following the /s/ of "It's," suggesting that even in fast speech there is a small but independent velic gesture associated with the oral sequence. This evidence, together with the clear case of separate "oral" and "nasal" velic movements in the normal rate utterances of Figure 3, suggests that the rapid utterance of "It's a lansal again" contains two gestures combined in overlapping fashion. Additionally supportive is the fact that in some of the more slowly produced "rapid" tokens of "It's a lansal again" vocalic and consonantal gestures showed two-stage lowering patterns (Bell-Berti, 1990). While the rapid productions of this sequence contained some instances of one-stage movements and some of two-stage movements, the slower "conversational rate" productions always showed two-stage movement patterns.

One major similarity between the factors of speaking rate and number of "neutral" segments is the fact that both can be modelled as a continuous function of gestural overlap. As pointed out in the discussion of Figure 5, small variations in the way the gestures overlap may have significant consequences in terms of whether a movement appears to follow a "one"- or "two-stage" pattern. Bell-Berti and Krakow, in fact, presented additional evidence supporting the continuous nature of the "one-" vs. "two-stage" phenomenon as a function of speaking rate and number of "neutral" segments. This is shown in Figure 8, which summarizes the results of manipulating the two factors, for the subject whose movement traces were shown in Figure 7. Increasing the duration of the "neutral" string preceding the nasal consonant, whether by adding vocalic segments, or by slowing the rate, led to an increase in the incidence of multi-stage patterns. (See also Löfquist (1989) and Munhall and Löfquist (1989) for similar results using rate change in laryngeal opening/closing movements.)

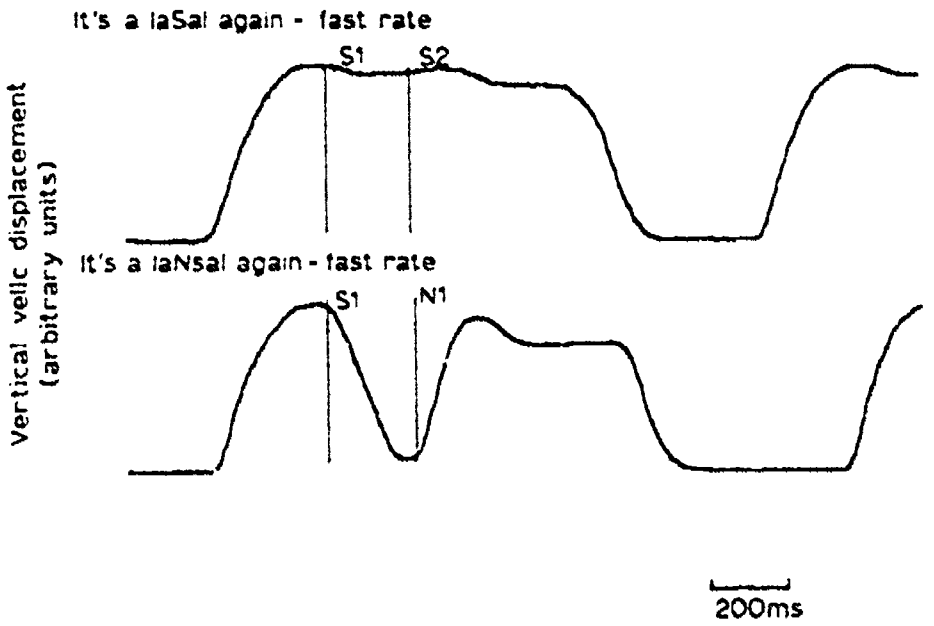


Figure 7. **Velotrace movements for oral and nasal versions of "It's a la(n)sal again," spoken at a fast rate. Vertical lines labelled S1 indicate the acoustic offset of the /s/ frication in "It's," S2 indicates the onset of the /s/ frication in "la(n)sal" and N1 indicates the onset of the nasal murmur for /n/. Tokens were aligned at S1.**

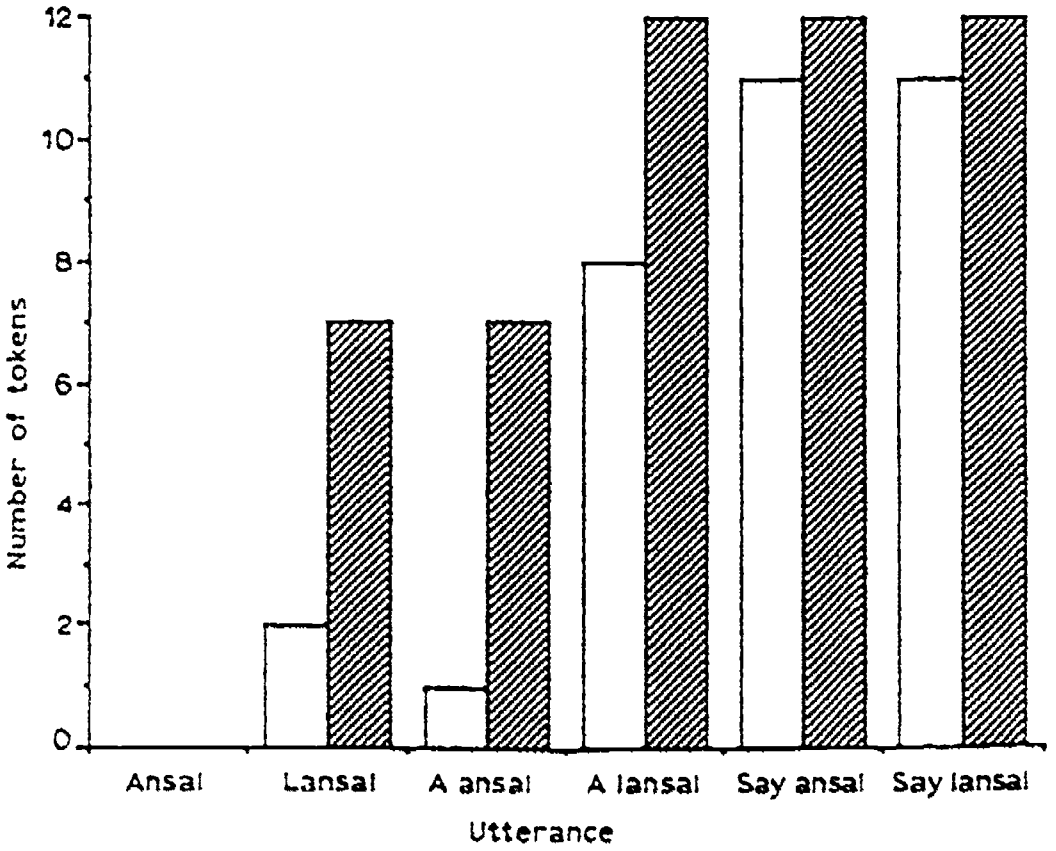


Figure 8. **Histogram showing number of tokens (out of a possible twelve) with a multi-stage pattern, for fast and normal rate productions of five utterances containing nasal consonants preceded by "neutral" segments. Utterances are arranged left to right according the number or duration of "neutral" segments. □: fast; ■: normal.**

Given these data, and additional data cited in Bell-Berti and Krakow (1988), we would like to suggest that variability in speech rate (whether due to fatigue, rushing, list effects, attitude, etc.) may lead to alternating patterns of one- vs. two-stage movements for different tokens of the same utterance containing either a nasal consonant or a rounded vowel. We would suggest, further, that such patterns can best be accounted for with the coproduction notion of overlap. In this model, even small variations of speaking rate within the categories of "fast" or "normal" would produce variation in the extent to which the gestures are overlapped, and thus variation in their apparent complexity of output , i.e., single- vs. multi-stage pattern of movement. (Note that, due to insufficient data, we are unable to determine if Bladon and Al-Bamerni's example of one- vs. two-stage movements is related to a difference in rate between tokens. That is, we were unable to consider either acoustic or articulatory duration characteristics in detail since only those two tokens were shown and no acoustic traces or durations were provided.)

In the framework we present here, the addition of "neutral" segments and a decrease in speaking rate both favor the occurrence of multi-stage articulatory patterns, because they allow individual gestures to emerge as distinct entities. As a note of interest, evidence that stress may have a similar effect can be found in Krakow (1986). In her data, multi-stage velic raising movements were more often observed in stressed syllables containing a nasal consonant than in matched unstressed syllables. It may be that something about stress, perhaps the increased duration that characterizes stressed sequences, creates an environment in which discrete gestures are more likely to emerge.

To summarize, in this paper we have re-examined some of the previous conflicting evidence concerning the domain of coarticulation and its underlying mechanisms of control. We have attempted to show (a) that multi-stage patterns may be analyzed (at least to some extent) as the product of multiple independent gestures, and (b) that much of the variability in the incidence of single- and multi-stage patterns may proceed from variations in numbers of segments or from suprasegmental factors. This is not to say that we have accounted for all sources of variability in speech movement data. For example, Kent et al. (1974) reported that when they asked two speakers to produce a sentence at normal and fast rates, two different strategies for speeding up were employed. One subject showed a pattern which we would consider to be consistent with a simple strategy of greater gestural overlap with increased rate. That is, for fast productions, this speaker never attained as extreme high or low velic positions as he did in his slower productions. The other subject managed to attain the same extent of displacement in normal and fast productions by increasing the velocity of velic movement in the faster rendition. (This is consistent with a strategy of minimizing the degree of overlap.) In spite of the difference in strategy, however, both subjects evidenced a high degree of stability in their productions; when the fast and slow trajectories were overlaid in a time-normalized manner for each subject, the peaks (for the oral consonants) and the valleys (for the nasal consonants) coincided. Thus, the co-production model must allow for alternative stable gestural patterns.

In this paper, we have attempted to show how, given the assumption of overlapping core gestures, many of the inconsistencies in previously reported results, as well as more fundamental problems in speech production theory, may be accounted for within the coproduction framework. In particular, combining the coproduction framework with the notion of variability in the phasing of gestures (as might, for example, be induced by rate changes) provides an account in which single- and multi-stage gestures can be derived from the same underlying articulatory control pattern.

REFERENCES

Bell-Berti, F. (1980). Velopharyngeal function: A spatial-temporal model. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 4, pp. 291-316.). New York: Academic Press.

Bell-Berti, F., Baer, T., Harris, K. S., & Niimi, S. (1981). Coarticulatory effects of vowel quality on velar function, *Journal of Phonetics, 36*, 187-193.

Bell-Berti, F., & Harris, K. S. (1974). More on the motor organization of speech gestures, *Haskins Laboratories Status Report on Speech Research, SR 37/38*, 73-77.

Bell-Berti, F., & Harris, K. S. (1979). Anticipatory coarticulation: Some implications from a study of lip rounding, *Journal of the Acoustical Society of America, 65*, 1268-1270.

Bell-Berti, F., and 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, *Journal of the Acoustical Society of America , 71*, 449-454.

Bell-Berti, F., & Krakow, R. A. (1990). Anticipatory velar lowering: A coproduction account. *Haskins Laboratories Status Report on Speech Research, SR-103/104*. (Also submitted, *Journal of the Acoustical Society of America.*)

Benguerel, A. P., & Cowan, H. A. (1974). Coarticulation of upper lip protrusion in French. *Phonetica, 30,* 41-55.

Bladon, R. A. W., & Al-Bamerni, A. (1982). One-stage and two-stage temporal patterns of velar coarticulation, *Journal of the Acoustical Society of America, 72,* S104(A).

Boyce, S. E. (1988). *The influence of phonological structure on articulatory organization in Turkish and in English: Vowel harmony and coarticulation.* Doctoral dissertation, Yale University Department of Linguistics.

Boyce, S. E., Krakow, R. A., & Bell-Berti, F. (in press). Phonological underspecification and speech motor organization. *Phonology.*

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

Brown, G. (1981) Consonant rounding in British English: The status of phonetic descriptions as historical data. In R. E. Asher & E. J. A. Henderson (Ed.), *Towards a history of phonetics.* Edinburgh: Edinburgh University Press.

Chiang, C. -M. (1987). *Models of coarticulation.* B. S. thesis, M. I. T., Department of Engineering and Computer Science.

Daniloff, R., & Moll, K. (1968). Coarticulation of lip rounding. *Journal of Speech and Hearing, Research, 11,* 707-721.

Delattre, P., & Freeman, D. (1968). A dialect study of American r's by x-ray motion picture. *Linguistics , 44,* 29-68.

Engstrand, O. (1981). Acoustic constraints of invariant input representation? An experimental study of selected articulatory movements and targets. *Reports from the Uppsala University Department of Linguistics, 7,* 67-94.

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

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.

Henke, W. L. (1966). *Dynamic articulatory model of speech production using computer simulation.* Doctoral dissertation, M. I. T., Department of Electrical Engineering and Computer Science.

Horiguchi, S., & Bell-Berti, F. (1987) The Velotrace: A device for monitoring velar position. *Cleft Palate Journal,* 24, 104-111.

Kay, B. A., Munhall, K. G., Vatikiotis-Bateson, E., & Kelso, J.A.S. (1985). A note on processing kinematic data: Sampling, filtering and differentiation, *Haskins Laboratories Status Report on Speech Research, SR 37/38,* 73-77.

Keating, P. A. (1988) Underspecification in phonetics, *Phonology, 5,* 3-29.

Kent, R. D., Carney, P. J., & Severeid, L. R. (1974). Velar movement and timing: Evaluation of a model for binary control. *Journal of Speech and Hearing Research, 17,* 470-488.

Krakow, R. A. (1986). Prosodic effects on velic movement. Paper presented at the meeting of the Linguistic Society of America, New York.

Leidner, D. R. (1973). *An electromyographic and acoustic study of American English liquids.* Doctoral dissertation, University of Connecticut.

Lindblom, B. (1983). The economy of speech gestures. In P. F. MacNeilage (Ed.), *The production of speech.* New York: Springer-Verlag.

Löfquist, A. (1989). Speech as audible gestures. Paper presented at the NATO ASI conference on speech production and speech modelling, Bonas, France.

Lubker, J. F. (1981). Temporal aspects of speech production, *Journal of Phonetics, 38,* 51-65.

Manuel, S. Y., & Krakow, R. A. (1984). Universal and language-particular aspects of vowel-to-vow coarticulation. *Haskins Laboratories Status Report on Speech Research, SR 77/78,* 69-78.

McAllister, J. (1978). Temporal asymmetry in labial coarticulation. *Working Papers, Stockholm Institute of Linguistics, 35,* 1-29.

Moll, K. (1962). Velopharyngeal closure on vowels. *Journal of Speech and Hearing Research, 5,* 30-37.

Moll, K., & Daniloff, R. G. (1971). Investigation of the timing of velar movements during speech. *Journal of the Acoustical Society of America, 50,* 678-684.

Munhall, K., & Löfquist, A. (1989). Gestural aggregation in speech, submitted to *Journal of Phonetics.*

Ohala, J. J. (1971). Monitoring soft palate movements in speech. *Project on Linguistic Analysis, 13,* University of California Berkeley, JO1-JO15.

Perkell, J. S. (1986). Coarticulation strategies: preliminary implications of a detailed analysis of lower lip protrusion movements. *Speech Communication, 5,* 47-68.

Perkell, J. S., & Chiang, C. -M. (1986). Preliminary support for a "hybrid model" of anticipatory coarticulation. In *Proceedings of the 12th International Congress of Acoustics,* July, A3-6.

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

Sussman, H. M., & Westbury, J. R. (1981). The effects of antagonistic gestures on temporal and amplitude parameters of anticipatory labial coarticulation, *Journal of Speech and Hearing Research, 46,*16-24.

Ushijima, T., & Sawashima, M. (1972). Fibersopic examination of velar movements during speech. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics,* 25-38.

FOOTNOTES

Journal of Phonetics, 18, 173-188 (1990).

†M. I. T., Cambridge, MA.

††Also Temple University, Philadelphia, PA.

†††Also St. John's University, Queens, NY.

††††William Patterson College, Wayne, NJ.

[1]Note that although the "two-stage" pattern is schematized in Figure 5 as an interaction between two gestures, as a practical matter the multi-stage nature of the trajectory may not be apparent unless at least two "neutral" segments precede the target segment. Presumably, this is because the adjacent target and "neutral" segment gestures overlap in time so closely that their merger is not apparent.

[2]The issue of how overlapping gestures interact with one another is outside the scope of this paper. For further discussion of this issue, see Boyce (1988), Munhall and Löfquist (1989), and Saltzman and Munhall (1989).

Haskins Laboratories Status Report on Speech Research
1990, SR-103/104, 51-66

Rotation and Translation of the Jaw During Speech*

Jan Edwards† and Katherine S. Harris††

A two-dimensional rigid-body model of jaw movement was used to describe jaw opening and closing gestures for vowels and for bilabial and alveolar consonants. Jaw movements were decomposed into three components: rotation about the terminal hinge axis, and the horizontal and vertical translation of that axis. Data were collected for three subjects in two separate recording sessions. Multiple regression analysis was used to examine the relationships among the three jaw movement components. For two subjects, but not for the third, an interdependence between jaw rotation and the first principle component of jaw translation (horizontal translation) was observed. For these two subjects, the first degree of freedom of jaw movement corresponded to a combination of rotation and the first principle component of jaw translation. For the third subject, the first degree of freedom of jaw movement corresponded to rotation alone. The results of this study, like those of Westbury (1988), indicate that an accurate description of jaw movement during speech requires the recording of two points of jaw movement.

Jaw movement during speech has generally been described as the pure translation (Kakita & Fujimura, 1977) or the pure rotation (Coker, 1976; Mermelstein, 1973) of a single point on the jaw. In the pure translation model, the jaw simply translates in some direction, usually defined as the principal component of jaw position variation. In the pure rotation model, the jaw rotates about a transverse axis that presumably passes through the mandibular condyles. With a few exceptions (e.g., Edwards, 1985; Gibbs & Messerman, 1972; Westbury, 1988), research on jaw movement during speech has examined the movement of a single point on the jaw.

However, the anatomy of the temporomandibular joint allows both rotation and translation. In the lower compartment of the temporomandibular joint, the mandibular condyle rotates against the inferior surface of the articular disc; in the upper compartment, the articular disc glides downward, forward, and sideward (Hjortso, 1955). The jaw is capable of rotating about a transverse or a vertical axis located through the condyles and of translating that axis in anterior-posterior, inferior-superior, and lateral-medial directions (Gibbs, Messerman, Resnick, & Derda, 1971). Therefore, a description of jaw position in terms of a single point on the jaw does not provide enough information to predict the position of every other point on the jaw. A rich literature on the physiology of mastication shows clearly that the simple translation and rotation models are anatomically inaccurate, at least for non-speech opening and closing gestures with displacements of comparable magnitude to those observed during speech (Hjortso, 1955; Posselt, 1968; Sarnat, 1964; Gibbs, et al., 1971). Furthermore, the results of Edwards (1985) and Westbury (1988) indicate that these single-point rotation and translation models are inaccurate for speech-related movements as well.

A comparison of the speech and dental literature suggests that, in many respects, jaw movement during speech appears to be more constrained than during mastication. It has consistently been

This article is based on a doctoral dissertation by the first author which was supported by NIH grant number 13617 to Haskins Laboratories. We would especially like to thank Tom Baer for his help on working through the geometry of the model and Osamu Fujimura for relating this problem to the analysis of the X-ray microbeam data, as well as for their careful reading of many versions of the dissertation. We would also like to thank Win Nelson, for his help on the curve-fitting algorithm; Nell Sedransk, for help with the statistical analysis; and David Kussovitsky, D.D.S., for constructing and attaching the dental appliances. Finally, we thank David Ostry, John Westbury, and a third anonymous reviewer for their thoughtful reviews of this paper.

observed that there is essentially no lateral movement of the jaw during speech (Gibbs & Messerman, 1972; Gentil & Gay, 1986). For example, Gentil and Gay (1986) observed less than .1 mm of jaw movement in the frontal plane during speech. These observations exclude condylar rotation about a vertical axis and lateral-medial translation of the articular disc during speech. Furthermore, the range of jaw opening and closing movements during speech is considerably less than during mastication. Gibbs and Messerman (1972) found that the maximal vertical opening of the jaw, measured at the central incisor, was two to four times greater for mastication than for speech. Thus, speech-related movements should uniformly lie within the range of vertical jaw position that involve a smooth combination of rotatory movements about a transverse axis and anterior-posterior, inferior-superior translation of this axis (Sarnat, 1964).

Thus, the results of previous studies of jaw movement during speech suggest that it is a combination of rotation about a transverse axis and the vertical and horizontal translation of this axis in a plane. Figure 1 illustrates the model we used to decompose speech-related jaw movements into three components: rotation about a transverse axis located approximately through the condyles (the terminal hinge axis) and the horizontal and vertical translation of this axis in the mid-sagittal plane. We developed this model because we wanted a description of jaw movement that was anatomically accurate. This model is, of course, equivalent to other two-point descriptions of jaw movement as a combination of rotation and translation such as Westbury (1988).

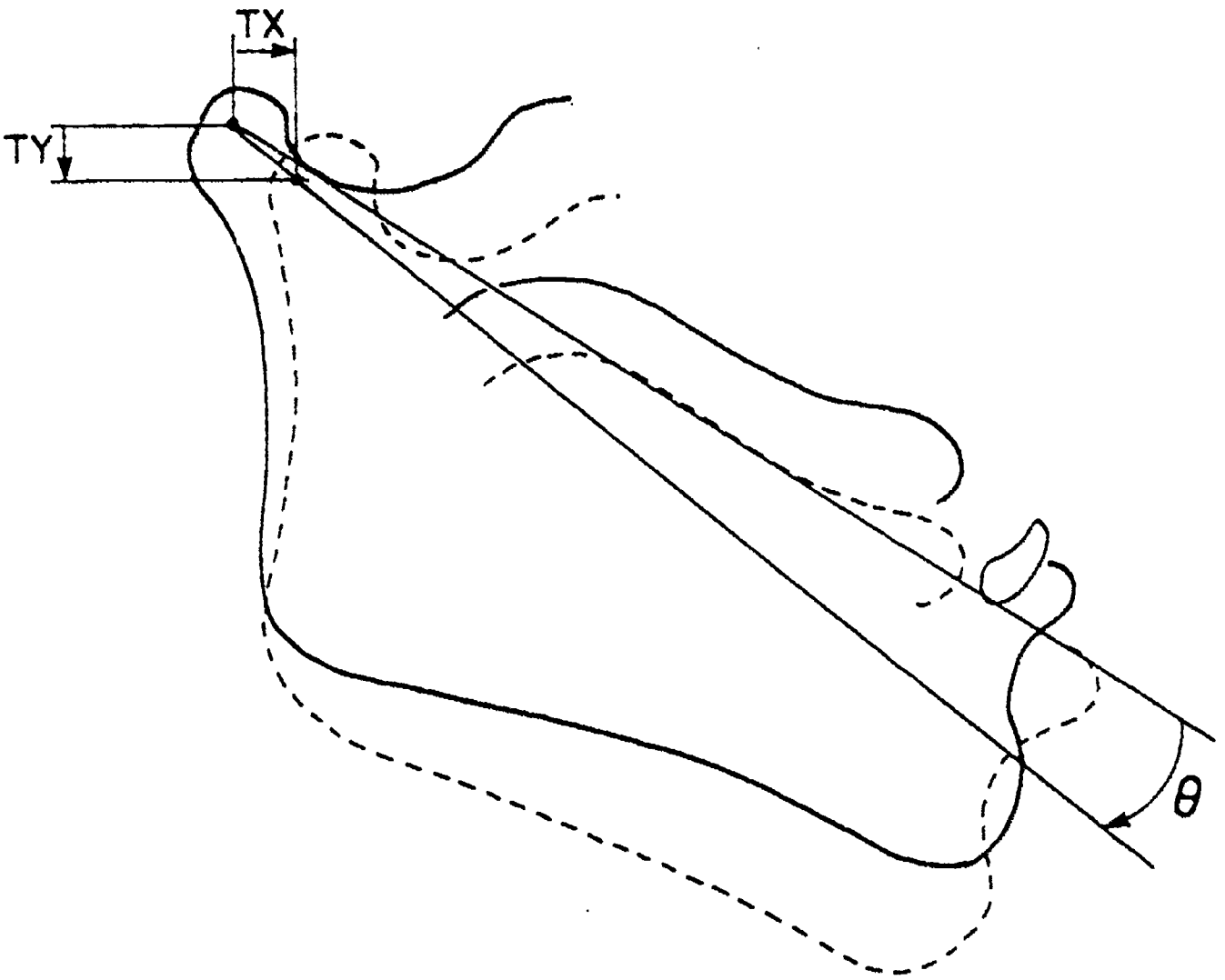


Figure 1. A two-dimensional rigid-body model of jaw movement during speech. Jaw movement is described as a combination of three components: θ, rotation; TX, horizontal translation of the axis of rotation; and TY, vertical translation of the axis of rotation. Solid lines show the rest position of the jaw; dashed lines show a more open position .

Multiple regression analysis was used to examine the relationships among the three components of jaw movement. The results of the multiple regression analyses were used to address two questions: (1) does jaw movement during speech-related opening and closing gestures utilize two or three degrees of freedom; and (2) how are these two (or three) degrees of freedom related to the three jaw movement components?

An accurate model of jaw movement is needed in order to relate jaw movement to jaw muscle activity. Moreover, an accurate model of jaw movement is also needed in order to relate tongue movement to tongue muscle activity. Because the tongue rests on the jaw, tongue movement includes both movement that is due to the tongue muscles and jaw-related movement. Another purpose of this paper was to compare the three models (the pure rotation model, the pure translation model, and our two-point model) with respect to their predictions of the contribution of the first degree of freedom of jaw movement to tongue displacement. This was done in order to determine which of the two simplified models was more accurate, what magnitude of error was introduced, and whether the error was consistent.

Methods

Subjects

The subjects were three normally dentate adult female native speakers of Standard American English (CG, JE, LF). All subjects were screened by a dentist to ensure that they did not exhibit any symptoms of temporomandibular joint disorder and also that they did not exhibit a midline shift during retrusive movements of the jaw. The same dentist, using Angle's (1907) classification determined that two subjects (CG and JE) have Class II occlusions and one subject (LF) has a Class I occlusion. Two of the subjects (CG and LF) were naive to the purpose of the experiment; the third subject (JE) was the experimenter.

Speech materials

The speech materials were 54 V1CV2 utterances. These VCV utterances were placed in a p____p frame for CG and JE and in a t___t frame for LF because it was observed that the jaw appliance (described below) appeared to interfere with bilabial closure for LF. All utterances were embedded in the carrier phrase "a _____ again," All nine combinations of [i], [a], [ae] were used for the V1-V2 context; the intervocalic consonant was a syllable-initial [p], [t], or [s]; lexical stress was placed on either V1 or V2. The speech materials were chosen so that the jaw opening gestures for the vowels could be clearly differentiated from the jaw closing gestures for the consonants.[1]

For CG and JE, the utterances were blocked in groups of six. Within each group, the first vowel (V1) and the intervocalic consonant (C) remained constant. The second vowel (V2) was varied in the order: [i], [a], [ae], [i], [a], [ae]. Within each block of six, primary stress alternated between V1 and V2. Whether V1 or V2 received primary stress on the first utterance within each block was chosen randomly. The order of presentation of the nine blocks of six utterance types was also randomized. Each of the nine blocks was presented to the subject on a 9 by 12 index card. The utterance types were also presented in blocks of six for LF, but the order of presentation of all four phonetic parameters (V1 identity, V2 identity, intervocalic consonant identity, stress pattern) was randomized. Five to seven tokens of each utterance type were produced sequentially. The first five correctly produced tokens were used for analysis.

Appliances

Each subject was individually fitted by a prosthodontist with two appliances. These appliances are illustrated schematically in Figure 2. The reference appliance consisted of a steel wire which was positioned to exit the mouth in the mid-sagittal plane directly between the labial margins of the upper and lower lips. It was bonded directly to an upper front tooth for two subjects (CG and LF) and attached by means of an orthodontic band for the third subject (JE). Two light-emitting diodes (LED's) were attached to this appliance in order to monitor head movement during the course of the experiment. The jaw appliance consisted of three parts: (1) a cast steel plate, molded to fasten onto the labial surfaces of the lateral incisor and the first and second premolars; (2) a cast steel rod which exited the mouth near the corner of the labial margins and another steel rod which extended back to the mid-sagittal plane from the corner of the mouth; and (3) a triangle on which three LED's were positioned. The jaw appliance was bonded directly to the labial surfaces of three lower teeth for all three subjects. The devices were attached by a dentist at least one hour before the data were recorded.

The appliances were designed with three considerations in mind: (1) no interference with intercuspation or terminal hinge position; (2) minimal interference with normal speech production; and (3) maximal stability of the appliance. All goals apparently were achieved.

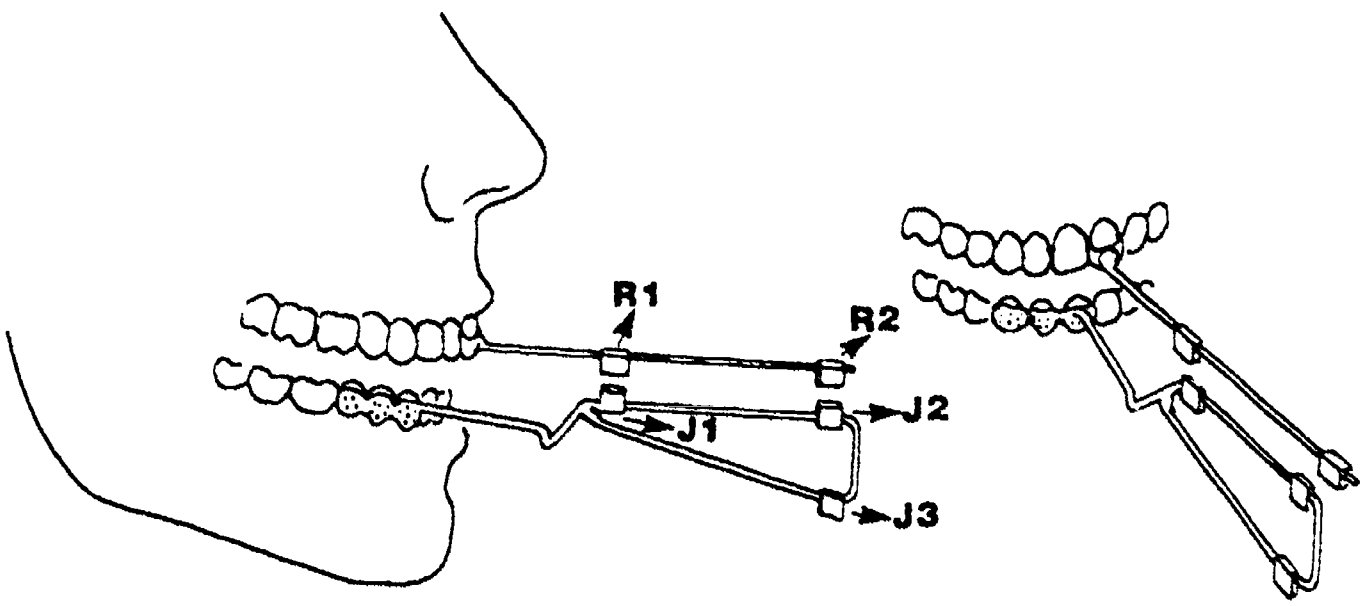


Figure 2. Schematic drawings of the reference and jaw appliances. R1 and R2: reference LEDs to record two points of head movement; J1, J2, J3: jaw LEDs to record three points of jaw movement. A. a sagittal view; B. a frontal view.

First, the subjects reported no interference with intercuspation or terminal hinge position. Second, the subjects reported and other observers noted only minimal interference with normal speech production. However, as mentioned above, the jaw appliance interfered with bilabial closure for LF. Because LF has a Class I occlusion, the distance between her upper and lower teeth in the anterior-posterior direction at intercuspation is smaller than for CG and JE. In order to avoid interference with centric occlusion, the jaw appliance for LF had to be placed somewhat lower on the labial surfaces of the lateral incisor and the first and second premolars. This positioning of the appliance resulted in noticeable interference with the production of bilabials and resulted in a change in the speech materials from pVCVp to tVCVt for this subject (see above). Third, there was no observable slippage of the appliances, which were custom-made to fit onto the labial surfaces of three teeth. The triangle construction in cast steel resulted in a light but stable appliance, with no visually perceptible vibration or yielding during speech.

Data acquisition

Jaw and head movements were recorded by means of an opto-electronic tracking system (Kay, Munhall, Bateson, & Kelso, 1985). A Selspot camera monitored the movement of infra-red light-emitting diodes; decoding electronics associated with the camera derived position data in x and y dimensions and represented them as analog voltages. These electrical signals were recorded on a multi-channel instrumentation tape recorder along with the speech acoustic signal. Calibration was achieved by moving one LED through a known distance (2 cm) in the field of view. The output of the Selspot optical system is linear plus or minus .05 cm for a 20 cm by 20 cm camera field, given a camera distance of 53 cm.

In order to assess intra-speaker variability, a second data recording session was run on all three speakers. The same procedure (including the order

of presentation of the speech material) was followed during the two data recording sessions for each speaker. The period of time between the first and second data recording session was one week, two weeks, and six months for LF, CG, and JE, respectively. A subset of the data from the second recording session (the V1t and the tV2 gestures) was selected for analysis.

Data processing

Both the acoustic and the movement data were digitized on a PDP 11/45 computer; the acoustic signal was sampled at a 10,000 samples per second rate and the movement signals were sampled at a 200 samples per second rate. Both were quantized with 12-bit precision. The simultaneously-recorded acoustic and kinematic waveforms were time-locked via a timing code generator/reader that was interfaced to the computer. Following analog-to-digital conversion, all of the data were transferred to a VAX 11/780 for further processing and analysis. The temporal alignment of the acoustic and kinematic waveforms is accurate within 1 sample (plus or minus 5 ms). The procedures for calibration and the correction for head movement are described in Edwards (1985) and Kay et al. (1985).

Each utterance token was divided into the two opening and the two closing gestures associated with : pV1, V1C, CV2, V2p for CG and JE; tV1, V1C, CV2, V2t for LF. Points of zero-velocity were used to determine onsets and offsets of the opening and closing gestures. Velocities were derived from the jaw displacement data by the application of a central difference algorithm and then smoothed, using a 25 ms smoothing window (Kay, et al., 1985).

Location of terminal hinge axis

In order to locate the terminal hinge axis, a series of non-speech, purely rotational gestures was also recorded for each subject. The terminal hinge position of the jaw is defined as the position in which the mandibular condyles are in their most posterior and superior position in the articular capsule. Most individuals can be taught to open and close their jaw a small amount while maintaining terminal hinge position (Sarnat, 1964). This purely rotational gesture is used by dentists to locate the terminal hinge axis.

Prosthodontists and orthodontists utilize a mechanical device such as a facebow or an adjustable articulator for axis location (Posselt, 1968). The device is attached to the jaw at two points: the lower front teeth and the mandibular condyles. The patient produces a purely rotational gesture and a stylus traces mandibular movement at the point of the condylar attachment. The dentist adjusts the location of the condylar attachment until it is directly on the axis of rotation so that the stylus tracing produces a point rather than a line. This condylar position is taken to be the terminal hinge axis.

A similar procedure was used in this experiment for axis location. except that a computational rather than a mechanical model, was used to find the location of the terminal hinge axis. Each subject was trained to perform a purely rotational maneuver and five of these movements were recorded during each data recording session before and after the recording of the speech data. An iterative optimization procedure (Chambers & Wilks, 1981) was used to fit curves to the selected data points of the purely rotational gestures. (See Edwards, 1985, for a detailed description of this procedure.)

Decomposition of jaw movement

Jaw movement during speech-related gestures was decomposed into rotation and horizontal and vertical translation, using the geometry illustrated in Figure 3. The terminal hinge position of the jaw was defined as the reference position. The angle Φ was defined as the angle that the line OJ1 made with the horizontal. The distance D was defined as the Euclidean distance between the point O (xo, yo) and the point J1 (x_{j1}, y_{j1}). The sine and cosine of Φ and the distance D were calculated using the previously determined coordinates of the axis of rotation (O) and the reference coordinates of one jaw LED (J1). Jaw rotation was defined as the angle of rotation θ (in degrees) formed by the two line segments J1-J2 and J1'-J2'. The sine and cosine of θ were calculated, using the reference and the new coordinates of two jaw LEDs (J1 and J2). Then, the sine and cosine of the new angle Φ' ($\theta + \Phi$) that the line OJ1 made with the horizontal was calculated. The new location of the axis of rotation (O') was calculated, using the angle Φ', the distance D (assumed to be constant), and the coordinates of point J1'. Finally, the x and y components of jaw translation (TX and TY), defined as the horizontal and vertical vectors from O to O', were calculated. The output of this procedure was a frame-by-frame description of jaw movement as a combination of three components: θ, the rotation in degrees about a fixed hinge axis relative to the reference position; TX, the horizontal translation of the terminal hinge axis; and TY, the vertical translation of the terminal hinge axis.

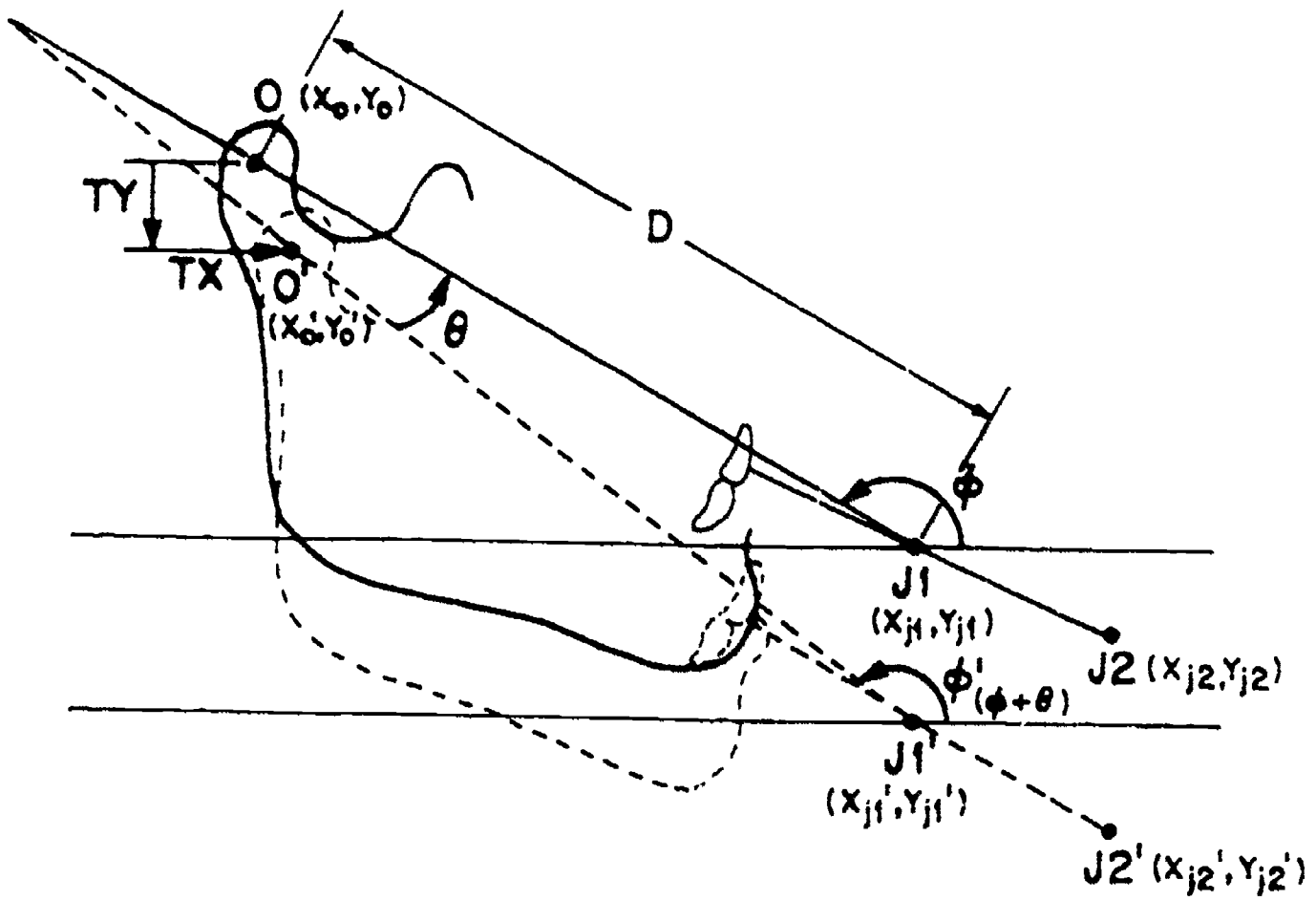


Figure 3. The geometry used for decomposition of jaw movement. Solid lines show the reference position of the jaw; dashed lines show the position of the jaw at some data frame.

Data analysis

Coordinate transformation. Multiple regression analysis was used to examine the relationships among the three components of jaw movement. First, however, a coordinate transformation was performed. For each data subset on which a multiple regression analysis was performed, the data points were rotated so that the first principal component of jaw translation was parallel to the x axis.

The coordinate transformations were performed for two reasons: first, so that the error terms in the regression analysis would be calculated using perpendicular rather than vertical distances to the best-fitting lines; and second, to permit meaningful discussion of inter-speaker differences by using coordinate systems which were defined with respect to the same functional criterion for all subjects. All subsequent discussion refers to the three components of jaw movement (TX, TY, θ) within the transformed coordinate systems.

Multiple regression analysis. Multiple regression analysis was used to determine whether any of the three components of jaw movement exhibited a functional dependence on any other component. Two functional relationships were considered as possibilities: (1) translation and rotation might be functionally interdependent; and (2) TX and TY, the first two principal components of jaw movement might be functionally interdependent. Therefore, two multiple regression analyses were performed. First, θ, the angle of rotation was analyzed as a function of TX, the first principal component of translation (hereafter, rotation analysis). Second, TY, the second principal component of jaw translation was analyzed as a function of TX (hereafter, translation analysis). TX was taken as the independent

variable in both analyses so that the results of the two analyses could be more easily compared. A disadvantage of this decision is that it leaves one pertinent question at least partially unresolved: i.e., can jaw rotation predict TX. This issue will be returned to below.

A quadratic model was used for the rotation analysis and a cubic model was used for the translation analysis for all three speakers. The regression equations for the rotation and translation analyses are given in (1) and (2) below, respectively.

(1) $\hat{\theta} = a_1X_1 + a_2X_1^2 + a_3$

(2) $\hat{Y} = b_1X_1 + b_2X_1^2 + b_3X_1^3 + b_4$.

The degree (i.e., the polymonial order of the equation) for each model was chosen by determining the highest degree for each model that appeared to fit the general shape of the data rather than specific data points for a given speaker. The same degree was used for the data of all three speakers so that results could be compared across speakers. For example, because a cubic model provided the best fit for the translation model for the data of subject JE, a cubic model was also used for the translation model for the data of all three subjects, although a quadratic model might have been adequate for fitting the data of CG and LF. The use of a higher order model than necessary will simply result in insignificant contributions to the squared multiple correlation of the higher order terms.

The data of the four gestures (pV1, V1C, CV2, V2p for CG and JE; tV1, V1C, CV2, V2t for LF) were analyzed separately. For each of the four gestures, the data were combined across three of the four phonetic parameters: (1) the stress pattern of the test syllable (either primary ("stressed") or secondary ("unstressed")); (2) the identity of the vowel in the test syllable ([i], [a], or [ae]): (3) the identity of the vowel in the non-test syllable (also [i], [a], or [ae]). The fourth phonetic parameter was the identity of the intervocalic consonant ([t], [p], or [s]). The data were analyzed separately for each intervocalic consonant because we found that combining across consonant identity resulted in substantially lower squared multiple correlations. The three phonetic parameters were coded as binary-valued covariates. In those cases for which a phonetic parameter could take on three values (e.g., the identities of the test and non-test vowels), two binary covariates were used. The regression equations used for the rotation and translation models for these analyses are given in (3) and (4) below, respectively.

(3) $\hat{\theta} = a_1X_1 + a_2X_1^2 + a_3X_2 + a_4X_3 + a_5X_4 + a_6X_5 + a_7X_6 + a_8$

(4) $\hat{Y} = b_1X_1 + b_2X_1^2 + b_3X_1^3 + b_4X_2 + b_5X_3 + b_6X_4 + b_7X_5 + b_8X_6 + b_9$

X and X3 specify test vowel identity; X4 and X5 specify non-test vowel identity; X6 specifies the stress pattern.

The effect of one additional covariate was also examined for subjects CG and JE because the order of presentation of utterance types was not fully randomized for these two subjects. It was assumed that the blocking of the utterance types would have no measurable effect on the relationships among the components of jaw movement. This assumption was tested by treating position within a block as an additional covariate with six values, each corresponding to a position between one and six. This covariate was added to the rotation and translation analyses for subjects JE and CG for the V1C and the CV2 gestures. Since the proportion of the variance accounted for by this variable was quite small (from 0 to 3%), it will be assumed that blocking the utterance types did not have a significant effect on the relationships among the three components of jaw movement.

Contributions of the three jaw movement components to resultant jaw displacement at the front teeth. In order to make quantitative comparisons among the three speakers, the amount of movement due to X and Y translation and the movement due to jaw rotation at a selected point on the mandible were calculated for the maximal opening position for the low vowels [a] and [ae] for the V1t and tV2 gestures for each speaker. The movement due to rotation was calculated using a straight line segment to approximate the arc which the angle subtended. The movement due to rotation will be referred to as "R."

Jaw opening at the front teeth was assumed to be the measurement of primary interest for speech production. Therefore, the radius was defined as the distance from the axis of rotation (O) (cf. Figure 3) to a lower front tooth. The x and y coordinate values of a point on a lower front tooth were calculated by measuring the distance between LED J1 (cf. Figure 3) and the tooth for each speaker and then extending the line segment connecting LED's J1 and J2 (cf. Figure 3) by this measured distance. All subsequent references to jaw displacement refer to the displacement of this point.

Projections of three jaw movement components onto resultant jaw displacement. These measurements were used tc calculate resultant jaw displacement at maximal opening for the V1t gestures. The three jaw movement components were added vectorially. In order to compare the relative contributions of the three jaw movement components, the projection of each component onto resultant jaw displacement was also calculated.

Jaw component of tongue displacement. The results of the translation and rotation analyses were used to determine which of the three jaw movement components (or combination thereof) corresponded to the first degree of freedom of jaw movement for each speaker. This empirically-determined model of jaw movement was then used to calculate the contribution of the first principal component of jaw movement to mid-tongue position at maximal jaw opening.

The predictions of the empirically determined model were compared to the predictions of the simplified pure rotation and pure translation models. Given the pure translation model, the jaw component of mid-tongue displacement is equal to the displacement of the lower front tooth along its first principal axis. Given the pure rotation model of jaw movement, the jaw component was calculated by multiplying total jaw displacement along its first principal axis by an appropriate proportional fraction, as shown in Figure 4. Given the relative positions of the mid-tongue and jaw pellets in data acquired with the Tokyo X-ray microbeam system (Kiritani, Itoh & Fujimura, 1975), it was estimated that approximately 60 percent of jaw rotation will be reflected in mid-tongue position. The predictions of the three models were compared by calculating the errors that the two simplified models introduce.

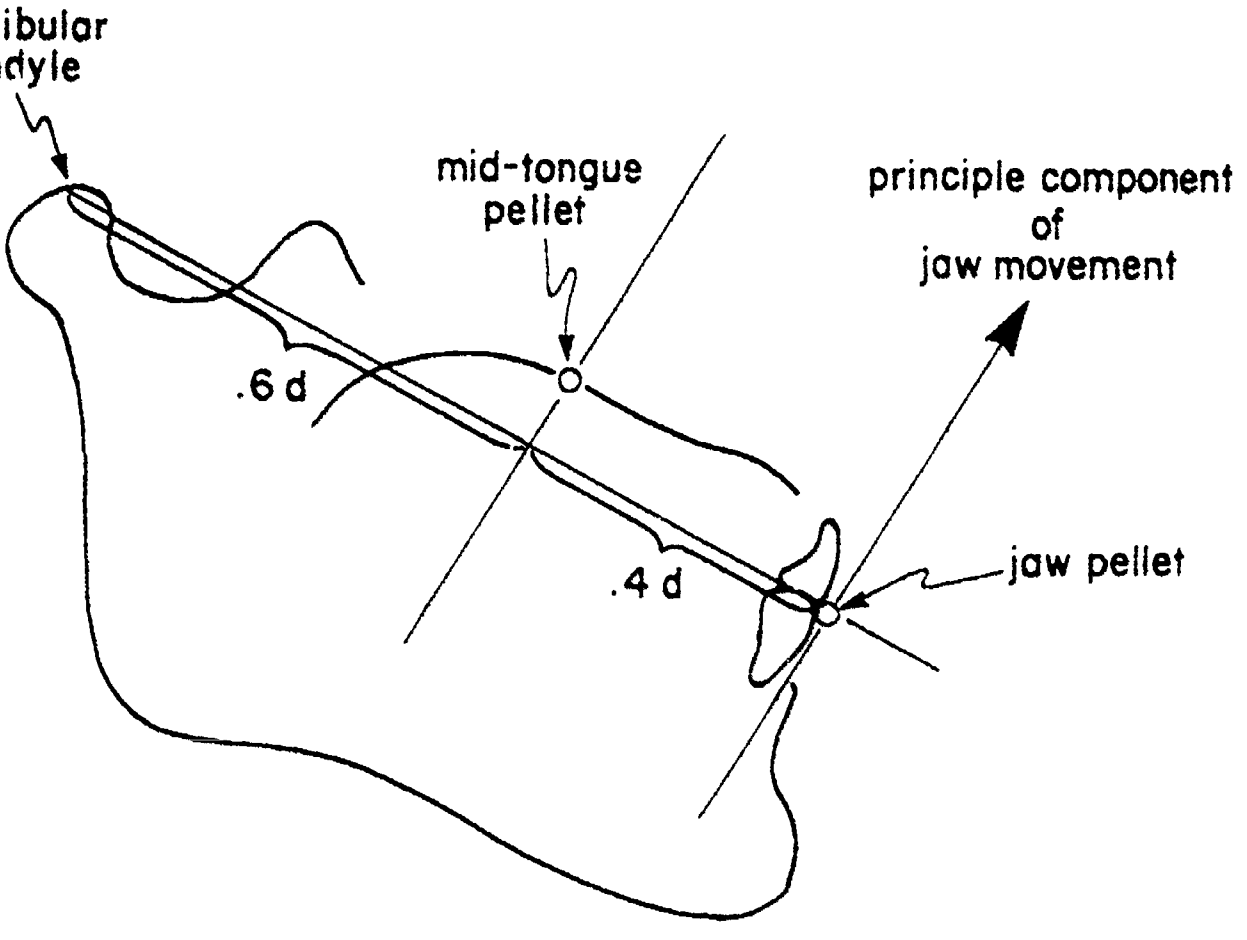


Figure 4. The jaw component of mid-tongue displacement, according to the pure rotation model. The displacement of the lower front tooth along its principal axis is multiplied by a proportional fraction. The proportional fraction is equal to the ratio of the distance of the mid-tongue pellet tc the axis of jaw rotation relative to the distance of the jaw pellet to the axis of rotation.

Results

Quantitative Relationships Among the Three Jaw Movement Components

The results for the three speakers are grossly similar in that the displacement of the three components of jaw movement was in the predicted direction for both opening and closing gestures. For all three speakers, the center of rotation moved down and front for jaw opening and moved up and back for jaw closing. Similarly, for all three speakers, the angle of jaw rotation became more open for jaw opening and less open for jaw closing. For all three speakers, rotation was the jaw movement component that contributed most to resultant jaw displacement at the front teeth. Another similarity among the three speakers was that the amount of resultant jaw displacement at the front teeth was generally greater for low vowels, as compared to high vowels, and for stressed vowels, as compared to unstressed vowels. The three speakers differed in that the amount of resultant jaw displacement was generally greatest for JE and least for LF.

Displacement of the three jaw movement components

Table 1 presents the mean displacements of the three jaw movement components from the minimum jaw position for V1 to the maximum jaw position for [t] for the vowels [a] and [ae] for all three speakers. The data for the other gestures showed similar patterns. The data for [i] are not included because all three speakers exhibited small and quite variable amounts of jaw displacement for this vowel. For the presented vowels, both similarities and differences among the three speakers can be observed. The speakers are similar in that all three exhibit greatest displacement of R and least displacement of TY. The speakers differ, however, in the amount of displacement of the three jaw movement components. JE consistently exhibits the greatest displacement of all three components; LF exhibits the least amount of TX; and CG exhibits a greater amount of TX than LF and roughly comparable amounts of TY and R.

Projections of the three jaw movement components onto resultant jaw displacement

The same pattern of inter-speaker similarities is observed for the projections of the three jaw movement components onto resultant jaw displacement, shown in Table 2. Again, all three speakers exhibit greatest displacement for the projection of R and least displacement for the projection of TY. The same pattern of inter-speaker differences is also observed, although the size of the inter-speaker differences decreases. JE exhibits the greatest displacements for the projections of all three components; LF exhibits the smallest displacement for the projection of TX; and CG exhibits greater displacement for the projection of TX than LF and roughly comparable displacements for the projections of TY and R.

Table 1. *Mean displacements (in mm) of the three jaw movement components for the V1t gestures (second recording session in parentheses).*

		CG			JE			LF		
vowel	stress	TX	TY	R	TX	TY	R	TX	TY	R
a	+	5.9 (3.6)	1.8 (1.6)	7.3 (6.2)	13.4 (6.1)	3.5 (2.1)	15.3 (15.1)	2.5 (2.4)	1.7 (1.3)	6.1 (9.2)
a	-	5.0 (2.7)	1.9 (1.4)	5.9 (5.1)	4.7 (3.9)	2.1 (2.0)	7.4 (9.7)	2.5 (1.8)	0.9 (1.0)	4.5 (6.2)
ae	+	6.0 (4.1)	2.0 (1.3)	7.4 (6.8)	14.4 (5.1)	4.2 (1.6)	18.3 (13.0)	3.4 (2.9)	1.8 (1.3)	8.4 (10.8)
ae	-	4.8 (3.8)	2.1 (1.2)	6.0 (6.8)	6.8 (4.5)	2.7 (2.0)	10.4 (12.2)	2.5 (1.7)	1.2 (0.8)	6.6 (8.6)

Note. TX = horizontal jaw translation; TY = vertical jaw translation; R = jaw rotation.

Table 2. *Mean projections (in mm) of the three jaw movement components onto resultant jaw displacement for the V1t gestures (second recording session in parentheses).*

		CG			JE			LF		
vowel	stress	TX	TY	R	TX	TY	R	TX	TY	R
a	+	4.5 (3.0)	1.2 (0.9)	6.9 (6.1)	6.0 (1.3)	3.2 (2.1)	11.3 (14.0)	1.7 (1.9)	1.4 (0.8)	6.1 (9.2)
a	-	3.8 (2.2)	1.2 (0.8)	5.6 (5.1)	1.4 (0.6)	2.0 (2.0)	6.2 (9.2)	1.8 (1.5)	0.6 (0.5)	4.4 (6.3)
ae	+	4.6 (3.4)	1.3 (0.7)	7.0 (6.7)	5.7 (1.0)	3.8 (1.6)	14.2 (12.2)	2.3 (2.4)	1.3 (0.8)	7.4 (10.8)
ae	-	3.6 (3.2)	1.4 (0.7)	5.8 (6.8)	2.0 (0.7)	2.6 (2.1)	8.7 (11.5)	1.7 (1.4)	0.9 (0.5)	6.6 (8.6)

Note. TX = horizontal jaw translation; TY = vertical jaw translation; R = jaw rotation.

Relative contributions of the three jaw movement components

The relative contributions of the three jaw movement components to resultant jaw displacement are presented graphically in Figure 5 and in Table 3 for the V1t gestures. For all three speakers, R consistently contributes most to resultant jaw displacement. For CG and LF, TY consistently contributes least to resultant jaw displacement, whereas for JE the contributions of TX and TY to resultant jaw displacement are more equal. It is not surprising that differences among the three speakers were observed in the absolute and relative contributions of the three jaw movement components, given that speakers are known to differ considerably in the amount of jaw movement during speech. The results of the second recording session revealed significant intra-speaker differences as well.

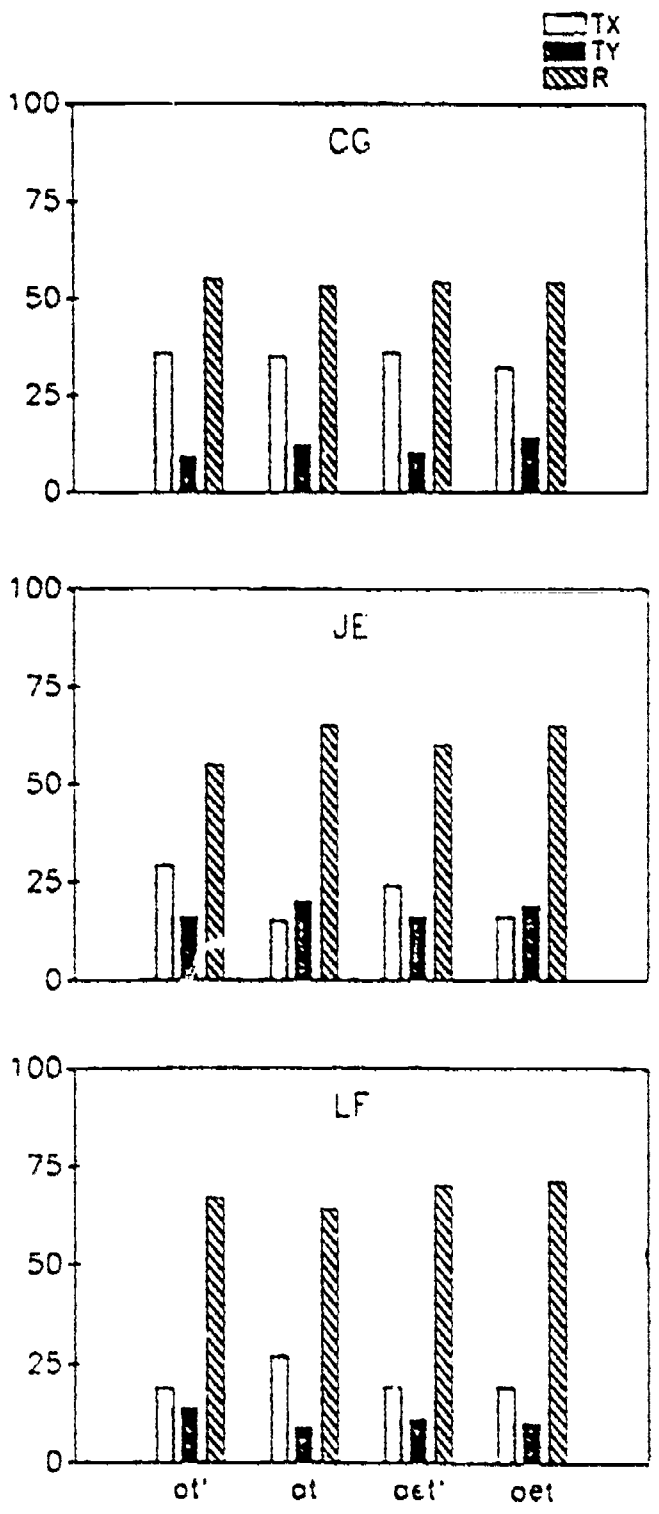


Figure 5. **Bar plots of the relative contributions (in percent) of the three jaw movement components to resultant jaw displacement for the V1t gestures in the first recording session.**

Table 3. *Relative contributions (in %) of the three jaw movement components to resultant jaw displacement for the V1t gestures (second recording session in parentheses).*

		CG			JE			LF		
vowel	stress	TX	TY	R	TX	TY	R	TX	TY	R
a	+	36 (30)	09 (09)	55 (61)	29 (10)	16 (09)	55 (81)	19 (16)	14 (07)	67 (77)
a	-	35 (28)	12 (10)	53 (62)	15 (05)	20 (17)	65 (78)	27 (18)	09 (07)	64 (75)
ae	+	36 (32)	10 (06)	54 (62)	24 (07)	16 (11)	60 (82)	19 (17)	11 (05)	70 (77)
ae	-	32 (30)	14 (06)	54 (64)	16 (05)	19 (15)	65 (80)	19 (13)	10 (05)	71 (82)

Note. TX = horizontal jaw translation; TY = vertical jaw translation; R = jaw rotation.

Intra-speaker differences

A second data-recording session was run on all three speakers so that the issue of within-speaker variability could be addressed. Speakers have been found to be quite variable with respect to the displacements of individual articulators for a particular speech segment within an experiment even if the phonetic context is held constant (Abbs, 1983; Ostry & Munhall, 1985; Vatikiotis-Bateson, 1988). Similar results were found in this study across two experiments. To our knowledge, there is no other published data on intra-speaker variability in articulator displacement across different recording sessions.

First, significant differences in the displacements of the three jaw movement components were observed for all three speakers. These data are shown in parentheses in Table 1. A comparison of these data with the parallel observations from the first recording session reveals different patterns of intra-speaker differences for each subject. CG exhibits significantly less displacement of all three jaw movement components in the second data recording session, as compared to the first: $t = 12.01$, $p < 0.00001$ for TX, $t = 5.12$, $p < 0.00001$ for TY; $t = 2.59$, $p < 0.05$ for R. JE exhibits significantly less displacement of TX and TY and no significant differences for R in the second data recording session, as compared to the first: $t = 8.27$, $p < 0.00001$ for TX; $t = 6.37$, $p < 0.00001$ for TY; $t = 1.71$, $p > 0.10$ for R. LF exhibits no significant differences for TX, significantly less displacement of TY, and significantly more displacement of R in the second data recording session, as compared to the first: $t = -.63$, $p > 0.10$ for TX; $t = 3.78$, $p < 0.01$, $t = -8.89$, $p < 0.00001$ for R.

The projections of the three jaw movement components onto resultant jaw displacement at the front teeth for the second recording session are shown in parentheses in Table 2. Again, a comparison of these data with their counterparts from the first data recording session reveals differences among the three speakers. CG exhibits significantly smaller projections of TX and TY onto resultant jaw displacement and no significant differences in R in the second data recording session, as compared to the first: $t = 4.21$, $p < 0.01$ for TX; $t = 5.20$, $p < 0.001$ for TY; $t = .38$, $p > 0.10$ for R. JE exhibits significantly smaller projections of all three jaw movement components onto resultant jaw displacement in the second data recording session, as compared to the first: $t = 4.97$, $p < 0.001$ for TX; $t = 3.63$, $p < 0.01$ for TY; $t = 2.34$, $p < 0.05$ for R. LF exhibits no significant differences for TX, significantly smaller projections of TY, and significantly greater projections of R onto resultant jaw displacement in the second data recording session, as compared to the first: $t = .42$, $p > 0.10$ for TX; $t = 3.38$, $p < 0.01$ for TY; $t = -7.53$, $p < 0.00001$ for R.

Significant intra-speaker differences were also observed with respect to the relative contributions of the three jaw movement components to resultant jaw displacement. These data are shown in parentheses in Table 3. These data were compared with the parallel observations from the first data recording session, with percentages transformed into arcsine units in order to stabilize the order variance (Brownlee, 1965). All three speakers exhibited significantly smaller relative contributions of TX and TY and significantly greater relative contributions of R to resultant jaw displacement in the second data recording session, as compared to the first. For CG, $t = 25.77$, $p < 0.00001$ for TX; $t = 15.4$, $p < 0.00001$ for TY; $t = -44.95$, $p < 0.00001$ for R. For JE, $t = 39.58$, $p < 0.00001$ for TX; $t = 9.23$, $p < 0.00001$ for TY; $t = -4.33$, $p < 0.01$ for R. For CG, $t = 30.88$, $p < 0.00001$ for TX; $t = 3.57$, $p < 0.01$ for TY; $t = -38.93$, $p < 0.00001$ for R.

Thus, intra-speaker as well as inter-speaker differences are observed. However, in spite of these quantitative differences within and across speakers, the multiple regression analysis revealed consistent similarities between CG and JE, as compared to LF.

Qualitative Relationships Among the Three Jaw Movement Components

As discussed above, multiple regression analysis was used to examine whether any of the three jaw movement components exhibited a functional dependence on any other component. In the rotation analysis, the angle of rotation was analyzed as a function of TX. In the translation analysis, TY was analyzed as a function of TX.

Rotation analysis

Table 4 gives the squared multiple correlations for all of the rotation analyses for all three speakers.[2] Both CG and JE exhibited a strong functional interdependence between R and TX, as indicated by the high squared multiple correlations for these two subjects. By contrast, LF did not consistently exhibit a functional interdependence between R and TX, as indicated by the low squared multiple correlations for this subject. These patterns of inter-speaker differences were preserved over the two recording sessions for each speaker, in spite of the significant intra-speaker differences described above. In both recording sessions, high squared multiple correlations are observed consistently for CG and JE and low squared multiple correlations are observed for LF.

The consistently high squared multiple correlations for CG and JE suggest that R and TX are functionally constrained to operate as a single degree of freedom during opening gestures for vowels and closing gestures for [t], [p], and [s] for both subjects. However, an examination of the regression coefficients, shown in Table 5 for CG and JE for the V1C and the CV2 gestures, indicates that the picture is somewhat more complicated. The regression coefficients are given for the linear component of the primary independent variable, TX. Pairwise tests of parallelism were performed to compare regression coefficients for the opening and closing gestures with consonant identity held constant (e.g., the regression coefficients of the V1t and the tV2 gestures were compared) and to compare regression coefficients for the different consonants with gesture type held constant (e.g., the regression coefficients of the V1t and the V1p gestures were compared). For both subjects, regression coefficients for the opening gestures were significantly different than regression coefficients for closing gestures. Furthermore, regression coefficients for both closing and opening gestures to and from different consonants were also significantly different. The fact that all pairwise tests of parallelism were significant is due, in part, to the large number of data points: it can be observed that the regression coefficients of CG, as compared to those of JE, are more clearly differentiated across the different consonants and across the opening and closing gestures. Nevertheless, these results suggest that even though TX and R are functionally interdependent for both CG and JE for the gestures under consideration, the precise nature of this relationship may vary with the phonetic context.

Table 4. *Squared multiple correlations for θ regressed on TX.*

		CG		JE		LF	
	Consonant	Session 1 r^2	Session 2 r^2	Session 1 r^2	Session 2 r^2	Session 1 r^2	Session 2 r^2
pV1	t	.67		.96		.20	
pV1	p	.74		.98		.28	
pV1	s	.62				.38	
V1C	t	.75	.85	.96	.92	.23	25
V1C	p	.84		.97		.21	
V1C	s	.56				.22	
CV2	t	.82	.90	.96	.90	.37	.23
CV2	p	.87		.98		.19	
CV2	s	.58				.07	
V2P	t	.83		.98		.45	
V2P	p	.80		.98		.55	
V2P	s	.65				.39	

Table 5. *Regression coefficients for the linear component of TX for the V1C and the CV2 gestures.*

	CG	JE
V1t	-3.21	-6.07
V1p	8.39	-6.25
V1s	-1.52	
tV2	-4.62	-4.97
pV2	-9.20	-5.20
sV2	-3.53	

Translation analysis

Table 6 gives the squared multiple correlations for the translation analysis for all three subjects for the two recording sessions. In contrast to the rotation analysis, the translation analysis did not reveal a consistent relationship between TX and TY for any of the three subjects across the two recording sessions. This can be observed in the low squared multiple correlations for all three subjects for the second recording session and the low squared multiple correlations for LF for the first recording session as well. In the first recording session, the squared multiple correlations for both CG and JE are relatively high for some of the utterance types. However, this relationship is not consistent across all the utterances, nor across the two recording sessions. The substantially lower squared multiple correlations for CG and JE in the second recording session as compared to the first are probably due to the significantly smaller amounts of translation in that session for these two subjects.

Decomposition of tongue position: A comparison of three models

Another purpose of this study was to compare the differences among three models of jaw movement—our two-point model, the pure translation model, and the pure rotation model in predicting the contribution of the first degree of freedom of jaw movement to tongue displacement. Table 7 presents the results of using each of these three models to calculate the contribution of the first degree of freedom of jaw movement to mid-tongue position at maximal jaw opening. The results are averaged across low vowels with primary stress.

Table 6. *Squared multiple correlations for TY regressed on TX.*

		CG		JE		LF	
	Consonent	Session 1 r^2	Session 2 r^2	Session 1 r^2	Session 2 r^2	Session 1 r^2	Session 2 r^2
pV1	t	.63		.42		.39	
pV1	p	.58		.67		.25	
pV1	s	.64				.35	
V1C	t	.51	.14	.41	.10	.38	.25
V1C	p	.70		.72		.26	
V1C	s	.58				.21	
CV2	t	.80	.28	.41	.11	.38	.12
CV2	p	.72		.70		.19	
CV2	s	.67				.20	
V2P	t	.80		.39		.51	
V2P	p	.69		.76		.21	
V2P	s	.59				.24	

Table 7. *Contribution of jaw displacement to mid-tongue displacement: Predictions of three models.*

	CG				JE				LF			
	Session 1		Session 2		Session 1		Session 2		Session 1		Session 2	
	V1t	tV2	V1t	tV2	V1t	tV2	V1t	tV2	V1t	tV2	V1t	tV2
Rotation & TX (mm)	9.1	9.9	7.3	7.8	14.8	15.2	9.3	9.6				
Pure Rotation (mm)	6.7	7.5	5.9	6.2	11.2	11.7	8.5	8.8	4.4	4.7	6.0	7.7
Pure Translation (mm)	11.6	12.5	9.8	10.4	18.6	19.5	14.2	14.6	7.3	7.9	10.0	12.8
Rotation Error (%)	26.0	24.0	19.0	21.0	24.0	23.0	9.0	8.0	0.0	0.0	0.0	0.0
Translation Error (%)	27.0	26.0	34.0	33.0	26.0	28.0	53.0	52.0	40.0	40.0	40.0	40.0

The predictions of the pure rotation and the pure translation models were calculated as described above. For our two-point model, the results of the regression analyses were used to determine what combination of the three components of jaw movement corresponded to the first degree of freedom of jaw movement for each subject. For CG and JE, the first degree of freedom of jaw movement is a combination of R and TX, since the rotation analysis revealed the functional interdependence of these two components. For LF, the first degree of freedom of jaw movement is R, since the regression analyses did not reveal a consistent functional interdependence between R and TX or between TX and TY. Therefore, for CG and JE, the contribution of the first degree of freedom of the two-point model was calculated by vectorial summation of X translation and 60 percent of jaw rotation. For LF, because the first degree of freedom of jaw movement corresponded to jaw rotation alone, the pure rotation model was used.

The results for LF are quite straightforward. Because the first degree of freedom of jaw movement corresponds to jaw rotation, no error was introduced by using the pure rotation model to calculate the contribution of the first principal component of jaw movement to tongue displacement. The errors introduced by using the pure translation model are, of course, always 40 percent of the predicted contribution, using the pure rotation model.

For CG and JE, the simplified models will result in two errors: an error in magnitude and an error in orientation. Because the jaw component of mid-tongue displacement contains a proportional fraction of jaw rotation, the principal component of jaw movement measured at the front teeth is not parallel to the principal component of jaw movement measured at mid-tongue. The errors in orientation ranged from 4 to 7 degrees for CG and from 9 to 15 degrees for JE across the two recording sessions. The errors in magnitude ranged from 8 to 53 percent of the contribution that was predicted by the combined rotation and translation model. In the first recording session, the differences between the two simplified models were quite small, although the pure rotation model was slightly more accurate for both speakers. For CG, the errors introduced by the pure rotation model averaged 24 to 26 percent of the contribution that was predicted by the combined rotation and translation model; the errors introduced by the pure translation model averaged 26 to 27 percent of the predicted contribution. For JE, the errors introduced by the pure rotation model averaged 23 to 24 percent; the errors introduced by the pure translation models averaged 26 to 28 percent.

In the second recording session, the predictions of the pure rotation model were systematically closer to the predictions of the combined rotation and translation model for these two speakers. For CG, the errors introduced by the pure rotation model averaged 19 to 21 percent of the contribution that was predicted by the combined rotation and translation model; the errors introduced by the pure translation model averaged 33 to 34 percent. For JE, the errors introduced by the pure rotation model averaged 8 to 9 percent of the contribution predicted by the combined rotation and translation model; the errors introduced by the pure translation model averaged 52 to 53 percent. Because the relative contributions of R and TX varied substantially across the two subjects, and across the two data recording sessions for each subject, it is not possible to develop a simple method to correct the errors introduced by using the pure rotation model for CG and JE.

DISCUSSION

Although jaw movement during speech has generally been modelled as a single degree of freedom system, the anatomy and physiology of the temporomandibular joint suggest that during opening and closing speech-related gestures, the jaw can move with up to three independent degrees of freedom. This study developed a more complex model of jaw movement as a combination of rotation about the terminal hinge axis and the simultaneous vertical and horizontal translation of that axis. The results of this study show clearly that the jaw both rotates and translates during opening gestures for vowels and closing gestures for [t], [p], and [s]. The anterior-inferior translation of the condyle during opening gestures is consistent with reports of activity of the inferior head of the lateral pterygoid for vowels (e.g., Tuller, Harris, & Gross, 1981). Posterior-superior translation of the condyle during closing gestures is consistent with reports of activity of the superior head of the lateral pterygoid for [t] and [p] (Tuller, et al., 1981). Since the lateral pterygoid is the only jaw muscle that attaches to the articular capsule and disc of the mandibular condyle, translation of the terminal hinge axis (i.e., movement of the articular disc) must be due to activity of this muscle.

The results of this study agree with those of Westbury (1988) in showing that a more complex

description of jaw movement during speech is necessary. However, this model is complicated since it requires the computation of the location of the terminal hinge axis and the derived trajectories for the translation of this axis. Westbury (1988) has developed an equivalent two-point model in which jaw movement is described as the angle formed by the intersection of the extensions of the maxillary occlusal plane and a line segment passing through the mandibular incisors and the horizontal and vertical translation of a point on the mandibular incisors. Westbury's model is simpler since two points of jaw movement on the incisors can be recorded directly and for most purposes it is equivalent. Let us consider some uses of any such two-point rigid body description of jaw movement.

Such a model can be used to determine the number of functional degrees of freedom of jaw movement during speech for a given speaker. (The model used in this study can also be used to relate these functional degrees of freedom to their anatomical components.) For example, in this study there were two speakers, CG and JE, who have two functional degrees of freedom of jaw movement during speech: the first degree of freedom corresponds to a combination of R and TX; the second degree of freedom corresponds to TY. For a third speaker, LF, three functional degrees of freedom of jaw movement during speech were observed: the first degree of freedom corresponds to R; the second degree of freedom corresponds to TX; the third degree of freedom corresponds to TY.

A model of jaw movement as a combination of rotation and translation is also needed in order to relate tongue muscle activity to tongue movement. Because the tongue rests on the jaw, tongue movement can be decomposed into movement that is due to the tongue muscles and movement that is jaw-related. This problem is becoming more important as the development of the X-ray microbeam system and the magnotometer make it possible for researchers to collect large quantities of tongue movement data. These results of this study show that using either a pure rotation or a pure translation model of jaw movement to estimate the contribution of the first degree of freedom of jaw movement to mid-tongue position will be inaccurate for some subjects. Of the two simplified models, the pure rotation model was more accurate across different subjects and across different experimental sessions. However, using the pure rotation model to estimate the contribution of the first principal component of jaw movement to tongue displacement will introduce errors both in magnitude and in orientation for some subjects.

A more accurate description of jaw movement as a combination of rotation and translation, may also provide insight into patterns of differences and similarities in mandibular movement across speakers. In this study, the similarities between CG and JE, as compared to LF, are of particular interest because they were observed in the face of quantitative inter-speaker differences among all three subjects and were preserved across two separate recording sessions. It is possible that these differences in jaw behavior are related to a structural difference between CG and JE, as compared to LF: CG and JE have Class II occlusions; LF has a Class I occlusion. It has been widely observed by dentists that speakers of different occlusal classes show systematic differences in their speech-related jaw movements (e.g., Pound, 1977). These differences are used by prosthodontists to determine the occlusal class of edentulous patients. (It should be noted that descriptions of these occlusal-dependent differences for speech are derived from visual observation and have not, to our knowledge, been studied quantitatively.) Dentists have observed that individuals with Class II occlusions generally exhibit relatively large amounts of anterior-posterior movement for opening and closing gestures into and out of alveolar consonants; individuals with Class I occlusions generally exhibit relatively small amounts of anterior-posterior movement in the same phonetic contexts; and individuals with Class III occlusions generally exhibit virtually no anterior-posterior movement in the same phonetic contexts. The results of this study are consistent with the occlusal class differences reported in the dental literature. CG and JE, with Class II occlusions, exhibited more anterior-posterior translation (TX) than LF and an interdependence between jaw rotation and TX. By contrast, LF, with a Class I occlusion, exhibited relatively little jaw translation and no functional relationship between jaw rotation and jaw translation.

REFERENCES

Abbs, J. H. (1983). Invariance and variability in speech production: A distinction between linguistic intent and its neuromotor implementation. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 202-218). Hillsdale, NJ: Lawrence Erlbaum Associates.

Angle, E. H. (1907). *Malocclusion of the teeth*. Philadelphia: S. S. White Dental Manufacturing Co.

Brownlee, K. A. (1965). *Statistical theory and methodology in science and engineering*. New York: John Wiley & Sons.

Chambers, J. M., & Wilks, A. R. (1981). Nonlinear model fitting. In AT&T Bell Laboratories Technical Memorandum. Murray Hill, NJ.

Coker, C. H. (1976). A model of articulatory dynamics and control. *Proceedings of the IEEE, 64*, 452-460.

Edwards, J. (1985). *Mandibular rotation and translation during speech.* Unpublished doctoral dissertation, CUNY.

Gentil, M., & Gay, T. (1986). Neuromuscular specialization of the mandibular motor system: Speech versus non-speech gestures. *Journal of Speech Communication, 5*, 69-82.

Gibbs, C. H., & Messerman, T. (1972). Jaw motion during speech. In *Orofacial function: Clinical research in dentistry and speech pathology*. (ASHA Reports, No. 7). Washington, DC: American Speech and Hearing Association.

Gibbs, C. H., Messerman, T., Reswick, J. B., & Derda, H. J. (1971). Functional movements of the mandible. *Journal of Prosthetic Dentistry, 26*, 604-620.

Hjortsjo, C., (1955). *Studies on the mechanics of the temporo-mandibular joint.* Lund, Sweden: C. W. K. Gleerup.

Kakita, Y., & Fujimura, O. (1977). A computational model of the tongue: A revised version. *Journal of the Acoustical Society of America, 62, S15*, (A).

Kay, B. A., Munhall, K. G., V.-Bateson, E., & Kelso, J. A. S. (1985). A note on processing kinematic data: Sampling, filtering, and differentiation. *Haskins Laboratories Status Reports on Speech Research, SR-81*, 291-303.

Kiritani, S., Itoh, K., & Fujimura, O. (1975). Tongue-pellet tracking by a computer-controlled X-ray microbeam system. *Journal of the Acoustical Society of America, 57*, 1516-1520.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America, 53*, 1070-1082.

Ostry, D. J. & Munhall, K. G. (1985). Coordination of lingual and laryngeal gestures in speech. *Journal of the Acoustical Society of America, 77*, (S1), S99 (A).

Pound, E. (1977). Let S be your guide. *Journal of Prosthetic Dentistry, 38*, 482-487.

Posselt, U. (1968). *Physiology of occlusion and rehabilitation.* Philadelphia: F. A. Davis Co.

Sarnat, B. G. (1964). *The temporomandibular joint.* Springfield, IL: Thomas.

Tuller, B., Harris, K. S., & Gross, B. (1981). Electromyographic study of the jaw muscles during speech. *Journal of Phonetics, 9*, 175-188.

Vatikiotis-Bateson, E. (1988). *Linguistic structure and articulatory dynamics.* Doctoral dissertation, Indiana University. [Distributed by the Indiana University Linguistics Club, Bloomington].

Westbury, J. R. (1988). Mandible and hyoid bone movements during speech. *Journal of Speech and Hearing Research, 31*, 405-416.

FOOTNOTES

Journal of Speech and Hearing Research, 33, 550-562 (1990).

†Hunter College of the City University of New York

††Also The Graduate Center of the City University of New York and Haskins Laboratories

[1]The speech materials were designed in order to examine the effects of stress and coarticulation on the relationships among jaw rotation and horizontal and vertical jaw translation. As it turned out, stress and coarticulation had very little effect on the relationships among the three components of jaw movement. This finding is discussed at length in Edwards (1985).

[2]All of the squared multiple correlations shown in Tables 4 and 6 are statistically significant because of the large number of data points. However, a squared multiple correlation is only taken to indicate a functional interdependence between two components of jaw movement if the independent variables account for at least 50% of the observed variance in the dependent variable.

Haskins Laboratories Status Report on Speech Research
1990, SR-103/104, 67-94

Linguistic Structure and Articulatory Dynamics: A Cross Language Study*

Eric Vatikiotis-Bateson and J. A. Scott Kelso†

In the study reported here, movement data were analyzed for paradigm representatives of the three most widely recognized temporal organization categories: English for stress timing, Japanese for mora timing, and French for syllable timing. Data from the three languages were elicited and analyzed as commensurately as possible, using the experimental methodology employed originally by Kelso, Vatikiotis-Bateson, Saltzman, and Kay (1985) for two speakers of English. The primary aim was to show the extent to which simple kinematic analysis of a primary articulator (the lower lip-jaw complex) can reveal universal and language-specific aspects of temporal organization and prosody.

Kelso et al. (1985) found that most of the spatiotemporal variability of the movement behavior can be accounted for in the highly linear covariation of peak velocity and displacement. They further found clear condition-specific effects of stress and speaking rate on displacement and on the slope of the linear relation between peak velocity and displacement. Kelso et al. interpreted these results in terms of a very simple and highly abstract second-order system such as an undamped linear spring-mass, which could characterize the spatiotemporal behavior of the system as a whole, as well as specific linguistic and performative distinctions in stress and speaking rate, based on the variable settings of a small number of underlying parameters.

1 INTRODUCTION

In an earlier study, Kelso, Vatikiotis-Bateson, Saltzman, and Kay (1985) examined the articulatory kinematics of lower lip-jaw motion recorded for two speakers of English during reiterant renditions of sentences taken from the Rainbow Passage (Fairbanks, 1960). Using /ba/ as the reiterant syllable, they found, as had many before them, that the individual kinematics of gestural duration, displacement, and peak velocity were consistently correlated with differences in emphatic stress and speaking rate. Stressed movement gestures were larger in displacement, longer in duration, and higher in peak velocity than unstressed gestures produced at the same speaking rate. Similarly, the kinematics associated with gestures produced at a conversational speaking rate were larger than those produced at a faster rate. Further, while there was a marked tendency for larger movements to take longer, this somewhat linear relation between a gesture's displacement and duration was not nearly as strong as that between gestural displacement and peak velocity. In particular, the displacement-duration (d-t) relation was quite noisy and showed spatial saturation for stressed gestures produced at the conversational speaking rate. On the other hand, the relation between peak velocity and displacement (Vp-d) had a strong linear component that accounted for 80-90% of the overall variance in the movement behavior.

Such linear covariation between peak velocity and displacement had been observed before in speech production (e.g., Kozhevnikov & Chistovich, 1965; Kuehn & Moll, 1976; Mermelstein, 1973; Ohala, Hiki, Hubler, &

We would like to thank Elliot Saltzman, Richard McGowan, Katherine Harris, and Vincent Gracco for their guidance and criticism through the course of this project. Research funds were provided by NIH grant DC-00121 and BRSG grant RR-05596.

Harshman, 1968; Sussman, MacNeilage, & Hanson, 1973), but it was not until the roughly contemporary studies by Ostry, Keller, and Parush (1983) and Kelso et al. (1985) that the Vp-d relation was investigated in any detail for running speech. In addition to the highly linear covariation between peak velocity and displacement throughout the data range, both groups found that variations in stress and, to lesser extent, speaking rate are characterized by systematic condition specific differences in the slope of the Vp-d relations. Also, both groups recognized that the Vp-d relation could be modeled as the behavior of a second order dynamical system such as a linear mass-spring. Kelso et al. (1985) went on to show in detail how the spatiotemporal behavior of the system could be simulated by adjusting the settings of just two underlying model parameters—equilibrium position and stiffness—easily inferred from mean gestural displacement and the slope of the Vp-d relation.

Approximating the observed, nearly sinusoidal motion of the speech articulators to the motion of a linear mass-spring has several advantages. First, the behavior of such systems is well-defined. A simple equation of motion, $m\ddot{x} + b\dot{x} + kx = 0$, with very few parameters can generate movements from an infinite number of initial conditions that will attain the same target or end point. Thus, different movement trajectories do not necessarily require different parametrizations of the movement equation. Second, experimental data may be mapped onto abstract dynamic parameters (e.g., Smith, Browman, & McGowan, 1988). That is, the mass (m), damping (b), and spring stiffness (k) parameters of the underlying dynamical model can be inferred from the observed values of and relations between the kinematic variables of acceleration ($\ddot{x}$), velocity ($\dot{x}$), and displacement (x). Third, despite the absence of an explicit time variable in the motion equation, the temporal as well as the spatial characteristics of motion are fully determined; movement duration is inversely proportional to and, therefore, recoverable from spring stiffness, k, which can be calculated from the Vp-d relation (see Vatikiotis-Bateson, 1988, Appendix A). Fourth, the same basic system can account for both rhythmic and discrete movement behavior with appropriate modification to the basic equation of motion—e.g., through addition of a nonlinear damping term (Kay, Kelso, Saltzman, & Schöner, 1987; Saltzman & Kelso, 1987). If, as appears likely, other biological movement behaviors besides speech, such as (discrete) reaching and (rhythmic) locomotion, could be successfully described as analogous second order systems, then this may provide a way to identify what is common to all biological movement behaviors (Kelso & Tuller, 1984).

Thus, this approach offers a promising framework for characterizing potentially complex movement behaviors in terms of the function-specific settings of just a few underlying parameters. In particular, the success of the simple model proposed by Kelso et al. (1985) in characterizing spatiotemporal behavior across changes in stress and speaking rate suggested that, with appropriate tuning, such a model might be applied universally across languages. However, because this model was based on data from only two English speakers repetitively producing one syllable, /ba/, it was clear that a somewhat broader corpus was needed. Consequently, Vatikiotis-Bateson (1988) analyzed comparable reiterant speech data using two reiterant syllables, /ba/ and /ma/, from at least three speakers each of English, French, and Japanese. These languages are generally accepted to differ substantially in their temporal organization and prosody (Abercrombie, 1967; Bloch, 1950; Pike, 1943). Kinematic variables associated with lower lip-jaw motion were analyzed within and across two prosodic conditions and two instructed speaking rates for each language. In the present paper, the results of that study are summarized, and are used to evaluate hypotheses concerning the existence and implementation of universal and language-specific constraints on supralaryngeal movement behavior.

2 METHODS AND PROCEDURES

Since much of the experimental and analytic methodology used in this study has been described in detail elsewhere (Kay, Munhall, Vatikiotis-Bateson, & Kelso, 1985; Kelso et al., 1985; Vatikiotis-Bateson, 1988), only the experimental aspects specific to this study are described in detail.

2.1 Subjects

Five native speakers of English, four speakers of French, and five speakers of Tokyo Japanese took part in the study. All but one speaker of Japanese (NK), who served as a pilot subject, were naive to the purposes of the experiment, had never been exposed to the reiterant speech task, and were paid for their participation. Speaker NK selected the Japanese stimuli and practiced them reiterantly before her experimental run.

2.2 Reiterant speech task

Reiterant, or mimicked, speech is a substitution task in which the speaker replaces each syllable of a target phrase or sentence with a test syllable such as /ba/ or /ma/, while trying to maintain the rhythmic and prosodic character of the original. The task was used extensively in early acoustic studies of metrics (Scripture, 1899a,b; Stetson, 1905; Wallin, 1901). More recently, it has been used for both acoustic (e.g., Larkey, 1983; Liberman & Streeter, 1978; Lindblom & Rapp, 1973) and articulatory (Kelso et al., 1985) studies.

2.3 Training procedure

The reiterant speech task was explained to English and French speakers as one involving syllable substitution and demonstrated using short phrases and sentences, such as "Mary had a little lamb," spoken at a conversational rate. Due to the controversy over whether the "unit" of timing in Tokyo dialect is the syllable or the mora (e.g., Higurashi, 1984; McCawley, 1978), no attempt was made to force Japanese speakers to use a unit that might be unnatural to them, especially in an already difficult task. Therefore, the reiterant speech task was simply demonstrated by speaker NK, using phrases containing only light, single mora syllables. Speakers practiced producing these phrases reiterantly using /ba/ and /ma/, first at a comfortable rate and then as fast as possible, until they could do so fluently with proper intonation and the right number of syllables at two distinct (experimenter determined) rates. Finally, subjects were instructed to memorize the two sentences to be used as experimental stimuli and told not to practice them reiterantly.

2.4 Stimuli

2.4.1 English. In order to maintain commensurability with the Kelso et al. (1985) study, the same two sentences from the Rainbow Passage (Fairbanks, 1960), the same stress assignment, and the same exclusion of sentence initial and phrase final syllables were used here. The one notable exception to this was that speakers in this study occasionally paused (shown below by vertical hash-marks) between the fourth and fifth syllables of sentence 1 (...sunlight || strikes...). Therefore, the fourth syllable, which in these cases becomes phrase-final, was excluded from analysis. In the sentences below, excluded syllables are within parentheses. The apostrophe indicates that the following syllable was marked for stress; all other syllables are treated as unstressed.

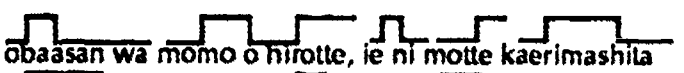
1. (When) the 'sun(light) || 'strikes 'raindrops in the ('air), || they 'act like a 'pris(m) || and 'form a 'rain(bow).
2. (There is), || ac'cording to 'leg(end), || a 'boiling 'pot of (gold), || at one (end).

2.4.2 French. Two sentences were chosen from the "Maximes" of LaRochefoucauld, 'written' in the mid to late seventeenth century. Although not all speakers were acquainted with these particular maxims, they were familiar with the genre. The two sentences were chosen because of their length and structure. The second sentence in particular, with its embedded relative clause, is grossly similar to English Sentence 2. Speakers accepted both sentences as non-archaic.

1. La ferocité naturelle || fait moins de cruel || que l'amour propre.
2. L'interêt, || qui aveugle les uns, || fait la lumière des autres.

Per convention (e.g., Anderson, 1982; Delattre, 1966; Selkirk, 1978), stress was assigned to all non-schwa (mute "e") word-final (in trisyllablic words) and phrase-final syllables. Generally, the final syllables of *aveugle* (elided with *les*), *propre* (unreleased),and *autres* (unreleased) were not mimicked in the reiterant productions. Finally, following the lead of Vaissiere (1983), who cites an increasing tendency for speakers to place stress phrase-initially due to the influence of the French telecommunications media, all phrase-initial syllables were treated as stressed. In the two sentences above, stressed syllables are underlined and phrase boundaries are marked by double vertical strokes.

2.4.3 Japanese. The two sentences used here were taken from a Japanese folk tale—the "Momotaroo"—well-known to all five speakers. They were chosen to be roughly of the same length (syllable count) as the two English sentences and to provide a range of multimora syllables (underlined syllables). The sentences are presented below, as transcribed and marked (lines above the word) for high and low tone by speaker NK. Accented morae such as the /wa/ of *kawa* are those on which a High-Low fall occurs within the accentual phrase (delimited by breaks in the tone level marking).



2.5 Experiment protocol

To be sure that talkers had memorized the two sentences and that they could produce them at two distinct rates, the experimental session began with five normal recitations of each sentence at each speaking rate. They were then instructed that for the remainder of the experiment they would produce normal-reiterant utterance pairs—that is, a normal recitation immediately followed by its reiterant rendition—for the specified sentence (1 or 2), speaking rate (conversational or fast), and syllable identity (/ba/ or /ma/). A balanced design was used to elicit 10 normal-reiterant utterance pairs for each condition. This resulted in 80 reiterant utterances (= 10 repetitions × 2 sentences × 2 speaking rates × 2 syllable types) for later analysis. Including errors and technical adjustments, experimental sessions lasted approximately 45 minutes.

2.6 Signal recording and conditioning

Vertical and horizontal movement of the lips and jaw were tracked midsagittally using a Huntspot (modified Selspot) system and recorded simultaneously with the acoustic output onto FM tape. A fourth LED was placed on the bridge of the nose as a reference for head movement. The analog movement and audio signals were digitized and stored to disk. Movement of the lips, jaw, and nose was sampled at 200 Hz. Speech acoustics were sampled at 10 kHz. After digitization, the movement signals were numerically corrected for vertical head movement, smoothed, and differentiated to obtain instantaneous velocity. [The complete processing sequence and hardware description is treated in detail in Kay et al., 1985].

2.7 Kinematic analysis

This study is concerned primarily with kinematic measures made from the vertical movements of the lower lip LED. The movement of this sensor is really that of an articulator complex, being composed of the jaw's contribution to lower lip movement as well as that of the lip alone. Figure 1 shows the vertical change of position of the lower lip LED over time (middle trace), the instantaneous velocity magnified ten times (top trace), and the audio waveform, for a portion of a reiterant production using /ba/.

The continuous motion of the lower lip-jaw complex was divided into successive opening (lowering from peak consonant closure position to peak vowel opening) and closing gestures (raising from peak vowel opening position to peak consonant closure).[1] Measures of displacement, duration, and peak velocity of motion were obtained for each gesture by means of an automated procedure that marks the position and time of waveform peaks and valleys. For position, the peaks and valleys correspond to the points of maximum consonant closure and vowel opening, respectively. From these the displacement and duration of opening (peak-to-valley) and closing (valley-to-peak) gestures were calculated. In the velocity trace, valleys (labeled Vp_O) denote the maximum, or peak, instantaneous velocity achieved during the opening gesture, and peaks (labeled Vp_C) the peak velocity for closing.

In addition to syllable identity (/ba/ or /ma/) and gesture type (opening or closing), data were coded by speaking rate (conversational or fast), sentence (1 or 2), syllable number (i.e., position in the utterance), and, for each language, a binary prosodic variable (stressed or unstressed for French and English, and high or low tone for Japanese).

As mentioned earlier, in order to avoid problems of phrase-final lengthening, well-documented for English and French, phrase- and utterance-final syllables were not analyzed. Other excluded data occurred when multiple peaks in a movement gesture's velocity profile made it difficult to define peak velocity for that gesture (less than 1% of the remaining gestures).

Finally, of the fourteen naive subjects in this study, ten (3 English, 3 French, and 4 Japanese) succeeded in producing prosodically intact reiterant renditions of the two target sentences appropriate to each language. In retrospect, use of untrained speakers was not a wonderful idea, but it did demonstrate the "all or nothing" character of speakers' ability to produce reiterant speech. No practice effect was observed other than the commonly observed tendency for speakers to produce utterances progressively faster over the course of an experiment.

One of the Japanese speakers (SM) often had trouble producing all phrases of an utterance correctly, especially when /ma/ was the reiterant syllable. His data were analyzed or not depending on whether or not he successfully mimicked two or more phrases within a given sentence rendition. A fifth Japanese speaker's data (ME) are included in the analysis of overall kinematic patterning, bringing the cross-language total to eleven speakers. Even though speaker ME did not produce the correct number of reiterant syllables to match the original utterances (by either mora or syllable count; for details, see Vatikiotis-Bateson, 1988), her productions are included because they were

judged to be prosodically plausible by a Japanese and a non-Japanese listener and because they show the same language-specific, overall spatio-temporal characteristics as those of the other Japanese speakers. In general, with these exceptions, measures of gestural displacement, duration, and peak velocity were made for approximately 3000 opening and closing gestures (i.e., 1500 syllables) per speaker.

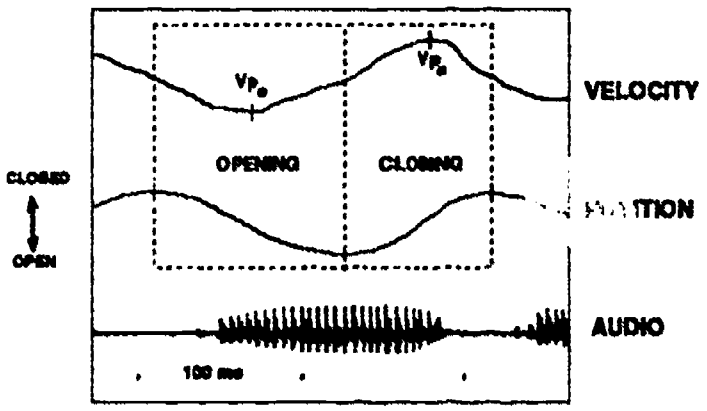


Figure 1. Time series representation of position, instantaneous velocity, and audio. Movements are divided into opening and closing gestures at peaks and valleys of position trace. Displacement and duration are computed between successive peaks and valleys. Each movement gesture has an associated peak velocity (Vp_o or Vp_c).

3 RESULTS AND DISCUSSION

Figure 2 shows times series representations of lower lip-jaw vertical position, instantaneous velocity, and associated speech acoustics for a representative reiterant production for a speaker of each language. Note that within a phrase, motion is continuous; there is no break or flattening of the valley between an opening (lowering) gesture and closing (raising) gesture that succeeds it. This is also true of the peaks of motion with the exception discussed below of geminate and homorganic clusters in Japanese.

In his thesis, Vatikiotis-Bateson (1988) found no systematic interaction between syllable identity and either the overall patterning of the kinematics or the condition-specific kinematic correlates of prosodic and speaking rate distinctions. In the main, there was a slight, but consistent, magnitude difference between /ba/ and /ma/. Therefore, in the current presentation, the results for the two syllable types are usually treated interchangeably.

A second [illegible] that was analyzed in detail by Vatikiotis-Bateson, but which is not treated here, concerned the ubiquitous covariation effects among the three individual kinematic variables. As summarized in Table 1, all speakers displayed a general tendency for mean values of the three kinematic variables to covary across changes of speaking rate and the linguistic variable. Kinematic means for movement gestures produced at faster speaking rates were smaller than those produced at the slower, conversational rates. Stressed gestures in English and French were larger in displacement, duration, and peak velocity than unstressed ones. In Japanese, there were consistent effects of accent related tone on the gestural kinematics: high tone gestures had smaller mean kinematics than low tone gestures.

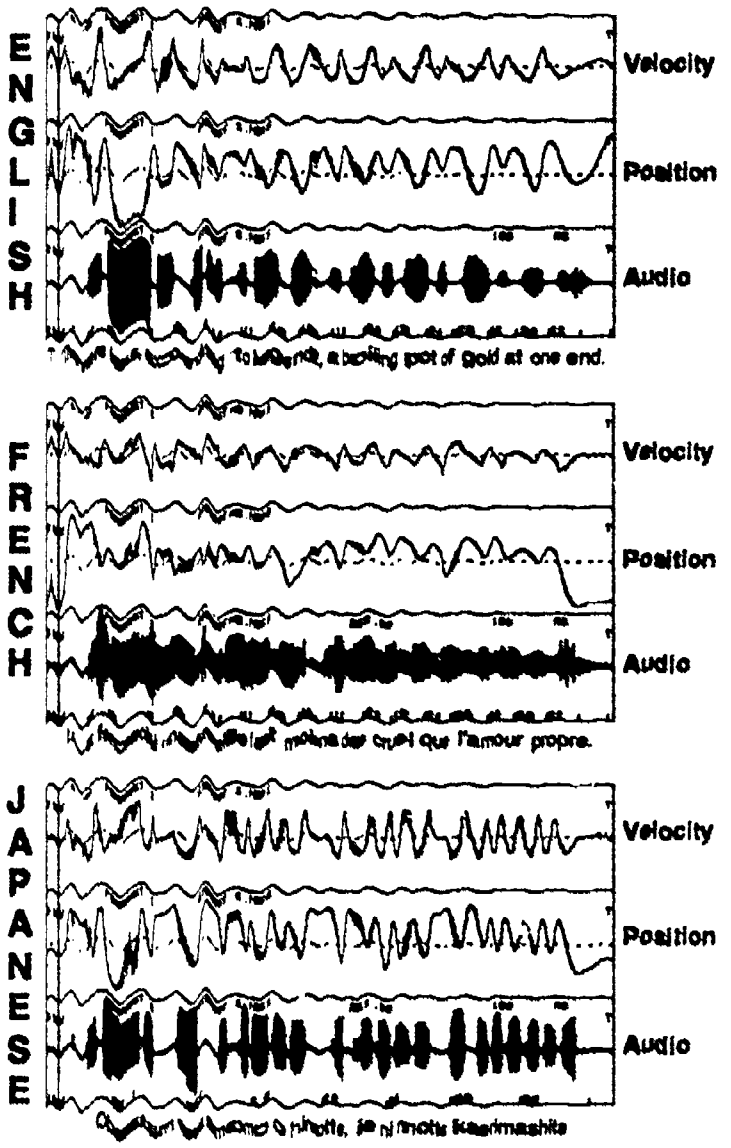


Figure 2. Position, instantaneous velocity, and audio traces for a representative reiterant sentence producted by a speaker of each language—English: Sentence 2, Speaker [illegible], /ma/; French: Sentence 1, Speaker CG, /ma/; Japanese: Sentence 2, Speaker NK, /ba/.

Table 1. *Within-language means and standard deviations (SD) for duration (DUR) in ms, displacement (DISP) in mm, and peak velocity (VP) in mm/s for opening and closing gestures. Coefficients of variation (CV = SD/DUR) are given for duration results.*

	ENGLISH		FRENCH		JAPANESE	
	Opening	Closing	Opening	Closing	Opening	Closing
DUR	96	88	80	73	76	71
SD	24	30	17	14	14	15
CV	0.25	0.34	0.21	0.19	0.18	0.21
DISP	7.02	6.94	5.33	4.99	6.83	6.36
SD	3.50	3.43	2.51	2.29	3.27	3.01
VP	132.6	147.1	123.3	121.8	147.2	147.8
SD	59.5	59.6	56.3	61.1	67.2	63.8

Finally, a third and perhaps the most important area of analysis not treated here, concerns Japanese multimora gesture sequences containing a geminate stop (e.g., /tt/ of *motte*) or homorganic cluster (e.g., /nt/ of *sentaku*) in which lip closure is held for a consistent period of time. That is, there is a break in the otherwise continuous alternation of opening and closing gestures. Because there is a sustained period during which there is no motion, this quite simply constitutes a portion of the speech production that cannot be accounted for within the dynamical scheme discussed here. Furthermore, the suggestion made originally by Kelso et al. (1985) that time need not be a controlled variable since it is recoverable from the relation between peak velocity and displacement, is further clouded by the additional finding that the stable period of silence during closure for the consonant cluster can be achieved via different configurations of the laryngeal and supralaryngeal articulators. These phenomena and their implications for the modeling discussed here are the focus of a detailed follow-up study of glottal-oral coordination in Japanese (Vatikiotis-Bateson, in preparation).

In what follows, lip-jaw movement data are considered in two ways. First, in § 3.1, the gestural kinematics are examined in order to assess the overall spatiotemporal character of different languages whose prosodic structures and perceived temporal organization are quite different. We demonstrate that differences observed among the language-specific data are commensurate with temporal differences we have come to expect through perceptual and other empirical observations. At the same time, we show that most of the observed differences can be described in terms of the scaling of a single abstract parameter, stiffness, inferred from the slope of the Vp-d relation. From this scaling we hypothesize that opening and closing movements of the lower lip-jaw complex may adhere to the constraints of an underlying, dynamical second-order system. The applicability of such a model to commensurate data from three languages so different in their temporal organization is, we argue, indicative of a universal constraint on speech movement behavior. The claim of universal applicability is bolstered further by observation of similar constraints in other biological movement systems (see Ostry, Cooke, & Munhall, 1987). Second, in § 3.2, language-specific parametrization of these potentially universal constraints is demonstrated by showing how language-specific prosodic distinctions and, to a lesser extent, differences in speaking rate are realized through similar means within the hypothesized dynamical system.

3.1 Overall patterning of gestural kinematics

3.1.1 Differences between opening and closing gestures. In their study of English, Kelso et al. (1985) found temporal but no spatial differences between opening and closing gestures. Specifically, opening and closing gesture displacements were the same, which was attributed to the fact that the lower lip-jaw complex achieved the same closure position on successive gestures. Durations, on the other hand, were consistently shorter for closing than opening gestures; accordingly, closing peak velocities generally were observed to be higher than opening ones.

The results of the study reported here corroborate the earlier findings for English.[2] Furthermore, there is a similar durational asymmetry for the French and Japanese data, as shown in Table 2. For all three languages, /ba/ productions were consistently more durationally asymmetrical than /ma/ productions. Within a

language, no tendency was observed for speakers having slower absolute rates to produce greater temporal asymmetries. Nor did within speaker changes of speaking rate have systematic effects on the gestural asymmetry. This is shown by the absence of interactions between speaking rate and gesture type for speakers of English and French, and by the inconsistency of the interaction among speakers of Japanese where the temporal asymmetry was actually larger at the faster speaking rate for two speakers (for details, see Vatikiotis-Bateson, 1988).

Thus, the three languages showed roughly the same durational asymmetry with opening gestures being consistently longer than closing gestures. However, the data for Japanese and French show a spatial asymmetry as well in which opening gesture displacement was slightly, but consistently, larger than closing (see Table 1). It is possible that this small net loss is a procedural artefact resulting from the fact that large phrase-initial opening gestures were included in the analyses, but that phrase-final closing gestures, which are also quite often large, were not. Nevertheless, the same criteria were applied to all three languages and point up a difference between English and the other two languages.

This finding has led to a subsequent investigation of the tendency for kinematic values to decline, incline, or remain level over the course of an utterance (Vatikiotis-Bateson & Fowler, 1988). In particular, English speakers showed roughly equal tendencies for all three outcomes. French and Japanese speakers, on the other hand, showed a distinct tendency toward spatial declination, but were more heterogeneous in their temporal behavior (Table 3).

A tentative conclusion from these findings is that the durational asymmetry between opening and closing gestures, observed for all speakers, is independent of the language-specific spatial asymmetry observed only for speakers of Japanese and French (Vatikiotis-Bateson & Fowler, 1988).

Table 2. *Within-language means for duration (DUR) in ms of opening and closing gestures as a function of syllable type, and durational differences between opening and closing gestures (O:C) in percent.*

	ENGLISH		FRENCH		JAPANESE	
	Opening	Closing	Opening	Closing	Opening	Closing
/ba/						
DUR	97	88	79	71	76	70
O:C	+10%		+11%		+9%	
/ma/						
DUR	95	88	80	75	77	72
O:C	+8%		+7%		+7%	

Table 3. *For each language, linear regression analysis was used to test articulatory declination (excluding pauses) as a function of utterance condition (2 sentences produced at 2 speaking rates for 3 or 4 speakers) for each gesture type. Each utterance condition is averaged over 20 repetitions (/ba/ and /ma/). The number of negative (NEG) and positive (POS) regressions are tabled for opening gestures by kinematic variable and language. The total possible for each kinematic variable-language cell is 12 (3 speakers × 4 conditions) for English and French, and 16 (4 speakers × 4 conditions) for Japanese.*

	DISPLACEMENT		DURATION		PEAK VELOCITY	
	NEG	POS	NEG	POS	NEG	POS
ENGLISH (Out of 12)	3	3	3	4	4	3
FRENCH (Out of 12)	7	0	1	2	7	0
JAPANESE (Out of 16)	8	1	9	2	8	1
TOTAL (Out of 40)	**18**	**4**	**13**	**8**	**19**	**4**

3.1.2 The relation between displacement and duration. A severe limitation of examining measures of displacement (a spatial measure) and duration (a temporal measure) individually—as has often been the case in production studies (e.g., Lindblom, 1963; Sussman et al. 1973)—is that the inevitable variance in each measure remains an unexplained residue. Examination of the relations between the spatial and temporal aspects of an articulatory event could reduce the ubiquitous variance found in the individual kinematics.

In addition to presenting the basic space-time view of the gestural data, the distance-time (d-t) relation provides a measure of the average speed (Vav = d / t) of articulatory gestures. A possible relation between displacement and duration is that they covary in a positive, linear fashion such that Vav is conserved; that is, as duration increases so too does displacement. This is more or less what has been observed for English (e.g., Kelso et al., 1985; Nelson, 1983; cf. Gay, 1981). In languages such as French and Japanese, whose temporal organizations are perceived to be much more regular, there might be less tendency for displacement and duration to covary simply because of reduced durational variability. Both of these possibilities are examined in the present study.

Figure 3 contains scatterplots of opening gestures (left) and closing gestures (right) for the /ma/ productions for representative speakers of English (top), French (middle), and Japanese (bottom).

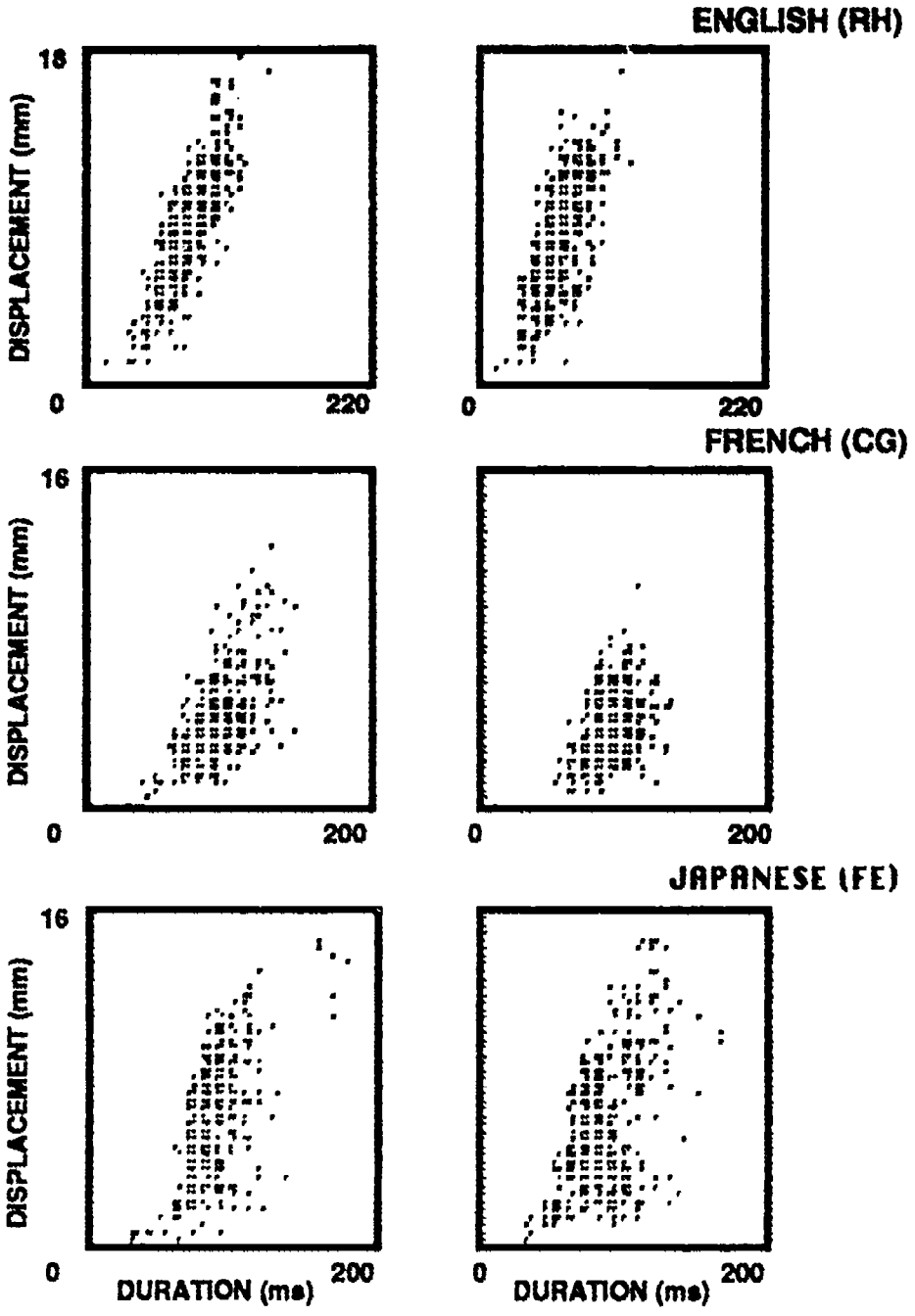


Figure 3. Scatterplots show the overall regression of gestural displacement (ordinate) on duration (abscissa) for the opening (left) and closing (right) gestures associated with /ma/ productions for one speaker of each language.

First, we observe a positive covariation between displacement and duration, which has a linear component that accounts for as much as 72% of the variability in the English data (Table 4). In general, the covariation is more linear for opening than closing gestures regardless of language. Second, the linear component of the regression is substantially higher for English than French or Japanese speakers. Indeed, there is only one instance (discussed below) in the French and Japanese data where the linear component accounts for more than 50% of the variability, and, in the majority of cases, it accounts for less than 25%. Third, while the covariation is positive and fast speaking rate gestures often have steeper slopes than gestures at longer duration conversational rates, there is little indication that average velocity is conserved. In part, this is due to the poor fit of the linear regression which accordingly underestimates the slope of the d-t relation. However, even in the best cases such as opening gestures for English, the slope values do not correspond well to the mean kinematics (see Vatikiotis-Bateson, 1988, Tables 2a and 4a).

Thus, the positive covariation of displacement and duration shows the expected tendency for larger movements to take longer, but is highly variable and usually accounts for less than half of the overall variance. It is unlikely, then, that conservation of average velocity plays a role in production of these movements. Indeed, for French and Japanese closing gestures, where durational variance is the smallest, there is little to no linear covariation between displacement and duration. Finally, the results for English are basically the same as those reported earlier by Kelso et al. (1985) for a much smaller sample restricted to /ba/ syllables.

Table 4. *Tabled below as a function of gesture type and syllable identity are number of observations (N), linear regression coefficient (r), and slope (m) for the overall regressions of gestural displacement on duration (d-t).*

			Opening Gestures			Closing Gestures		
			N	r	m	N	r	m
ENGLISH								
	RH	/ba/	544	.85	138.2	544	.71	129.7
		/ma/	529	.84	151.5	529	.66	140.0
	MP	/ba/	556	.62	48.6	554	.51	35.0
		/ma/	540	.64	69.2	538	.45	34.0
	JK	/ba/	538	.73	71.1	537	.75	69.4
		/ma/	540	.65	72.6	540	.75	59.9
JAPANESE								
	FE	/ba/	471	.67	109.6	468	.48	55.2
		/ma/	489	.49	87.8	487	.39	50.1
	ME	/ba/	682	.55	137.0	644	.64	121.6
		/ma/	670	.60	141.5	632	.65	116.7
	SM	/ba/	340	.46	95.2	340	.38	79.5
		/ma/	355	.36	74.1	355	.26	46.1
	MY	/ba/	809	.72	96.6	813	.42	65.6
		/ma/	810	.63	91.7	810	.46	85.1
	NK	/ba/	564	.89	118.0	564	.82	98.8
		/ma/	523	.87	105.6	520	.79	86.6
FRENCH								
	BA	/ba/	467	.70	95.4	468	.40	77.7
		/ma/	465	.68	115.0	465	.28	51.5
	DP	/ba/	517	.65	93.1	517	.34	56.8
		/ma/	486	.49	71.2	486	.18	33.5
	CG	/ba/	449	.50	55.4	447	.58	49.7
		/ma/	467	.61	73.0	467	.41	48.1

3.1.3 The relation between displacement and peak velocity. The relation between peak velocity and displacement (Vp-d) is of interest because it is consistent with the behavior of many physical systems that can be modeled using second-order dynamics. If we restrict our second-order model to the slightly idealized, but very simple, undamped linear mass-spring, then spring stiffness, k, is proportional to the slope of the Vp-d relation, whose units are temporal (w).[3] To the extent that such modeling may be appropriately applied to speech production, two predictions should hold. First, differences in mean duration should be matched by differences in the slope of the Vp-d relation, indicative of corresponding stiffness differences. Second, differences in temporal variability should be reflected by differences in the degree of linear covariation between peak velocity and displacement. Both of these predictions are born out by the data as illustrated in Figure 4.

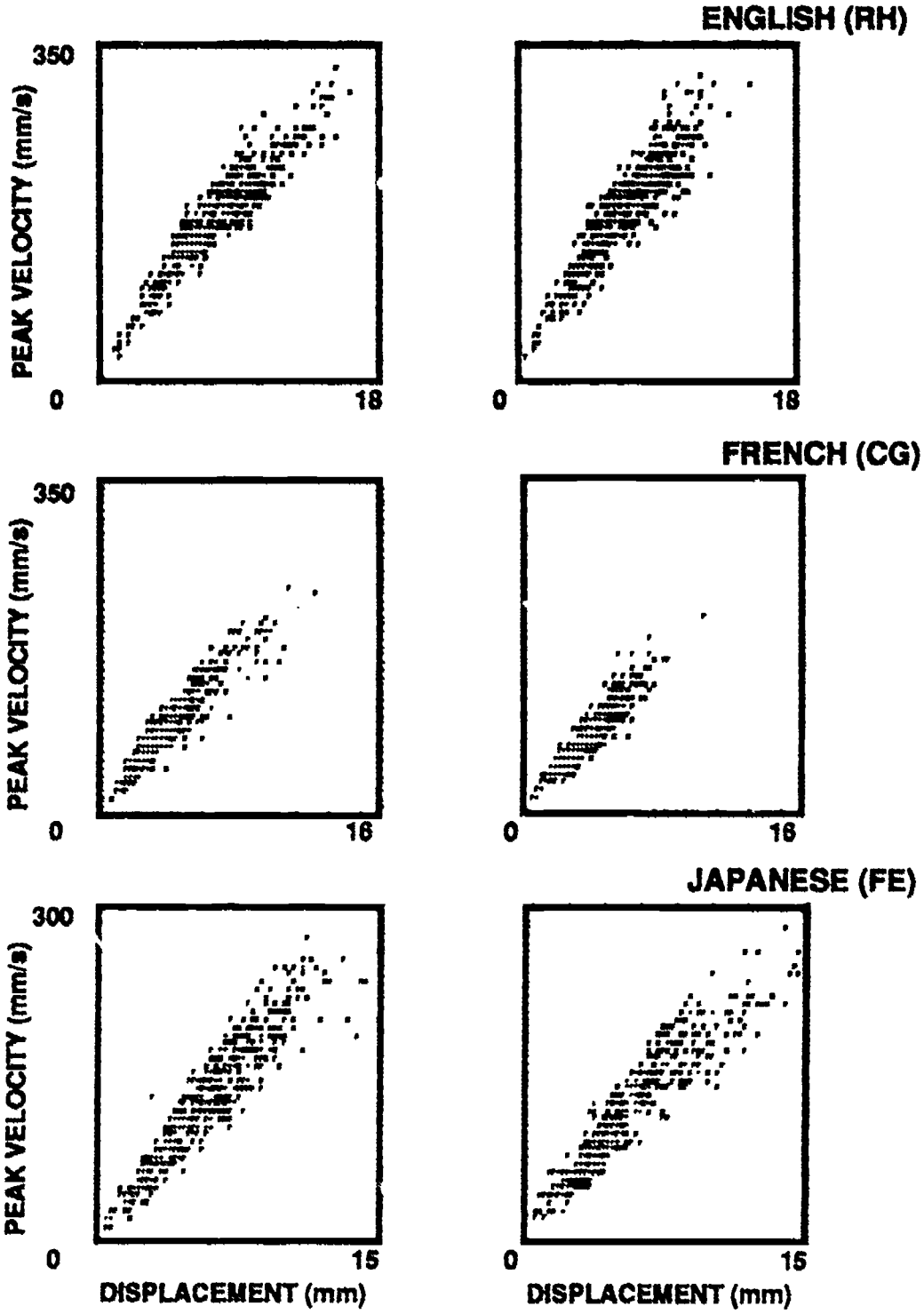


Figure 4. Scatterplots show the overall regression of gestural peak velocity (ordinate) on displacement (abscissa) for the opening (left) and closing (right) gestures associated with /ma/ productions for one speaker of each language.

The figure contains scatterplots of the covariation between peak velocity and displacement on a gesture-by-gesture basis for the opening and closing /ma/ gestures for one speaker of each language. It should be immediately clear that the Vp-d relation accounts for substantially more of the spatiotemporal variability than the d-t relation. In keeping with the shorter movement duration of closing than opening gestures, we observe steeper Vp-d slopes for closing gestures. This pattern holds generally for all speakers, regardless of language (Table 5). The one exception to this, which demonstrates well the hypothesized correlation between Vp-d slope and duration, is English speaker JK, discussed in § 3.1.1 above. For /ma/ productions, closing gestures were longer in duration than opening ones. For this case and contrary to the results for the other speakers, the slope of the Vp-d relation was steeper for opening than closing gestures (Tables 4, 5), which is exactly what the hypothesized model predicts.

Correlation coefficients, with the exception of speaker JK, are consistently higher for closing than opening gestures, which in terms of the hypothesized dynamical model should indicate differences in temporal stability. To the extent that opening and closing gestures differ in temporal stability, there are matching differences in Vp-d regression coefficients (for within-language results, see Table 1; for speaker-specific results, see Vatikiotis-Bateson, 1988).

Appropriate to the prediction of a second order spring system, quantitative comparison of the data shows differences in Vp-d slope and degree of linear covariation. Vp-d slopes are steeper for the French and Japanese movement data, whose mean durations are shorter than those of the English data. Slope differences were tested and t-values are given in Table 5.[4]

Correlation coefficients are higher for the French and, with one exception discussed below, the Japanese data, while both show temporal variability lower than that of English. Indeed, the linear component of the Vp-d relation consistently accounted for about 90% of the spatiotemporal variance for French and Japanese speakers. For English speakers, on the other hand, the linear regression was substantially weaker, accounting for only 79% of the variance; still it was stronger than that observed for the d-t relation.

The data of the exceptional Japanese speaker, NK, help demonstrate the relevance of the second-order description of articulatory motion as shown by the two relations among kinematic variables examined in this study. The patterning of her data is more like that of English than that of the other French and Japanese speakers in that there is a relatively strong linear component in the covariation of displacement and duration and a relatively weak one in the covariation of peak velocity and displacement. Indeed, the linear components of the d-t and Vp-d regressions account for equal portions of the overall spatiotemporal variance, 71%. That is, to the extent that displacement and duration covary linearly and that there is spatial variability, the system should not show a high degree of temporal stability.[5]

Basically, then, the same differences are seen within and across the three languages, appropriate to the hypothesized setting of an underlying dynamic stiffness parameter. Differential settings of stiffness, more reliably inferred from differences in Vp-d slope than from gestural duration itself, could be all that's necessary to distinguish not only universal differences between opening and closing, but also language-specific differences in speaking rate and temporal variability. That is, French and Japanese are perceived to be temporally more regular than English. When interpreted in terms of an abstract underlying dynamical system, this perception may be corroborated for production through observation of the simple relation between peak velocity and displacement.

3.1.4 Relation to movement cycle duration in three languages. The patterning shown above for the Vp-d relation is commensurate with the observation that syllable- and mora-timed languages such as French and Japanese are produced at higher syllable rates than stress-timed languages such as English (Dauer, 1983). As far as we know, prior to this study, cross-language comparison of syllable production rates have relied on temporal acoustic measures of syllable duration, such as those collated by Dauer (ibid). Table 6 shows that motion of the lip-jaw complex provides a reliable articulatory measure of absolute speaking rate and temporal variability. Instead of acoustic syllable duration measured from burst release or vowel onset, we measured movement cycle duration, defined between successive peak closure positions of the lower lip-jaw. Each movement cycle contains an opening-closing gesture sequence. This measure, while clearly different from acoustic syllable duration, is isomorphic in that each movement cycle encompasses only one vowel bounded on either side by some portion of the adjacent consonant closure.

Table 5. *Mean gestural duration (DUR), linear regression coefficient (r), and linear slope (m) are tabled as a function of gesture type and syllable identity for the regression of peak velocity on displacement (Vp-d). Vp-d slope differences between opening and closing gestures were tested (one-tailed t). Asterisks indicate probabilities at the .05, .01, and .001 levels. 'X' indicates the difference is counter to that predicted.*

			Opening Gestures			Closing Gestures			Slope
			DUR	r	m	DUR	r	m	Comparison
ENGLISH									
	RH	/ba/	86	.94	14.3	76	.91	17.7	***
		/ma/	86	.94	15.7	74	.91	18.4	***
	MP	/ba/	89	.89	15.5	80	.88	19.1	***
		/ma/	90	.93	19.7	79	.87	19.6	ns
	JK	/ba/	113	.77	10.4	105	.87	13.5	***
		/ma/	107	.90	16.4	107	.74	10.6	*** X
JAPANESE									
	FE	/ba/	76	.97	18.6	74	.95	18.8	ns
		/ma/	79	.95	21.2	79	.93	18.6	***
	ME	/ba/	78	.94	18.2	74	.89	16.4	*** X
		/ma/	78	.91	18.2	74	.84	15.4	*** X
	SM	/ba/	69	.93	20.5	63	.91	22.0	*
		/ma/	68	.92	21.6	65	.86	21.1	ns
	MY	/ba/	78	.94	17.6	66	.95	23.4	***
		/ma/	77	.92	19.6	70	.94	22.9	***
	NK	/ba/	77	.88	13.3	64	.88	14.8	***
		/ma/	77	.86	13.1	67	.75	10.9	*** X
FRENCH									
	BA	/ba/	75	.94	18.7	67	.90	22.2	***
		/ma/	76	.95	20.4	70	.90	22.9	***
	DP	/ba/	73	.95	21.8	70	.92	21.9	ns
		/ma/	76	.94	23.5	70	.94	22.5	* X
	CG	/ba/	91	.95	17.6	81	.93	16.2	*** X
		/ma/	89	.94	17.9	86	.93	18.0	ns

Table 6. *Mean closure-to-closure duration (DUR) in ms, standard deviation (SD), frequency (FREQ = 1/DUR) in Hz, and coefficient of variation (CV = SD/DUR) are tabled as a function of syllable type and speaking rate.*

			NORMAL				CONVERSATIONAL			
			DUR	SD	FREQ	CV	DUR	SD	FREQ	CV
ENGLISH										
	RH	/ba/	179	29	5.59	0.16	147	29	6.80	0.20
		/ma/	175	25	5.71	0.14	146	23	6.85	0.16
	MP	/ba/	185	38	5.41	0.21	153	23	6.54	0.15
		/ma/	181	36	5.52	0.20	157	24	6.37	0.15
	JK	/ba/	237	54	4.22	0.23	199	45	5.03	0.23
		/ma/	234	58	4.27	0.25	195	43	5.13	0.22
JAPANESE										
	NK	/ba/	157	18	6.37	0.11	124	20	8.06	0.16
		/ma/	160	22	6.25	0.14	130	24	7.69	0.18
	FE	/ba/	155	26	6.45	0.17	145	26	6.90	0.18
		/ma/	161	22	6.21	0.14	158	21	6.33	0.13
	ME	/ba/	156	16	6.41	0.10	141	13	7.09	0.09
		/ma/	156	18	6.41	0.12	142	12	7.04	0.08
	SM	/ba/	135	15	7.41	0.11	128	12	7.81	0.09
		/ma/	136	11	7.35	0.08	132	14	7.58	0.11
	MY	/ba/	157	23	6.37	0.15	131	16	7.63	0.12
		/ma/	160	20	6.25	0.13	134	16	7.46	0.12
FRENCH										
	BA	/ba/	153	22	6.54	0.14	130	11	7.69	0.8
		/ma/	155	19	6.45	0.12	135	12	7.41	0.09
	DP	/ba/	147	18	6.80	0.12	138	15	7.25	0.11
		/ma/	149	15	6.71	0.10	142	15	7.04	0.11
	CG	/ba/	177	26	5.65	0.15	161	25	6.21	0.16
		/ma/	183	23	5.46	0.13	166	22	6.02	0.13

As shown by the mean duration (DUR) and frequency (FREQ) values, there is a clear difference in absolute speaking rate, with English being slower than the other two languages, regardless of which instructed speaking rate was used—i.e., conversational or fast. The standard deviation values (SD) reveal a concomitant difference in temporal variability in that variability increases with mean duration. Furthermore, because coefficients of variation (CV) are higher for English than the other two languages, it is evident that the scaling exhibited here is not linear.

Finally, there is a fairly clear correspondence between language-specific differences in absolute speaking rate and the relation between peak velocity and displacement: Steeper slopes of the Vp-d relation and higher correlation coefficients are observed for languages produced at faster absolute speaking rates. This can be seen indirectly in Figure 5 in which the upper panels show the Vp-d regression lines for each language as a function of gesture type. Although the small slope difference between French and Japanese opening gestures is reliable ($t = 7.25$, $p < 0.001$), the difference between those two languages and English is visibly much larger.[6] In the lower panel of the figure, the average gestural duration for each language are shown as a function of gesture type. Certainly in keeping with the Vp-d regressions for opening gestures, durational differences are relatively small between French and Japanese, but quite large between them and English.

3.1.5 Brief summary of overall patterning. To summarize, we have considered two relations between easily measured kinematic variables of gestural displacement, duration, and peak velocity for a single dimension of movement within a single articulator complex. They have proved useful in providing an articulatory characterization of differences between languages that are perceived readily enough, but which previously have been difficult to identify quantitatively in production. Although highly variable, the relation between displacement and duration showed more linear covariation for those gestures demonstrating the higher durational variability. This finding is exactly what we expect when viewed in conjunction with the extremely linear covariation of peak velocity and displacement and the inference of an underlying second order dynamical system. That is, linear covariation of displacement and duration should be weakest for those cases of greatest temporal stability—i.e., relatively uniform gestural durations regardless of gestural displacement—as seen in closing gestures for French or Japanese. Conversely, these are exactly the cases where the highest degree of linear covariation of peak velocity and displacement is observed.

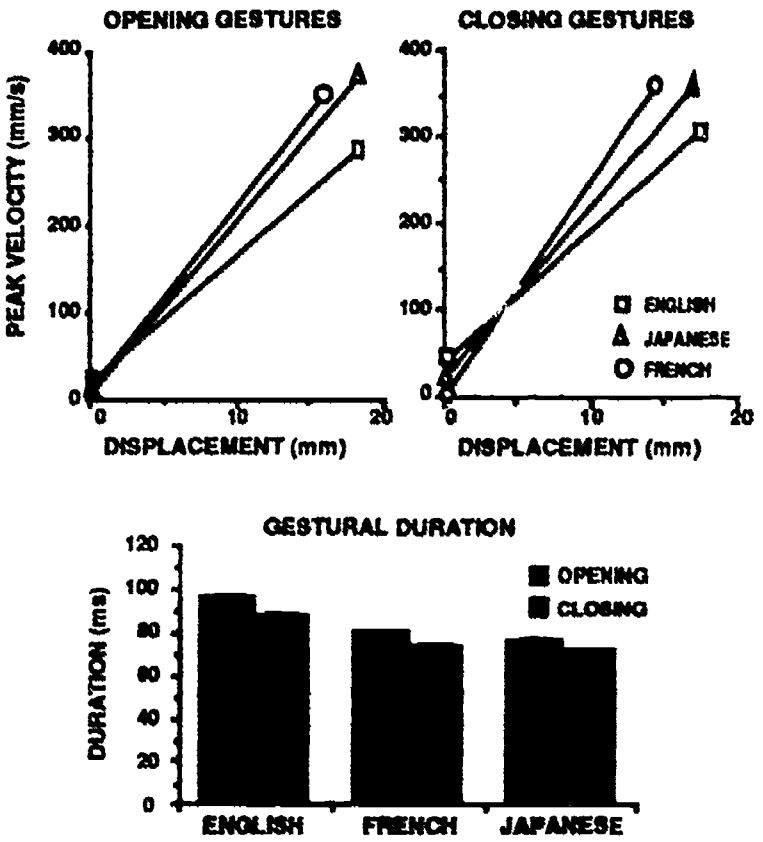


Figure 5. In the upper panels, best-fit lines denote the overall (pooled across speakers) Vp-d regression for the opening and closing gestures of each language. The lower panel shows mean duration (pooled across speakers) for each language as a function of gesture type.

3.2 Kinematic analysis of condition-specific effects

Having found the relation between peak velocity and displacement to be useful in characterizing and distinguishing the overall movement behavior of three languages, we examine now whether the Vp-d relation can be used to capture the language-specific prosodic and performance distinctions of stress (or accent-related pitch) and speaking rate, respectively.

3.2.1 Kinematic correlates of stress in English and French. We begin by examining the kinematic correlates of stress distinctions in French and English. As noted earlier and shown in Table 1, stress distinctions in both languages showed consistent correlates in the lip-jaw kinematics. Stressed gestures were larger in displacement, longer in duration, and higher in peak velocity than unstressed gestures. The biggest difference in behavior of the individual kinematics between the two languages is in the absence of stress effect on the duration of closing gestures in French. That is, the duration of closing gestures in French is the same regardless of stress and its effect on displacement and peak velocity. This can be seen in Figure 6 in which the condition means (marginals) for duration are plotted as a function of whether the gesture is opening or closing for two speakers of each language.

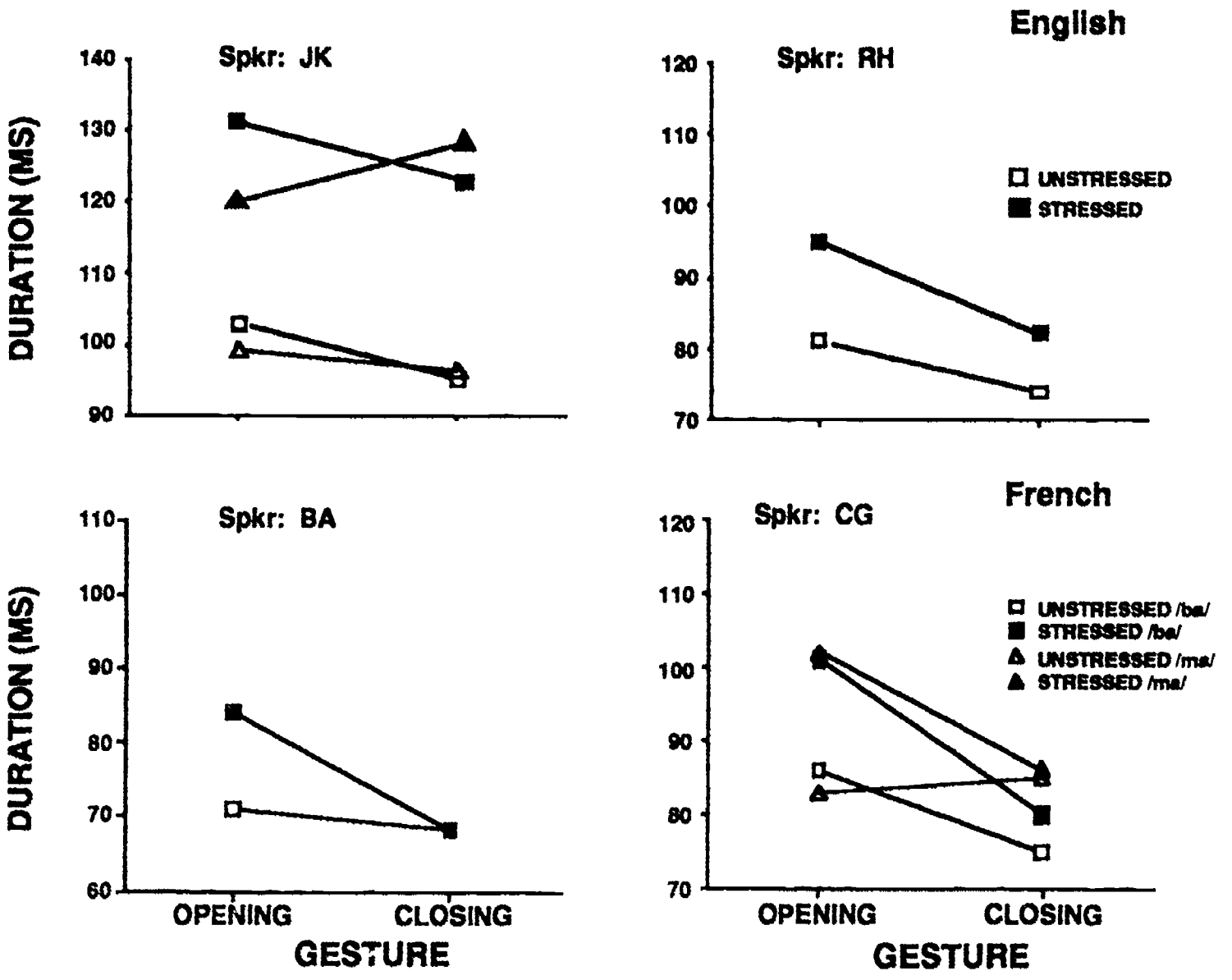


Figure 6. Mean gestural duration is plotted to show the interaction either of gesture type and stress or of gesture type, stress, and syllable type for two speakers each of English (top) and French (bottom). Filled and unfilled boxes (and triangles) denote stressed and unstressed gestures, respectively.

The question we consider next is whether the Vp-d relation can be used to uncover dynamic correlates of condition-specific distinctions in stress. Linear regressions of peak velocity against displacement were computed for each stress condition as a function of syllable identity, gesture type, and speaking rate. Similar to the findings of Kelso et al. (1985) for English, the condition-specific regressions had a very strong linear component for both stressed and unstressed gestures, accounting for most of the condition-specific spatiotemporal variability. As shown in Table 7, the linear component is even stronger for the French data.

Figure 7 shows best-fit regression lines for each stress condition plotted as a function of gesture type and speaking rate. Two things should be gleaned from the figure. First, while there is quite a bit of overlap in the distribution of the data, stressed and unstressed gestures tend to occupy different regions of the overall distribution. Second, regardless of language-specific differences in absolute slope, the Vp-d relation is generally steeper for smaller unstressed gestures than larger stressed gestures.

Slope comparisons show that speakers consistently differentiated Vp-d slopes for at least one gesture type and almost without exception in the direction predicted by the hypothesized relation between Vp-d slope and gestural duration. That is, steeper slopes were observed for that condition having the shorter mean duration (see Table 8).

Table 7. *Tabled below as a function of gesture type and syllable identity are number of observations (N), linear regression coefficient (r), and linear slope (m) for the condition-specific regressions of peak velocity on displacement (Vp-d).*

		Opening /ba/			Closing /ba/			Opening /ma/			Closing /ma/		
		N	r	m	N	r	m	N	r	m	N	r	m
ENGLISH													
RH	-str/N	154	.95	15.2	154	.89	18.6	154	.94	14.9	152	.90	19.9
	+str/N	119	.86	13.3	119	.79	14.8	120	.90	15.7	112	.79	17.2
	-str/F	151	.97	16.6	151	.96	22.6	150	.96	17.6	147	.94	23.0
	+str/F	120	.89	14.1	120	.89	19.6	118	.89	17.1	118	.89	20.6
MI	-str/N	157	.91	15.2	157	.91	20.1	158	.88	17.2	158	.89	21.4
	+str/N	120	.88	16.0	119	.81	17.2	117	.88	19.9	117	.69	15.4
	-str/F	159	.91	17.6	158	.88	21.3	151	.94	19.8	149	.91	24.1
	+str/F	120	.89	17.1	120	.87	18.9	114	.91	21.5	114	.[illegible]	20.2
JK	-str/N	153	.78	11.1	153	.86	14.0	146	.92	18.1	146	.86	13.4
	+str/N	115	.74	12.0	115	.64	9.8	117	.84	17.0	117	.56	7.2
	-str/F	155	.88	14.7	154	.87	16.6	149	.89	17.2	149	.86	16.7
	+str/F	115	.83	14.0	115	.81	14.1	114	.88	18.7	114	.56	9.3
FRENCH													
BA	-str/N	153	.93	21.5	153	.87	22.5	154	.92	21.3	154	.85	23.5
	+str/N	78	.94	18.1	78	.94	21.7	78	.94	20.9	78	.93	21.2
	-str/F	155	.91	21.7	156	.90	29.2	145	.91	23.6	145	.85	26.2
	+str/F	81	.94	21.1	81	.91	26.1	73	.94	21.0	73	.91	27.7
DP	-str/N	186	.93	24.6	186	.89	20.9	175	.95	27.3	175	.93	22.1
	+str/N	80	.92	20.0	80	.90	20.1	78	.95	21.3	78	.94	21.5
	-str/F	172	.94	26.3	172	.94	23.1	158	.91	26.4	158	.92	22.1
	+str/F	79	.94	21.3	79	.92	21.7	75	.92	23.5	75	.95	22.2
CG	-str/N	152	.94	20.3	152	.91	16.2	160	.94	21.1	160	.93	16.2
	+str/N	77	.93	15.4	76	.94	15.6	73	.91	15.6	73	.93	17.3
	-str/F	145	.89	19.2	143	.90	16.9	161	.93	22.4	161	.94	18.2
	+str/F	75	.95	17.2	76	.93	16.7	73	.94	17.4	73	.95	21.1

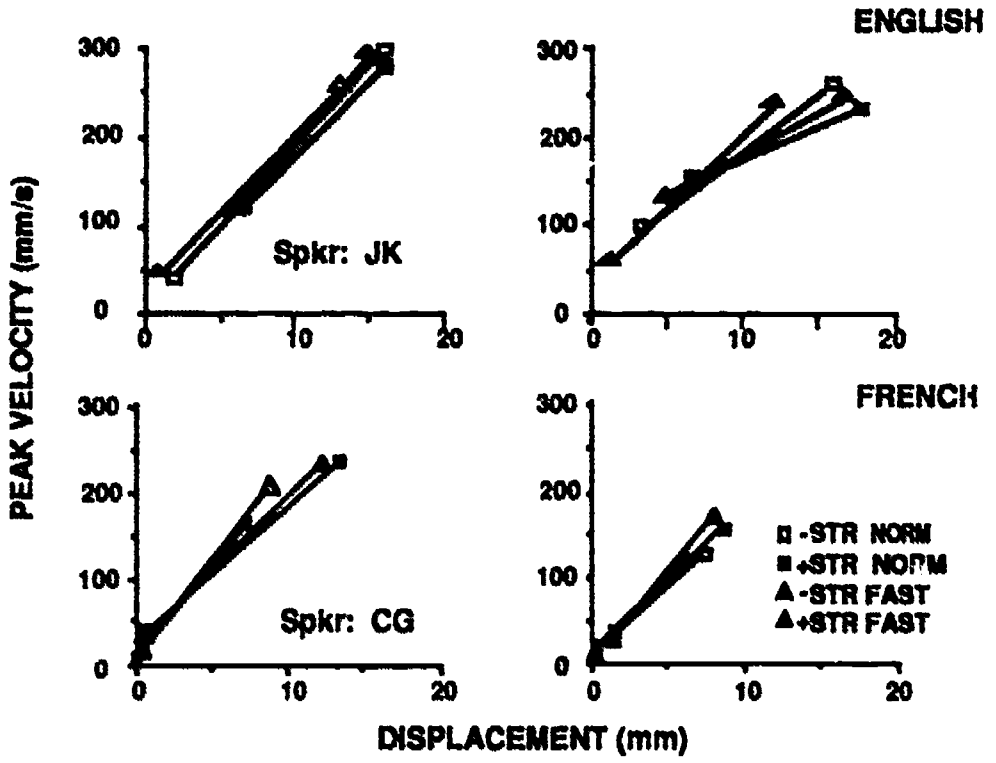


Figure 7. **Representative data are shown for opening (left) and closing (right) gestures of one syllable type for representative speakers of English (top) and French (bottom). Regressions of peak velocity (ordinate) on displacement (abscissa) for the four stress-rate conditions are depicted by best-fit lines whose lengths denote the range of variation on displacement.**

Table 8. *Mean gestural duration is given for each stress condition as a function of gesture type and syllable identity. Vp-d slopes were tested (one-tailed t) for stress differences. Asterisks indicate probabilities at the .05, .01, and .001 levels. 'X' indicates the difference is counter to that predicted.*

			Opening			Closing		
		Rate	Duration		Slope	Duration		Slope
			+str	-str	Comparison	+str	-str	Comparison
ENGLISH								
RH	/ba/	Norm	104	88	**	91	79	***
		Fast	88	69	***	76	64	***
	/ma/	Norm	103	88	ns	88	76	**
		Fast	86	72	ns	73	64	**
MP	/ba/	Norm	109	88	ns	94	83	**
		Fast	88	75	ns	75	71	*
	/ma/	Norm	105	90	** X	93	79	***
		Fast	89	79	ns	78	71	***
JK	/ba/	Norm	143	110	ns	134	99	***
		Fast	119	91	ns	115	83	**
	/ma/	Norm	132	105	ns	138	102	***
		Fast	107	90	ns	119	81	***
FRENCH								
BA	/ba/	Norm	96	73	***	73	72	ns
		Fast	73	67	ns	61	61	*
	/ma/	Norm	94	75	ns	74	74	*
		Fast	72	69	**	64	66	***
DP	/ba/	Norm	88	69	***	71	73	ns
		Fast	82	66	***	68	67	ns
	/ma/	Norm	87	72	***	70	73	ns
		Fast	86	69	*	67	68	ns
CG	/ba/	Norm	107	90	***	84	80	ns
		Fast	96	82	*	77	72	ns
	/ma/	Norm	105	88	***	90	90	ns
		Fast	98	79	***	82	80	*** X

Thus, two languages that differ in temporal organization and in the etiology and function of a prosodic distinction, which we are calling stress, show the same patterns of behavior in and among the kinematic variables of the lip-jaw complex. In particular, the highly linear condition-specific covariation of peak velocity and displacement accounts for the bulk of the within-condition spatiotemporal variability and shows that, despite variability in the absolute kinematic magnitudes, the relation between variables is stable within stress conditions. The potential universality of this phenomenon is further demonstrated in the next section for the data of Japanese in which the observed prosodic distinction is not perceived as even remotely similar to stress in French or English.

3.2.2 Kinematic correlates of accent related pitch in Japanese. In this section, we consider the kinematic correlates of high-low accent related tone distinction in Japanese. As described in detail by Vatikiotis-Bateson (1988), analysis of the individual kinematics revealed a surprising supralaryngeal instantiation of a tone distinction previously assumed to be strictly laryngeal. In particular, high tone gestures were generally associated with smaller lip-jaw kinematics than low tone gestures, despite differences in speaking rate, gesture type, and syllable identity. Though the magnitude of the kinematic difference was not as pronounced for the Japanese tone distinction as for French and English stress (see Table 1, above), the consistent patterning across speakers suggests the phenomenon is real.[7]

The condition-specific covariation of peak velocity and displacement was extremely linear (see Table 9). Best-fit lines depicting the slope and extent of the regression of peak velocity on displacement are given for two speakers in Figure 8. A major difference between these data and the stress condition data for English and French is the relatively small difference in mean displacement for the different tone conditions. This can be seen in the much larger overlap of the two conditions for Japanese.

Table 9. *Tabled below as a function of gesture type and syllable identity are number of observations (N), linear regression coefficient (r), and linear slope (m) for each condition-specific regression of peak velocity on displacement (Vp-d).*

	/ba/						/ma/					
	Opening			Closing			Opening			Closing		
JAPANESE	N	r	m	N	r	m	N	r	m	N	r	m
NK												
High/N	131	.71	12.7	131	.72	14.3	116	.77	15.5	116	.58	10.9
Low/F	151	.89	15.1	151	.83	13.0	146	.81	14.0	144	.59	8.6
High/F	130	.92	16.6	130	.87	18.0	128	.82	15.6	128	.68	14.6
FE												
Low/N	91	.94	18.2	92	.93	16.4	93	.92	20.8	93	.85	14.6
High/N	97	.98	18.9	97	.97	25.6	104	.97	21.9	104	.96	23.8
Low/F	75	.94	16.6	75	.96	18.0	92	.95	20.8	92	.93	16.6
High/F	67	.98	20.0	66	.96	25.6	94	.92	17.9	93	.95	21.3
ME												
Low/N	180	.89	17.5	150	.80	16.0	176	.84	17.0	158	.69	12.8
High/N	179	.88	19.3	168	.73	13.9	178	.91	19.3	177	.74	12.6
Low/F	164	.95	18.4	139	.92	19.6	158	.93	19.1	141	.87	18.0
High/F	160	.96	21.0	150	.96	22.7	156	.95	23.1	156	.91	17.8
SM												
Low/N	77	.90	20.2	77	.91	22.7	45	.94	22.1	45	.91	24.0
High/N	91	.89	20.4	91	.86	20.4	56	.95	29.1	56	.83	23.3
Low/F	49	.91	19.7	49	.93	21.7	83	.87	19.8	83	.82	20.0
High/F	57	.92	22.1	57	.94	23.8	107	.93	23.1	107	.86	21.2

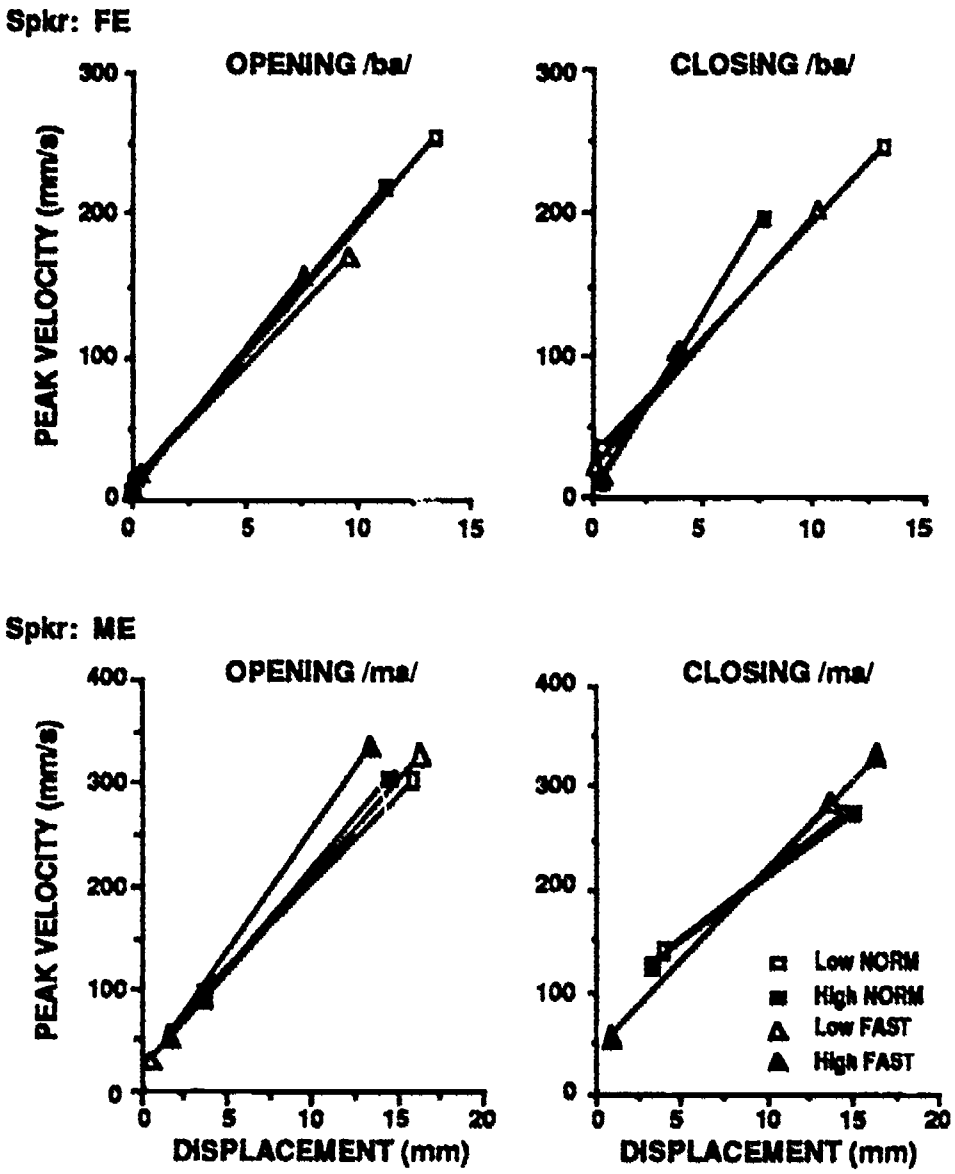


Figure 8. **Best-fit lines for the four tone-rate conditions depict the regression of peak velocity on displacement for two Japanese speakers' opening and closing gestures.**

Slope comparisons for the different tone level conditions are given as a function of speaking rate, gesture type, and syllable identity in Table 10. Although slope differences are not found for every comparison that we might predict, there is only one case of a difference contrary to the model prediction of steeper Vp-d slope for shorter mean duration. That is, slopes tend to be steeper for the condition (high tone) having the shorter mean duration, consistent with the modeling of articulatory behavior in terms of an hypothesized second order system with a variable stiffness parameter.

3.2.3 Kinematic analysis of speaking rate in three languages. In general, consistent kinematic correlates of speaking rate distinctions have been much harder to identify than those of stress (e.g., Gay, 1981; Gay, Ushijima, Hirose, & Cooper, 1974; Kelso et al., 1985; Kuehn & Moll, 1976; Ostry & Munhall, 1985; Ostry et al., 1983). The results of this study follow that tradition: movement gestures produced at faster speaking rates tended to be smaller in displacement, duration, and peak velocity, but the effect on the kinematics was neither as marked nor as consistent as that of the prosodic variable. This can be seen especially clearly in the relatively large percentage of slope differences among the rate-specific Vp-d relations that were contrary to the prediction of steeper slope for the faster rate condition (see Table 11).

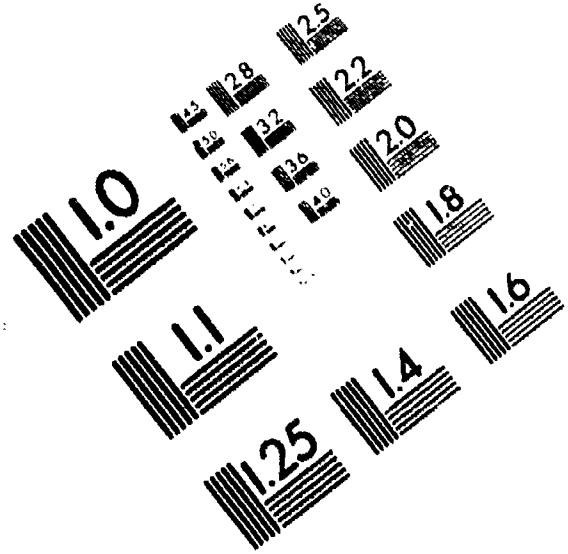
What is interesting about these results is their uniformity across the three languages, indicating that language-specific differences in absolute speaking rate do not influence how speaking rate distinctions are realized kinematically. That is, the temporal compression of Japanese and French compared to English does not seem to worsen the correlation between movement behavior and instructed speaking rate distinctions.

Table 10. *Mean gestural duration is given for each tone level as a function of gesture type and syllable identity. Vp-d slopes were tested (one-tailed) for tone level differences. Asterisks indicate probabilities at the .05, .01, and .001 levels. 'X' indicates the difference is counter to that predicted.*

			Opening			Closing		
		Rate	Duration		Slope	Duration		Slope
			Low	High	Comparison	Low	High	Comparison
NK								
	/ba/	Norm	89	83	ns	73	68	*
		Fast	72	64	ns	59	53	***
	/ma/	Norm	90	81	ns	77	72	ns
		Fast	72	64	ns	65	58	***
FE								
	/ba/	Norm	84	77	ns	77	72	***
		Fast	81	67	***	72	68	***
	/ma/	Norm	8	79	ns	80	78	***
		Fast	8	76	*** X	76	79	**
ME								
	/ba/	Norm	84	75	*	80	81	ns
		Fast	78	69	***	71	70	***
	/ma/	Norm	84	77	**	78	79	ns
		Fast	79	70	***	71	71	ns
SM								
	/ba/	Norm	73	68	ns	65	66	ns
		Fast	70	65	ns	60	61	ns
	/ma/	Norm	75	65	***	66	68	ns
		Fast	60	61	**	62	65	ns

Table 11. *Vp-d slopes were tested for speaking rate differences. T-values (one-tailed) and probabilities are given for each linguistic variable as a function of gesture type and syllable identity. 'X' indicates the difference is counter to that predicted.*

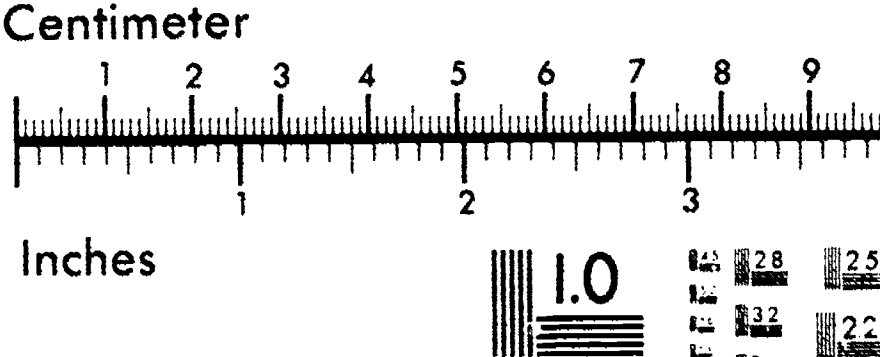
		Opening /ba/		Closing		Opening /ma/		Closing	
		t	p	t	p	t	p	t	p
ENGLISH									
RH	-str	2.63	***	4.26	***	4.26	***	4.27	***
	+str	0.79	ns	3.36	***	1.23	ns	2.54	***
MP	-str	2.76	***	0.86	ns	1.77	*	2.10	**
	+str	0.90	ns	1.13	ns	1.13	ns	2.52	**
JK	-str	3.70	***	2.37	***	0.88	ns	3.13	***
	+str	1.45	ns	2.93	***	1.21	ns	1.27	ns
FRENCH									
BA	-str	0.15	ns	4.26	***	2.03	**	1.51	ns
	+str	2.44	***	2.67	***	0.06	ns	3.63	***
DP	-str	1.63	ns	2.14	**	0.73	ns	0.06	ns
	+str	1.00	ns	1.01	ns	1.58	ns	0.51	ns
CG	-str	1.14	ns	0.73	ns	1.39	ns	2.80	***
	+str	1.72	*	1.11	ns	1.15	ns	3.14	***
JAPANESE									
NK	Low	0.18	ns	1.15	ns	0.22	ns	0.91	ns
	High	3.20	***	2.42	***	0.11	ns	1.85	*
FE	Low	1.63	ns	1.63	ns	0.03	ns	1.61	ns
	High	1.60	ns	0.05	ns	4.02	***	-2.25	**
ME	Low	1.08	ns	3.03	***	1.98	**	4.96	***
	High	2.29	***	7.53	***	4.14	***	5.42	***
SM	Low	0.28	ns	0.57	ns	1.22	ns	-1.67	* X
	High	0.92	ns	1.80	*	-3.47	***	0.69	ns



Association for Information and Image Ma

1100 Wayne Avenue Suite 1100
Silver Spring, Maryland 20910

301/587-8202



3.2.4 Summary of condition-specific effects. In summary, three languages of very different temporal and prosodic structure are remarkably uniform in the way linguistically relevant distinctions are reflected in the relation between kinematic variables of peak velocity and displacement. In each case, slope of the Vp-d relation tends to be steeper for the prosodic condition having the smaller mean duration and displacement. Thus, unstressed gestures in French and English and high tone gestures in Japanese had smaller mean kinematics and steeper Vp-d slopes than corresponding stressed and low tone gestures. Although significant slope differences were not always found for both opening and closing gestures, they were usually seen for at least one gesture type or the other. Furthermore, there were very few instances of reliable slope differences in the direction counter to that predicted—i.e., steeper slopes for the condition having longer mean duration; more often there was a failure to show a slope difference corresponding to an observed mean duration difference.

Speaking rate distinctions were realized in similar fashion; smaller, fast rate gestures tended to have steeper Vp-d slopes. However, and in keeping with previous work with English, the correlation was not as consistent as that observed for the prosodic distinction.

4 GENERAL DISCUSSION: CROSS-LANGUAGE COMPARISON

The main theoretical goal of the current study was to see the extent to which both universal constraints on movement behavior and language-specific scaling of those constraints could be characterized according to a simple dynamical model of articulatory behavior, such as that proposed by Kelso et al. (1985). In what follows, we first discuss the major empirical similarities and differences observed among the three languages. Then, we interpret these observations, in the context of the particular dynamical model proposed by Kelso et al. (1985).

4.1 Universal similarities

4.1.1 Durational asymmetry. The three languages showed approximately the same degree of asymmetry between the duration of opening and closing gestures: opening gestures were consistently longer in duration than closing ones, regardless for the most part of syllable identity, stress or tone, and instructed speaking rate.[8] Durational asymmetry, then, appears to be unrelated to differences in absolute speaking rate.

A similar asymmetry between opening and closing movements (i.e., faster closing than opening gestures) has been observed for other activities involving repetitive jaw movements such as chewing. Hiiemae and Crompton (1985) show for a variety of animals that the asymmetry persists despite differences in rate of chewing and substances chewed (e.g., liquids vs. solids). Durational asymmetry in these cases, then, may be simply a consequence of the anatomical and neurophysiological structure of the temporomandibular region. However, other anatomical structures also show temporal asymmetry between the two phases of repetitive movements, such as arm movements or foot-tapping in marking the beat of a metronome (Stetson, 1905). Whether asymmetry in these cases results from differences analogous to those of the jaw or from other constraints on rhythmic behavior is beyond the scope of this discussion. The point remains that the observed asymmetry may not be special to speech and may reflect dispositional constraints on the basic spatiotemporal patterning of successive movement gestures.

4.1.2 Relations among kinematic variables. Movement gestures that were consistently larger in displacement (e.g., stressed gestures in French and English) were almost always found to be longer in duration and higher in peak velocity. In addition to the covariation of kinematic means for each speaking rate and stress or tone level condition, the tendency for kinematic measures to covary was observed also on a gesture-by-gesture basis. Positive correlation among the kinematic variables was demonstrated across the full range of movement gestures through observation of both the highly linear covariation of peak velocity and displacement and the consistently less linear covariation of movement duration and displacement.

Condition-specific regressions of peak velocity on displacement were also highly linear and generally occupied distinct regions of the overall distribution. This included within-condition regressions for stress in French and English, tone level in Japanese, and speaking rate in all three languages.

In most cases, the slope of the linear regression relating peak velocity and displacement was reliably steeper for the condition having shorter mean movement duration. Of the remaining cases, there were almost no instances where the relation between Vp-d slope and duration were reversed—e.g., where steeper Vp-d slope accompanied longer movement duration. While the relation between peak velocity and displacement was highly linear and durational variance was quite small within

each condition, there were usually marked slope differences associated with differences in mean duration between conditions. Similarly and in keeping with the positive covariation of displacement and duration, steeper condition specific Vp-d slopes were associated with smaller mean displacements, hence the tendency for different conditions to occupy distinct regions of the overall regression function.[9]

The same condition-specific effects are observed in the kinematics and their interrelation for all speakers of the three languages investigated here, regardless of differences in syllable identity, speaking rate, or the value of the linguistic variable. That is, the same sort of difference is seen in both the mean gestural kinematics and in the slope of the Vp-d relation for: stress distinctions in English and French (while the stress distinctions themselves are arguably quite different); tonal distinctions in Japanese; and speaking rate distinctions in all three languages.[10]

The strength of the similarity observed among languages suggests that linguistically relevant movement behavior, for all its apparent diversity, is realized within fairly narrow limits. Observation of a fairly uniform durational asymmetry between opening and closing gestures, on the one hand, and the very consistently constrained relations among kinematic variables across different linguistic and performance demands, on the other, point to the existence of underlying universal constraints in speech production. To what extent these limits may be attributed to anatomical, neurophysiological, and/or 'higher order' cognitive mechanisms of rhythmical movement behavior is an open question.

4.2 Language specific differences

4.2.1 Absolute speaking rate. When measured as the number of opening-closing gesture pairs per second, English speakers had the slowest absolute speaking rates, while French and Japanese speakers had the fastest, with the Japanese slightly faster than the French. The speaking rate values of this study agree well with those reported by Dauer (1983) for these three and other languages of similar temporal organizations. Furthermore, language-specific differences in absolute speaking rate are matched fairly well by differences in the language specific slopes of the overall relation between peak velocity and displacement.

Dauer also noted a possible correlation among the languages she considered between absolute speaking rate, perceived temporal organization, and phonotactic constraints on syllable structure (cf. Scott, Isard, & de Boysson-Bardies, 1985). That is, the languages with the simplest canonical syllable structure have the fastest speaking rates and are most likely to be perceived as syllable- or mora- timed.

Evidence that produced and perceived syllable structure can be affected by speaking rate was provided long ago by Stetson (1951). In a series of frequency scaling tasks, he showed that CVC syllables produced repetitively are perceived as CV's at a certain speaking rate (between 3.5 and 4.5 Hz). Similarly, he showed that repetitively produced VC syllables are perceived as CV's at a certain rate. Thus, in the latter case, a sequence of /ip/ syllables would be heard as /pi/ at around 4 Hz. Although different syllable structures are not tested in this study, it is possible that speakers in the experimental situation will pick preferred speaking rates which conform to the syllable structural constraints of their language (For more recent treatments of this phenomenon, see Kelso, Saltzman, & Tuller, 1986a,b; Tuller & Kelso, 1990).

In non-speech perception tasks, it is well-known that, when presented with an otherwise featureless sequence of uniformly repeated sounds, we invariably impose an alternating rhythmic pattern (e.g., Householder, 1957; Morton, Marcus, & Frankish, 1976). In 1894, Bolton showed that the pattern imposed depends on the repetition rate of the stimuli. An isochronous sequence of telephonic blips resulted in a simple alternation when presented at 3-4 Hz. However, at higher repetition rates (6-8 Hz), listeners heard either large isochronous groups containing six to eight blips or alternating groups containing three to four blips each.[11] Thus, in addition to production constraints on syllable structure, it is quite likely that the perceptual system constrains particular temporal organizations to be best realized at specific absolute speaking rates (see below).

4.2.2 The peak velocity-displacement function. A second major difference observed between the three languages is that they showed different degrees of overall linear covariation among peak velocity and displacement; correlation coefficients were highest for Japanese and French and lowest for English. That is, the linear Vp-d regression, which accounts for most of the spatiotemporal variability of English, accounts for even more of the variability in the two languages produced at faster absolute speaking rates. Since the Vp-d relation is expressed in temporal units,

this finding corroborates the measured difference in durational variability observed among the three languages (see Table 1). Furthermore, it is consistent with temporal pattern differences that are implicit in the traditional categorizations of these languages as stress-, syllable-, and mora-timed. Thus, a somewhat poorer linear correlation between peak velocity and displacement (r = .89 for English), indicative of greater temporal variability, is just what the temporal alternation of stressed and unstressed syllables in English would predict. Similarly, Japanese and French, whose temporal organizations traditionally have been ascribed to the regular succession of the mora and syllable, respectively, show less temporal variability and higher values of the correlation coefficients (rs = .94).

We have seen that differences in the slope and goodness-of-fit of the Vp-d regression for the three languages do indeed match other aspects of their temporal behavior. We now ask whether these differences are related and, if so, what bearing do they have on the possibly analogous constraints on perception discussed above? That is, absolute speaking rate and temporal variability may be linked and plausible constraints on perception identified. If so, then it might be possible to rationalize the relation between language-specific differences in temporal organization and phonotactic structure by hypothesizing that absolute speaking rates are set in accordance with perceptual constraints that are linked to the complexity of a given language's temporal patterning.

The temporal results of this study show that absolute speaking rate and temporal variability are correlated within and across languages. It is commonly observed that mean duration and standard deviation are positively correlated and display fairly constant coefficients of variation. This can be attributed to the statistical nature of Poisson distributions—non-normal distributions that characterize a wide range of durational measures in speech production (Crystal & House, 1986; also, see Vatikiotis-Bateson, 1988). As speakers increase speaking rate (decreasing mean duration), temporal variability will decrease simply as a function of the shape of the temporal distribution of the data. Such skewing of the distribution is certainly observed in this study. In fact, the coefficients of variation (the proportional relation between mean and standard deviation; see Table 1) decrease somewhat across the different languages as mean duration decreases. In addition to obeying the statistical constraint that temporal mean and variance decrease together, this nonlinearity suggests that languages may carve out temporal niches in keeping with the specific distinctions being made and the general constraints on the system.

Perceptual constraints could condition the effect of speaking rate on perceived temporal organization. It is not yet clear what the limits are on how small a temporal discrepancy we can hear or produce in sequence, since they have been studied sporadically in a variety of contexts (e.g., Espinoza-Varas & Watson, 1986; Fujisaki, Nakamura, & Imoto, 1975; Lehiste, 1977; Morton et al., 1976). However, the evidence does suggest that temporal discrepancies must be at least 30-40 ms to be useful in rhythmic contrasts. Therefore, the duration differences between one- and two-mora syllables in Japanese (for details, see Vatikiotis-Bateson, 1988) and between stressed and unstressed syllables in English, which are about the same (35-40 ms), are large enough to be easily perceived. On the other hand, the overall temporal variability for one-mora gestures in Japanese is only 15-20 ms and, thus, unlikely to support a bimodal temporal distinction—e.g., between high and low tone syllables.[12]

Though necessarily speculative in many respects, this account provides a means of bringing together under one abstract scheme the traditionally disparate domains of phonetic implementation and linguistic structure (i.e., temporal organization and, perhaps, phonotactic constraints on syllable structure). It has been hypothesized that motion of the articulators can be described on a gesture-by-gesture basis in terms of a largely linear second order dynamical system (Kelso et al., 1985; Ostry et al., 1983). In such a framework, movement duration and hence speaking rate are reflected in the relation between peak velocity and displacement, which is expressed in temporal units proportional to stiffness. The greater the linearity of that relation, the greater the temporal stability (more uniform stiffness); and the steeper the Vp-d slope the faster the speaking rate (higher stiffness).

What this means is that the basic dichotomy in temporal organization among the three languages—viz., syllable and mora vs. stress timing—could result at least in part from scaling the dynamic parameter, stiffness, that controls gesture duration across a boundary imposed by constraints on perception. On the slow side of the speaking rate boundary, successive syllables would tend to alternate, while on the high side, they would clump into alternating groups (á la

Bolton, 1894). Without ascribing causality, syllable structure should be compatible with perceptual constraints on temporal organization; longer syllable durations (as in English) should accompany more complicated syllable structures, while faster rates of production should accompany simpler syllable structures (as in Japanese).

French and Japanese arguably could be in the midst of structural changes that have consequences for their temporal patterning and syllable structure constraints. As suggested by Vaissiere (1983) as well as by the results of the current study, French may moving away from perceptually isochronous (to English ears) syllable-timing towards an alternating stress pattern. Syllable structure constraints for French, while more restrictive than for English, allow fairly complex syllable structures. We would predict that as the shift towards stress-timing continues, speaking rate will slow accordingly.[13]

Similarly, Japanese is almost certainly moving towards more complicated syllable structures at the surface phonetic level and perhaps at other levels as well. As the incidence of heavy (two-mora) syllables increases, the overall speaking rate will slow down and any strictly temporal relation to light (one-mora) syllables is likely to give way to other styles of organization. For example, durations of light and heavy syllables could become compensatorily related as duration of heavy syllables become more variable due to increased diversity of syllable structures. In fact, there is a growing body of evidence from studies of acoustic duration that the mora as a measurable temporal unit has given way to inter-syllabic compensation constrained by fixed word duration (see Beckman, 1982; Dalby & Port, 1981; Port, Dalby, & O'Dell, 1986; Port, Maeda, & Al-Ani, 1980).

4.2.3 Displacement and its relation to duration. In English, the stronger linear covariation between displacement and duration reflects a gestural bimodality that is temporal (i.e., longer stressed vs. shorter unstressed gesture durations) as well as spatial. In French, however, this covariation is weaker, reflecting the fact that there was always a stress effect on displacement regardless of whether there was an accompanying effect on duration (Vatikiotis-Bateson, 1988). The overall temporal stability of movement cycle duration in French is partially due to the absence of a stress effect on the duration of closing gestures. In fact, unstressed closing gestures compensated somewhat for the durational effect of stress in opening gestures for two of the three speakers (ibid). Thus, for French, it is mean gestural displacement, from which we infer control of equilibrium position, that consistently varied with stress. It is possible, then, that equilibrium position is the primary controlled parameter for stress distinctions in both French and English.

The finding that tone level distinctions in Japanese have supralaryngeal effects similar to those of stress in English and French, does not necessarily argue against the conventional wisdom that such a distinction is primarily laryngeal. As discussed in the previous section, the production rate of one-mora (light) syllables is probably too fast to support a perceptible, bimodal temporal distribution. This adds to the unlikelihood that the supralaryngeal spatiotemporal behavior of single mora gestures is being intentionally exploited (see Footnote 8). Rather, heavy (two mora) syllables form the perceptible durational contrast with light syllables. Concomitant spatial bimodality, in Japanese reiterant speech at least, appears to be just one way of achieving the contrast—"held" gestures is another way (see Vatikiotis-Bateson, 1988). As with stress distinctions in French and English, equilibrium position could be the parameter that is varied to produce intentional spatial distinctions in articulatory kinematics.

Summarizing the situation for the three languages, we see stress distinctions marked for English by clear differences in both duration and displacement, while for French we see consistent differences only in displacement. In Japanese, we suggest that tone level effects on the supralaryngeal articulators may be a side effect of a primarily laryngeal event, but that intentional differences between heavy and light syllables, while primarily durational, may also be instantiated by modulation of gestural displacement. These variations in displacement and duration may be characterized in terms of an hypothesized model of gestural control in which modulation of underlying equilibrium position relates to displacement effects and modulation of stiffness to durational effects (Kelso et al., 1985). Inference of these underlying dynamic parameters of motion has proved useful in characterizing various aspects of articulator behavior common, perhaps, to all languages as well as language specific attributes of rhythm and prosody. Thus, stiffness indexes temporal organization differences among the three languages. The alternating rhythm of English is inseparable from the alternation of stresses. Longer gestures are also stressed,

therefore there is a clear difference in both parameters. In French, the rhythmic tendency is toward stable non-final syllable durations; but, we see stress effects on displacement that do or do not have concomitant effects on duration depending on gesture type. If the rhythmic pattern of French becomes one of alternation then the results for French and English should become less distinct. Stiffness differences are minimal, but clear, among single mora productions of Japanese, showing only the slight tendency to take longer to go farther (see below). Differential settings of equilibrium position are hypothesized to underlie intentional distinctions in all three languages regardless of temporal organization differences.

4.2.4 Limitations of the model. A main goal of this study has been to characterize universal and language specific aspects of rhythm and prosody in terms of hypothesized dynamical constraints on a single dimension of articulator motion. Gestural kinematic behavior was compared to that of an undamped linear mass-spring, and values were inferred for two underlying dynamic parameters, equilibrium position (spatial) and stiffness (temporal). In such a linear system, the slope of the Vp-d function is indicative of spring stiffness, k, and is inversely proportional to duration (T^{-2}). A strictly linear Vp-d relation would indicate constant duration and uniform stiffness. Comparing different Vp-d slopes as we have done in this study, steeper slopes would indicate greater average stiffness and shorter duration. Furthermore, if the system could be modeled completely as a linear mass-spring, there would be no relation whatsoever between displacement and duration.

Although this simple model does describe the data very well, it does not account for the consistently observed inverse relation between Vp-d slope and mean gestural displacement. Therefore, in this section, we consider two modifications of the model system that might better account for the observed tendency for Vp-d slope to decrease as displacement increases: the addition of either a nonlinear stiffness or a linear damping term.

Because stiffness, inferred from the slope of the Vp-d relation, tends to decrease as mean displacement increases, the system's behavior resembles that of a nonlinear "soft" spring. Furthermore, this "softening" of spring stiffness appears to vary among the languages in a lawful way. The effect of the nonlinearity is greater at the slower absolute speaking rates of English than French or Japanese as shown by the fact that the overall linear covariation of peak velocity and displacement accounts for a smaller percentage of the spatiotemporal variance in the English data than that of the other two languages. It is possible that the relative strengths of the linear and nonlinear stiffness components for the different languages vary according to changes in the linear component alone. That is, the observed correlation between absolute speaking rates and the degrees to which the overall distributions adhere to a linear stiffness function might reflect a language-specific variation of linear stiffness, but a language-independent nonlinear stiffness component.

In their analysis of English reiterant speech, Kelso et al. (1985) also observed this relation between gestural duration and displacement. However, when the relation between acceleration (second temporal derivative of position) and displacement was examined around the spatial midpoint (inferred equilibrium position) of the movement gestures, condition-specific differences in slope of the $\ddot{x}/x$ relation were observed. They concluded from this that the overall distribution of the data was composed of different, condition-specific linear spring functions, whose stiffness was actively modulated according to displacement.

Even though we used a very different approach here to assess the overall system stiffness, we still were unable to identify a stable nonlinear function for the data of this study, either for individual speakers or pooled across speakers within a language.[14] Thus, while there may be an overall nonlinearity that might be consistent for all speakers, we are left with the conclusion that a linear system of the sort originally proposed by Kelso et al. (1985) still best describes the articulatory motion, both within and across specific, intentionally defined subcomponents.

Another possible account for the inverse dependency observed between gestural displacement and Vp-d slope is to add a small linear damping term to the original mass-spring model. In particular, the system behaves as though it might be underdamped, especially for the production of larger (e.g., stressed in English) movement gestures. In an underdamped linear system, the effect of damping is to reduce peak velocity for a given displacement and decrease observed frequency, relative to the undamped system's behavior. Such an underdamped system could generate data, in keeping with the results of this study (i.e., slope of the Vp-d relation decreases and observed duration increases at larger displacements), provided that the system damping increased (decreased) with increases (decreases) in displacement extent.[15]

Possible physiological factors that might contribute to such covariation between damping and movement extent are increased viscosity of the mandibular joint and increased resistance to stretch in the muscles of the face and lip region for larger displacements of the jaw and lips. While our measured displacements were not extremely large (relative to possible non-speech displacements), our transduction of only the vertical movement of the lip-jaw complex could have increasingly underestimated the extent, but not necessarily the duration, of the true trajectory for our larger displacements. This would follow from the possibility that the combination of translational and rotational components of jaw motion (see Edwards, 1985; Edwards & Harris, 1990) might be quite different biomechanically at larger than at smaller displacements; specifically, the contribution of the rotational component could increase with displacement.

Although damping has been used to model speech articulator motion, e.g., the examination of sequences of Japanese nonsense syllables by Kiritani's group (Imagawa, Kiritani, Masaki, & Shirai, 1985; Masaki et al., 1985), modeling of the movement dynamics is yet to be done in terms of the effects of translation and rotation of the jaw on the movement trajectories. Once such results become available (Gracco, unpublished observations), we should be better able to judge whether the overall movement behavior is best described by condition-specific active changes of stiffness (Kelso et al., 1985) and/or possible changes of damping.

5 CONCLUSION

In the foregoing, we have shown that much can be learned about the spatiotemporal organization of speech articulation from simple kinematic analysis of unidimensional articulator motion during reiterant speech production. For the speakers of each language, it was seen that the highly linear relation between gestural displacement (d) and peak velocity (Vp) accounts for most of the overall spatiotemporal variance of the system. Similarly, when specific conditions of a linguistic variable and speaking rate were considered, the condition-specific distributions demonstrated the same highly linear relation between peak velocity and displacement and occupied overlapping but usually distinct regions of the overall distribution. It was further shown that slope of the Vp-d relation, whether for a specific condition or for the entire data set, reflected measured mean gestural duration. Finally, we noted that there is an inverse relation between condition-specific mean displacement and slope of the Vp-d function. This is in keeping with the observed tendency for larger movement gestures to take longer. From these facts, we conclude that the motion of the system can be characterized in terms of an abstract second order dynamical system, whose underlying parameters, stiffness and equilibrium position, can be quantitatively inferred from the observed discrete kinematic measures (duration, displacement, and peak velocity) and their interrelations—specifically, the Vp-d and displacement-duration functions.

By comparing the French, Japanese, and English data, it was shown that the results for all three languages are qualitatively the same, yet quantitatively differ in accordance with independently demonstrated differences in temporal organization and prosody. It was suggested further that the temporal organization differences observed among languages may be based primarily on the severely constrained interaction of absolute speaking rate, production constraints on syllable structure, and inherent constraints on the perception of temporal distinctions. Finally, although further analysis is required to more adequately characterize the slight but consistently observed nonlinearity of the system, we conclude from these results that articulatory motion can be modeled in terms of a small number of universal underlying dynamic parameters whose values can be appropriately set to meet language-specific criteria.

REFERENCES

Abercrombie, D. (1967). *Elements of general phonetics*. Chicago: Aldine.

Anderson, S. R. (1982). The analysis of French shwa: Or how to get something for nothing. *Language, 58*, 534-573.

Beckman, M. (1982). Segment duration and the mora in Japanese. *Phonetica, 39*, 113-135.

Bloch, B. (1950). Studies in colloquial Japanese IV: Phonemics. *Language, 26*, 86-125.

Bolton, T. L. (1894). Rhythm. *American Journal of Psychology, 6*, 145-238.

Boring, E. O. (1950). *A history of experimental psychology* (2nd ed.). New York: Appleton-Century-Crofts.

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

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Crystal, T. H., & House, A. S. (1986). Variation of timing control: Maturational or statistical? *Journal of the Acoustical Society of America, 79*, Suppl. 1, S54.

Dalby, J., & Port, R. (1981). Temporal structure of Japanese: Segment, mora and word. *Research in Phonetics* (Indiana University Phonetics Laboratory), *2*, 149-172.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics, 11*, 51-62.

Delattre, P. C. (1966). A comparison of syllable length conditioning among languages. *International Review of Applied Linguistics, 4*, 183-198.

Edwards, J. (1985). *Mandibular rotation and translation during speech.* Unpublished doctoral dissertation. CUNY.

Edwards, J., & Harris, K. (1990). Rotation and translation of the jaw during speech. *Journal of Speech and Hearing Research, 33*, 550-562.

Espinoza-Varas, B., & Watson, C. S. (1986). Temporal discrimination for single components of non-speech auditory patterns. *Journal of the Acoustical Society of America, 80*, 1685-1694.

Fairbanks, G. (1960). *Voice and articulation drillbook.* New York: Harper and Row.

Ferguson, G. A. (1981). *Statistical analysis in psychology and education.* New York: McGraw-Hill.

Fujisaki, H., Nakamura, K., & Imoto, T. (1975). Auditory perception of duration of speech and non-speech stimuli. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 197-220). New York: Academic Press.

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

Gay, T. J., Ushijima, T., Hirose, H., & Cooper, F. S. (1974). Effect of speaking rate on labial consonant-vowel articulation. *Journal of Phonetics, 2*, 47-63.

Higurashi, Y. (1984). *The accent of extended word structure in Tokyo standard Japanese.* Tokyo: Educa.

Hiiemae, K. M., & Crompton, A. W. (1985). Mastication, food transport, and swallowing. In M. Hillebrand, D. Bramble, D. Kiem, & D. Wake (Eds.), *Functional vertebrate morphology.* Cambridge, MA: Belknap.

Householder, F. W. (1957). Accent, juncture, intonation, and my grandfather's reader. *Word, 13*, 234-245.

Imagawa, H., Kiritani, S., Masaki, S., & Shirai, K. (1985). Contextual variation in the jaw position for the vowels in /CVC/ utterances. *Annual Bulletin (Research Institute of Logopedics and Phoniatrics, Tokyo), 19*, 7-19.

Jordan, D. W., & Smith, P. (1977). *Nonlinear ordinary differential equations.* Oxford: Oxford University Press.

Kay, B. A., Kelso, J. A. S., Saltzman, E. L., & Schöner, G. (1987). The space-time behavior of single and bimanual rhythmical movements. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 178-192.

Kay, B. A., Munhall, K. G., V.-Bateson, E., & Kelso, J. A. S. (1985). Processing movement data at Haskins: Sampling, filtering, and differentiation. *Haskins Laboratories Status Report on Speech Research, SR-81*, 291-303.

Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986a). The dynamical perspective in speech production: Data and theory. *Journal of Phonetics, 14*, 29-59.

Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986b). Intentional contents, communicative context, and task dynamics: A reply to the commentators. *Journal of Phonetics, 14*, 171-196.

Kelso, J. A. S., & Tuller, B. (1984). A dynamical basis for action systems. In M. S. Gazzaniga, (Ed.), *Handbook of cognitive neuroscience.* New York: Plenum.

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

Kozhevnikov, V. A., & Chistovich, L. A. (1966). *Rech, Artikulyatsiya, i vospriyatiye, [Speech: Articulation and perception] (30*, p. 543, originally published 1965). Washington, DC: Joint Publications Res. Service.

Kuehn, D. P., & Moll, K. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics, 4*, 303-320.

Larkey, L. S. (1983). Reiterant speech: An acoustic and perceptual evaluation. *Journal of the Acoustical Society of America, 73*, 1337-1345.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics, 5*, 253-264.

Liberman, M. Y., & Streeter, L. A. (1978). Use of nonsense-syllable mimicry in the study of prosodic phenomena. *Journal of the Acoustical Society of America, 63*, 231-233.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America, 35*, 1773-1781.

Lindblom, B., & Rapp, K. (1973). Some temporal regularities of spoken Swedish. *Papers from the Institute of Linguistics* (University of Stockholm), *21*, 1-59.

Masaki, S., Shirai, K., Imagawa, H., & Kiritani, S. (1985). Differences in jaw opening for vowels due to speaking rate and word-internal position in the production of vowel sequence words. *Annual Bulletin (Research Institute of Logopedics and Phoniatrics, Tokyo), 19*, 29-46.

McCawley, J. (1978). What is a tone language? In V. A. Fromkin (Ed.), *Tone: A linguistic survey.* New York: Academic Press.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America, 53*, 1070-1082.

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review, 83*, 405-408.

Nelson, W. L. (1983). Physical principles of economies of skilled movements. *Biological Cybernetics, 46*, 135-147.

Ohala, J. J., Hiki, S., Hubler, S., & Harshman, R. (1968). Photoelectric methods of transducing lip and jaw movements in speech. *UCLA Working Papers in Phonetics, 10*, 135-144.

Ostry, D. J., Cooke, J. D., & Munhall, K. G. (1987). Velocity curves of human arm and speech movements. *Experimental Brain Research, 68*, 37-46.

Ostry, D. J., Keller, E., & Parush, A. (1983). Similarities in the control of speech articulators and the limbs: Kinematics of tongue dorsum movement in speech. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 622-636.

Ostry, D. J., & Munhall, K. G. (1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America, 77*, 640-648.

Pike, K. L. (1943). Phonetics. *Language and literature (Vol. 21).* Ann Arbor: University of Michigan Press.

Port, R. F., Al-Ani, & Maeda, S. (1980). Temporal compensation and universal phonetics. *Phonetica, 37*, 235-252.

Port, R. F., Dalby, J., & O'Dell, M. (1986). Evidence for mora timing in Japanese. *Research in Phonetics (Indiana University Phonetics Laboratory), 5*, 1-36.

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

Scott, D., Isard, S. D., & de Boysson-Bardies, B. (1985). Perceptual isochrony in English and French. *Journal of Phonetics, 13*, 155-162.

Scripture, E. W. (1899a). Researches in experimental phonetics. *Yale Psychological Studies, 7*, 1-101.

Scripture, E. W. (1899b). Observations on rhythmic action. *Yale Psychological Studies, 7*, 102-108.

Selkirk, E. O. (1978). The French foot: On the status of 'mute' e. *Journal of French Linguistics, 1*, 141-150.

Smith, C. L., Browman, C., & McGowan, R. S. (1988). Applying the program NEWPAR to extract dynamic parameters from movement trajectories. *Journal of the Acoustical Society of America, 84*, S128.

Stetson, R. H. (1905). A motor theory of rhythm and discrete succession. II. *Psychological Review, 12*, 293-350.

Stetson, R. H. (1951). *Motor phonetics: A study of speech movements in action* (2nd ed.). Amsterdam: North Holland. (First ed. 1928 in *Archives Neerlandaises de phonetique experimentale, 3*).

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

Tuller, B., & Kelso, J. A. S. (1990). Phase transitions in speech production and their perceptual consequences. In M. Jeannerod (Ed.), *Attention and performance XIII.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Vaissiere, J. (1983). Language-independent prosodic features. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and measurement.* New York: Springer-Verlag.

Vatikiotis-Bateson, E. (in preparation). Remote and local correlates of Japanese lexical accent in supralaryngeal articulation.

Vatikiotis-Bateson, E. (1988). *Linguistic structure and articulatory dynamics.* Bloomington: Indiana University Linguistics Club.

Vatikiotis-Bateson, E., & Fowler, C. (1988). Kinematic analysis of articulatory declination. *Journal of the Acoustical Society of America, 84,* S128.

Wallin, J. E. W. (1901). Researches on the rhythm of speech. *Yale Psychological Studies, 9,* 1-142.

Wenk, B. J., & Wioland, F. (1982). Is French really syllable-timed? *Journal of Phonetics, 10,* 193-216.

FOOTNOTES

**Journal of Phonetics,* in press.

†Center for Complex Systems, Florida Atlantic University, Boca Raton, FL.

[1]It must be remembered that the tendency to think of movements as cyclical or composed of opening and closing gestures such as those analyzed in this study is rather arbitrary, albeit conventional. For example, instead of defining movement gestures from peaks and valleys of the position trace, they could have been defined from peaks and valleys of velocity as suggested by Browman and Goldstein (1986). In fact, in ordinary (i.e., non-reiterant) speech, where the normal variety of articulators is involved, we are more likely to see bilabial consonant-to-low-vowel sequences that consist of closing-opening movements of the lower lip-jaw centered around the consonant. Also, with the exception of the production of underlying Japanese geminates, there are no gestural forms other than simple opening and closing gestures—i.e., there are no plateaus or broad valleys in which time passes but the articulator doesn't move (see Vatikiotis-Bateson, 1988).

[2]The one exception to this is speaker JK's /ma/ productions in which there was no durational difference between opening and closing gestures due to a stress interaction. Specifically, her stressed opening gestures were shorter in duration and had smaller displacements than stressed closing gestures and, for that matter, stressed opening /ba/ gestures. This anomaly has consistent (anomalous) consequences for the correlation of Vp-d slope and duration in the comparisons of both opening and closing gestures and condition-specific stress correlates (see Tables V and VII and the discussion in § 3.1.3).

[3]Specifically, in an undamped linear mass-spring system, k is expressed in temporal units, ω_0^2, and $Vp = \omega A$, where ω is angular frequency and A is half the peak-to-valley displacement. Therefore, within this framework, it should be clear that the relevant variable of time control need not be duration *per se*, since duration may be recovered from the slope of the Vp-d relation. For example, the linear scaling of displacement and peak velocity, which accounts for the bulk of the observed spatiotemporal variability of these data, could be the result of physiologically specifying different levels of muscle activity, but constrained by an abstract stiffness setting.

[4]Slopes of linear regressions may be compared using a test for parallelism, which results in t-values, conservatively adjusted for different N's (Cohen & Cohen, 1975). We are indebted to Randy Flanagan and David Ostry of McGill University for passing on the following formula.

$$t = \frac{M_2 - M_1}{\sqrt{\left(\frac{(N_1 - 2)(RMS_1) + (N_2 - 2)(RMS_2)}{(N_1 + N_2 - 4)}\right)\left(\frac{1}{(N_1 - 1)(VAR_1)} + \frac{1}{(N_2 - 1)(VAR_2)}\right)}}$$

This test is used for slope comparisons throughout this study. Since specific predictions are made concerning the direction of the difference, a one-tailed test is appropriate (Ferguson, 1981). Although one-tailed tests make it easier to obtain reliable differences in favor of the claims being made, reliable differences contrary to those claims are also more easily obtained. Given the large N's involved, difference will be considered reliable at the 5% level if t = 1.645.

[5]What is particularly interesting about NK's data, but at this stage uninterpretable, is that they may be indicative of a different production strategy. This speaker is a highly trained speech therapist who, unlike the other speakers of the study, had an opportunity to practice the stimuli reiterantly before the experiment. The relatively tight linear covariation between displacement and duration, despite values of temporal variability and absolute speaking rate typical of the other Japanese and French speakers, could be the result of a more practiced speaking style.

[6]Note that the model prediction (e.g., Kelso et al., 1985) that steeper Vp-d regressions accompany higher movement frequencies breaks down for the Japanese and French data in that the mean half-cycle frequency is slightly higher for Japanese but the slope of the Vp-d regression is less. Although the anomaly is at this point unresolved, it does point up an interesting statistical problem facing corpora of this sort. There are a large number of data points for each speaker resulting in very large within-language N's. Even when the effect of highly correlated samples are partialed out, the N's are so large (3000-5000 per language) that minuscule differences in slope or mean duration are highly significant. On the other hand, the small numbers of speakers analyzed for each language comprise very small samples of the population. Possibly, then, data for more speakers would provide a better balanced within-language spread of data, thus offsetting the statistical power of the large N per speaker. Of course, it is quite possible that these results point up other limitations of the analysis. For example, there might be subtle language-specific differences in the relation between opening and closing gestures, which has been investigated only temporally in this study. Or, the correct equation of motion might require specification of additional terms, which themselves may be susceptible to language specific parametrization (see § 4.2.4). The situation is less clear for closing gestures in that there is a larger Vp-d slope difference between Japanese and French closing gestures, and less of a durational difference, than there is for opening gestures. We believe this to be due at least partly to the reduction of Vp-d slope for Japanese closing gestures, stemming from the confounding effect of including pregeminate movements for speakers NK and FE.

[7]When distinct pitch registers are produced, it is often observed that, in addition to greater tension of the vocal folds, high tones are produced with a concomitant raising of the larynx. This effectively shortens the vocal tract and raises formant

frequencies. Since formant height is proportional to jaw position in open tube vowels such as /a/—the lower the jaw the higher the formant values—it is possible that speakers may try to counteract the raising of formant values. Japanese speakers, then, could reduce or perhaps eliminate the difference in formant frequencies caused by the tone level distinction by not lowering the jaw so far during production of high tone vowels. [We are grateful to Arthur Abramson for helpful discussion of this issue.]

[8]The small observed effect of syllable identity on temporal asymmetry could belie aerodynamic or other articulatory differences in the production of oral and nasal stops. Both this and the finding that the mean kinematics are consistently larger for /ma/ than /ba/ productions need further investigation—e.g., observation of the upper lip kinematics and their timing relative to those of the lower lip-jaw complex reported here.

[9]This is exclusively a within-speaker phenomenon. There is no evidence of a correlation between absolute magnitude of kinematic variables and Vp-d slope across either languages or speakers within a language.

[10]In the case of Japanese, the ubiquity of this patterning is further supported by the two-mora productions of underlying /baa/ in *obaasan* (Vatikiotis-Bateson, 1988). These productions are longer in duration, larger in displacement, and higher in peak velocity than one-mora gestures. Thus, they occupy a distinct region of the spatiotemporal distribution. Furthermore, the relation between peak velocity and displacement is highly linear despite the restricted range of the distribution. Finally, the slope of the Vp-d relation is shallower for two-mora than one-mora gestures.

[11]Bolton used these findings to support Wundt's notion that attention is rhythmic, occurring in "waves" of fixed temporal span (discussed in Boring, 1950). Therefore, the effect of increased rate is to capture more blips in each part of the alternating wave. This notion may have some merit especially in the face of criticism (e.g., Wenk & Wioland, 1982) that the linguistic timing categories accepted for the last few decades are the peculiar result of English linguists' ears. That is, we are used to hearing the alternating sequence appropriate to relatively slow speaking rates. It would be interesting to see how the same study would turn out using listeners of the three languages discussed here.

[12]This is consistent with Delattre's analysis of non phrase-final stress contrasts in French as primarily due to differences in pitch and/or amplitude, rather than duration (Delattre, 1966). Also, as shown in §§ 3.2.1-2 above, there is a definite tendency for stress in French to be durationally distinguished only in opening gestures.

[13]While we have no evidence from colloquial French, it is not unreasonable to speculate that with the French language increasingly in the grip of broadcasters, further attention-getting contrasts will be introduced that will effect changes in the perceived temporal organization.

[14]Specifically, we computed nonlinear regressions for both individual speaker data and pooled data for each language. A Newton-Gaussian algorithm was used to fit the nonlinear function, $Vp^2 = \omega_0{}^2A^2 - \delta A^4$, in which A = 1/2 the peak-to-valley displacement and $-\delta A^4$ defines the contribution of stiffness nonlinearity to the overall relation between stiffness and amplitude. This Vp-d function is derived from a nonlinear spring function of the form $F_s = -\omega_0{}^2\Delta x + \delta\Delta x^3$; where F_s = spring force, $\omega_0{}^2$ = (angular frequency)^2 = k = the mass-normalized, linear stiffness coefficient, and δ = the nonlinear stiffness coefficient. This Vp-d relation was derived (with the invaluable assistance of Richard McGowan) for the mass-normalized nonlinear spring system using the harmonic balance method (e.g., Jordan & Smith, 1977). Note that when $\delta = 0$, this reduces to the familiar linear case of $Vp = \omega_0 A$. The regression analysis estimated ω_0 and δ, which resulted in reasonable estimates of ω_0. However, in every case, the asymptotic correlation of ω_0 and δ was extremely high, indicating that the solution was unstable for these two parameters. Given that the linear Vp-d function, which estimates ω_0, already accounts for the bulk of the system's spatiotemporal variance, we conclude that the instability of the nonlinear function is due to the coupling of ω_0 and δ (i.e., active variations in one are rigidly linked to variations in the other).

[15]In the undamped linear system, the relation between peak velocity and displacement is simply $Vp/A = \omega_0$, where A is half the peak-to-valley displacement. In the underdamped linear system, we assume the ratio of exponential growth to natural frequency, ε, to be small and derive the following expressions: a) $\omega \approx \omega_0 (1 - \varepsilon)$, where ω is the observed (damped natural) frequency; and b) $Vp/A \approx \omega_0 [1 + (\pi/2 - 1)\varepsilon] + O(\varepsilon^2)$, where the final term denotes the second-order error term. Therefore, for a given displacement, the peak velocity and, hence, the slope of the Vp-d relation will be less than in the undamped system. It is also important to note that, because the system is linear, the condition-specific regressions of peak velocity on displacement will still be linear. Again, we are indebted to Richard McGowan for his patient assistance.

Haskins Laboratories Status Report on Speech Research
1990, SR-103/104, 95-110

Gestural Specification Using Dynamically-defined Articulatory Structures*

Catherine P. Browman and Louis Goldstein†

1 INTRODUCTION

In recent years, we have been pursuing an approach to phonetics and phonology that invokes dynamically-defined articulatory gestures as the basic units. In other papers we have outlined the theoretical motivation for this approach,[1] pursued some implications for historical change and casual speech,[2] discussed how distinctiveness could be captured within gestural structures,[3] and explored the relation between a phonology of articulatory gestures and other nonlinear phonologies.[4] One basic tenet in all these papers has been that much is missed when the line between phonological patterning and physical processes is drawn too firmly. The strong form of our view proposes that phonological structure resides in the organization of the physical actions involved in speaking. Thus, we call the approach we have been pursuing an "articulatory phonology."

However, we also think that characterizing speech in terms of dynamical gestures has much to offer regardless of one's hypotheses about the nature of phonology. Such a characterization uses a form of description that has proven useful in other domains of action (e.g., Cooke, 1980; Kelso, Holt, Rubin, & Kugler, 1981; Kelso & Tuller, 1984a; Kugler & Turvey, 1987; Saltzman & Kelso, 1987), and thus relates speech activity to more general issues in motor behavior as well as drawing upon principles and techniques from this area. Moreover, it provides a framework that allows analytical and rigorous investigations of articulatory structure to be conducted in a way that makes direct contact with linguistic issues. From this perspective, then, the gestural framework could be pursued from within different phonological approaches as a phonetics of dynamically-specified articulatory gestures.

Within the framework being developed, the basic units are dynamically-defined articulatory gestures. These gestures are coordinative structures (Turvey, 1977) modeled in terms of task dynamics (Saltzman, 1986; Saltzman & Kelso, 1987). Task dynamics captures two important properties of gestures. First, the gestures are defined in terms of speech *tasks*, the formation and release of various constrictions such as bilabial closure (for [b]). Such tasks typically involve the coordinated motions of several articulators rather than the independent motions of individual articulators (such as the lower lip, upper lip, and jaw, in the example of [b]). Second, the gestures are defined in terms of the underlying *dynamics* that serve to characterize the motions. Such a dynamical description provides a representation that is itself time-free, and yet characterizes the articulatory movements through space and over time, as a function of the system's dynamical parameters. Thus, a dynamical description simplifies the relation between categorical and continuous characterizations of articulation, which is desirable from both a practical and a theoretical perspective. (Further discussions of the application of dynamics, and specifically task dynamics, to speech can be found in Browman & Goldstein, 1985; Fowler, Rubin, Remez, & Turvey, 1980; Hawkins, in press; Kelso & Tuller, 1984b; Ostry & Munhall, 1985; Saltzman & Kelso, 1987; Vatikiotis-Bateson, 1988).

To aid in making the gestural framework as rigorous and testable as possible, we are developing a computational system in conjunction with our colleagues Elliot Saltzman and Philip Rubin at Haskins Laboratories. Figure 1 portrays

Our thanks to Caroline Smith and Elliot Saltzman for critiquing earlier versions of this paper, to Joshua Katz for helping analyze the microbeam data, and to Diana Matson for aid in data collection and figure preparation. This work was supported by NSF grant BNS 8820099 and NIH grants HD-01994 and NS-13617 to Haskins Laboratories.

the components of the system: the linguistic gestural model specifies a gestural score (see below) given some input, the task dynamic model generates articulator movements given the gestural score, and the vocal tract model generates an acoustic signal given the articulator movements. The task dynamic and vocal tract models are fairly completely described in Saltzman and Munhall (1989) and Rubin, Baer, and Mermelstein (1981), respectively. The best description of the linguistic gestural model to date is in Browman and Goldstein (1987).

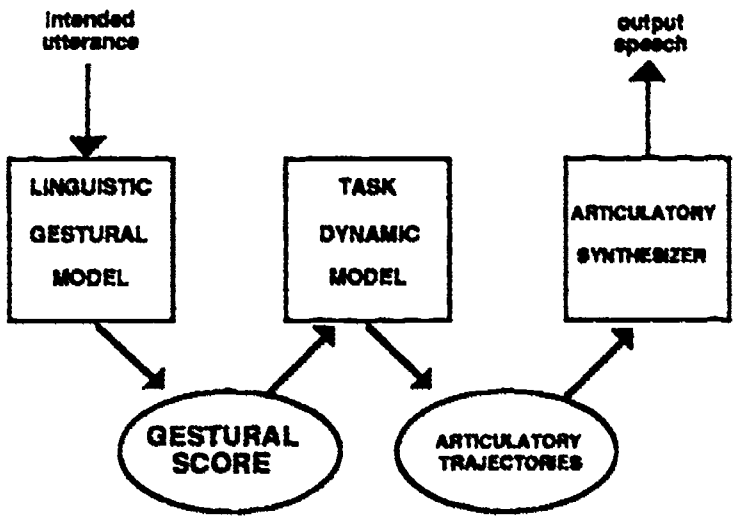


Figure 1. **Gestural computational model.**

In the current computational model, there are eight variables, called tract variables, that can be used to specify speech tasks. These variables, and the articulators that are coordinated to achieve the speech task for each variable, are shown in Figure 2, on the left and right respectively. Note that for oral gestures, the tract variables are paired, with a separate tract variable for each of the two dimensions of constriction formation: CL, the constriction location (i.e., the place along the wall of the oral cavity where the constriction is formed), and CD, the constriction degree (i.e., the size of the constriction). The gestures for an utterance are organized by phasing (and other) statements in the linguistic gestural model into a gestural score (see Figure 3) that contains the activation intervals (domain of active control) for each gesture, and the values of the dynamic parameters for each of the gesture's tract variables. (These parameters are identified and explained in § 2 below). Within each of the boxes in the figure, the values of the dynamical parameters are fixed, and serve to define the particular gesture in question. This gestural score is input to the task dynamic model, which uses the information about gestural activation and parameter values to generate the movements of the tract variables (shown superimposed on the gestural score in Figure 3).

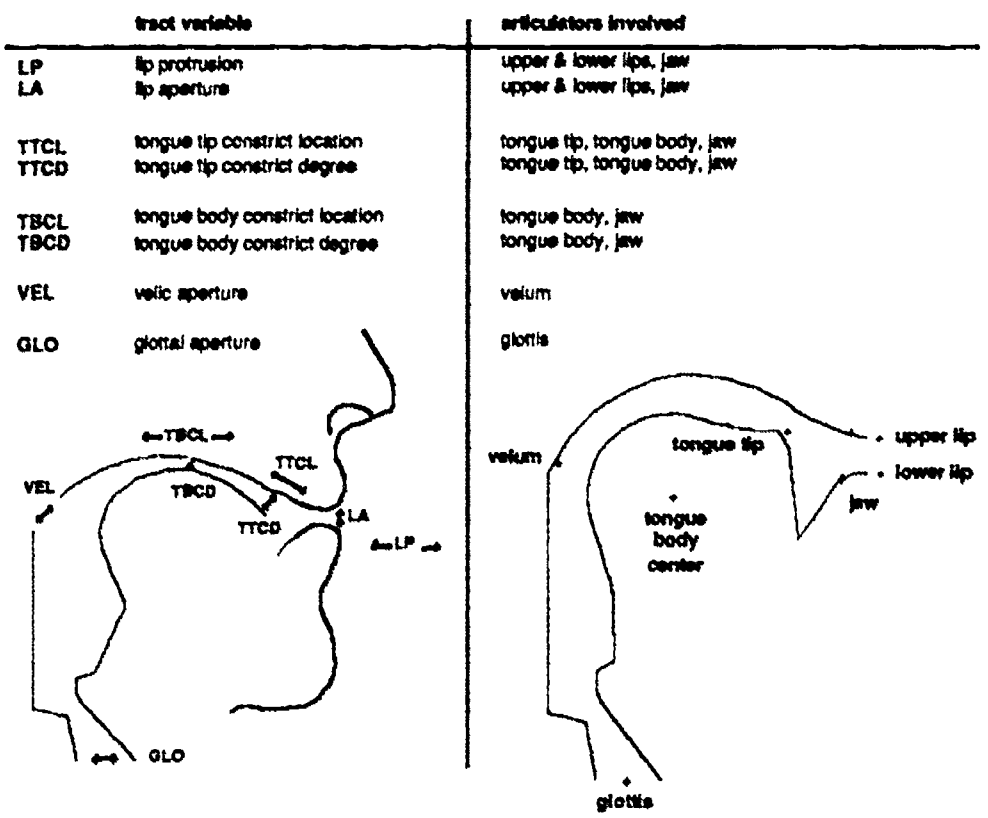


Figure 2. **Tract variables and the articulators involved.**

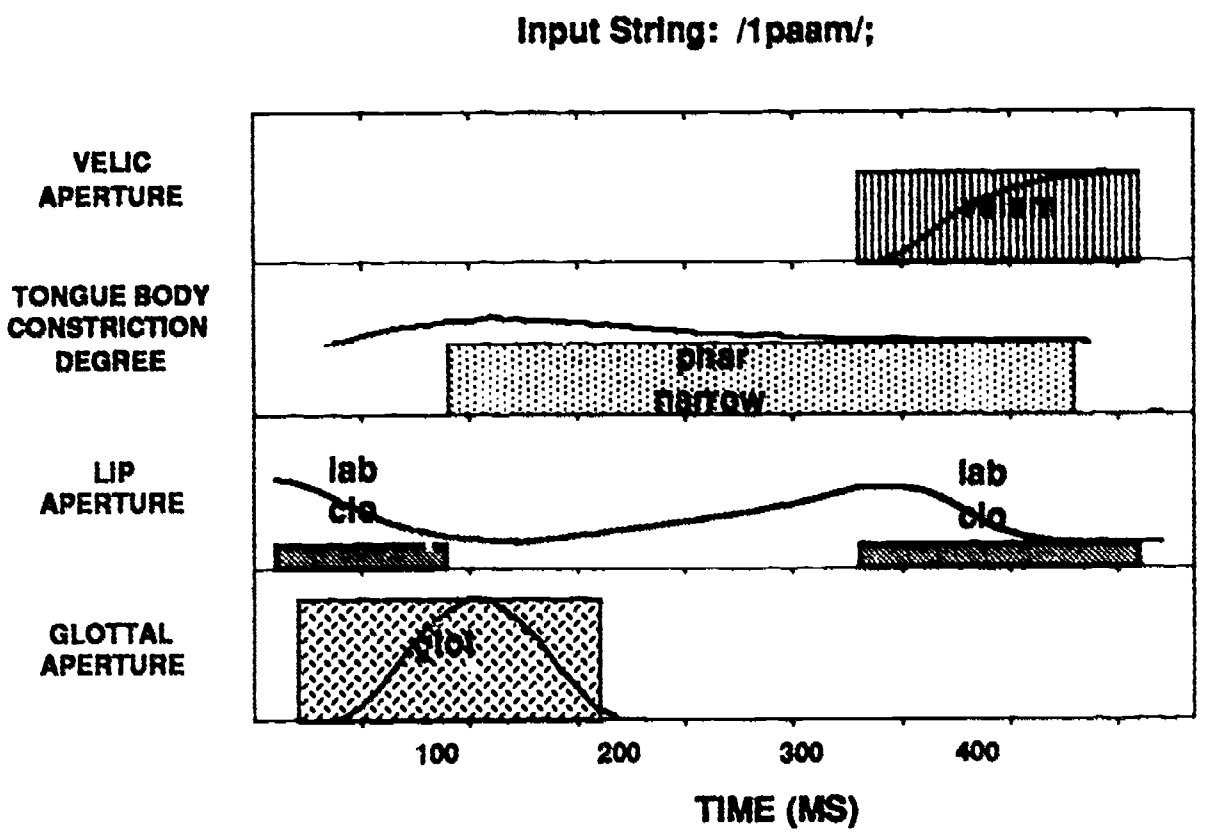


Figure 3. **Gestural score and generated tract variable motions for "palm" (pronounced [pɑm]). The input is specified in ARPAbet, so IPA /pɑm/ = ARPAbet input string /paam/. The boxes indicate gestural activation, and the curves the generated tract variable movements. Within each panel, the height of 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.**

The gestural score serves as an *input* specification, from which the movement of the vocal tract articulators and the resulting acoustic *output* unfold in a lawful fashion, as currently simulated by the task dynamic and vocal tract models. This clear separation between input and output can do useful work. That is, one of the benefits of gestural specification is that certain acoustic (and perceptual) properties of utterances can be accounted for as by-products or side-effects of their gestural organization. An example of a possible "by-product" perceptual effect will be considered in some detail in § 3, after gestural specification is discussed in § 2. We will begin, however, by clarifying the distinction between input and output further.

1.1 Input-output relations

Consider the case of a single gesture in isolation. Within the model, each gesture is a dynamical control regime that regulates the formation (and/or release) of a characteristic constriction within the vocal tract. The gesture's dynamical parameters include a "target" (equilibrium position) specification of the values for the location and degree of a particular vocal tract constriction (e.g., bilabial). When the gesture's regime is active, the associated articulatory synergy (for a bilabial, the upper lip, lower lip, and jaw) displays a characteristic time-varying response to this gestural input (calculated by the task dynamic model). Eventually, assuming the gesture is active long enough, the constriction targets are achieved. The articulator motion determines, in turn, the time-varying shape of the vocal tract and thus the acoustic output (computed by the vocal tract model). In this simple one-gesture universe, a given constriction will always yield the same acoustic output (ignoring the non-involved articulators), and thus the information captured by parameterizing the input (constriction) will be in 1:1 relation to the output (acoustics). If the nature of multi-gestural structures in speech were such that the gestures were produced in strict, non-overlapping, sequence, the choice of input or output would still not make a lot of difference. By the end of a given gesture, roughly the same acoustics would always be achieved (although the path to get there would depend on the state of the system left by the preceding gesture).

In real speech, of course, gestures overlap in time: they are co-produced (Fowler, 1980). Thus, the acoustic output associated with a given gesture will vary as a function of other concurrently active gestures. The overlapping of invariant articulatory gestures can account, among other things, for the varying acoustic frequency characteristics of stop consonant bursts and formant transitions in the environment of different vowels (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). A demonstration of how invariant articulations might yield varying acoustics is given in Öhman (1967), who shows midsagittal X-ray profiles in which the tip of the tongue reaches a relatively invariant target position for the consonant in /idi/, /ada/, and /udu/, while the shape of the tongue body and lips at the time of this consonant target are determined primarily by the vowel. That is, an invariant consonant gesture (produced with the tongue tip) is overlapping different neighboring vowel gestures (produced with the tongue body and lips). The acoustic properties of the stop closure and release will reflect the combined simultaneous effects of lips, tongue tip, and tongue body gestures. This kind of overlapping production ("coproduction") has been viewed as the cause of other kinds of "coarticulation" (e.g., vowel-to-vowel effects—Fowler, 1981), and of allophonic variation (e.g., nasalization of vowels before nasal consonants—Krakow, 1989).

The above discussion touched on the utility of specifying phonetic information in terms of gestural input. Turning now to the role of the acoustic output, in our view the output associated with a set of gestures is relevant to the "tuning" of the parameter values associated with individual gestures and their organization into assemblies (Browman & Goldstein, 1989, 1991; Goldstein, 1989). This tuning occurs during the development of language within an individual talker, and is also relevant to establishing the patterns of contrastive gestures that languages come to employ. While the development of gross constriction gestures of the lips, tongue tip, and tongue dorsum is a universal part of language development (as can be seen in babbling, e.g., Locke, 1983), the language-specific values of constriction location and degree associated with each of these gestures must be acquired by the child from listening to the acoustic output. (For example, the parameters of the tongue tip closure gesture are different in English and French, specifically in constriction location). The child's job may be made easier by the fact that languages tend to favor certain patterns of parameter values for gestures, as well as certain patterns of gestural organization. Contrastive values for constriction location and degree tend to evolve in such a way that the acoustic properties associated with a given set of parameter values are relatively stable (Stevens, 1972, 1989) and tend to differ sufficiently from the parameter values for other contrasting gestures (Lindblom, MacNeilage, & Studdert-Kennedy, 1983). Thus, there may well be systemic preferences for how gestures are parametrized that take into account their output, at least in ideal, careful speech contexts.

Output considerations do not, however, appear to actively constrain the processes of variation in speech production. Once the pattern of gestures for a given language is acquired, we argue, variation with respect to speaking style and prosodic context follows from very general principles of gestural overlap and magnitude that are blind to their acoustic consequences. The extreme case of this blindness is when gestures increase their overlap to the extent that one becomes completely hidden by others. For example, as discussed in Browman and Goldstein (1987, 1989) increase in overlap between (invariant) input gestures can lead to apparent (i.e., acoustic and perceptual) deletions and assimilations in the output. One gesture can be acoustically hidden by other concurrent ones. An example presented in these papers is the deletion of the final /t/ in "perfect" when in the phrase "perfect memory." X-ray evidence showed that the alveolar closure for the /t/ was still being produced in the fluent phrase, even though its acoustic consequences were completely hidden by the preceding velar closure and the following bilabial closure. From the point of view of the acoustic output, such changes in the production of a word are drastic, deleting all the criterial output properties of segments. However, all the input gestures are present; only their organization has been changed. Moreover, this change is hypothesized to be a very general characteristic of casual, fluent speech—gestures tend to show increasing overlap. These acoustic changes, and many other superficially unrelated ones, follow automatically from the gestural structures and this principle of variation. Thus, by specifying gestural input, rather than acoustic output, certain types of variation can be accounted for in general, explanatory ways.

2 GESTURAL SPECIFICATION

In the preceding section, we argued that it is possible to gain insight into various phenomena

by using a conceptual framework that describes phonetic structure in terms of overlapping input gestures (spatiotemporal articulatory units), and that distinguishes carefully between input and output. In many cases, relevant observations can be made (as in the case of "perfect memory") without detailed analyses of the dynamical characteristics of the gestures or their relationships. Nonetheless, to apply the framework more widely, and to see how (or whether) the quantitative aspects of the articulatory structure of speech and its variability can be accounted for, it is necessary to understand the details of exactly what is involved in specifying gestures using task dynamics.

A task dynamic specification of speech gestures is a constrained, reduced degree-of-freedom description compared to the continuous movement trajectories of multiple articulators that are observable. This can be seen in two ways, one relating to the notion of "task" and one to the dynamical aspects of the specification. First, the concept of the tract variable means that not all the articulators need to be analyzed individually, but rather articulatory actions can be combined into the linguistically significant task variables. Second, a dynamic description provides a way of characterizing all the points in a trajectory using only a few numbers (the values of the dynamic parameters).

Despite the constrained nature of task dynamics, it is often desirable to ease the analysis process by reducing the degrees of freedom further wherever possible. We do this by making simplifying assumptions (such as a constant damping ratio); the analysis of articulatory data in the light of these assumptions appears to lead to acceptable preliminary gestural specifications. We further assume that the results of a given data analysis apply as generally as possible, and thus treat results of simple analyses as hypotheses to be tested for general applicability. These simplifying assumptions and hypothesized generalizations are sure to be wrong in many respects, and ultimately need to be challenged and modified as appropriate. Nevertheless, we will include them in the overview below so that the reader may better evaluate this approach to reducing the dimensionality of articulatory description. In addition the overview will include indications of some directions of research into gestural specification that seem particularly promising to us.

In § 2.1, the dynamical control parameters will be described. In § 2.2, we will briefly look at the coordination of gestures into larger assemblies or constellations in terms of relative phasing parameters. In § 2.3. some possible types of prosodic gestural variation will be explored.

2.1 Individual gestures

2.1.1 Dynamical specification. Dynamical specification of a gesture requires first choosing the type of dynamical regime that will govern the motion of a particular tract variable (and its associated articulators). In the current formulation of the task dynamic model, the regime is always specified as a damped mass spring ("point attractor") model with constant mass, as shown in (1):

(1) $m\ddot{x} + b\dot{x} + k(x-x_0) = 0$

where

- m = mass (currently fixed at 1.0 in the model)
- b = damping
- k = stiffness
- $\ddot{x}$ = instantaneous tract variable acceleration
- $\dot{x}$ = instantaneous tract variable velocity
- x = instantaneous position of the tract variable
- x_0 = rest position of the tract variable

Specification of a gesture further requires that the values of the dynamical parameters in (1) (for the appropriate tract variable or variables[5]) be set, and that the (temporal) domain of active control for the tract variable(s) be delimited. The three parameters whose values must be specified for each tract variable are (a) the rest position x_0, (b) the stiffness k, and (c) the damping ratio ($b/(2[mk]^{1/2})$, from which the damping, b can be computed). Since oral gestures have CD and CL tract variables (see Figure 2), these three parameters must be specified for each. Each of the boxes in Figure 3 is defined by constant values for each of these three parameters. We will briefly explain what the parameters mean, and present some of the issues that must be resolved when using dynamically characterized gestures.

(a) The *rest position*, x_0, is related to the notion of "target"—it determines the tract variable position towards which the system moves. The system approaches the rest, or target, value most closely, without overshoot, in the case of critical damping. For oral gestures, questions involving specification of x_0 are related to the familiar characterization of phonetic units in terms of manner and place: constriction degree (CD) and constriction location (CL) respectively (cf. Fant, 1960; Stevens, 1989; Stevens & House, 1955).

The x_0 value for the gestures in the current model are based on available articulatory descriptions, tuned so that the vocal tract model output sounds appropriate.

(b) It is in questions of stiffness and damping ratio that a dynamic approach is distinguished from other approaches. *Stiffness, k*, determines (in conjunction with damping ratio) the durational characteristics of tract variable motion associated with the gesture: the stiffer the tract variable, the less time it will take to achieve its rest position, everything else being equal. In an undamped system, stiffness determines the frequency of oscillation. (It is important not to confuse the durational effects of stiffness with acoustic duration. As will be discussed in § 2.3, stiffness is only one of several possible determinants of acoustic duration). Stiffness does not have a long history as a phonetic descriptor; therefore, basic questions about it still need to be addressed. In the current formulation of the linguistic gestural model, only two underlying stiffness values are used, both derived from articulatory analyses—one for consonants and one for vowels. While this works quite well as a first approximation, further work is required to determine whether stiffness co-varies with x_0, and whether the stiffnesses of the two related tract variables (CD and CL) should differ from each other. Another question concerning stiffness involves the possibility that it could form the basis for natural classes. For example, gestures for glides might differ from those for vowels primarily in their stiffnesses (glides being stiffer); similarly, gestures for stops (and affricates) might be stiffer than those for fricatives.

(c) *Damping ratio* determines what happens when the tract variable approaches its rest position—whether it overshoots this value (underdamping), approaches it as a limit without overshooting (critical damping), or never approaches it very closely at all (overdamping). In the current model, all non-laryngeal gestures are assumed to be critically damped. However, this assumption needs to be further investigated. As with stiffness, it is also of interest to ask if the damping ratio helps define phonetic natural classes. For example, it is possible that flaps might be less highly damped than other gestures.

2.1.2 Parameter estimation. In order to address the above issues, it is necessary to estimate the values of the parameters from analyses of observed articulatory data. However, in so doing, it is important to realize that such analyses can only provide approximations to the gestural specification since gestures are comparatively abstract—they are not the articulatory movements themselves, but rather the functions underlying the observed movements. In some cases, the relationship between the observed movements and the underlying gestural regimes will be particularly opaque, such as when two simultaneously active gestures are affecting the same tract variables, e.g., the velar closure gesture and the vowel gesture in "key." In this example, both gestures are defined in terms of TBCL and TBCD, and are also partially overlapping in time. Thus, the observed tract variable motions will be affected by both gestures, making it difficult to separate out the contributions of the individual gestures. This suggests a strategy of first analyzing utterances in which co-active gestures involve distinct tract variables (e.g. bilabial consonants and unrounded vowels), and then generalizing to the more difficult cases. A further contributor to the difficulty of relating observed motions to underlying gestural regimes lies in the fact that the task dynamic framework currently provides no analytical procedure for dealing with the effects of physical mass and biomechanical constraints. However, bearing these caveats in mind, let us see how the various parameters can be estimated.

For x_0, a "target" value for location and degree of a constriction can, at least in principle, be estimated from examining articulatory data (such as lateral X-rays or X-ray microbeam data), no matter how difficult this is in practice. The stiffness and damping ratio of gestures must be determined by a mathematical analysis of articulatory movement data, preferably movement data from which it is possible to calculate an approximation to tract variable motion. If the damping ratio of the movements being analyzed is known (e.g., if there is reason to think that it is close to zero), then it is possible to estimate the stiffness as a function of the movement duration, or alternatively by the ratio of peak velocity to maximum displacement. The first technique has been employed by Browman and Goldstein (1985), the latter by Kelso, Vatikiotis-Bateson, Saltzman, and Kay (1985), Vatikiotis-Bateson (1988), and Beckman, Edwards, and Fletcher (in press), among others. If the damping ratio of the data is unknown, then it is possible to estimate parameter values for both stiffness and damping ratio using parameter estimation methods. One such method is used in a program currently under

development at Haskins Laboratories (McGowan, Smith, Browman, & Kay, 1988; McGowan, Smith, Browman, & Kay, 1990); this program assumes that a sequence of observed values was generated by a damped mass-spring system, and computes a least-squares estimate of the parameter values that could give rise to that sequence.

In order to employ these parameter estimation techniques, it is necessary to choose some stretch of time of a tract variable's motion during which the tract variable is assumed to be under active control of the dynamical system being fitted. This amounts to a hypothesis about the domain of active gestural control. Articulatory analyses have typically assumed that displacement extrema (in the articulator's or tract variable's time function) demarcate the edges of active gestural control (e.g., Kelso, Vatikiotis-Bateson, Saltzman, & Kay, 1985; Vatikiotis-Bateson, 1988). Recently this assumption has begun to be questioned (see Browman & Goldstein, 1985; McGarr, Löfqvist, and Story (submitted); Smith, Browman, McGowan, & Kay, submitted). In the current formulation of the linguistic gestural model, active gestural control is bounded by the edges of comparatively flat displacement "plateaus" (with the plateaus themselves being "uncontrolled").

The question of the domain of gestural activity raises issues that must eventually be resolved in tandem with other specification issues. For example, while the current computational model uses step functions to indicate regions of gestural control, it has been suggested that ramped onsets and offsets should be used instead (Perrier, Abry, & Keller, 1988). In such an approach, the generated shape of the tract variable motion would be influenced by the ramping function as well as the other parameters. Such additional degrees of freedom would be undesirable because of the additional complexity, but may prove necessary in the end to accurately model tract variable motions. Another question about the domain of gestural activity is perhaps of more immediate linguistic interest. In the current model, the three superficially distinct components of a gesture (formation of constriction, a "holding" phase [McGarr, Löfqvist, & Story, submitted], and release) are generated using separate domains of activation. However, it might be preferable to use a single domain of activation to encompass them all (or some intermediate possibility). The choice here interacts with the choice of damping ratio, in that a single critically damped regime could not generate all three components. Ultimately, the solution to this may need to also consider a wider range of dynamical regime types. For example, a periodic attractor (see Abraham & Shaw, 1982) might be a more appropriate regime for combining constriction and release components.

2.2 Gestural constellations

Since a typical utterance consists of more than a single gesture, the relations among gestures must be characterized in addition to the individual gestures themselves. Ultimately some kind(s) of dynamical self-organizing principles may be found that would serve to narrow the range of coordinative possibilities to a few distinct modes (e.g., Kay, Kelso, Saltzman, & Schöner, 1987; Turvey, Rosenblum, Kugler, & Schmidt, 1986). Preliminary attempts to find distinct coordinative modes among speech gestures have shown similarities to other motor tasks. For example, Kelso, Saltzman, and Tuller (1986) described a phase transition in the coordination of syllable-final bilabial closure and glottal opening gestures in /ip/. As repetition rate increased, there was an abrupt transition from syllable-final coordination to syllable-initial coordination that was similar to the kind of phase transition observed in repetitive finger-waggling (Haken, Kelso, & Bunz, 1985). Such research may provide avenues for future characterizations of gestural relations. At present, however, we find it necessary to pursue simpler approximations to the question of gestural organization in order to get an empirical handle on it.

In the linguistic gestural component of the present computational system, gestural coordination is specified in terms of the relative phasing of gestures (Browman & Goldstein, 1987; see Kelso & Tuller, 1987, and Nittrouer, Munhall, Kelso, Tuller, & Harris, 1988, for general discussions of using phasing in the specification of speech organization). Two gestures are coordinated by specifying two points, one in each gesture, that must coincide temporally. The points (in a gesture) are defined in terms of the phase of a "virtual" cycle whose duration (i.e., period) is determined by only the stiffness (natural frequency) ascribed to the gesture. In addition to phasing, this virtual cycle is used in the specification of gestural activation: A gesture is defined as beginning at the 0 degree point of this virtual cycle, and it remains active until some later phase in the virtual cycle. Thus, as the stiffness of a gesture changes, the amount of time it is activated will automatically change as well as

its temporal relation (as a whole) with other gestures, including those it is phased with respect to.

Given this general approach to specifying how the gestures constituting an utterance are coordinated with respect to each another, a number of issues arise. (a) First, for a given pair of gestures that are to be coordinated, the actual phases of the synchronized points must be determined. (b) Second, not every gesture in an utterance is phased with respect to every other gesture, so the sets of gestures to be coordinated must be determined. (c) Finally, in some cases phasing may occur with respect to larger gestural collectives rather than with respect to individual gestures. We now discuss these issues in turn.

(a) The decision as to which phases to synchronize (between two gestures) can be made by examining movement data and observing what pattern of synchronization seems most characteristic (or most invariant) across multiple tokens of the particular gestural structure (see Tuller, Kelso, & Harris, 1982, for an example of such an approach). An important question is whether there is a limited subset of points that languages use for phasing. In investigations to date using this model (see, for example, Browman & Goldstein, 1987), satisfactory results have been obtained by using only a few different points: the achievement of target and the onset of movement (either towards or away from a target), where points are based on intervals of active control defined using the edges of extrema plateaus rather than single extrema points. This is also consistent with the observations of Krakow (1989), who found that the phasing of the velum-lowering gesture for nasal consonants with respect to the oral constriction is "bistable"—it is phased either with respect to oral gesture onset or with respect to achievement of target (depending on syllable position). Much remains to be done on this important question, however. Moreover, it might ultimately be preferable to state an overall phasing between two gestures rather than to synchronize particular points. This could perhaps be done in terms of a coupling function between the two control regimes (e.g., Kay et al., 1987).

(b) A related issue is the choice of which gestures to coordinate with each other. Browman & Goldstein (1987) proposed that vocalic gestures are phased with respect to preceding (syllable-initial) consonantal gestures and that (syllable-final) consonantal gestures are phased with respect to preceding vocalic gestures. The phasing of V to C and C to V appears to work for English, at least as a first approximation, but the gestures that are coordinated may differ in different languages—for example, some languages may coordinate vocalic gestures directly. It is possible that the choice of gestures to be coordinated may be correlated with the prosodic nature of the language. For example, Smith (1988) has proposed (on the basis of acoustic evidence) that coordination might be C-V in languages such as Japanese that have been described as mora-timed but V-V in languages such as Italian that have been described as syllable-timed.

(c) In the discussion so far, it has been assumed that individual gestures are phased with respect to other individual gestures. It may, however, be the case that some gestures are organized into larger collectives or constellations for the purposes of coordination. It is to be hoped that such constellations would correspond to linguistically significant units such as segments, syllable onsets, or syllables—but this need not be the case. Note that the investigation of the relation of gestures to linguistically significant units can proceed in two (related) ways: some particular unit can be assumed, and gestural correlates searched for; or articulatory movements can be parsed and gestural units established on the basis of criteria such as cohesiveness and variability.

An excellent example of a study proceeding from linguistic unit to articulation can be found in Sproat & Fujimura (1989), in which both light and dark /l/ were discovered to be complex segments—gestural constellations—differing primarily in the relative timing of the tongue tip and tongue body gestures. An example of a study arriving at units beginning from an analysis of movement data can be found in Browman and Goldstein (1988), who suggested that gestures in the syllable onset form a unit for the purposes of coordination with the syllable nucleus, whereas coda gestures do not but rather are timed individually. Further questions can be posed relating, for example, syllable structure to variability in phasing and the amount of overlap between gestures, with tautosyllabic gestures expected to show more overlap and less variability than heterosyllabic gestures.

2.3 Prosodic variation of gestures

Thus far, the discussion has focussed on how the canonical forms of utterances are specified using gestures. However, it is also important to understand how the proposed gestural structures may be flexibly adjusted to yield the differences in

articulation (and acoustics) that are observed in different prosodic environments. Since the specification includes both the parameters of individual gestures and the phasing parameters that define intergestural organization, variation in either (or both) of these parameter sets is available as a way of characterizing prosodic differences.

One of the major aspects of prosodic variation involves temporal variation: for example, the acoustic duration of stressed syllables is typically longer than that of unstressed syllables, and phrase-final syllables are often lengthened acoustically relative to other syllables. In the dynamic approach, quantitative temporal information is provided not by specifying time directly but by specifying the parameters of the gestural regimes and their phasing. The pattern of relationships among the abstract dynamical coefficients of the model's equations automatically generates the quantitative temporal information as an inherent feature of the motion of the articulators. Thus, variation in acoustic duration could be a result of changes either in gestural parameters or in intergestural phasing. Consider the example "add." Changing the stiffness of the vocalic gesture for [ae] would (everything else being equal) change the amount of time required for the articulators to move to the configuration for [ae], and hence the acoustic duration associated with it. Changing the relative phasing between the vocalic gesture for [ae] and the following alveolar closure gesture for [d] would also change the acoustic duration of the [ae].

Given these two aspects of gestural specification, either of which will be associated with changes in acoustic duration, it is of interest to ask how these mechanisms are related to different types of prosodic variation. Stress, for example, has been associated with a decrease in the stiffness of the stressed gestures (Browman & Goldstein, 1985; Kelso et al., 1985). However, McGarr et al. (submitted) have suggested that the difference between stressed and unstressed vocalic gestures lies in the duration of a comparatively steady state position, rather than in the movement towards this position as would be expected with stiffness changes. Such a pattern would be consistent with a change in the phasing between the vowel and the following consonant, with the longer steady state portion resulting from the consonant's delay (decreased overlap). Alternatively, or in addition, the steady state portion might involve an increased interval of active control of the vowel gesture itself, or of its "holding" phase (if separate).

Beckman, Edwards, and Fletcher (in press) have investigated these various factors in English. Unlike previous studies, Beckman et al. explicitly compared the two types of gestural variation. They analyzed jaw movements in syllables that were either phrase-final or non-phrase-final, and also either accented or unaccented, in order to investigate possible gestural correlates of both phrase-final lengthening and accentual lengthening. They found that the increases in acoustic duration associated with final lengthening and with accent were articulatorily quite different, and suggested that these differences might be modelled by the two types of gestural variation. Specifically, subjects slowed down the movement of the jaw into the phrase-final closure, without concomitant changes in amplitude of movement. However, accented jaw movements had larger amplitudes compared to unaccented movements, and the movement itself lasted longer but was not slower (i.e., did not have smaller peak velocity). They suggested that the phrase-final lengthening might be modelled by reducing the stiffness of the final closure (at least at normal speech rates), whereas the accent effect might be modelled by increasing the intergestural spacing (phasing) for the accented syllables as opposed to unaccented syllables.

It is difficult to compare the results of the various studies on stress and accent, since the criteria being used to identify the hypothesized domain of gestural control are often different, or not clearly laid out. More research is clearly needed to resolve the apparent conflicts. However, the possibility that two different dynamic mechanisms might be associated with two different prosodic phenomena is quite intriguing. That is, the dynamic characterization might reveal physical differences between different prosodic phenomena that are lost in an description based on acoustic duration, in which the consequences of the different dynamic mechanisms merge.

3 REDUCED SYLLABLES

We are now in position to examine in more detail the kind of analyses that can be performed in a gestural framework. The primary example we will discuss in this section involves reduced syllables in English. The "vowel" portion of reduced syllables in English is difficult to characterize, and typically shows great contextual

variation. For example, consider the first syllable in the word "beret." It is sometimes produced in such a way as to be transcribed with the vowel [ə] (schwa) preceding the [ɹ] of the following syllable. In other cases, it may be represented as a syllabic "r": [ɻ] or [ɚ]. Finally, in casual speech, it may cease to be syllabic altogether, the tendency to do so being a "graded" one, dependent on a number of contextual factors (e.g., Dalby, 1984).

This variation can be thought of in one of two ways. First, the speaker might be selecting one of the three distinct variants, with the choice of variants being a probabilistic function of the style of speaking and the local environment. Alternatively, the speaker might be maintaining a single input gestural structure, with the different variants resulting from the independently motivated casual speech processes—increase in overlap and reduction in magnitude (Browman & Goldstein, 1987). From this latter perspective, the different variants are not selected from the lexicon, but rather represent different acoustic and perceptual consequences of completely continuous variation in talking processes—increasing overlap and reduction. To see how this would work, a gestural specification for such reduced "nuclei" that could yield this kind of behavior needs to be identified.

A candidate gestural structure is one in which the nucleus of a reduced syllable does not have any explicit vowel gesture associated with it. In this structure, reduced syllables would be characterized by an organization of the consonants preceding and following the nucleus such that the consonants show no overlap and thus produce an acoustic interval for the nucleus that is gesturally unspecified. (Note that this structure is the gestural equivalent of identifying a unit by a skeletal x-slot—timing information—but no melodic information). According to this hypothesis, the shape of the vocal tract during the interval between the two non-overlapping gestures would be determined both by the positions in which the articulators were left by the preceding gesture and by the movement of the articulators to their own specific neutral positions when they are not involved in any active gesture.

Given such a gestural structure, utterances such as "beret" and "bray" each contain the same gestures. The distinction between them is in the organization of the initial consonant gestures (bilabial closure and rhotic). In "bray," the bilabial closure and rhotic gestures should be tightly (and closely) organized with respect to one another, and to the vowel, according to the C-center hypothesis (for syllable onsets) outlined in Browman & Goldstein (1988). In "beret," we hypothesize that the C-center organization does not hold for these gestures, but rather the bilabial closure is set off from the rhotic and vowel gestures, showing no overlap with them. In casual speech, then, it would be possible for the degree of overlap to increase to the value more usually seen for "bray." This would then be perceived as a vowel deletion, or a loss of syllabicity.

To test the plausibility of this hypothesized difference in gestural structure, a series of gestural scores was constructed that differed only in the amount of overlap between the initial bilabial and rhotic gestures. If the gestural organizations for "bray" and "beret" differ only in overlap then we would expect listeners' percepts to shift from one to the other when this parameter is varied. The scores were generated, not by analyzing articulations of "bray" and "beret," but by using the existing generalized statements in the linguistic gestural model (with one exception to be discussed below) that had been derived from other articulatory analyses. The statement specifying the phase relation between the bilabial closure gesture and the rhotic gestures was then manipulated to produce the stimulus series. Figure 4 shows partial gestural scores for the two ends of the overlap continuum. The boxes show the activation intervals for the rhotic and bilabial closure gestures, and the superimposed curves show the vertical motions of the tongue tip and lower lip.

The rhotic was generated as a "complex" gestural constellation consisting of two simultaneous gestures, one controlling the tongue tip tract variables (producing a retroflex constriction on the hard palate) and one controlling the tongue body tract variables (producing an upper pharyngeal constriction).[6] Figure 5 shows the midsagittal outline of the vocal tract model when both of the gestures in the rhotic constellation have reached their targets. The complete gestural scores contained no overlap between the rhotic and the vocalic gesture. This differs from what would be produced by our generalized phasing statements and is unlikely to be correct, but we did not want to introduce any additional assumptions about how the tongue body gesture for the rhotic would blend with the tongue body gesture for the vowel (see Browman & Goldstein, 1989; Saltzman & Munhall, 1989 for a discussion of within-tract-variable blending).

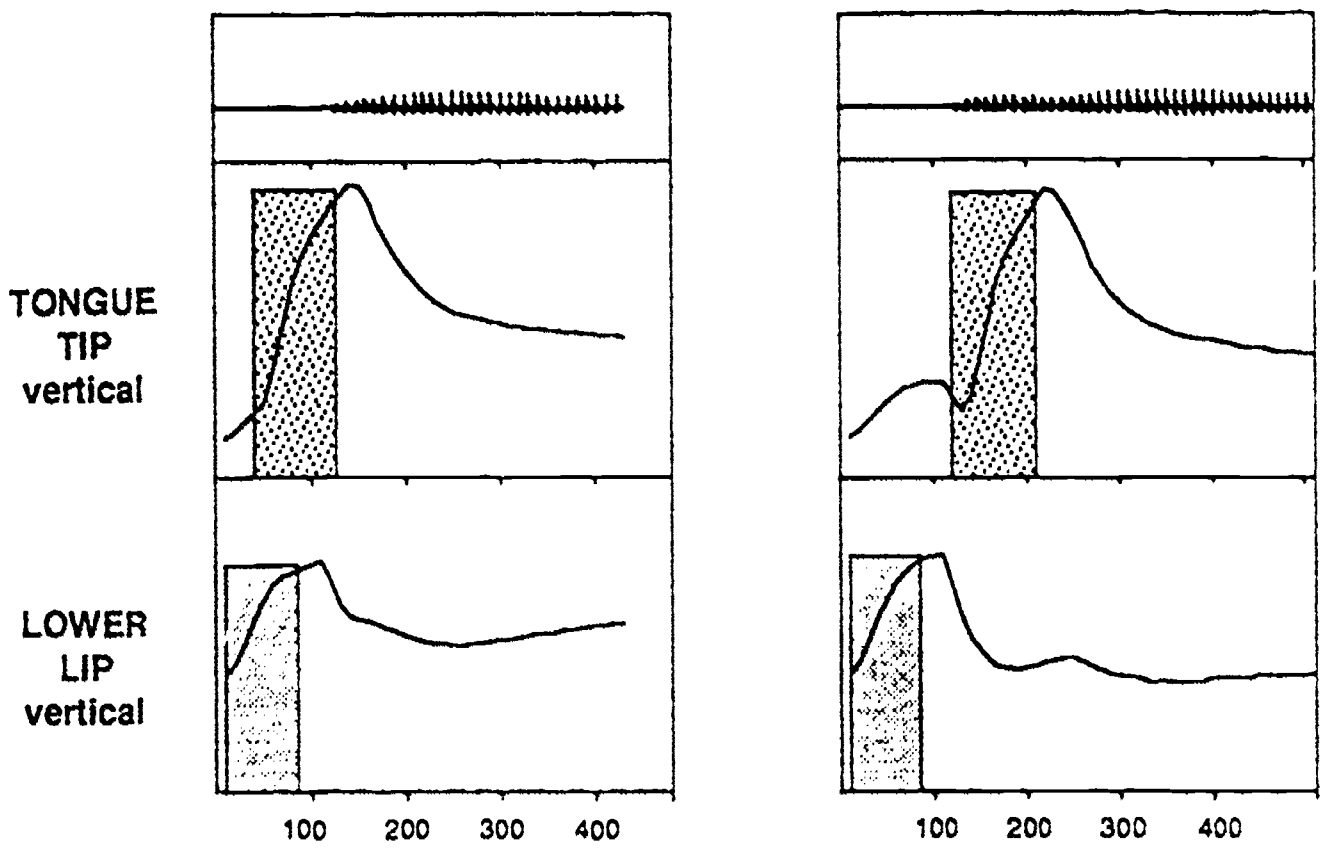


Figure 4. Gestural scores and articulator motions for the initial bilabial closure and tongue tip rhotic gestures in "beret" for the ends of overlap continuum. To facilitate comparison with X-ray data, vertical motions of the articulators are displayed, rather than the generated tract variable motions. The higher the curve, the higher the articulator in space. Boxes indicate gestural activation; (a) maximum overlap (40 ms) (b) maximum separation (40 ms).

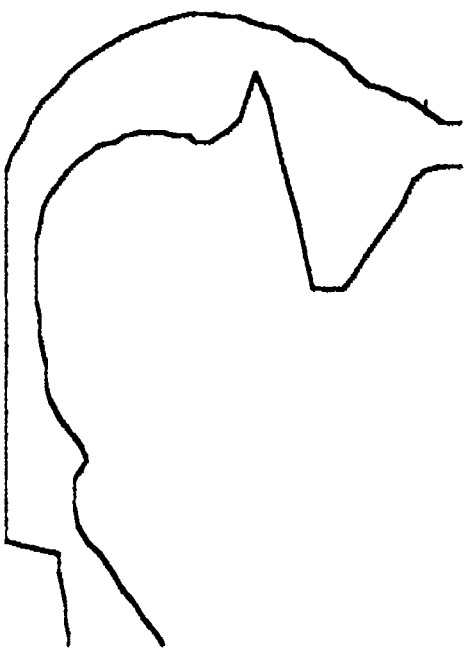


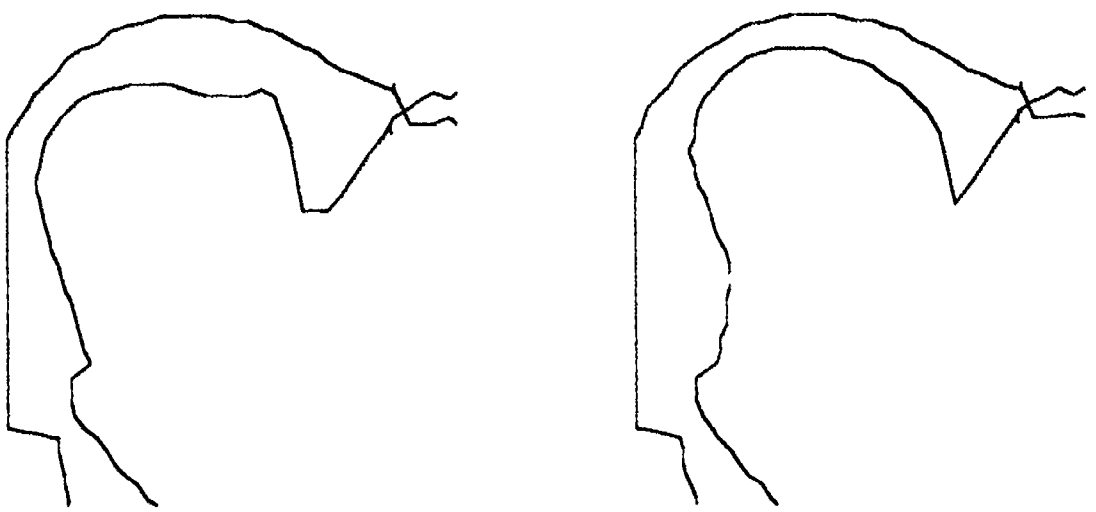
Figure 5. Midsagittal vocal tract model shape for [ɹ].

At one endpoint of the overlap continuum (Figure 4a) there is 40 ms overlap of the control regimes for the rhotic and bilabial closure gestures, while at the other endpoint (Figure 4b) the gestures' control regimes are separated by 40 ms. As a way of visualizing the effect of differential overlap, Figure 6 shows the midsagittal outline of the vocal tract model when active control for the bilabial closure gesture turns off, for the two endpoint stimuli. For the maximal overlap configuration (Figure 4a), Figure 6a shows that the tongue shape has already begun to look like that for the rhotic at the point at which active control for the bilabial closure gesture turns off. However, for the maximal separation configuration (Figure 4b), Figure 6b shows there is no r-shape visible when the bilabial gesture turns off.

A total of nine gestural scores were created, with the phasing for the intermediate stimuli chosen so that, going from extreme overlap to extreme separation, the overlap decreased by exactly one synthesis frame (10 ms) for each step. The nine gestural scores were input to the task dynamic and vocal tract models to generate synthetic speech stimuli. The maximally separated endpoint stimulus was played informally to listeners to satisfy ourselves that the

various gestures could be accurately perceived in a completely open-response format. Then a stimulus tape was created by randomizing ten repetitions of each of the nine stimuli; six listeners were asked to identify each token in this set as "bray" or "beret."

Figure 7 shows the percentage of "bray" or "beret" responses for the stimuli, totalled over all six subjects. As a group, the listeners switched from 67% "bray" responses at 10 ms overlap to 83% "beret" responses at the next step (which had zero overlap). The first "beret" point was at 0 ms overlap for four of the six subjects, and one 10 ms step earlier for the other two subjects. Five of the six subjects switched responses in a single 10 ms step: the average "bray" response from the frame just before the individual's crossover was 80%, while the average "beret" response one frame later was 80%. Thus, there was an abrupt switch from a percept of one syllable to a percept of two syllables caused solely by changing the amount of overlap. A first conclusion, then, is that changing overlap alone is effective in distinguishing pairs like "bray" and "beret."



(a) (b)

Figure 6. Midsagittal vocal tract model shapes when the bilabial closure gesture is turned off (a) for gestural score with maximal overlap (b) for gestural score with maximal separation.

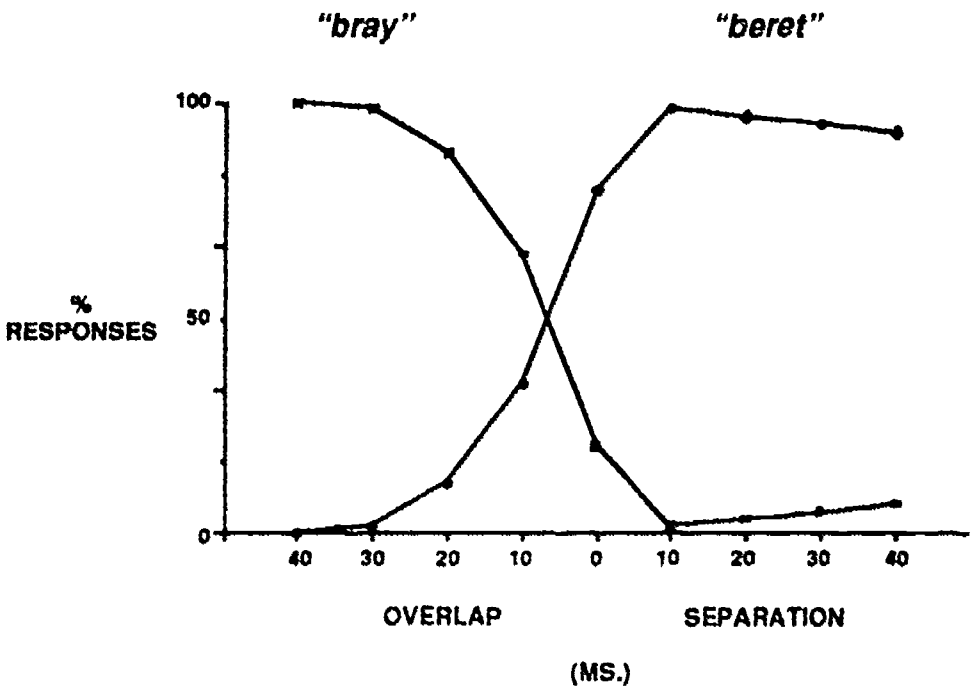


Figure 7. Percentage of "bray" vs. "beret" responses, totalled across all subjects, as overlap between bilabial closure and rhotic gestures decreases. 100% = 60 responses.

A more surprising aspect of the results involved the location of the perceived boundary. For the first stimulus perceived as "beret," there was 0 ms overlap between the bilabial closure control regime and the rhotic control regimes. Thus, "beret" responses were made to stimuli with no overlap between control regimes and "bray" responses were made to stimuli that did show overlap of control regimes. It may be, of course, that this was serendipitous. It is also true that the cross-over stimulus was the middle one in the continuum, as is to be expected in such experiments. Nonetheless, coincidence of the theoretical notion of overlap with the particulars of the results is encouraging, and lends plausibility to the hypothesis that the underlying difference in gestural structures for pairs like "bray" and "beret" could reside in the phasing of the consonant gestures. As noted earlier, this hypothesis would automatically account for the perceived loss of syllabicity in casual speech: with increased overlap, the gestural structure for "beret" would be the same as that for "bray."

While these results with model-generated speech are encouraging, tokens of natural speech must be examined to see whether their articulation is consistent with this proposed gestural difference. We have recently begun to collect X-ray microbeam data relevant to this issue at the NIH facility at the University of Wisconsin. Preliminary analyses appear consistent with the hypothesis of differing gestural overlap. As can be seen in Table 1, the pairs "braid"/"bereted" and "prayed"/"parade" differ in the length of the interval between the articulatory bilabial release and the achievement of the rhotic target. (The bilabial release was defined to be the point at which the lower lip lowering out of the bilabial closure first increased to a velocity of 0.1 mm/sec. The achievement of target for the rhotic was defined as the point at which tongue tip raising into the rhotic first decreased to a velocity of 0.1 mm/sec.)

Table 1. *Interval between articulatory bilabial release and rhotic target.*

		ms				ms	
		n	mean	s.d	n	mean	s.d.
accented	braid	4	13	12	bereted 6	116	13
	prayed	5	-1	21	parade 6	158	17
unaccented	braid	5	-1	11	bereted 3	73	25
	prayed	5	-3	17	parade 5	90	20

The difference in the bilabial-rhotic interval for the mono- and bi-syllables analyzed is consistent with a difference in overlap. However, before claiming that the sole difference between such pairs resides in the overlap, it would be necessary both to determine the gestural onsets and to show that there is no extra tongue body movement (for schwa) in the bisyllables. Although the analysis has not proceeded to the point that definite conclusions can be made, the representative tokens in Figure 8 suggest that the difference between the mono- and bi- syllables might indeed be ascribable solely to a difference in overlap.

Figure 8 shows the movement of pellets placed on the tongue dorsum, lower lip, and tongue tip for "braid" (solid lines), overlaid with data for "bereted" (dotted lines). Both words were produced in the phrase "I say_today," with "braid" (or "bereted") accented. Note that the lower lip is relatively high during the rhotic, indicating that a rounding gesture accompanies the tongue tip gesture, as is regularly seen for American English /r/ (Delattre & Freeman, 1968). Two separated labial raising movements can be observed for "bereted." The two utterances clearly differ in the relative timing of the rhotic tongue tip raising motion and the initial bilabial gesture (the points used in the measurements for Table 1 are marked with arrows). Moreover, the tongue dorsum movements in the two utterances are quite similar. If "bereted" were to contain a separate vowel gesture (a schwa) that is absent in "braid," we would expect to see a difference between the two utterances in the behavior of this pellet (see Browman & Goldstein, in press).

However, another articulatory study does not support the strong form of the claim that perceived instances of reduced syllables are solely the result of non-overlap of successive consonant gestures. Browman and Goldstein (in press) examined the nature of medial schwa vowels in utterances of the form /'pV1pəpV2pə/ to test the hypothesis that there would be no explicit vocalic gesture associated with the schwa, and therefore that the tongue would move in a continuous fashion from the position for V1 to that for V2, passing through some intermediate position during the acoustic interval for the schwa. This was tested both by statistical analyses of X-ray microbeam data and by simulations using various hypothesized gestural scores. The analyses showed that the strong form of the hypothesis could not be maintained for these utterances, at least in the environment where V1=V2=/i/ or V1=V2=/u/. In these cases, the position of the tongue during the schwa was lowered compared to the high vowels on either side, whereas the hypothesis predicted that the tongue should remain high.

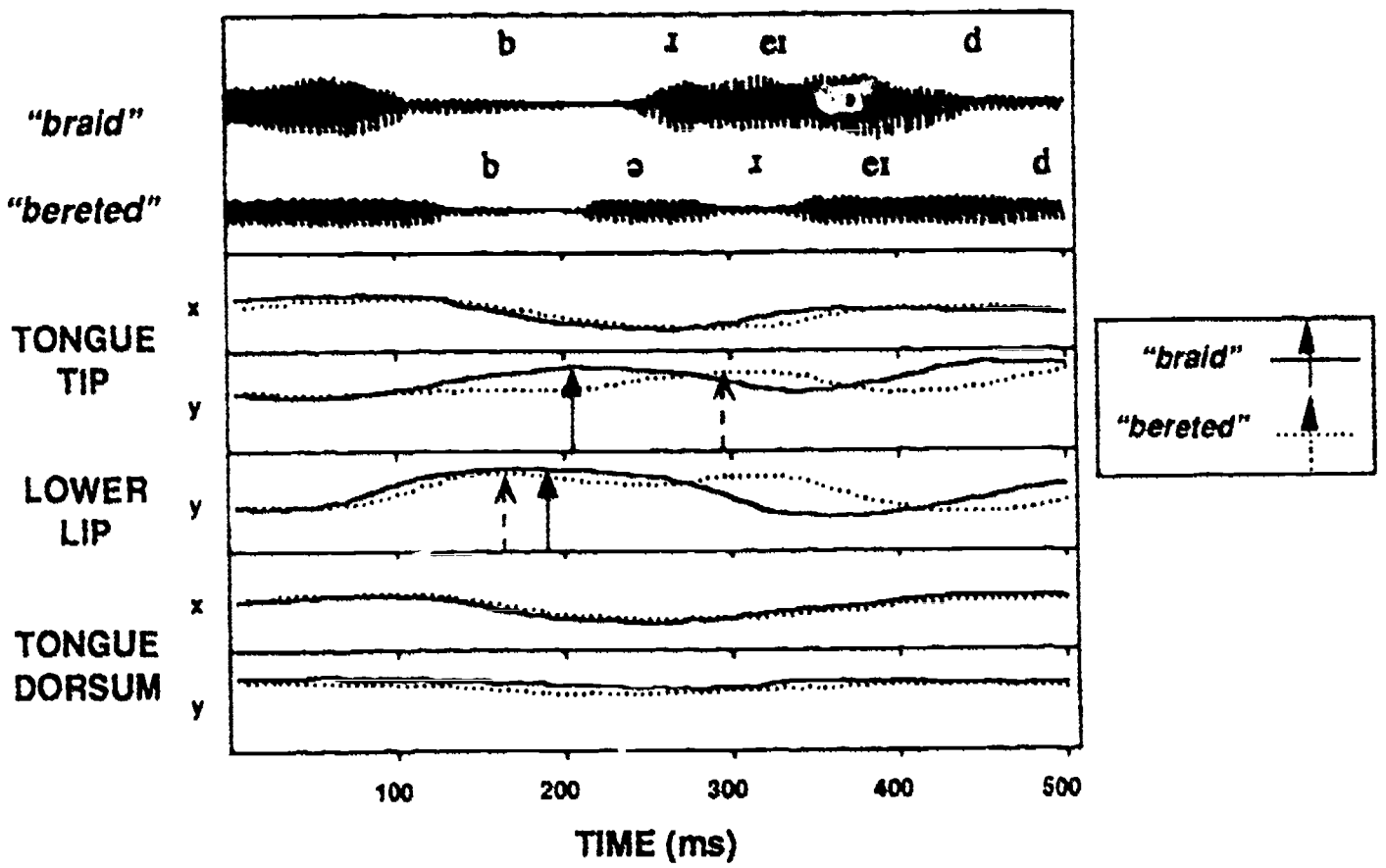


Figure 8. X-ray data for "braid" and "bereted," showing horizontal and vertical movement of pellets on tongue tip and tongue dorsum and vertical movement of pellet on lower lip. The vertical extent of each panel is 30 millimeters. The arrows mark the achievement of target for the rhotic (tongue tip) and the onset of the opening movement for the bilabial (lower lip). The two sets of data are lined us such that the achievement of target for the bilabial closures coincide.

While the strong form of the gestural overlap hypothesis was not supported by this study, viewing schwa in terms of gestural overlap led to an interesting and viable form of conceptualizing the data. The target for the X-ray tongue pellets for the schwa turned out to be completely "colorless:" it was the mean of the targets for all the full vowels. Overlapping V2 (but not V1) during the entire time domain of this colorless schwa proved to correctly capture patterns of systematic articulatory variation during the schwa. The identity of the schwa, then, was very weak, both in its colorless nature and in its being completely overlapped by a full vowel. Moreover, it was possible to obtain a percept of schwa in a simulation of /pipəpipə/ in which there was no active schwa gesture. This gestural score differed from the gestural score that encoded the data analyses only in having the schwa gesture removed and the consonants on either side phased closer together so that the acoustic schwa duration was shorter. For acoustics generated from this gestural score, a schwa percept was obtained in an informal listening test, even though V1 and V2 were /i/.

Although gestural overlap played an important role in both these studies, and although the schwa was weak in the second study, the results of the two studies with regard to the hypothesis of a totally unspecified reduced syllable nucleus were conflicting. The conflict might be resolved in several different ways. It might be, for example, that further investigations would discover that all reduced syllables contain a gesture for the nucleus. In such a case, an increase in overlap between the surrounding consonants could still be the source of the different variants of "beret." Another possibility is that reduced syllables in different phonetic and/or morphological environments (e.g., stop-stop vs. stop-liquid, or "Rosa's" vs. "roses") might be associated with different gestural structures. On the basis of the above results, we would expect that reduced syllables in stop-stop environments and also in words such as "Rosa('s)" might contain a vocalic gesture for schwa, while those in stop-liquid environments and words such as "roses" might have no separate gesture. This further suggests that investigations conducted within the gestural framework might lead to interesting typologies of

reduced syllables, and more generally of processes involving changes in syllabicity.

One such possible typology involves the development of epenthetic vowels in languages. Matson (in preparation) has hypothesized that such vowels develop from the perception of the interval between two consonant gestures, as the consonants spread apart (in time) and overlap less. The fact that these intervals may later be identified with one of the full vowel gestures of the language could be considered a subsequent, 'listener-based,' sound change, along the lines of Ohala's (1981) model. Matson proposes that different types of epenthesis occur depending on whether the separating consonants are in the same syllable or not. On the one hand, if the consonants are either heterosyllabic or "unsyllabifiable," she finds that languages insert a constant vowel, and that such vowels are universally non-low, as would be expected of an interval resulting from consonant gesture separation, in which the tongue body would frequently be high due to the surrounding consonant constrictions. On the other hand, if the separating consonants are tautosyllabic, she finds, following Steriade (in press), examples in which the epenthetic vowel is identical with the (original) syllable nucleus. This is predicted by a gestural analysis, because the separation (sliding) in this case uncovers the overlapping vowel gesture that underlies the entire original syllable. Steriade (in press) further shows that slightly different amounts of sliding in these cases can yield the metatheses that are found, in place of epenthesis, in related languages. While the details of such a typology are still sketchy, it further demonstrates the kinds of questions one can ask within a framework in which gestural overlap is directly characterized and input structures are clearly distinguished from output consequences.

REFERENCES

Abraham, R. H., & Shaw, C. D. (1982). *Dynamics—The geometry of behavior*. Santa Cruz, CA: Aerial Press.

Beckman, M. E., Edwards, J., & Fletcher, J. (in press). Prosodic structure and tempo in a sonority model of articulatory dynamics. Proceedings of the Second Conference of Laboratory Phonology, Edinburgh, Scotland, 28 June 1989 - 3 July 1989.

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

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

Browman, C. P., & Goldstein, L. (1986b). Dynamic processes in linguistics: Casual speech and historical change. *PAW Review*, 1, 17-18.

Browman, C. P., & Goldstein, L. (1987). Tiers in Articulatory Phonology, with some implications for casual speech. *Haskins Laboratories Status Report on Speech Research*, *SR-92*, 1-30. To appear 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. (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. (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. (in press). "Targetless" schwa: An articulatory analysis. Proceedings of the Second Conference of Laboratory Phonology, Edinburgh, Scotland, 28 June 1989 - 3 July 1989.

Cooke, J. D. (1980). The organization of simple, skilled movements. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 199-212). Amsterdam: North-Holland.

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

Delattre, P., & Freeman, D. C. (1968). A dialect study of American r's by X-ray motion picture. *Linguistics*, *44*, 29-68.

Fant, G. (1960). *Acoustic theory of speech production*. Mouton: 's Gravenhage.

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

Fowler, C. A. (1981). A relationship between coarticulation and compensatory shortening. *Phonetica* 38, 35-50.

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* (pp. 373-420). New York: Academic Press.

Goldstein, L. (1989). On the domain of the quantal theory. *Journal of Phonetics*, 17, 91-97.

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

Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, *51*, 347-356.

Hawkins, S. (in press). An introduction to task dynamics. Proceedings of the Second Conference of Laboratory Phonology, Edinburgh, Scotland, 28 June 1989 - 3 July 1989.

Kay, B. A., Kelso, J. A. S., Saltzman, E. L., & Schöner, G. (1987). Space-time behavior of single and bimanual rhythmical movement: Data and limit cycle model. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 178-192.

Kelso, J. A. S., & Tuller, B. (1984a). A dynamical basis for action systems. In M. Gazzaniga (Ed.), *Handbook of cognitive neuroscience* (pp. 321-356). New York: Plenum.

Kelso, J. A. S., & Tuller, B. (1984b). Converging evidence in support of common dynamical principles for speech and movement coordination. *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology*, *246*, R928-R935.

Kelso, J. A. S., & Tuller, B. (1987). Intrinsic time in speech production: Theory, methodology, and preliminary observations. In E. Keller & M. Gopnik (Eds.), *Motor and sensory processes of language* (pp. 203-222). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kelso, J. A. S., Holt, K. G., Rubin, P., & Kugler, P. N. (1981). Patterns of human interlimb coordination emerge from the properties of nonlinear limit cycle oscillatory processes: Theory and data. *Journal of Motor Behavior*, *13*, 226-221.

Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*, *14*, 29-59.

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., & 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.

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

Kugler, P. N., & Turvey, M. T. (1987). *Information, natural law, and the self-assembly of rhythmic movement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Liberman, A., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431-436.

Lindau, M. (1978). Vowel features. *Language, 54*, 541-563.

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*. The Hague: Mouton.

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

Matson, D. (in preparation). *Epenthesis in articulatory phonology*.

McGarr, N. S, Löfqvist, A., & Story, R. (submitted). Jaw kinematics in hearing-impaired speakers. *Journal of the Acoustical Society of America*.

McGowan, R. S., Smith, C. L., Browman, C. P., & Kay, B. A. (1988). Extracting dynamic parameters from articulatory movement. *Journal of the Acoustical Society of America, 83*, S113. Paper presented at the 115th meeting of ASA, Seattle.

McGowan, R. S., Smith, C. L., Browman, C. P., & Kay, B. A. (1990). Methods for least-squares parameter identification for articulatory movement and the program PARFIT. *Haskins Laboratories Status Report on Speech Research, 101/102*, 220-230.

Nittrouer, S., Munhall, K., Kelso, J. A. S., Tuller, B., & Harris, K. S. (1988). Patterns of interarticulator phasing and their relation to linguistic structure. *Journal of the Acoustical Society of America, 84*, 1653-1661.

Ohala, J. J. (1981). The listener as a source of sound change. In C. S. Masek, R. A. Hendrick, & M. F. Miller (Eds.), *Papers from the parasession on language and behavior* (pp. 178-203). Chicago: Chicago Linguistic Society.

Ohala, J. J. (1985). Around 'flat.' In V. Fromkin (Ed.), *Phonetic linguistics*. New York: Academic.

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

Ostry, D. J., & Munhall, K. (1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America, 77*, 640-648.

Perrier, P., Abry, C., & Keller, E. (1988). Vers une modelisation des mouvements du dos de la langue. *Bulletin du Laboratoire de la Communication Parle'e, 2*, 45-63.

Rubin, P., T. Baer, & P. Mermelstein (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America, 70*, 321-328.

Saltzman, E. (1986). Task dynamic coordination of the speech articulators: A preliminary model. In H. Heuer & C. Fromm (Eds.), *Generation and modulation of action patterns (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.

Smith, C. (1988). A cross-linguistic contrast in consonant and vowel timing. *Journal of the Acoustical Society of America, 86*, S84. Paper presented at the 116th meeting of the ASA, Honolulu.

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. Presented at the LSA Annual Meeting, Washington DC, Dec 1989.

Steriade, D. (in press). Gestures and autosegments: Comments on Browman and Goldstein's "Gestures in Articulatory Phonology." In J. Kingston & M.E. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and the physics of speech*. Cambridge: Cambridge University Press.

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

Stevens, K. N., & House, A. S. (1955). Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America, 27*, 484-493.

Tuller, B., Kelso, J. A. S., & Harris, K. S. (1982). Interarticulator phasing as an index of temporal regularity in speech. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 460-472.

Turvey, M. T. (1977). Preliminaries to a theory of action with reference to vision. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing* (pp. 211-265). Hillsdale, NJ: Lawrence Erlbaum Associates.

Turvey, M. T., Rosenblum, L. D., Kugler, P. N., & Schmidt, K. G. (1986). Fluctuations and phase symmetry in coordinated rhythmic movements. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 564-583.

Vatikiotis-Bateson, E. (1988). *Linguistic structure and articulatory dynamics: A cross-language study*. Bloomington: Indiana University Linguistics Club.

FOOTNOTES

*Appears in *Journal of Phonetics, 18*, 299-320 (1990).

†Also Yale University Department of Linguistics.

[1]Browman and Goldstein (1986a)

[2]Browman and Goldstein (1986b, 1987, 1991)

[3]Goldstein and Browman (1986), Browman & Goldstein (1986a, 1991)

[4]Browman and Goldstein (1989)

[5]In addition to specifying the particular tract variables associated with the gesture, the relative contributions (called weights) of the associated articulators (see Figure 2) must also be specified. These relative contributions hold only under "everything else being equal" conditions. The actual articulatory contributions will depend on the ensemble of concurrently active gestures. See Saltzman and Munhall (1989) for discussion.

[6]Lindau (1978) argues that it is the tongue body gesture that constitutes an articulatory invariant for American English /r/. The tongue tip gesture may be replaced in some speakers and in some environments by a 'bunched' tongue body gesture—a constriction formed by the tongue body at the margin between the hard and soft palates (Delattre & Freeman, 1968). The fact that these two gestures regularly cooccur may be an example of how gestural structures may be tuned by their acoustic effects. As Delattre and Freeman (1968) point out, the acoustic effect of each of the constrictions is to lower F3 and bring it closer to F2. An account of this in terms of the standing wave pattern for F3 is presented in Ohala (1985).

Haskins Laboratories Status Report on Speech Research
1990, SR-103/104, 111-124

Stimulus Order Effects in Vowel Discrimination*

Bruno H. Repp and Robert G. Crowder†

In same-different discrimination tasks employing isolated vowel sounds, subjects often give significantly more "different" responses to one order of two stimuli than to the other order. Cowan and Morse (1986) proposed a *neutralization hypothesis* to account for such effects: The first vowel in a pair is assumed to change its quality in memory in the direction of the neutral vowel, schwa. We conducted three experiments using a variety of vowels and obtained some initial support for the hypothesis, using a large stimulus set, but conflicting evidence with smaller stimulus sets. Rather than becoming more similar to schwa, the first vowel in a pair seems to drift towards the interior of the stimulus range employed in a given test. We discuss several possible explanations for this tendency and note its relation to presentation order effects obtained in other psychophysical paradigms.

INTRODUCTION

The perception of vowels has long been of central interest to speech researchers (see Nearey, 1989, for a recent review). Isolated vowels, being the simplest instantiation of speech, provide a common testing ground for theories of auditory psychophysics and of speech perception. For the former, they offer the challenge of complexity, for the latter the advantage of simplicity. Even though these sounds are far removed from the connected speech that speech perception theories ultimately need to be concerned with, they are not entirely unnatural: Some isolated vowels occur as exclamations, fillers, or even as real words; all are readily elicited as pronunciation "prototypes" from native speakers; and their presentation in a perceptual test usually engages the listeners' mechanisms of phonetic categorization unless stimulus uncertainty is minimized (see Macmillan, Goldberg, & Braida, 1988).

To study the identification and discrimination of isolated vowels, many researchers have used stimuli drawn from an acoustic continuum spanning two or three phonetic categories. Although the perception of isolated vowels is n strongly categorical (i.e., discrimination perfo - mance within phonetic categories is well above chance), there is usually a contribution of phonetic categorization to discrimination performance (i.e., discrimination is most accurate in the category boundary regions). This was demonstrated, for example, by Pisoni (1973, 1975) in several discrimination paradigms, including a "same-different" task. This simple task has been employed in a number of later studies concerned with the role of auditory memory in vowel discrimination.

In one of these studies, we (Repp, Healy, & Crowder, 1979) presented subjects with pairs of stimuli from a 13-member synthetic /i/-/ɪ/-/ɛ/ continuum, obtained by stepwise linear interpolation between the formant frequencies of /ɪ/ and /ɛ/. Our most important finding was that a substantial part of the discrimination performance could be accounted for by contrast effects between the members of stimulus pairs, as revealed in a labeling task, though it remained unclear whether these contrast effects were the cause or the consequence of heightened discriminability. (See also Healy & Repp, 1982.) We also observed, in agreement with earlier results of Shigeno and Fujisaki (1980), that retroactive contrast (the effect of the second vowel in a pair on the labeling

This research was supported by NICHD Grant HD01994 to Haskins Laboratories and by NSF Grant GB86 08344 to Robert G. Crowder. We are grateful to William Flack for assistance in running experiments and tabulating data, and to Nelson Cowan for helpful comments.

of the first) was larger than proactive contrast (the converse) when both vowels in a pair had to be classified, presumably because the first vowel in a pair had to be held longer in memory and thus was less stable when the second vowel arrived. In addition to these contrast effects, however, the subjects' responses revealed an unexpected effect of stimulus order: For pairs of nonidentical stimuli from the /i/-/ɪ/ region of the continuum, a higher percentage of correct "different" responses was obtained when the more /i/-like stimulus came second in a pair than when it came first.[1] At the /ɛ/ end of the continuum, however, this order effect was absent or even reversed.

Earlier authors employing the same-different paradigm had not paid any attention to such order effects and had simply combined the responses for the two orders of each pair. The order effects attracted our attention because they were quite large in some stimulus pairs and, apparently, different in nature from the contrast effects, whose occurrence among vowel stimuli has been known for a long time (e.g., Eimas 1963; Thompson & Hollien, 1970). Whereas contrast effects extended across the whole vowel continuum, order effects were most pronounced at the /i/-end. More importantly, contrast effects virtually disappeared when the interstimulus interval was lengthened and filled with an intervening irrelevant vowel, but stimulus order effects survived such interference. Thus they seemed to be caused by a different mechanism. At the time, we did not pursue the explanation of this secondary finding any further.

An article by Cowan and Morse (1986) drew renewed attention to these order effects and suggested that they reveal a hitherto unnoticed property of memory for vowels. These authors used stimuli ranging from /i/ to /ɪ/, corresponding to one half of our earlier continuum. Again, discrimination accuracy was higher when the more /i/-like stimulus came second in a pair. However, the effect also interacted with interstimulus interval: It grew *larger* as the (empty) interval was increased from 250 to 2000 ms, due to a more rapid decrease in discriminability for those stimulus pairs in which the more /i/-like stimulus came first. On the basis of these results, Cowan and Morse proposed that the perceived quality of the first vowel in a pair changes gradually while it is held in memory.[2] Specifically, they suggested that it changes towards a more neutral quality—that its internal representation drifts toward the center of the acoustic-phonetic vowel space (henceforth, the *neutralization hypothesis*). They further speculated that this drift may be strongest for vowels such as /i/, which are near the periphery of the vowel space.

Thus, according to this hypothesis, an /i/-like vowel held in memory becomes more like /ə/ and hence more similar to /ɪ/ (/ɪ/ being more central than /i/ in the vowel space; see Figure 1 below), whereas an /ɪ/-like vowel held in memory becomes even more central and hence more dissimilar from /i/. Therefore, an /i/-like vowel is difficult to discriminate from a following, more /ɪ/-like vowel, while the reverse order is easy to discriminate. The neutralization hypothesis also predicts a reduction of the order effect at the /ɛ/-end of an /i/-/ɪ/-/ɛ/ continuum, though not a reversal (as mistakenly claimed by Cowan and Morse), since /ɛ/ is somewhat more central than /ɪ/ (Figure 1).

The neutralization hypothesis is interesting because it suggests a speech-specific memory mechanism. However, at this point its supporting evidence rests entirely on high front vowels. Cowan and Morse stressed the need for studies of order effects in other regions of the vowel space. The purpose of the present experiments was to fill this gap, and thereby to assess the validity of the neutralization hypothesis.

I. EXPERIMENT 1

In this experiment we employed nine groups of stimuli from all over the vowel space, arranged in a way that enabled us to make clear predictions about the direction and magnitude of order effects.

A. Methods

1. Stimuli

Figure 1 represents the stimuli schematically as points in the two-dimensional acoustic space defined by the frequencies of the lowest two formants (F1 and F2). Eight monophthongal vowels, /ɪ,ɛ,æ,ɑ,ɔ,ʊ,u/, were selected from the Peterson and Barney (1952) norms for adult male speakers of American English.[3] In addition, the neutral vowel /ə/, which was not included in the study of Peterson and Barney and is less well defined phonetically, was assumed to have F1 and F2 frequencies of 500 and 1500 Hz, respectively (i.e., the resonance frequencies of a uniform tube having the length of the average male vocal tract; see Chiba & Kajiyama, 1941). These nine *prototype* vowels are located at the centers of the crosses in Figure 1.

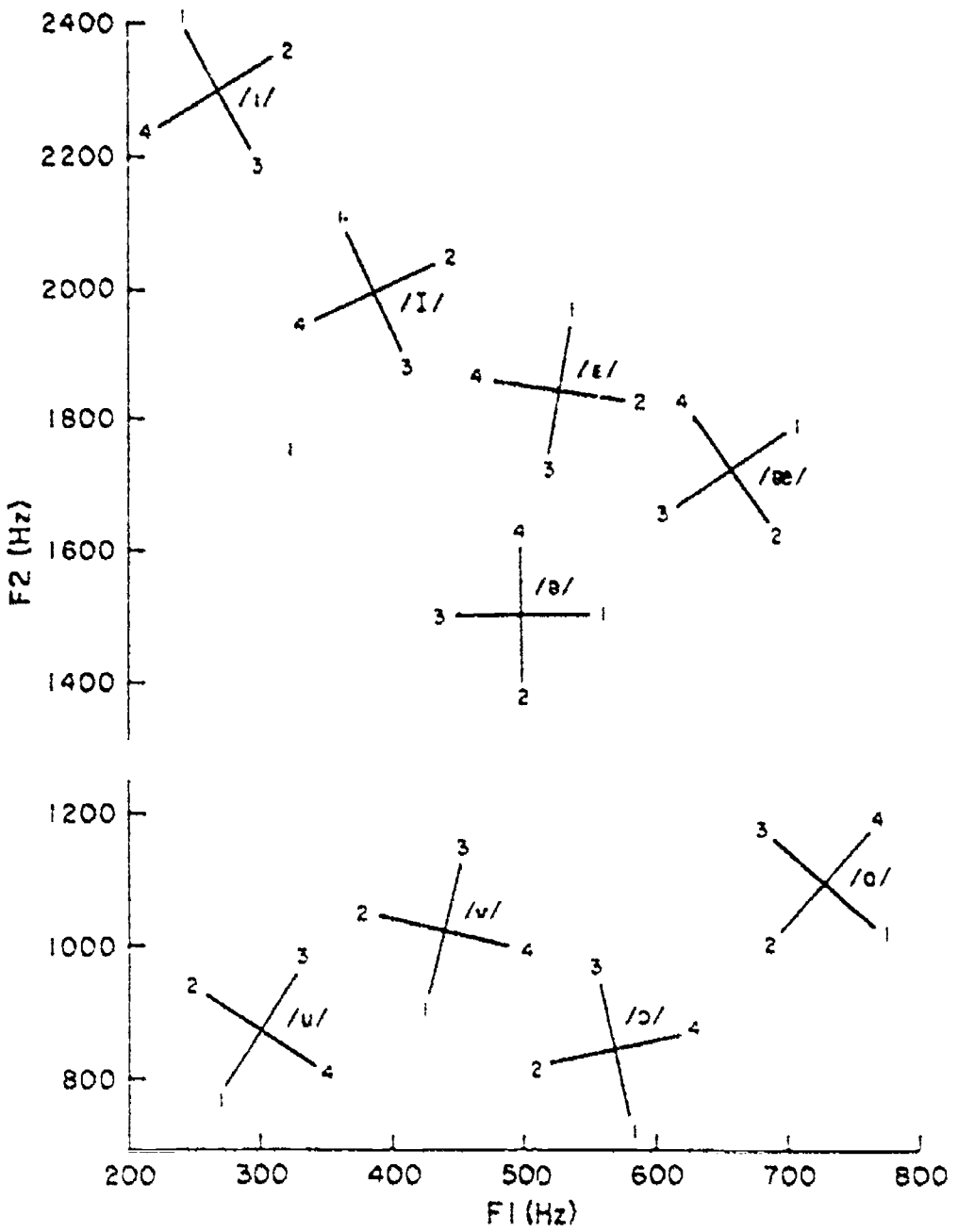


Figure 1. Positions of the stimuli of Experiment 1 in F1-F2 space.

For each of the prototype vowels, four *neighbors* in vowel space were chosen as indicated by the endpoints of the cross arms in Figure 1. Each cross was oriented so that one of its arms pointed directly to the neutral /ə/ vowel. The four neighbors of each prototype were numbered in a clockwise fashion starting with the vowel farthest from /ə/ (cf. Figure 1). The neighbors of the /ə/ prototype were arbitrarily determined by a cross whose arms were parallel to the F1-F2 coordinates, and they were numbered arbitrarily. The fixed arm length of all crosses was chosen on the basis of pilot observations, so as to make the neighbors fairly difficult to discriminate from the prototypes in a high-uncertainty task.

The formant frequencies of all the vowels are listed in Table 1. The third formant of all stimuli was fixed at 2440 Hz, except for /i/ and its neighbors, which had an F3 of 3010 Hz. The stimuli were synthesized on the Haskins Laboratories serial resonance software synthesizer with a duration of 250 ms and a linearly falling fundamental frequency contour (100-80 Hz). An experimental tape was recorded containing four blocks (replications) of 117 randomly ordered stimulus pairs each. The 117 pairs resulted from each of the nine prototypes being paired with each of its four neighbors in both temporal orders (72 pairs), and each vowel being paired with itself (45 pairs). The ratio of "different" to "same" pairs thus was 8:5. The interstimulus intervals were 500 ms within pairs, 2 s between pairs, and 5 s after each group of 13 pairs.

Table 1. *Formant frequencies (in Hz) of the stimuli used in Experiment 1. (P = prototype, N = neighbor)*

Stimulus		F1	F2
/i/	P	270	2290
	N1	245	2380
	N2	315	2340
	N3	295	2200
	N4	225	2240
/ɪ/	P	390	1990
	N1	370	2080
	N2	435	2030
	N3	410	1900
	N4	345	1950
/ɛ/	P	530	1840
	N1	538	1940
	N2	580	1825
	N3	522	1740
	N4	480	1855
/æ/	P	660	1720
	N1	700	1780
	N2	690	1640
	N3	620	1660
	N4	630	1800
/ɑ/	P	730	1090
	N1	767	1020
	N2	695	1010
	N3	693	1160
	N4	765	1170
/ɔ/	P	570	840
	N1	580	740
	N2	520	820
	N3	560	940
	N4	620	860
/ʊ/	P	440	1020
	N1	425	920
	N2	390	1045
	N3	455	1120
	N4	490	995
/u/	P	300	870
	N1	272	785
	N2	258	925
	N3	328	955
	N4	342	815
/ə/	P	500	1500
	N1	500	1600
	N2	550	1500
	N3	500	1400
	N4	450	1500

2. Subjects and Procedure

Twenty-four undergraduate subjects participated in this study for course credit. Each subject listened to the tape once, and then to the first three blocks again, so that seven blocks were presented in all. The first block was considered practice and was not scored. Presentation was over loudspeakers in a quiet room. The task was to respond "same" or "different" to each pair by circling the appropriate response on an answer sheet.

B. Results and Discussion

The predictions of the neutralization hypothesis were as follows: Pairings of prototype (P) vowels with neighbors N1 and N3, which lie on the axis pointing towards /ə/, should yield large order effects of opposite sign. N1 pairs should show a positive order effect (defined as more correct "different" responses when P comes first than when it comes second), whereas N3 pairs should show a negative order effect, perhaps of smaller absolute size because of their more central location in vowel space. N2 and N4 pairs, on the other hand, should not show any significant order effects. For pairs involving the /ə/ prototype there were no clear predictions, except that order effects should be small. Any large order effects obtained in this region would suggest that the true /ə/ prototype is located elsewhere.

The results are shown in Table 2. It is evident, first, that discrimination accuracy was not very high but obviously above chance: Hit rates ("different" responses to nonidentical pairs) were uniformly higher than false-alarm rates ("different" responses to identical pairs). This performance level was optimal for observing large order effects. Pairs involving /i/ received markedly fewer "different" responses than the rest; otherwise, performance did not vary substantially across the vowel space.

The results of interest, the order effects, are shown at the bottom of the table. These effects were computed by subtracting the percentage of "different" responses for pairs in which P came second from that for pairs in which P came first. It can be seen that a number of stimulus pairs showed large (> ± 10%) order effects, but that pairs involving /ə/ showed only small effects, as predicted. In the following statistical analyses, these latter pairs were excluded because they did not follow the general stimulus design.

Two separate ANOVAs were conducted, one on N1 and N3 pairs, and the other on N2 and N4 pairs. Each had the factors Vowel (8), Neighbor (2), and Order (2). For the first analysis, the neutralization hypothesis predicted a significant Neighbor by Order interaction, due to positive order effects in pairs involving N1 and negative order effects in pairs involving N3. This interaction was indeed highly significant, $F(1,23) = 34.48$, $p < .0001$, although there was also a significant triple interaction involving Vowel, $F(7,161) = 4.19$, $p = .0003$. Two-way follow-up analyses were therefore conducted on N1 and N3 pairs separately.[4]

Table 2. *Average percentages of "different" responses for all stimulus pairs in Experiment 1, and order effects (prototype first minus prototype second).*

	Prototype vowel (P)								
Pair	/i/	/I/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʊ/	/u/	/ə/
				Identical pairs					
P-P	6.9	17.4	16.0	15.3	11.8	13.9	22.2	26.4	4.2
N1-N1	11.1	19.4	13.2	9.7	11.1	13.2	18.8	12.5	16.0
N2-N2	12.5	13.9	15.3	15.3	10.4	18.8	21.5	22.9	6.3
N3-N3	9.0	16.7	18.8	19.4	13.2	18.1	19.4	19.4	6.3
N4-N4	8.3	16.7	16.0	13.1	9.7	16.0	16.7	18.1	10.4
				Nonidentical pairs: prototype first					
P-N1	22.2	39.6	70.1	63.2	64.6	72.2	54.9	59.7	43.8
P-N2	37.5	63.9	52.8	38.2	46.5	56.3	71.5	66.0	35.4
P-N3	15.3	34.0	54.9	61.8	50.7	56.9	48.6	35.4	35.4
P-N4	43.1	74.3	49.3	49.3	54.9	60.4	64.6	40.3	34.0
				Nonidentical pairs: prototype second					
N1-P	10.4	27.8	46.5	35.4	43.8	53.5	61.8	27.1	36.8
N2-P	41.0	70.1	49.3	39.6	54.9	48.6	46.5	32.6	36.8
N3-P	20.1	53.5	69.4	65.3	54.9	54.9	49.3	56.3	28.5
N4-P	23.6	68.1	54.9	52.8	38.2	55.6	66.7	62.5	31.9
				Nonidentical pairs: order effects (difference scores)					
N1	11.8	11.8	23.6	27.8	20.8	18.7	-7.1	32.6	7.0
N2	-3.5	-6.2	3.5	-1.4	-8.4	7.7	25.0	33.4	-1.4
N3	-4.8	-19.5	-14.5	-3.5	-4.2	2.0	-0.7	-20.9	6.9
N4	19.5	6.2	-5.6	-3.5	16.7	4.8	-2.1	-22.2	2.1

For N1 pairs, there was a highly significant positive Order effect, $F(1,23) = 54.78$, $p < .0001$, but also a significant Vowel by Order interaction, $F(7,161) = 4.75$, $p = .0001$. As can be seen in Table 2, seven of the eight vowels showed large positive order effects, though their magnitude varied considerably; one vowel (/u/), however, showed a small negative effect. For N3 pairs, there was the predicted negative Order effect, $F(1,23) = 8.62$, $p = .0074$, as well as a weak Vowel by Order interaction, $F(7,161) = 2.73$, $p = .0106$. Actually, only three vowels showed large negative effects; all other effects were of negligible size. Although the neutralization hypothesis predicted smaller absolute order effects in N3 than in N1 pairs (which held for seven of the eight vowels), this large variability was unexpected.

For the analysis of the N2 and N4 pairs, the neutralization hypothesis predicted an absence of order effects. There was, however, a significant (positive) main effect of Order, $F(1,23) = 8.46$, $p = .0079$, and although the Neighbor by Order interaction was not significant, there was a highly significant triple interaction, $F(7,161) = 7.28$, $p < .0001$. A separate follow-up analysis of N2 pairs again revealed a main effect of Order, $F(1,23) = 8.60$, $p = .0075$, and a strong Vowel by Order interaction, $F(7,161) = 5.38$, $p < .0001$. As can be seen in Table 2, two vowels (/u/ and /ʊ/) unexpectedly showed large positive order effects; all other

vowels showed small effects, as predicted. A separate analysis of N4 pairs did not yield a significant main effect of Order but again a significant Vowel by Order interaction, $F(7,161) = 4.53$, $p = .0001$. Table 2 shows that two vowels (/i/, /ɑ/) yielded sizeable positive order effects, and one (/u/) a negative effect, with the rest being negligible.

On the whole, these results confirm the main predictions of the neutralization hypothesis: Positive order effects for N1 pairs, negative effects for N3 pairs, and mostly negligible effects for N2 and N4 pairs. There are a number of local deviations from the predictions, however, which are too large to be ignored. Some of these deviant results could be explained by the ad hoc assumption that back vowels changed in memory not towards /ə/ but towards a quality close to the N3 of /ɔ/ (see Figure 1). This would predict positive order effects for pairings of /u/ with its N1 and N2, of /ʊ/ with its N2, of /ɔ/ with its N1, and of /ɑ/ with its N1 and N4, all of which were in fact obtained; also, negative order effects for pairings of /u/ with its N3 and N4, of /ʊ/ with its N4, of /ɔ/ with its N3, and of /ɑ/ with its N2 and N3, which were clearly realized only in the case of /u/ but were not strongly contradicted elsewhere; and no order effects for pairings of /ʊ/ with its N1 and N3, and for /ɔ/ with its N2 and N4, which was confirmed (cf. Table 2). Front vowels and /ə/, on the other hand, must still be assumed to decay towards a quality near /ə/; otherwise, order effects would have to be predicted for /ə/ paired with its N2 and N4, and for /æ/ paired with its N4, none of which were obtained. Thus, if different neutral points are assumed for front and back vowels, only one large order effect remains unaccounted for (/i/ paired with its N4).

Unfortunately, we have no independent justification for assuming different neutral points for front and back vowels. It was also surprising that pairs including /i/ and its N3, which are similar to stimulus pairs that had yielded large negative order effects in the studies of Repp et al. (1979) and Cowan and Morse (1986), showed only a negligible order effect here. This suggested to us the possibility that the pattern of order effects is not fixed but depends on the stimulus ensemble used in an experiment. Experiment 2 was conducted to address this question.

II. EXPERIMENT 2

In this study we reused four of the vowel sets of Experiment 1, those grouped around the /ɪ/, /ɛ/, /æ/, and /ə/ prototypes. These are precisely the stimulus sets that yielded results supporting the neutralization hypothesis. The critical set was /ɛ/, which was adjacent to each of the other three in vowel space (see Figure 1). The stimulus pairs from that set were presented in three separate tests, each time intermixed with the stimulus pairs from one of the other three sets. If order effects were sensitive to stimulus context, they should follow significantly different patterns for the same /ɛ/ stimuli in the three different tests. The pattern of order effects for the context stimuli should also be changed in comparison to Experiment 1. Specifically, we suspected that the hypothetical neutral point might be in different locations in different contexts, perhaps closer to the centroid of the stimulus ensemble used in a particular test.

As an additional manipulation, we included two different ISIs in our design. Cowan and Morse (1986) found that order effects in the /i/-/ɪ/ region increased with ISI, due to a more rapid decline in discrimination performance for pairs in which the more /i/-like stimulus came first. They pointed out that this provides important support for the neutralization hypothesis, whose main assumption is that the memory representations of vowels change over time. In the present study we intended to replicate their finding by using two ISIs (200 ms and 1 s) that straddled the ISI of 500 ms used in Experiment 1.

A. Methods

1. Stimuli

The stimuli were the /ɪ/, /ɛ/, /æ/, and /ə/ sets of Experiment 1. Three separate test tapes were recorded, the first containing /ɛ/ and /ɪ/ stimuli, the second /ɛ/ and /æ/ stimuli, and the third /ɛ/ and /ə/ stimuli. Each tape contained six randomized sequences of 52 stimulus pairs. These consisted of 10 pairs of identical stimuli (each of the two prototypes and each of the eight neighbors paired with itself once) and 16 pairs of nonidentical stimuli (each of the two prototypes paired with each of its four neighbors, in both orders), each presented with two ISIs: 200 ms and 1 s.

2. Subjects and Procedure

Eighteen subjects from the same general pool participated. Each subject listened to each stimulus tape, in a balanced order. The procedure was identical to that of Experiment 1.

B. Results and Discussion

The results are presented in Table 3, with the order effects at the bottom. Consider first the results for the /ɛ/ stimuli, shown in the last three

columns. A 4-way ANOVA was conducted on these data, with the factors Context (/ɪ/, /æ/, /ə/), Order (P first, second), ISI (short, long), and Neighbor (1, 2, 3, 4). Several significant effects did not involve Order: The main effect of ISI, $F(1,17) = 8.15$, $p = .0109$, reflected better discrimination performance at the shorter ISI, which is not surprising; the main effect of Neighbor, $F(3,51) = 11.18$, $p = .0001$, was due to much better performance for N3 pairs than for the other pairs, a surprising finding; and the Context by Neighbor interaction, $F(6,102) = 3.07$, $p = .0084$, reflected a tendency towards reduced discrimination performance for pairs involving the neighbor stimulus most dissimilar to the context (N2 in the /ɪ/ context, N4 in the /æ/ context, N1 in the /ə/ context), which suggests a contrastive influence of the context on the perceptual structure of the /ɛ/ category.

Table 3. *Average percentages of "different" responses for all stimulus pairs in Experiment 2, and order effects (prototype first minus prototype second).*

Pair	ISI	Prototype vowel (P) /ɪ/	/æ/	/ə/	/ɛ/ɪ	/ɛ/æ	/ɛ/ə
				Identical pairs			
P-P	200	13.9	5.6	15.7	7.4	9.3	15.7
	1000	5.6	9.3	9.3	11.1	13.0	13.9
N1-N1	200	13.0	3.7	7.4	7.4	5.6	9.3
	1000	13.0	5.6	16.7	7.4	10.2	5.6
N2-N2	200	7.4	3.7	9.3	5.6	7.4	13.9
	1000	11.1	9.3	11.1	6.5	7.4	8.3
N3-N3	200	4.6	7.4	6.5	4.6	6.5	7.4
	1000	8.3	5.6	9.3	5.6	8.3	9.3
N4-N4	200	3.7	10.2	18.5	7.4	4.6	13.9
	1000	9.3	5.6	13.0	12.0	6.5	13.0
				Nonidentical pairs: prototype first			
P-N1	200	62.0	71.3	74.1	63.9	63.0	63.9
	1000	55.6	67.6	73.1	53.7	51.9	52.8
P-N2	200	60.2	52.8	48.1	51.9	51.9	61.1
	1000	57.4	45.4	31.5	40.7	38.0	62.0
P-N3	200	46.3	72.2	68.5	72.2	75.0	74.1
	1000	32.4	68.5	55.6	80.6	78.7	71.3
P-N4	200	75.9	66.7	74.1	59.3	62.0	71.3
	1000	75.0	64.8	65.7	58.3	57.4	63.9
				Nonidentical pairs: prototype second			
N1-P	200	35.2	54.6	33.3	53.7	53.7	56.5
	1000	27.8	39.8	19.4	38.9	44.4	32.4
N2-P	200	72.2	49.1	60.2	46.3	59.3	55.6
	1000	74.1	24.1	46.3	48.1	66.7	47.2
N3-P	200	49.1	71.3	75.0	71.3	75.9	76.9
	1000	38.9	62.0	72.2	42.6	61.1	58.3
N4-P	200	63.0	63.0	46.3	59.3	53.7	60.2
	1000	41.7	45.4	33.3	50.9	33.3	56.5
				Nonidentical pairs: order effects (difference scores)			
N1	200	26.9	16.7	40.7	10.2	9.3	7.4
	1000	27.8	27.8	53.7	14.8	7.4	20.4
N2	200	-12.0	3.7	-12.0	5.6	-7.4	5.6
	1000	-16.7	21.3	-14.8	-7.4	-28.7	14.8
N3	200	-2.8	0.9	-6.5	0.9	-0.9	-2.8
	1000	-6.5	6.5	-16.7	38.0	17.6	13.0
N4	200	13.0	3.7	27.8	0.0	8.3	11.1
	1000	33.3	19.4	32.4	7.4	24.1	7.4

Effects involving Order were of primary interest: There was a highly significant main effect of Order, $F(1,17) = 25.54$, $p = .0001$, which indicated a positive stimulus order effect overall. The interaction of Order and Neighbor fell short of significance. Several other interactions were significant, however. One was between Order, ISI, and Neighbor, $F(3,51) = 8.08$, $p = .0002$, indicating that an Order by Neighbor interaction emerged at the longer ISI. Another significant interaction involved Context, Order, and Neighbor, $F(6,102) = 4.05$, $p = .0011$, indicating that the pattern of order effects did change with test context. This was particularly true at the longer ISI; the quadruple interaction was just significant, $F(6,102) = 2.24$, $p = .0449$.

Since /ɛ/ stimuli showed no really large order effects at the shorter ISI, the results at the longer ISI are of primary interest. Table 3 reveals that N1 pairs showed a positive order effect, as predicted by the neutralization hypothesis, though the effects were smaller here than in Experiment 1, especially in the /æ/ test context, despite the longer ISI. N3 pairs, on the other hand, showed results highly discrepant from those of Experiment 1. Instead of the negative effect obtained there and predicted by the neutralization hypothesis, these pairs exhibited positive order effects, with the effect in the /ɪ/ context condition being more than twice as large as the effects in the other two context conditions. Furthermore, N2 and N4 pairs, which had not shown any order effects in Experiment 1 (as predicted by the neutralization hypothesis), showed some large effects here that depended on context: N2 pairs showed a large negative effect in the /æ/ context, but a positive effect in the /ə/ context. N4 pairs showed a large positive effect in the /æ/ context.

We turn now to the results for the contextual stimuli, which are shown in the first three columns of Table 3. We dispense with statistical analyses here, which presumably would show many complex interactions, and instead discuss the pattern of substantial order effects in relation to Experiment 1. Even more consistently than with the /ɛ/ stimuli, order effects increased in absolute magnitude at the longer ISI, but there were also a number of large order effects at the short ISI here. The results for the /ɪ/ stimuli were not unlike those in Experiment 1, though the effects differed in size and, overall, were less compatible with the neutralization hypothesis: A large positive effect for N1 pairs, but only a negligible negative effect for N3 pairs; a moderate negative effect for N2 pairs, and a large positive effect for N4 pairs. The results for /æ/ stimuli at the shorter ISI were quite similar to the Experiment 1 results, showing only a positive effect for N1 pairs. At the longer ISI, however, positive effects emerged for N2 and N4 pairs as well. The most discrepant results were obtained for /ə/ stimuli, which in Experiment 1 had not exhibited any large order effects at all. In this experiment, all pairs showed order effects. Those for N1 and N4 pairs were extremely large and positive, those for N2 and N3 pairs smaller and negative.

These data provide strong indications that changes in the test environment affected the pattern of order effects. Overall, the results are much less favorable to the neutralization hypothesis than the results of Experiment 1. The reason for this may be that the "neutral point" that vowels in memory drift towards is specific to each stimulus ensemble. If such a point exists, it should be possible to infer its location from the patterns of order effects for the two stimulus sets in a given context condition: Arms of a stimulus cross that are associated with negative order effects point outward, towards the "neutral point," whereas arms associated with positive effects point inward. If the results for each of the two stimulus crosses are internally consistent in that they point in a particular direction, then the neutral point is located at the intersection of these two directions.

In the test containing /ɪ/ and /ɛ/ stimuli, the pattern of order effects for the /ɪ/ stimuli (N2 and N3 negative, N1 and N4 positive) points towards /ɛ/; the pattern for the /ɛ/ stimuli in that context is not internally consistent (both N1 and N3 positive) but is most compatible with a neutral point at the /ɛ/ prototype. Thus these data suggest a neutral point in the vicinity of the /ɛ/ prototype, which incidentally is consistent with the data of Repp et al. (1979) and of Cowan and Morse (1986).

In the test containing /æ/ and /ɛ/ stimuli, the results for the /æ/ stimuli (all positive) point inwards towards the /æ/ prototype, whereas the results for /ɛ/ stimuli (N2 negative, N3 and N4 positive) point "north of" /æ/. Thus the neutral point here may have been located near the /æ/ prototype.

Finally, in the test containing /ə/ and /ɛ/, the results for the /ə/ stimuli (N2 and N3 negative, N1 and N4 positive) point quite clearly to the "southwest" (i.e., away from /ɛ/), whereas the /ɛ/ order effects are all positive and therefore indicate a "neutral" point at the /ɛ/ prototype. Thus, the data from this test are contradictory and do not suggest a unique neutral point.

Although these results do not provide very clear support for stimulus range specific "neutral points," they are even less supportive of the range-independent neutralization hypothesis of Cowan and Morse (1986). They replicate only their finding that order effects, regardless of their direction, increase with interstimulus interval. It is noteworthy, however, that the results for *individual* vowel sets (i.e., for each cross in Figure 1) are nearly always internally consistent and thus "point" in a particular direction. Opposite neighbors *never* yielded large negative order effects. Thus we need to consider the possibility that each vowel set has its own individual neutral point, as it were. This seems unparsimonious, but there is in fact a plausible rationale. Each vowel category has a best exemplar or prototype (see, e.g., Grieser & Kuhl, 1989) which may or may not coincide with the prototype suggested by the data of Peterson and Barney (1952). Moreover, the location of a vowel prototype is likely to be sensitive to stimulus context. The pattern of order effects may tell us something about the actual locations of vowel prototypes and their shifts with changes in context. Rather than becoming more neutral in memory, vowels may become more prototypical.

Although this seems a very reasonable hypothesis, there is a serious problem with it: It makes just the opposite predictions of the neutralization hypothesis. The data of Experiments 1 and 2 are not at all compatible with the idea that vowels in memory become assimilated to prototypes, because they would imply that these prototypes are often located centrally in vowel space, which does not make sense. It is still possible that prototypes play a role, but that this role is not assimilative but contrastive in nature. Before discussing this idea further, we report the results of a third experiment, in which we employed the methodology of our original study (Repp et al., 1979)—viz., stimulus continua spanning two vowel categories—to partially replicate Experiment 2 and to conduct another specific test of the predictions of the neutralization hypothesis. Replication of the Experiment 2 results seemed desirable because of their striking inconsistencies with Experiment 1.[5]

III. EXPERIMENT 3

In Experiment 3 we employed two vowel continua, one ranging from /ɛ/ to /æ/, and the other from /ɛ/ to /ə/. The first stimulus series continued where the /i/-/ɪ/-/ɛ/ continuum used by Repp et al. (1979) ended. Since the two endpoint vowels, /ɛ/ and /æ, are about equally peripheral in the vowel space (see Figure 1), the neutralization hypothesis of Cowan and Morse (1986) predicts no pronounced order effects along this continuum. Indeed, in Experiment 1 pairings of /ɛ/ with its N2 and of /æ/ with its N4, which lie approximately on this continuum (see Figure 1), yielded no order effects. In Experiment 2, however, large order effects (negative and positive, respectively) emerged for these very same pairs at the longer ISI. Since Experiment 3 used the same long ISI and an even more restricted stimulus context, it was expected to replicate the results of Experiment 2.

The second continuum ranged from /ɛ/ to /ə/. According to the neutralization hypothesis, strong negative order effects should be obtained at the /ɛ/ end of this continuum, but none at the /ə/ end. These predictions were upheld in Experiment 1 for pairings of /ɛ/ with its N3 and of /ə/ with its N4, which lie almost exactly on the /ɛ/-/ə/ continuum (see Figure 1). Again, the results of Experiment 2 were contradictory: The very same pairs yielded a small and a large positive order effect, respectively. We wondered whether Experiment 3 would replicate this curious pattern.

A. Methods

1. Stimuli

The formant frequencies for the /ɛ/, /æ/, and /ə/ prototypes, which served here as continuum endpoints, were the same as in Experiments 1 and 2 (see Table 1). Five additional vowels were interpolated linearly between /ɛ/ and /æ/, and between /ɛ/ and /ə/, to obtain two 7-member vowel continua. Other stimulus characteristics were the same as previously.

The stimuli of each continuum were recorded in pairs on separate experimental tapes. The interstimulus interval was 1 s within pairs and 2.4 s between pairs. The pairs varied in the degree of stimulus separation (measured in steps on the continuum). For each continuum, there were 37 pairs: 7 identical pairs, 12 one-step pairs (6 stimulus combinations, 2 orders), 10 two-step pairs, and 8 three-step pairs. Ten blocks of these 37 pairs were recorded for each continuum, with different random orders in each block.

2. Subjects and Procedure

Twenty-four undergraduate students served as subjects. Each participated in two sessions, in each of which the same stimulus tapes were presented. In one session, they were asked to give same-different responses; in the other, they

identified the second vowel in each pair. The identification data will not be reported here in detail.[6] The order of these two conditions, and of the two stimulus sets within sessions, was counterbalanced across subjects. The tapes were played back monaurally in a quiet room using a tape recorder and earphones of good quality. Half the subjects listened to the stimuli in their right ear, and the other half in their left ear; no significant ear differences were observed.

B. Results and Discussion

1. /ɛ/-/æ/ continuum

Figure 2, left panel, shows the results for nonidentical pairs from the /ɛ/-/æ/ continuum as a function of location on the continuum, step size, and stimulus order. Predictably, the percentage of "different" responses increased as the step size increased. Scores also tended to be highest in the middle of the continuum, which is in agreement with the previously demonstrated tendency for isolated vowels to be perceived in a semi-categorical fashion (Repp et al., 1979). Effects of stimulus order are represented by the difference between the closed and open symbols. There were large order effects for one- and especially two-step pairs; for three-step pairs, a ceiling effect probably prevented order effects from emerging. At the /ɛ/ end of the continuum discrimination performance was much better when the more /ɛ/-like stimulus occurred second than when it occurred first; this difference is particularly large for stimulus pair 1-3. In the middle of the continuum there were no pronounced order effects, but at the /æ/ end a reversal occurred: Correct responses were more frequent when the /æ/ endpoint stimulus occurred second.

Analyses of variance were conducted on 1-step and 2-step pairs separately, with the factors stimulus pair and order. The stimulus pair by order interaction, which reflects the change in magnitude and direction of the order effect across the continuum, was highly significant for 1-step pairs, $F(5,110) = 6.79$, $p < .0001$, and for 2-step pairs, $F(4,88) = 39.09$, $p < .0001$. In addition, there was a main effect of stimulus pair for 1-step pairs, $F(5,110) = 13.17$, $p < .0001$, and for 2-step pairs, $F(4,88) = 27.20$, $p < .0001$, which reflects the aforementioned performance peak in the middle of the continuum, as well as the fact that discrimination was better at the /æ/ end than at the /ɛ/ end. For 2-step pairs, there was also a main effect of order, $F(1,22) = 22.03$, $p = .0001$, due to the exceptionally large order effect for the 1-3 stimulus pair.

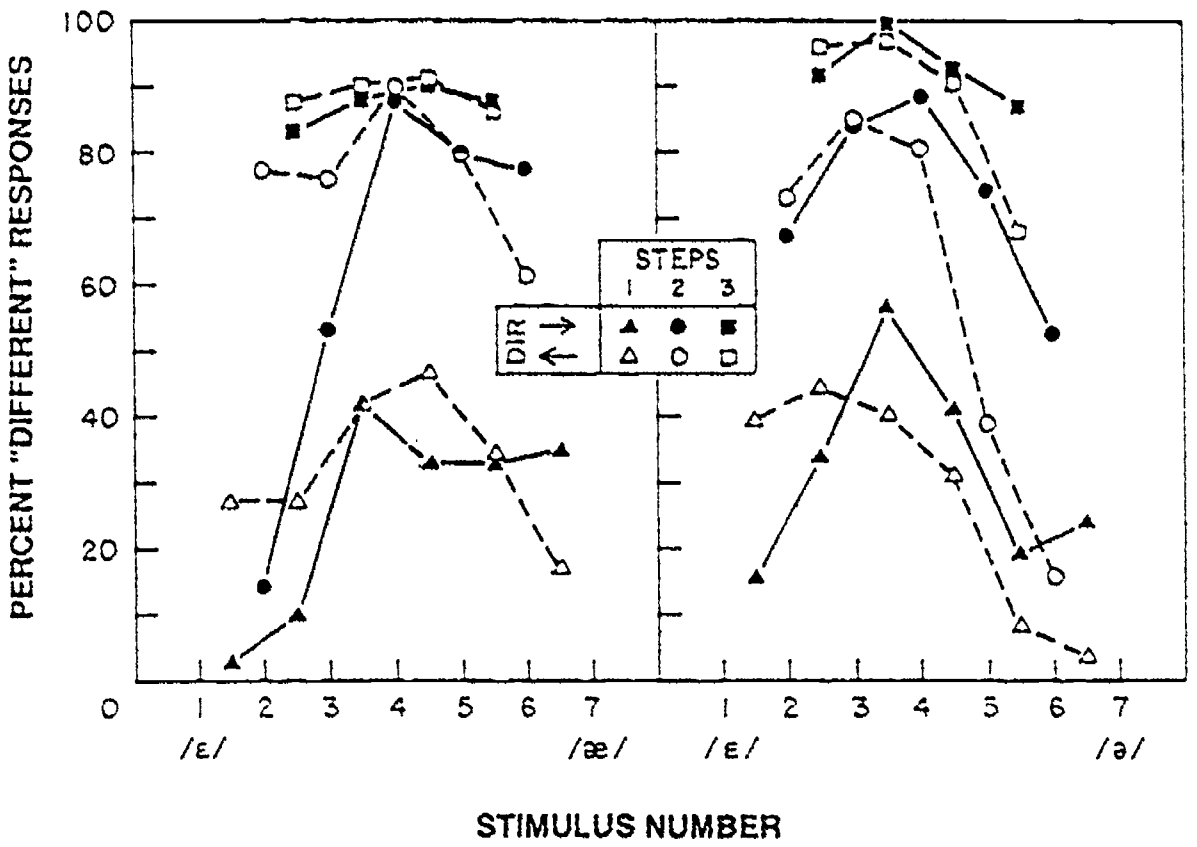


Figure 2. Results of Experiment 3: Percentages of "different" responses to pairs of nonidentical stimuli from /ɛ/-/æ/ (left panel) and /ɛ/-/ə/ (right panel) continua. Parameters are stimulus order (ascending versus descending direction) and separation (one, two, or three steps).

In the terminology of the earlier experiments, these results show negative order effects at both continuum endpoints, which is inconsistent with the neutralization hypothesis and with the results of Experiment 1. The results are more similar to those of Experiment 2, where a large negative effect was obtained on the /ɛ/ side, but a small positive effect on the /æ/ side. There, a "neutral point" near the /æ/ prototype was suggested. Here, a neutral point is defined by the point on the continuum where no order effect is obtained (i.e., the point at which the functions for the two stimulus orders in Figure 2 cross each other). That is somewhere between stimuli 4 and 5, which is closer to /æ/ than to /ɛ/. It is worth noting that the labeling data obtained from the same subjects showed the /ɛ/-/æ/ category boundary to be in the same location. These data, then, are reasonably consistent with Experiment 2; the differences may be attributed to the changes in stimulus range and frequency (the prototype stimuli occurred more often than other stimuli in the earlier experiments, but not in Experiment 3).[7]

2. /ɛ/-/ə/ continuum

The results for this continuum are displayed in Figure 2, right panel. Again, there were stimulus order effects that reversed direction along the continuum. At the /ɛ/ end, order effects emerged only for 1-step pairs and favored pairs in which the more /ɛ/-like stimulus came second. Effects at the /ə/ end were larger and in the opposite direction. The crossover point was closer to /ɛ/ than to /ə/.

The statistical analyses showed the pattern of results to be very reliable. The stimulus pair by order interaction was highly significant for 1-step pairs, $F(5,110) = 9.71$, $p < .0001$, 2-step pairs, $F(4,88) = 12.71$, $p < .0001$, and even for 3-step pairs, $F(3,66) = 12.91$, $p < .0001$. In addition, there was a significant main effect of stimulus pair for 1-step pairs, $F(5,110) = 14.36$, $p < .0001$, 2-step pairs, $F(4,88) = 32.91$, $p < .0001$, and 3-step pairs, $F(3,66) = 11.96$, $p < .0001$, due to better discrimination performance in the center of the continuum, plus higher scores at the /ɛ/ end than at the /ə/ end. A main effect of order was obtained for 2-step pairs, $F(1,22) = 29.96$, $p < .0001$, and for 3-step pairs, $F(1,22) = 7.96$, $p = < .0099$, due to the large order effects at the /ə/ end of the continuum.

The /ɛ/-/ə/ continuum thus yielded negative order effects at both endpoints, with the larger effects at the /ə/ end. The negative order effect at the /ɛ/-end is consistent with the neutralization hypothesis and with the data of Experiment 1. However, the large negative order effect at the /ə/-end is in strong contradiction to both. Unfortunately, it also contradicts the findings of Experiment 2 which showed a large positive effect for /ə/ as well as a small positive effect for /ɛ/. These data, it will be recalled, were inconsistent in that they did not suggest a single neutral point; they remain mysterious. The present data suggest a neutral point somewhere between stimuli 3 and 4 on the continuum (i.e., closer to /ɛ/ than to /ə/). Again we note that the category boundary obtained in the labeling task fell there also.

IV. GENERAL DISCUSSION

The purpose of the present series of experiments was to test the generality of the neutralization hypothesis proposed by Cowan and Morse (1986). We found many deviations from the predictions of this hypothesis, such as the large stimulus order effects in the vicinity of /ə/ obtained in Experiments 2 and 3. Only Experiment 1 yielded data that, on the whole, seemed to support the hypothesis. Although that experiment may seem to have been the strongest test because it included the largest variety of stimuli, it may actually have been the weakest: If stimulus order effects depend on the distribution of the stimuli in vowel space, then the most representative distribution has the neutral vowel at its center and therefore may yield data that seem to support the neutralization hypothesis. Only by using more limited stimulus distributions can the range-specific nature of the order effects be revealed.

The discrepancies among the results of Experiments 1-3 provide ample evidence of such range-specific changes, though it must be admitted that the pattern of effects obtained cannot always be rationalized. On the whole, however, our data suggest that vowels change in memory not necessarily towards the neutral vowel /ə/, but towards a quality that lies within the stimulus range of a given experiment. What could be the reason for this?

It is well known, and the data from our Experiment 3 confirm, that the perception of isolated vowels is weakly categorical: Discrimination tends to be best around the category boundaries. These discrimination peaks suggest that covert categorization plays a role in the "same-different" task. Almost certainly, the first vowel in a stimulus pair is remembered in a dual code, one categorical and the other continuous (Fujisaki & Kawashima, 1970; Pisoni, 1973, 1975). While, at a short ISI, subjects can utilize the auditory stimu

lus trace for comparisons, at longer ISIs they must rely increasingly on the category label assigned to the first vowel in a pair. Since the size of order effects increases with ISI, it seems likely that these effects are a phenomenon related to the covert categorization of the vowel stimuli.[8]

We already noted, however, that simple phonetic classification does not predict the order effects that were in fact obtained. Phonetic categorization amounts to an assimilation to the prototype, so that *positive* order effects would be predicted at the ends of stimulus continua. The negative order effects obtained suggest that vowels held in memory were assimilated towards some standard(s) located *between* prototypes. The only "special" point in that ambiguous region is the category boundary—the point of maximum uncertainty. Indeed, we found in Experiment 3 that the "neutral point" suggested by the order effects coincided with the category boundary. It seems, therefore, that the category boundary somehow "attracts" vowels in memory; at the same time, however, such a process cannot be reconciled with the idea of covert phonetic categorization. Also, it is far from clear why the perceptually most stable vowels (the prototypes, and others near them) should exhibit the largest changes in memory.

There is a way, however, of accounting for these data on the basis of phonetic categorization. The apparent changes in the remembered quality of the first stimulus of a vowel pair may not occur autonomously during the silent interstimulus interval, but rather may be caused by the arrival of the second stimulus, which interacts with the memory trace of the first. This suggestion is supported by three solid findings from earlier research. First, it is well known that successively presented vowels engage in contrastive interactions (Thompson & Hollien, 1970; Repp et al., 1979), provided the interstimulus interval is not too short (Shigeno & Fujisaki, 1980) and they can be perceived as belonging to different phonetic categories (Shigeno, 1986). Contrast, of course, facilitates discrimination. Second, vowels that are unambiguous representatives of a phonological category exert larger contrast effects than do more ambiguous vowels; it is the latter that are pushed around in the context of less ambiguous neighbors (Crowder, 1982). Third, it has also been shown that, when pairs of vowels are to be judged, retroactive contrast is larger than proactive contrast (Repp et al., 1979; Shigeno & Fujisaki, 1980), presumably because a memory trace is less stable than a newly arrived stimulus. These three observations together predict stimulus order effects of the kind found in Experiment 3 and earlier: At either endpoint of a vowel continuum, discrimination should be easier *when the more ambiguous vowel comes first and the less ambiguous vowel comes second in a pair,* because the retroactive contrast effect in such a pair will be larger than any proactive contrast effect obtained in the opposite arrangement. An increase in the order effect with temporal separation between the stimuli is also consistent with this explanation: As the memory trace of an initially stable vowel becomes weaker over time, its proactive contrast effect on a following unstable vowel will decrease, as shown by Crowder (1982). In fact, the labeling data in Experiment 3 revealed no significant proactive contrast effects at all (footnote 6). On the other hand, when an unstable vowel is followed by a stable vowel, the retroactive contrast effect exerted by the latter on the former will stay the same or even increase with temporal separation. Thus, according to this interpretation, order effects do not occur because prototypical vowels become less stable in memory, but because unstable vowels shift away from following stable vowels.

Attractive as this explanation seems, there is a problem with it. Repp et al. (1979) found that an interfering vowel sound eliminated retroactive contrast effects but left stimulus order effects intact. Similarly, reanalysis of data from an unpublished vowel discrimination experiment by one of us (RGC), in which interfering sounds were used together with a long ISI, revealed large stimulus order effects. These findings suggest that contrast and order effects are unrelated. The present experiments provide no additional information on that point. Since no interfering sound was present, it is possible that retroactive contrast was operating. However, since retroactive effects are only slightly larger than proactive effects (Repp et al., 1979), the total absence of proactive contrast effects in Experiment 3 suggests that retroactive contrast, if present at all, was not very strong. Thus it seems that the retroactive contrast explanation may not be correct, after all.

A possible solution to this dilemma is suggested by the psychophysical theory of Durlach and Braida (1969; Braida, Durlach, Lim, Berliner, Rabinowitz, & Purks, 1984), which has been applied to vowel resolution by Macmillan et al. (1988). These authors distinguish between a sensory trace and a more stable "context code." The context code is not limited to the phonetic

labels listeners can apply; rather, it reflects their maximum labeling capacity. The context code thus is a subphonetic, quasi-categorical representation. It is also unstable and, as its name indicates, subject to influences of stimulus context. Although listeners may assign a phonetic category to the first stimulus in a pair when it arrives, they may use its richer context code to compare it to a following stimulus. This context code may be subject to retroactive contrast, even when the phonetic category assigned to the first stimulus in a pair remains unaffected (i.e., is not revised by the listener).[9]

This interpretation receives independent support from other psychophysical studies involving nonspeech, even non-auditory, stimuli. Effects of presentation order in the method of constant stimuli have been noted since the earliest days of psychophysics (see Needham, 1934; Hellström, 1985). A recent demonstration was provided by Masin and Fanton (1989) who used vertical lines in a visual length discrimination task. They concluded that subjects used a quasi-categorical code (i.e., a context code) to compare successive stimuli, and that "the categorical comparison is accompanied by an additional inferential decision process that uses only the category relative to the second stimulus because more weight is given to that category, or because the category relative to the first stimulus is momentarily forgotten" (p. 485). They do not assume a change in the memory code of the preceding stimulus through retroactive contrast, but the same net effect is achieved by a hypothetical weighting process favoring the more recent stimulus, which is really just another metaphor for memory degradation of the earlier stimulus. The argument is easily transferred to stimuli such as vowels, as long as it is assumed that their context code is always in terms of abstract labels that reflect the relative location of stimuli in the range of all stimuli employed. Within the context code, continuum endpoints do not function as anchors (Macmillan et al., 1988) and therefore can plausibly degrade in memory.

It seems, therefore, that stimulus order effects in vowel discrimination represent an instance of more general presentation order effects in psychophysical judgment, not a phenomenon specific to the memory coding of speech sounds. The fact that there remain a number of unexplained irregularities in our results may be attributed to the acoustic complexity of vowels, compared to the simple unidimensional stimuli used in most psychophysical studies of "time order errors." In most general terms, such time order errors are due to a contraction of the effective range for remembered stimuli, a consequence of gradually substituting generic information for specific information that is lost (Hellström, 1985). The generic information reflects the recent stimulus history (Helson's, 1964, "adaptation level"). The neutralization hypothesis of Cowan and Morse (1986) may be seen as a specific application of these general principles to certain sets of vowels whose adaptation level happens to be in the neutral region.

REFERENCES

Braida, L. D., Durlach, N. I., Lim, J. S., Berliner, J. E., Rabinowitz, W. M., & Purks, S. R. (1984). Intensity perception. XIII. Perceptual anchor model of context coding. *Journal of the Acoustical Society of America, 76*, 722-731.

Chiba, T., & Kajiyama, M. (1941). *The vowel, Its nature and structure*. Tokyo: Kaiseikan.

Cowan, N., and Morse, P. A. (1986). The use of auditory and phonetic memory in vowel discrimination. *Journal of the Acoustical Society of America, 79*, 500-507.

Crowder, R. G. (1982). Decay of auditory information in vowel discrimination. *Journal of Experimental Psychology: Learning Memory, & Cognition, 8*, 153-162.

Durlach, N. I., & Braida, L. D. (1969). "Intensity perception. I. Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America, 46*, 372-383.

Eimas, P. D. (1963). The relation between identification and discrimination along speech and non-speech continua. *Language and Speech, 6*, 206-217.

Fujisaki, H., & Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanics.. *Annual Report of the Engineering Research Institute* (University of Tokyo), *29*, 207-214.

Grieser, D., & Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology, 25*, 577-588.

Healy, A. F., & Repp, B. H. (1982). Context independence and phonetic mediation in categorical perception. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 68-80.

Helson, H. (1964). *Adaptation-level theory: An experimental and systematic approach to behavior*. (Harper & Row, New York).

Hellström, A. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin, 97*, 35-61.

Kewley-Port, D., & Atal, B. S. (1989). Perceptual differences between vowels located in a limited phonetic space. *Journal of the Acoustical Society of America, 85*, 1726-1740.

Macmillan, N. A., Goldberg, R. F., & Braida, L. D. (1988). Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua. *Journal of the Acoustical Society of America, 84*, 1262-1280.

Masin, S. C., & Fanton, V. (1989). An explanation for the presentation-order effect in the method of constant stimuli. *Perception & Psychophysics, 46*, 483-486.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85*, 2088-2113.

Needham, J. G. (1934). The time error in comparison judgments. *Psychological Bulletin, 31*, 229-243.

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.

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

Pisoni, D. B. (1975). Auditory short-term memory and vowel perception. *Memory & Cognition, 3*, 7-18.

Repp, B. H., Healy, B. H., & Crowder, R. G. (1979). Categories and context in the perception of isolated steady-state vowels. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 129-145.

Shigeno, S. (1986). The auditory Tau and Kappa effects for speech and nonspeech stimuli. *Perception & Psychophysics, 40*, 9-19.

Shigeno, S., & Fujisaki, H. (1980). Context effects in phonetic and non-phonetic vowel judgments. *Annual Bulletin RILP* (Tokyo) *14*, 217-224.

Thompson, C. L., & Hollien, H. (1970). Some contextual effects on the perception of synthetic vowels. *Language and Speech, 13*, 1-13.

FOOTNOTES

Journal of the Acoustical Society of America, 88, 2080-2090 (1990).

†Also Department of Psychology, Yale University.

[1]We note, with some embarrassment, that the effect is incorrectly described in Experiment 2 of Repp et al. (1979, p. 138). We are confident that this is a mistake in the text, and that the data conformed to the description given here and in Experiment 1 of Repp et al. (1979, p. 134).

[2]They generously credit us (Repp et al., 1979) with this idea, though we did not state it explicitly.

[3]We follow common practice in referring to the Peterson-Barney for the synthesis of isolated vowels, even though these data derive from vowels produced in a /h_d/ context.

[4]We are aware of the dangers of conducting multiple analyses on the same data without adjusting the p levels. However, these follow-up analyses serve the sole purpose of clarifying complex interactions, and the significance levels are generally so high as to make adjustments superfluous.

[5]We are taking the liberty of describing Experiment 3 in these terms for expository reasons. Actually, Experiment 3 was conducted before Experiments 1 and 2.

[6]The purpose of this condition was to assess proactive contrast effects. Somewhat surprisingly (see, e.g., Crowder, 1982), no significant effects were found. Note that the occurrence of retroactive contrast effects is not precluded by these findings (see General Discussion).

[7]Another demonstration that order effects change with stimulus range is obtained from a comparison with the old data of Repp et al. (1979). They showed a small negative order effect at the /ɛ/ end of the /i/-/ɪ/-/ɛ/ continuum. In those pairs, however, the stimulus paired with the /ɛ/ prototype was more /ɪ/-like, while in the present pairs it was more /æ/-like; hence the present negative order effect at the /ɛ/ end is contrary to the effect obtained previously.

[8]Accordingly, stimulus order effects should be smaller in tasks that force subjects to rely more on the stimulus trace. Kewley-Port and Atal (1989) conducted experiments with four sets of vowel stimuli (/i/-/ɪ/, /ɛ/-/æ/, /u/-/ʊ/, and /æ/-/ɑ/-/ɔ/) arranged in prototype-neighbor configurations, but the task required numerical dissimilarity judgments for the two vowels in a pair. We re-analyzed their raw data (kindly provided by Diane Kewley-Port) and found stimulus order effects to be small and following a consistent pattern in only one of the sets, /u/-/ʊ/. That pattern agreed with that obtained in our Experiment 1, suggesting that /u/-like vowels drifted towards /ʊ/.

[9]Macmillan et al. also noted that points of perceptual stability serve as "anchors" for the context code. Although anchors are often located at the ends of stimulus continua, Macmillan et al. deduced from their vowel discrimination data that *boundary stimuli* served as anchors on their /i/-/ɪ/-/ɛ/ continuum. This surprising (and somewhat tentative) conclusion is in agreement with the order effects obtained in our experiments, though it leads back to the neutralization metaphor and should perhaps be regarded with caution.

Haskins Laboratories Status Report on Speech Research
1990, SR-103/104, 125-136

The Haskins Laboratories' Pulse Code Modulation (PCM) System

D. H. Whalen, E. R. Wiley, Philip E. Rubin, and Franklin S. Cooper

The Pulse Code Modulation (PCM) method of digitizing analog signals has become a standard both in digital audio and in speech research, the focus of this paper. The solutions to some problems encountered in earlier systems at Haskins Laboratories are outlined, along with general properties of A/D conversion. Specialized features of the current Haskins Laboratories system, which has also been installed at more than a dozen other laboratories, are also detailed: the Nyquist filter response; the high frequency pre-emphasis filter characteristics; the dynamic range; the timing resolution, for single and (synchronized) dual channel signals; and the form of the digitized speech files (header information, data, and label structure). While the solutions adopted in this system are not intended to be considered a standard, the design principles involved are of interest to users and creators of other PCM systems.

INTRODUCTION

The Pulse Code Modulation (PCM) system of digitizing analog waveforms, in which amplitude samples are taken at frequent, regular intervals, can accurately represent continuously varying signals as binary digital numbers (cf. Goodall, 1947). In the years since its introduction, PCM has become the standard technique for the digital sampling of analog signals for research purposes (in preference to such alternatives as delta-modulation or predictive coding of various sorts; cf. Heute (1988)). PCM systems are now available for almost any computer, and the recording industry's digital CD's have surpassed analog formats in sales.

Although PCM systems are now commonplace, this has not always been the case. When Haskins Laboratories needed an interactive, multi-channel system in the mid 1960's, such systems simply were not available. A design was devised, and implemented in an unconventional way, to meet the needs of our researchers. Much of our speech research at that time was concerned with perceptual responses to different words or syllables arriving at the two ears simultaneously or with small temporal offsets. Stimulus tapes for such experiments could be made by tape splicing (a separate tape for each ear) and rerecording the signals onto a dual-track tape, but the method was both error-prone and laborious. Moreover, each change in stimulus condition—different pairings of the overlapping words, differences in relative onset time, or in relative level—required doing the whole job over again. Hence, the design objective was to store all the stimulus words in the computer, then convert them back to analog form, and bring them out in real time to a listener or, in the usual case, to a dual-track recorder in whatever combination of stimuli, offsets and levels the experimenter might choose.

The system that resulted is still in use, but its very singularity makes it mostly of historical interest. Certain aspects of that system, however, are incorporated into newer systems based on current, commercially available hardware. These newer systems are in place at Haskins Laboratories and at more than a dozen other sites

The writing of this article was supported by NIH contract N01-HD-5-2910 to Haskins Laboratories. We thank Michael D'Angelo, Vincent Gulisano, Mark Tiede, Ignatius G. Mattingly, Patrick W. Nye, Tom Carrell, David B. Pisoni, and two anonymous reviewers for helpful comments. We also thank Leonard Szubowicz for the time and care spent designing and implementing the original version of the Haskins PCM software.

in the United States and abroad. These will be described in detail so that current and future users of Haskins-based systems can have easy reference to them, and so that designers of other systems can see the reasoning that went into the choices made. The basic principles of A/D conversion will be outlined along the way.

Early Problems and Solutions

In 1964, when the earliest PCM system at Haskins Laboratories was begun, the challenge for our designers was simply to create a system where none could be bought. Although PCM was common in telephony, there were no commercial systems available for programmable computers. We therefore designed a system to be interfaced with a Honeywell DDP24 computer (and later on a DDP 224) with 8K of memory. Although only brief stretches of speech could be digitized directly into core memory, double buffering allowed the system to deal with continuous speech input; that is, the incoming digital stream was stored alternately in one of two buffer areas of memory while earlier samples were being read out from the other buffer and written to digital tape. For output, 2.8 seconds of speech could be called up at will, directly from core memory. Longer sequences could be compiled onto digital tape, and then read off from the tape in near-real time. For two-channel synchronized output, the samples stored on the tape alternated between the two speech channels. Later, faster disks became available, so that long, one channel sequences could be output with it going to tape. The same might have happened for the two-channel output, except that technology passed this system by, and it disappeared when the DDP 224 was liquidated for its gold content in 1982.

The next challenge was to meet the growing demands of an increasing research field by adding more channels which could access a set of common disks, avoiding both the recording on digital tape and the limitation to one user at a time. The result was a multi-channel PCM system, which was designed by Leonard Szubowicz, Rod McGuire and E. R. Wiley and implemented with the collaboration of Richard Sharkany. It consists (it is still in use) of four output channels and two input channels, controlling DMA (Direct Memory Access) boards and filled continuously in FIFO (First In, First Out) circuits. Memory is dynamically allocated to each active channel; the amount is trimmed back as other requests come in, or expanded as other channels become inactive. The advantage of this memory management is that large memory areas make the rare FIFO shutdown (i.e., data did not arrive in time) even rarer. The advantage of FIFO organization is that buffers can be filled with less concern for time-critical disk accesses. A drawback is that the system does not know exactly where in the output it is, since only the DMA has that information, so that the controlling computer cannot receive an exact reading of how far the sequence has gone.

While the speech waveform is the primary signal of interest at this laboratory, other analog signals such as the output of transducers measuring the speech articulators and muscles (electromyographic (EMG) signals) are also used. Many such signals are more restricted in the frequency domain, and thus can be represented adequately at slower sampling rates. The lower the rate, the less disk space is used. Even for speech, some purposes are well-served by the 10 kHz rate, while others need the information between 5 kHz and 10 kHz which is preserved at the 20 kHz rate. Each of the six channels can be used at a 10 kHz or 20 kHz sampling rate. One input channel and one output channel also support the rates of 100, 200, 500, 1000, 2000, 4000, 5000, 8000, and 16000 samples per second. If necessary, these two channels can be connected to an external clock which can be run at any rate up to 50 kHz.

When the system was designed, computer memory was quite limited, so the simplest, memory intensive solutions to real-time output were not available. To obtain a large throughput from a small system, our design undid the major advance in computation, von Neumann's use of data and instructions in the same area. Given the small address area of our platform, the PDP 11/04, there was very little room to write a program and extremely little left over for data. To overcome this limitation, additional memory was attached, even though the processor could not access it. However, the DMA's could, and the program was capable of telling them how to do so. In this way, an adequate amount of memory was available to sustain a throughput of about 40,000 data samples per second, divided among up to four channels.

A continuing concern was the synchronization of any two PCM channels. This was accomplished by setting any two channels to wait for the same clock. When the clock is started, the two channels begin at exactly the same time. The primary goal of this feature was the easy creation of stimulus tapes for dichotic listening procedures (Cooper & Mattingly, 1969). It also allowed the simultaneous input of two analog channels, e.g., speech and

laryngograph. Further, an output and input channel could also be synchronized, allowing for such features as resampling a file with different characteristics (such as sampling rate).

Sometimes, it is convenient to have an arbitrary audio signal which marks the passage of a certain portion of the main signal. An example is the use of a tone to start a clock for a timed response from a subject. While this could be accomplished by having a second, synchronized channel outputting such a "mark tone," the lack of variation in the signal allows for a simpler solution, and one which would allow marktones to accompany two-channel output. Each output channel is thus associated with a marktone channel, which allows the output of an unvarying audio signal (a 1 kHz tone, in this case) without any increase in processing load. Whenever a sample is output which has the second highest bit set, a 1 kHz tone 4 ms in duration is simultaneously output on the marktone channel. This tone can be recorded along with the main signal, allowing (for example) the synchronization of the main signal with other devices, such as a reaction timer. Since the marktone is essentially part of the data stream, it does not impose any further load on the system: The second highest bit is part of the 16-bit word that is stored in the computer, but not part of the 12 bits of data. Thus, marktones can be freely intermixed with either or both channels of synchronized output.

While the PCM system just described is still in use at Haskins Laboratories, it is no longer the only system in use there. Input and output (A/D and D/A) boards from Data Translation, Inc., have been added to several VAXstations (from Digital Equipment Corp., or DEC) and made compatible with the file and data formats from the older system. Such features as the file format, the synchronization of channels, and the characteristics of the filters have been maintained. So while the convenient features of the old system can be included in the new systems, these systems, unlike the original, can be duplicated at other laboratories.

Computer Environments

The main Haskins PCM system, with its four output channels and two input channels, consists of a PDP 11/04 (Digital Equipment Corp.) which shares disks with a VAX 11/780 (DEC) via a Local Area VAX Cluster. These disks contain the computer files which store the digitized samples of the PCM system. The VAX and the 11/04 communicate via two 16-bit parallel programmable I/O interfaces. Control parameters, such as disk addresses and start or stop signals, are passed from the VAX to the 11/04, and status words are passed back to the VAX. When input or output is being performed, the 11/04 has priority on the disks, allowing it the best chance of completing its time-sensitive tasks. For both input and output, the disk files must be contiguous, rather than being spread across several segments as an ordinary file would be. If the file were not contiguous, computing an address for a file extension and repositioning the heads would often take longer than the amount of time used to output the data obtained on the previous disk access.

The newer systems use Data Translation A/D and D/A boards installed in MicroVAXes or VAXstations. In contrast with the older system, the PCM data must pass through the main CPU. This requires the process performing the input or output to be set to real-time priority, but does not automatically exclude other jobs from running on the computer. Having only the PCM job, however, reduces the chance that the data cannot be read off the disk within the time allowed. Also unlike the older system, the new systems support only a single user. And though there are two output boards on most of the new systems, they both demand the same CPU resources, so only one signal, or two synchronized signals, can be processed at a time.

Dynamic Range

Dynamic range is the ratio of the maximum to minimum amplitude difference in the signal which can be accurately represented. Thus, the primary limitation on this is the number of bits of resolution used for representing the data. The Haskins PCM format for data consists of 12 bits of digitization, which can represent 4096 distinct values. These are stored in 2 byte (16 bit) words, with the upper four bits, the ones not used for data, contain output control information (see § 6.2). 16 bit systems are quite common, and form the basis of digital audio systems. 8 bit systems, which can represent 256 distinct values, are used in many personal computers, but they do not have adequate resolution for many research purposes. The coding itself is simply a binary representation of the quantized voltage. Most systems, including the Haskins one, avoid having a sign bit by adding a dc offset half as large as the dynamic range. For a 12 bit system, this means that the original representations of -10 V to +10 V as -2048 to 2047 will be stored machine-internally as values ranging from 0 to 4095. (Thus the dynamic range

is, more accurately, -10 V to +9.995 V, since one value of the coding scheme must be used for zero, thus leaving the range one value off center; for the rest of this paper, the value +10 V will be used, even though 9.995 V is meant.) In the Haskins system, each value is represented as a 16-bit number.

With a 12 bit system, the theoretical dynamic range is 72.2 dB. This is calculated from the formula 20 log 2^n, where n is the number of bits in the system. Conveniently, this reduces to 6.0206n. Machine-internal noise effectively reduces this by one bit, yielding a more realistic estimate of 66.2 dB. By contrast, a 16 bit system has a theoretical range of 96.3 dB, and an 8 bit system, 48.2.

When digitizing, the system cannot differentiate between signals which reach the upper or lower quantization limits and those which exceed them and thus fall outside the dynamic range. Any signal which exceeds either of the limits will therefore be truncated to the limiting value, resulting in "peak clipping." While the clipping of a single sample will have relatively benign consequences, many successive peak clipped samples will result in an obnoxious noise and unreliable frequency analysis of the clipped region. The only remedy for peak clipping is avoiding it by re-inputting the signal at a lower level.

Any PCM system has inherent limits on the size of differences in the input voltage that can be represented accurately. Analog values which fall within the range of one bit will be given a single digital value. The divergence from the original signal due to these limits is called "quantization error." Since the voltages of -10 V to +10 V are covered by 12 bits in the Haskins system, the quantization error is 4.88 mV (or 0.0244%) for signals using the entire dynamic range. For low amplitude sounds using less of the dynamic range, the quantization error will be larger, in terms of percent.

Timing Resolution

The frequency at which the system examines the analog signal and codes it into a digital number is the "sampling rate." This rate imposes a limit on the frequencies within the original signal which can be accurately represented. If there is an input signal which has a frequency higher than half of the sampling rate, its samples will be indistinguishable from those of a lower frequency signal. This shift in apparent frequency is called "aliasing," and the frequency above which the effect occurs is called the Nyquist frequency (see § 6.1).

The sampling rate also imposes limits on the accuracy of frequency measurements for some aspects of the speech signal—formants and, most noticeably, the fundamental frequency (F0). For a file sampled at 10 kHz, an F0 of 100 Hz will be limited in accuracy to +/- .5%. While this is usually quite acceptable, there are times when greater accuracy is desirable. For higher F0's, however, the error due to temporal quantization is much larger. For a typical female F0 of 200 Hz, the accuracy is +/- 1%, and for a high (but not exceptional) child's F0 of 500 Hz, it goes to +/- 2.5%. All these figures can be cut in half for files sampled at 20 kHz, but even +/- 1.3% is variable enough to obscure some effects. The most clear-cut instance in which these differences become important is in the measurement of vocal jitter (e.g. Baken, (1987), pp. 166-188). That is, the difference in F0 between adjacent pitch periods. Here, the differences add up, because a half-sample excluded from one period will be added into the next, increasing apparent jitter, when there may in fact be none. The cost of higher accuracy, in this case, is the larger storage space required. Doubling the sampling rate doubles the amount of disk storage needed.

Another timing relationship is that between two channels which are started at the same time. For synchronized channels in the Haskins system, whether on input or output, the time difference between the two channels is nonexistent. Both channels read the same clock, and thus they both start at exactly the time that the clock starts. When digitizing, there is a minuscule amount of *amplitude* decay for the second channel, since the signals will be read off the sample-and-hold circuits after the 20 microseconds it takes for the first channel to perform its coding. However, since the decay for these circuits is measured in seconds, and the coding occurs at a delay which is considerably less than half of the sampling rate, the reduction in amplitude is truly negligible. The important fact is that the two channels are triggered at exactly the same time, rather than half a sample apart.

The absolute simultaneity of the two channels has been preserved in our more recent systems based on commercially available boards. The input and output boards from Data Translation, Inc. have two channels available on each, but our system ignores the second channel and uses a second board instead. One consequence is that the two channels are completely simultaneous rather than slightly offset, as they are when the two channels of one board are used. A more practical

consequence is that the samples from the two files do not have to be interleaved as they are read into memory. This saves a considerable amount of overhead for the system, allowing a much more flexible approach to the capture and presentation of simultaneous signals. Files of any length can be played together in any combination with no more processing time than for a single file.

Filter Characteristics

Every analog signal that is to be digitized, and every conversion of a digital signal into an analog one, benefits from the use of filters. Unfiltered digital output can produce severe "digitization noise," due to the sharp edges of the pulses that are produced by the digital samples. On input, frequencies which cannot be accurately represented must be filtered out so that they do not contaminate the signal with aliased sounds (see the end of the previous section). (Even if we are not interested in the nature of the signals above the Nyquist frequency, they must be filtered out to avoid contaminating the spectral content below the the Nyquist frequency.) Since the limit is called the Nyquist frequency, the filters are called Nyquist filters.

A more specialized filter, which aids in the representation and analysis of high frequency sounds, is the high frequency pre-emphasis filter.

In creating a PCM file, the combination of filters to be used is specified in the program, and that combination is stored in the header of the new file. For outputting a PCM file, the program determines the appropriate filters based on information in the file header. Once these are selected, they cannot be changed. Resetting the filters usually results in an audible click, which would be unacceptable in the midst of an output.

Nyquist Filters

The filters that Haskins systems use to eliminate frequencies above the Nyquist frequency are hardware filters with the response shown in Figure 1. Components below 4.8 kHz (or 9.6 for the 20 kHz system) emerge with only minor reduction in amplitude, while those above are are severely attenuated. At 5 kHz (or 10 kHz), the attenuation is at a maximum, approximately 50 dB. Most filters are described in terms of the number of db per octave that the attenuation attains. Since the attenuation here is accomplished in much less than an octave, it is misleading to describe this cutoff in a db/octave formula. Stated in those terms, these filters have a 1200 db/octave attenuation, which is over 16 times larger than the entire dynamic range. Since it is theoretically impossible to attenuate a signal more than the dynamic range allows, this number is impossibly large. Instead, the filters should be described as sharply tuned and reaching the 3 db attenuation level at 4.8 (or 9.6) kHz. In any event, the sounds above the Nyquist frequency have virtually no chance of affecting the signal any more than the background noise does.

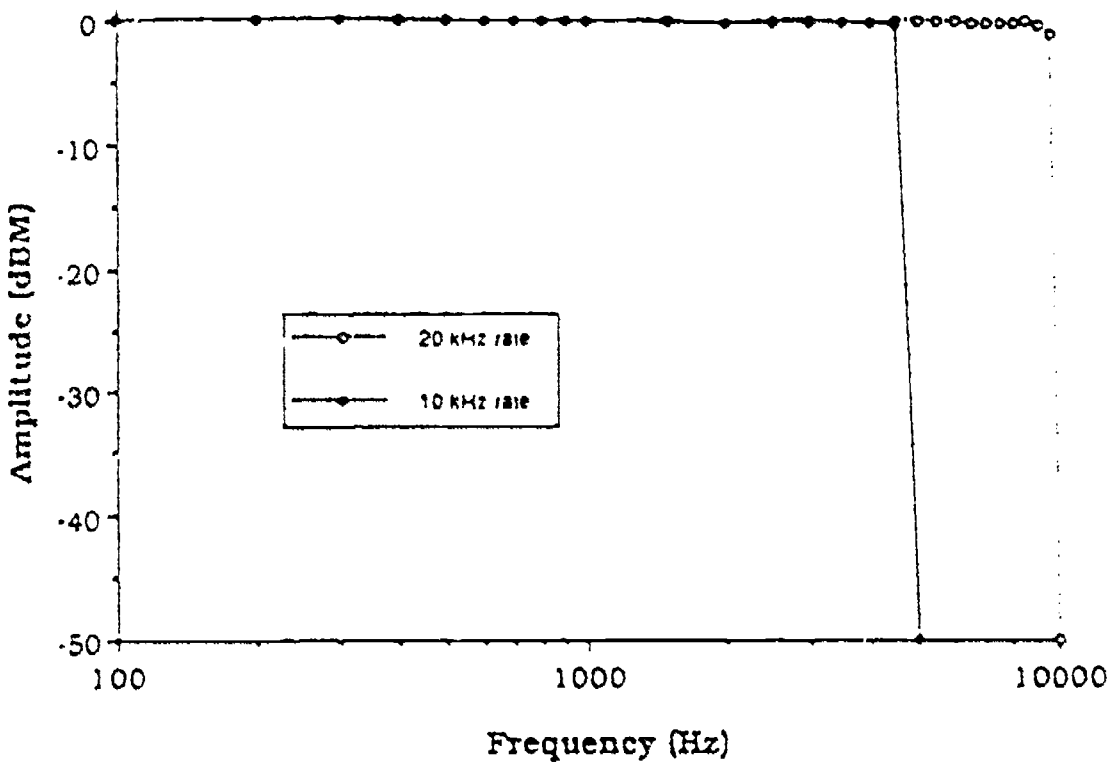


Figure 1. **Resultant amplitude of 0 dBM test signals of differing frequencies after passing through the Nyquist filter. Measurements shown are for one system, but similar results obtain for other Haskins systems.**

High Frequency Pre-emphasis Filters

For signals such as speech which are primarily driven by low frequency sources, the high frequency components generally have lower amplitude than the low frequency ones. Of course, high frequency signals of a given amplitude, being more intense, will sound louder than low frequency signals of the same amplitude, so that in a sense the high frequency signals are more perceptually salient than their amplitude would suggest. Nonetheless, early researchers found that the high frequencies, especially of speech, were difficult to measure or even detect when input at their natural level. In order to rectify this situation, a hardware filter was selected which could boost the high frequencies (before digitization) by a reliable and known amount. A complementary filter could then reduce their amplitudes by the same amount when the digitized signal was played out. There is a slight gain in accuracy of the digitization, since the quantization error will be a smaller proportion for a signal which uses more of the dynamic range. For the /f/ noise to be examined in Figure 3, for example, the quantization error is about 0.488% for for the non-pre-emphasized signal while it is about 0.029% for the pre-emphasized signal. Although this difference is sizable, the improvement in quality may not be very noticeable to the naked ear [though see Whalen (1984) for a demonstration of perceptual effects of differences that are not consciously detectable].

Figure 2 shows the pre-emphasis function used with the 20 kHz sampling rate. The response is fairly linear up to 1 kHz, then rises exponentially, shown as a straight line in Figure 2, where frequency is represented in a log scale. On output, a filter with exactly the reverse characteristics is used. Thus if the amplitude value is read as a decrement, this figure can be used to represent the de-emphasis filter as well. The same filter is actually used for the 10 kHz rate, but since the Nyquist filter (which in this case functions as an anti-digitization noise or "anti-imaging" filter) follows it, there will be nothing left above 5 kHz.

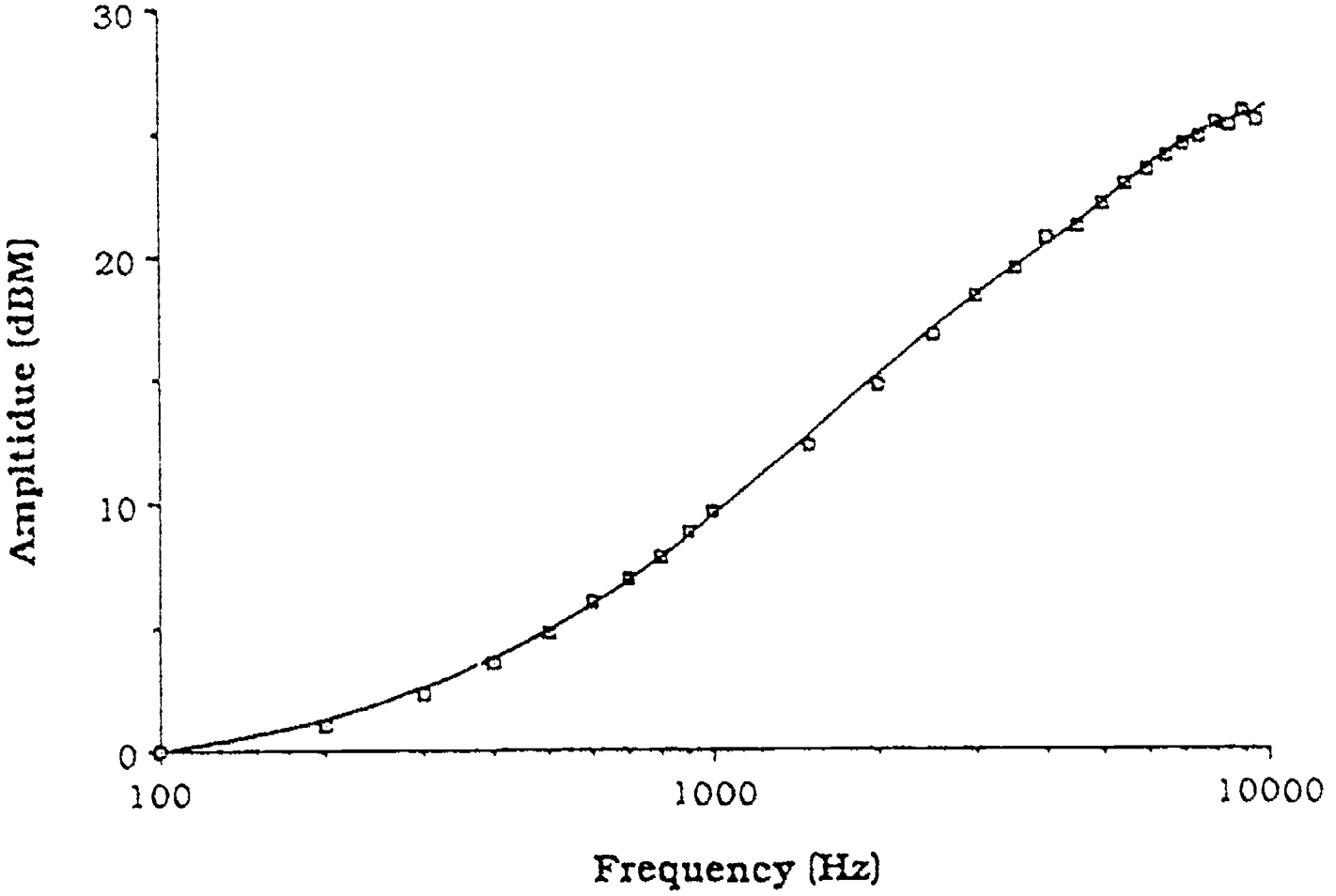


Figure 2. Resultant amplitude of 0 dBM test signals of differing frequencies after passing through the high frequency pre-emphasis and Nyquist filters. Symbols represent measurements for one system, and the line is a fitted polynomial. Because of the Nyquist filter, the output level drops steeply at 10 kHz (not shown).

Ideally, the pre-emphasis filter should equalize the long term speech spectrum so that the maximum use of the dynamic range is achieved for each frequency region. Clearly, no one filter shape can serve this function, since different speakers, and even the same speaker at different times, will generate different long-term spectra. The shape of the pre-emphasis function is a compromise based on the sorts of long term spectra encountered in the early research. The function is not based on properties of the human auditory system, though it bears a superficial resemblance to the ear's increase in sensitivity between 1500 and 4500 Hz (e.g., Robinson & Dadson, 1956). There is also some resemblance to the historically later Dolby noise reduction systems. Although Dolby systems have become standard in the recording industry, there are good reasons not to use them as part of a PCM system. While the Dolby system greatly increases the separation of low intensity, high frequency signals from the noise encountered on playback from audio tape, it would be inappropriate to use it as a front-end to a digitizer, since digitized signals are not subject to media noise. (Even for signals which are simply recorded on audio tape for later digitization with a PCM system, Dolby noise reduction may be inappropriate. The net effects of the Dolby filters may be benign in terms of intelligibility, but finer acoustic measurements, e.g., the bandwidths of formants which happen to lie at the edge of one of the four Dolby bands, may be affected. In addition, having the tape noise at a constant level makes it easier to take into account when comparing the amplitude of speech sounds. Reducing the tape noise for high frequency sounds would reduce their amplitude compared with low frequency sounds which included the noise.) Similarly, there are digital techniques such as first-differencing which can have similar effects without requiring the hardware filters. However, such digital filters are neither sharp enough nor linear enough for many of the measurements that are made in the speech field. So, for consistency and reproducibility, the hardware filter approach has the most benefits. This system does have the drawback that the PCM representation of these signals cannot be played back faithfully on other systems unless the other systems have the same filter. (They can be played back without the de-emphasis filter, and the speech is usually quite recognizable, just distorted by the additional amplitude in the high frequencies.) For many purposes, such representations are adequate.

Figure 3 shows the effect of this pre-emphasis filtering system. In the top panel is the waveform of the word "fast," with the high frequencies pre-emphasized. The characteristically weak /f/ fricative noise is easy to discern in the first 100 ms. In the bottom panel, exactly the same signal (input synchronously on the second input channel) is shown in its non-pre-emphasized version. The onset of the /f/ noise is very difficult to discern at this level of resolution. The middle panel of Figure 3 shows the result of magnifying the display of the bottom panel by a factor of three. The shape of the fricative noise is now somewhat clearer, though the gradualness of the beginning of the noise is still somewhat hard to make out, but the vocalic segment (/æ/) is now (visually) peak-clipped. Along with the fricative noise, the low-frequency, dc air flow noise can also be seen. Such information is useful for recognizing less than optimal recordings, but it is not part of the speech signal. With pre-emphasis, the shape of both the fricative noise and the vocalic segment are evident, and there is no need to use separate magnifications to make them so. While the /f/ noise could have tolerated much greater pre-emphasis, the /s/ noise (around 375-450 ms), which also contains high frequencies, could not.

Pre-emphasis is not without its cost in other regards, however. Although the frequency analysis of the high frequencies is more accurate, the amplitude values of those frequencies relative to low frequency components are inflated. While the amount of change is predictable, it is not terribly convenient for humans looking at the display to calculate. When many comparisons of, say, the amplitude of F4 to that of F1 are to be made, pre-emphasis is definitely a drawback. If F5 is in question, however, it may be that the structure of the formant itself is not discernible without the pre-emphasis, so that the translation of the amplitude is a necessary evil. Such comparisons are relatively rare, however, and most researchers take advantage of the greater resolution in pre-emphasized digitization.

One other cost deserves mention, since it has already caused a certain amount of confusion in the literature (Fowler, Whalen, & Cooper, 1988; Howell, 1988; Tuller & Fowler, 1981). In that work, the amplitude of various speech signals was equated without the complete destruction of the speech information by a technique called infinite-peak-clipping (Licklider & Pollack, 1948). For each sample of the signal, positive values are amplified to the maximum level and negative values to the minimum.

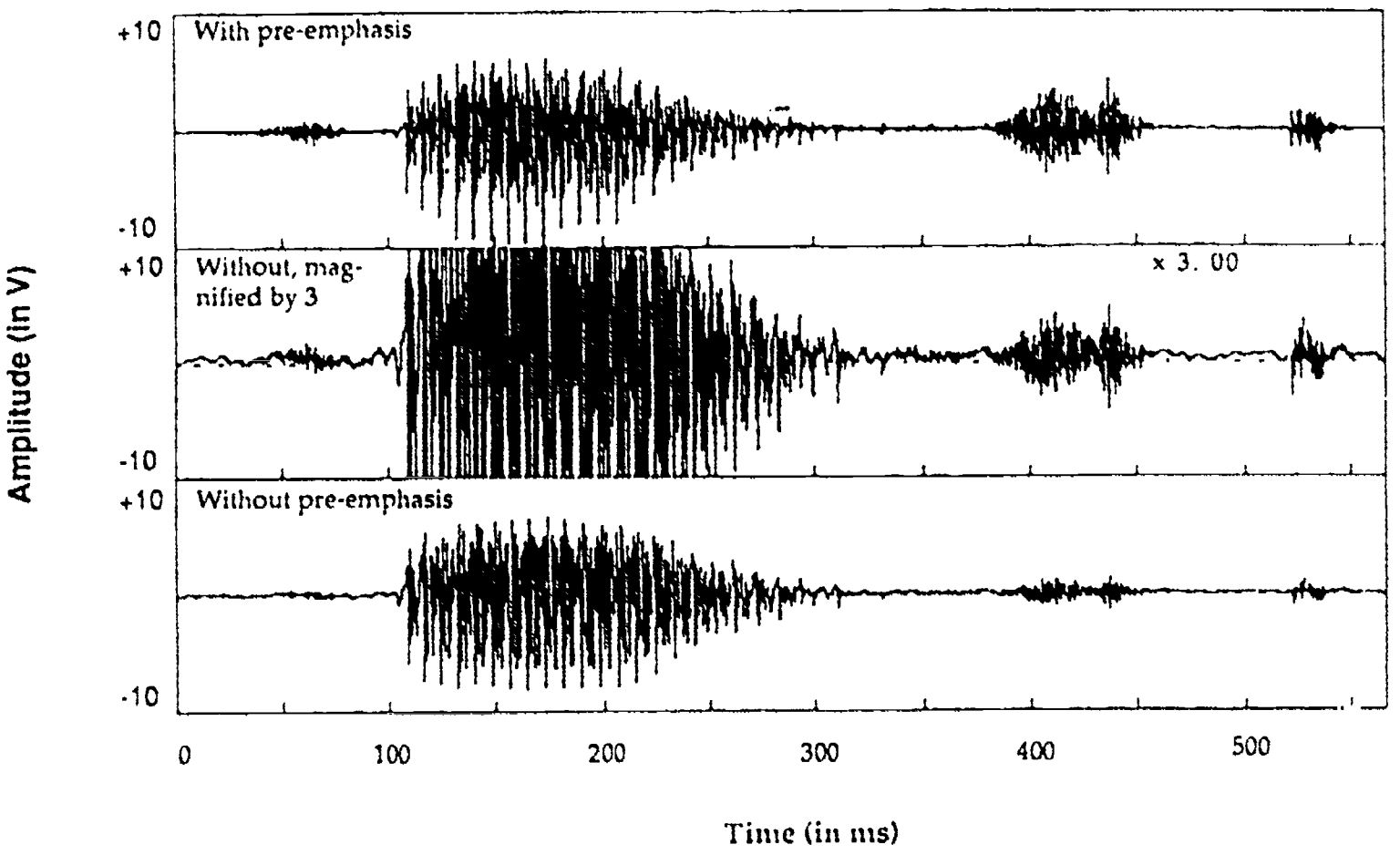


Figure 3. **Waveforms of the word "fast" under two sampling and two display conditions. The top and bottom panels represent the syllable with and without pre-emphasis, respectively, at original amplitude. The middle panel is the non-pre-emphasized signal magnified by a factor of 3.**

The result is an irritatingly noisy, though usually recognizable, utterance. If the original file was pre-emphasized, however, it would normally go through the de-emphasis filter. When output through the de-emphasis filter, the high frequencies are lowered in amplitude, so that signals with different frequency components would once again have different amplitudes, despite the infinite-peak-clipping. If the de-emphasis filter is avoided (which can be done by changing the PCM file header), the intended result is obtained even for pre-emphasized files. (The pre-emphasis filter rarely changes the sign of a sample, though it can happen when an intense high frequency sound occurs with a simple low frequency sound.)

Another technique, which results in a sound called "signal-correlated noise" (Schroeder, 1968), interacts with the pre-emphasis function. Signal correlated noise retains the amplitude contour of the source sound but has a flat spectrum. The samples of approximately half the digitized source have their signs changed at random while the magnitude remains the same. The overall energy remains the same, since the same amount of deviation from the baseline is present. But, since the direction the wave takes is randomly related to its original direction, the spectrum of the signal is flat. For a pre-emphasized original signal, however, the spectrum of the signal-correlated-noise is flat only machine-internally. If the noise passes through the de-emphasis filter, the high frequencies will fall off by the amount specified in Figure 2. This does not restore any of the spectral structure of the original, but the spectrum is not perfectly flat either. Avoiding the de-emphasis filter will not salvage the noise, since that would maintain the flat spectrum but change the amplitude contour. For sounds which are going to have signal correlated noise stimuli created from them, a non-pre-emphasized original is preferable. Alternatively, a brief description of the deviation from a flat spectrum (the high frequency roll-off) is necessary.

Haskins PCM File Formats

The information in this section is quite detailed, and will be of interest primarily to users of the Haskins system. The kinds of information included, though, may be of interest to users of

other PCM systems. The format of digitized files takes advantage of the special features of the Haskins PCM hardware (such as marktones) and of in-house programs (such as the labels of the waveform editor WENDY). For third party software, modifications are required. For example, the ILS package of Signal Technology Inc. is a large set of programs for doing signal analysis. By default, these programs expect a header format in PCM files that is contains some of the same information as Haskins headers but puts them in different locations. The input and output routines have been changed so that ILS can put its information at an otherwise unused part of the header, leaving the rest in the Haskins format. Another alternative that is employed by some newer Haskins programs is to translate from one header format to the other, and create two versions of a file if needed.

These features will be discussed in the order in which they appear in the computer file. The first component of the file is a header block of 512 bytes, which contains information about the characteristics of the data. The next is the data itself, taking up as many 512 byte blocks as are needed to accommodate the number of samples in the file. The final, optional portion is a section of up to four trailer blocks containing labels of locations within the file. (This label format is in the process of being superseded by separate label files.)

The conventions presented here are not intended as a standard (cf. Mertus, 1989), since there are many concerns which are not adequately addressed by this format. Just to give one example, there is currently a word in the header to indicate the number of bits of resolution (always 12 for current Haskins systems), but this format may not be optimal for a more broadly defined standard. The present discussion is intended to make the information more accessible for those laboratories which do use the format already, and to bring the Haskins conventions to the attention of those devising their own systems.

PCM Headers

The initial portion of each PCM file consists of a "Header" which contains attributes of the sampled data within the file. For some files, especially those from the Haskins Physiological Speech Processing (PSP) system, the header also establishes a correspondence between time and sample position within the file. The first file block of the PCM file (512 bytes on DEC systems) is used, though for speech files much of it is simply zero-filled. Physiological files contain more information (see below).

PCM Data

The PCM data begin in the first block immediately following the header block. Samples are stored as fixed length 128 byte records of 64 words, and are usually input into contiguous files, though the files do not have to be contiguous for analysis programs which do not do real-time output. To output a section of a sampled data file with the older system, it must be contiguous. The newer systems can read noncontiguous files into memory sufficiently fast to keep the real-time output going.

One 12-bit sample is stored in the low order bits of each 16-bit word. This 12-bit sample represents a bipolar analog voltage that ranges from endpoints set near -10 and 10 volts. The four high order bits in each 16-bit word form a control field that is utilized by the audio output system. When samples are read for analysis within the computer, this control field must be cleared before subtracting the midline. That is, if one of the control bits is set, it will appear to the general computer as a legitimate part of a number, even though it would be far outside the dynamic range. Normally, these bits should also be cleared when samples are written out to a PCM file. Programs that generate speech files must truncate the samples to avoid overflow into the control field.

The following is the format of the data word:

bit position	description
1 - 12	data field
13	if set, data field is an inter-stimulus-interval value
14	if set, something is wrong
15	if set, a mark tone will be generated a that sample
16	if set, something is wrong

To conform with the conventions used by the A/D and D/A converters at Haskins Laboratories, the signal voltage levels are encoded digitally in excess-2048 form, that is:

-10 volts is encoded to 0
0 volts is encoded to 2048
10 volts is encoded to 4095

Thus, a 16-bit bipolar digital value that ranges from -2048 to 2047 can be obtained by subtracting 2048 from the 12-bit encoded sample value.

Haskins PCM Labels

Labels are used to record the position, and optionally the range, of user-defined portions of the PCM file. Each label consists of a string of alphanumerics (beginning, by convention, with a letter) which is a file-unique name for the label; a location, given in milliseconds from the beginning of the file; a left range and a right range, which can be set in terms of milliseconds in relation to the label; and a code to determine whether there is a mark tone or not.

The length of a single label is 32 bytes. The older style maximum number of labels was 64. (In the older style of programs, labels were stored in trailer block(s) of the PCM file immediately following the data blocks within the file.) If there are old-style labels stored in the file, the number is contained in a field in the header block (word 7). Many of our own programs currently change automatically from old to new style any time a PCM file is accessed.

The old format for labels in a Haskins PCM file:

byte position	length of field	description
1	4	label left range
5	4	label right range
9	4	label location (time value of label)
13	1	label mark tone flag
14	19	name of label

The unit for time representation is one 20,000th of a second, and the scope of a label is defined to extend from its time value minus its left range, to its time value plus its right range.

The new format consists of separate ASCII files containing label information coded by keywords, of which many are common but some are specific to one program. This allows for greater flexibility in the number of labels that can be maintained, convenient correction or even creation of labels with a text editor, and compact sharing of labels across several related files (such as physiological measurements of one event which might end up in a dozen different files). The implementation of this system is in progress, and eventually it will be the only one used by Haskins programs.

Summary

The Haskins PCM system is a combination of standard techniques and unique features. Copies have been built with custom-made hardware and, more recently, with commercially available boards. Some salient features are: convenient input and output of signals of any length (dependent on the system's disk rather than on the PCM system constraints); exactly simultaneous synchronization of two channels (either two output, two input, or an input and an output) without the need for interleaving the samples; consistent pre-emphasis of high frequencies for easier analysis, and converse de-emphasis for accurate reproduction; the capability of having any number of marktones associated with a file without any added load on the system. This system has been used in generating the data for dozens of papers over the last twenty years, and will continue to be used both at Haskins Laboratories itself and at the growing number of laboratories which are using the system.

REFERENCES

Baken, R. J. (1987). *Clinical measurement of speech and voice*. College-Hill: Boston.

Cooper, F. S., & Mattingly, I. G. (1969). A computer-controlled PCM system for the investigation of dichotic speech perception. *Journal of the Acoustical Society of America, 46*, S115 (A).

Fowler, C. A., Whalen, D. H., & Cooper, A. M. (1988). Perceived timing is produced timing: A reply to Howell. *Perception & Psychophysics, 43*, 94-98.

Goodall, W. M. (1947). Telephony by pulse code modulation. *The Bell System Technical Journal, 26*, 395-409.

Heute, U. (1988). Medium-rate speech coding—trial of a review. *Speech Communication, 7*, 125-149.

Howell, P. (1988). Prediction of the P-center location from the distribution of energy in the amplitude envelope: I. *Perception & Psychophysics, 43*, 90-93.

Licklider, J. C. R., & Pollack, I. (1948). Effects of differentiation, integration and infinite peak clipping upon the intelligibility of speech. *Journal of the Acoustical Society of America, 25*, 375-388.

Mertus, J. (1989). Standards for PCM files. *Behavior Research Methods, Instruments, & Computers, 21*, 126-129

Robinson, D. W., & Dadson, R. S. (1956). A redetermination of the equal-loudness relations for pure tones. *British Journal of Applied Physics, 7*, 166-181.

Schroeder, M. R. (1968). Reference signal for signal quality studies. *Journal of the Acoustical Society of America, 44*, 1735-1736.

Tuller, B., & Fowler, C. A. (1981). The contribution of amplitude to the perception of isochrony. *Haskins Laboratories Status Report on Speech Research, SR-65*, 245-250.

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

APPENDIX

Information Stored in the Haskins PCM File Headers.

There are seven main header entries which occupy the first eight words of the header block. They are:

start pos.	# of words	description
1	1	DATA TYPE INDICATOR, a 1 in this field indicates a sampled data format file that will be recognized as such by Haskins software.
2	2	SAMPLED DATA SIZE, double precision integer representation of the size of the file (number of samples). The first word is the low order part of the count.
4	1	SAMPLING RATE, expressed as samples taken per second.
5	1	ATTRIBUTES, format of word: - bit 0 is the pre-emphasis flag, if 0, the data were pre-emphasized during sampling (the level of higher frequencies were boosted) and should be de-emphasized when output; if 1, the data were not pre-emphasized - bit 1 is the filtering flag if 0, the data were filtered during sampling at the Nyquist frequency. if 1, the data were not Nyquist filtered the remainder of the word (14 bits) are unused.
6	1	NUMBER OF ADDITIONAL HEADER BLOCKS. No longer implemented.
7	1	NUMBER OF LABELS, if greater than zero, then the file contains labels that are stored in the trailer blocks of the file, each label is of 32 byte length. The remaining 249 words of the header block code the following:

start pos.	# of words	description
8	1	Revision level (indicates which version of the arrangement of information in the header is used).
9	2	Virtual block number of first trailer block (where old style labels are kept.)
11	1	Number of trailer blocks (for old style labels).
12	1	Data source (currently, either VAX (1) or unknown (0))
13	1	Number of bits of resolution (only 12 is implemented)
14	1	Source (no longer implemented)
15	50	Filler words PSP (Physiological Signal Processing) information
65	1	Datel hardware input mode: (0 = EMG data, which is already filtered and integrated; 1 = speech, which must be at 10 kHz to be synchronized with physiological measurements; 2 = LED (usually movement) data, in which the x and y values each take up a channel; 3 = Electropalatagraph data, where each word represents the on/off state of the 63 contact points in the false palate.)
66	5	Filler words
71	1	PSP header version number
72	1	Samples per frame
73	1	Channel map (a sixteen bit word which serves as a bitmap representation of which of the sixteen possible input channels are actually being used)
74	1	Data file record size

75	2	M calibration constant: Together with the B constant, this allows the machine units in the file to be interpreted as physical units. The physical value = M*(sample value) + B. So M is a scaling factor and B is an offset.
77	2	B calibration constant
79	6	Calibration units (12 characters): A description of the units that result from the application of the calibration constants (e.g. "millimeters").
85	16	Index file name (32 characters): Name of a file which contains a catalog of the number of samples associated with each octal code (a time marker on the analog tape) for all the other PCM files which were created in the same input pass as this one. This information allows for the compensation for minor speed changes in the analog tape system.
101	2	Smoothing constant: If the file was smoothed (as is usual for EMG signals), this is the size (in milliseconds) of the base of the triangular averaging filter.
103	2	Line up point: location of an event chosen by the experimenter to coordinate the displays across PCM files. If PSP header version number = 0, then the line up point is in samples, if PSP header version number 0 then the line up point is in 1/20,000th of a second.
105	2	Graphics scaling - Y min
107	2	Graphics scaling - Y max
109	20	Filler words
129	128	Filler words

Note that the set of filler words of the header may contain the ILS header information if the file has been analyzed with the Haskins-modified version of ILS.

Haskins Laboratories Status Report on Speech Research
1990, SR-103/104, 137-152

Factors Contributing to Performance on Phoneme Awareness Tasks in School-aged Children*

Anne E. Fowler†

To examine factors potentially responsible for the robust association between phoneme awareness and reading ability, a novel pair of tasks was designed a) to control for nonlinguistic task variables and metacognitive skill; b) to minimize demands on working memory and verbal production; and c) to assess the role of reading experience and spelling strategies. With these factors taken into account, phoneme awareness remained significantly and specifically associated with decoding ability in children aged 71/2 to 10 years. Results on the new measure also corresponded to performance on an existing, widely used, measure of phonological awareness. In contrast, scores on the task selected as a closely parallel nonverbal analogue was unrelated to reading or to phonological awareness. These results, including comparisons of good and poor readers matched on reading level, but differing in age, suggest that the ability to isolate and identify phonemes continues to be an important determinant of reading aptitude during the school years.

Two decades of research has placed phonological awareness, and specifically phoneme awareness, squarely at the center of our efforts to understand the sources of reading disability. Children who have difficulty in reading also have difficulty in analyzing spoken language into the phoneme units to which letters roughly refer, or in performing operations upon these elements. Performance on phoneme analysis tasks involving phoneme counting, deletion, reversal or substitution has been shown to account for as much as 40 to 70% of the variance between skilled and unskilled readers (e.g., Mann, 1984; Mann & Liberman, 1984; Pratt & Brady, 1988; Rosner & Simon, 1971; Stanovich, Cunningham, & Feeman, 1984; Tunmer & Nesdale, 1985). However, recent research studies and the complex structure of phoneme awareness tasks leave some ambiguity about the locus of difficulty. Although training studies suggest that gaining access to the segmental nature of speech is important in first learning to read (e.g., Ball & Blachman, 1988; Bradley & Bryant, 1983; Lundberg, Frost & Peterson, 1988), a number of other factors common to phoneme awareness tasks and reading may contribute to the strong and continued association once reading instruction has begun.

In disentangling some of the many factors contributing to poor performance on phoneme awareness tasks, three areas merit close attention. First, it has been suggested that failure on phoneme awareness tasks may stem from a more general difficulty at the *metacognitive* level (e.g., Tunmer, 1988) that should be evident in nonlinguistic tasks as well. Second, it has been noted that those phoneme awareness tasks that correlate most strongly with reading beyond Grade 2 make heavy demands on the phonological

The data reported here were originally presented at the Orton Dyslexia Society National Meeting, Tampa, November 1988. I am indebted to Susan Brady, Virginia Mann, Donald Shankweiler, and most especially, Isabelle Liberman, for their assistance in the design of this study and for comments on an earlier draft of the paper. I am grateful as well for the work of Susan M. LaBrecque, who created the picture stimuli and collected and coded the data; and Michael Escobar, who performed preliminary data analyses. Special thanks are due to the many teachers and children of the second, third and fourth grades in the East Haven Public Schools and St. Francis and St. Bernadette Parochial Schools in New Haven for their patience and interest. Funding for the research was provided by NICHHD Research Program Award #5PO1 HD21888. This work was completed during the author's tenure as a Science Scholar with the National Down Syndrome Society.

processor, often requiring the child to hold a phonological string in *working memory*, to manipulate this string, and to produce a verbal response. It may be argued that these demands extend well beyond the simple awareness of the segmental structure of speech, making it difficult to determine whether poor performance stems from a lack of awareness or from difficulties with more basic aspects of phonological processing (Brady, in press; Yopp, 1988). Third, and perhaps most significantly, it has been argued that successful performance on phoneme awareness tasks may result from rather than predict success in acquiring written language. Many tasks taken as evidence of the child's access to phonology could involve *spelling strategies,* in which the child counts or manipulates letters rather than sounds. If spelling knowledge could lead to a correct response, then it is difficult to rule out the possibility that spelling leads and shapes the representations being tapped in phonological awareness tasks, rather than vice versa (Ehri, 1989; Morais, Cary, Alegria, & Bertelson 1979; Treiman, 1985).

In the present study, each of these concerns was given full consideration. To tease apart the contribution of each factor to poor performance on phoneme awareness measures, a need was recognized for a new, more analytic measure, with a parallel nonverbal control. When the variables discussed above have been taken into account, does awareness of the phonemic structure of words continue to be significantly and specifically associated with reading ability in school-aged children? Before describing the present study, designed to address this question, the concerns about each variable will be more fully discussed.

Metacognitive factors in phonological awareness. The original hypothesis generating an interest in phonological awareness focussed on the unnatural demands of reading (Gleitman & Rozin, 1977; Liberman, 1971; Rozin, 1975). Reading, in contrast to speaking or listening, requires the learner to become consciously aware of abstract phonological units embedded in a perceptually continuous speech stream. Although phoneme units may be available implicitly to guide the universal, early, and untutored development of speaking and listening, explicit awareness is available only to the cognitively mature child and then usually only in the context of direct instruction (Liberman, 1989). One hypothesis about the achievement of phoneme awareness is that it should be associated with metacognitive awareness in other tasks.

Such a possibility seems plausible when one considers the rather formidable cognitive requirements which characterize individual phonological awareness tasks. Almost any phonological awareness measure requires the child to isolate, identify, count, and order the abstract elements of the language. In more complex measures strongly associated with reading, children have been asked to select which of three choices does *not* match a target item in regard to the final segment (e.g., Bradley & Bryant, 1983); or to relate colored blocks to abstract phonological elements and to track transpositions of phonemic segments with exchanges of the appropriate blocks (Lindamood & Lindamood, 1971).

A small number of studies have directly addressed the possibility that the difficulty in performing metalinguistic tasks may derive from a more general failure to understand and cope with the nonlinguistic task requirements. Cognitive factors that have been investigated include following instructions, counting, sequencing, and isolating smaller units. As a visual analog to typical phonological awareness tasks, Lundberg (1978; Lundberg, Olofsson, & Wall, 1980) asked children to locate simple shapes within more complex meaningful pictures. Similarly, Mann (1986) asked children to count angles in a picture just as they counted phoneme segments. As auditory controls, Pratt and Brady (1988) and Morais, Bertelson, Cary, and Alegria (1986) developed xylophone tapping tasks to parallel some of the requirements of a phoneme segmentation test (Lindamood & Lindamood, 1971) and a phoneme deletion test (Rosner & Simon, 1971), respectively. The finding that none of these roughly comparable nonlinguistic measures is associated with reading minimizes the likelihood that the extraneous task demands assessed are critical factors in accounting for reading group differences in these studies.

On the other hand, several investigators have suggested that children must attain a minimum level in general cognitive development in order to be able to reflect upon and manipulate the structural features of language. In particular, it has been argued that phoneme awareness requires the ability to handle part-whole relations and to shift attention from meaning to form (i.e., to decenter); these achievements characterize the Piagetian stage of concrete operations (onset at 5 to 7 years). The most compelling evidence in support of this position derives from a prediction study in which children who lacked these general metacognitive abilities in kindergarten were

unlikely to achieve phoneme awareness when reading instruction was introduced during the next year (see Tunmer, Herriman, & Nesdale, 1988, for a review).

Similarly, Treiman and Baron (1981) suggested that an insensitivity to the internal structure of syllables (i.e., to phonemes) may be related to a general cognitive disposition of young children to focus on global rather than analytic aspects of stimuli (Smith & Kemler, 1977). However, in a direct test of this hypothesis, Mann, Tobin and Wilson (1988) compared kindergarten children's classification of nonsense syllables with their classification of geometric figures; in each case, a global and an analytic choice was available. They interpreted their failure to find a correlation between these two measures as arguing against the view that a common factor underlies the shift from a holistic to an analytic strategy.

To summarize, the evidence that general cognitive factors contribute importantly to performance on phoneme awareness tasks is suggestive, but limited. Because each new phoneme awareness measure introduces its own requirements, the present goal was to include an appropriate control task to assess and control for the contribution of general metacognitive skill to performance on both phoneme awareness and reading.

The role of working memory in phonological awareness tasks. Seeking to better understand the source of metaphonological problems in poor readers, a number of researchers have focussed their attention on more basic language processes in these individuals. There is growing evidence that the phonological structures underlying all of language processing are less well-developed in the poor reader (Liberman & Shankweiler, 1985; Stanovich, 1982). Among the abilities investigated, the most striking and pervasive characteristic of the poor reader is a difficulty in maintaining verbal material in memory; efficient operation of working memory appears to depend on strengths within the phonological domain (for reviews, see Brady, 1986; Jorm, 1983: and Torgesen, 1985). The poor reader is also less able to decode speech in noise (Brady, Shankweiler, & Mann, 1983; Palley, 1986), to accurately articulate tongue twisters and multisyllabic words (Rapala & Brady, 1990), or to access the phonological representation of words in the lexicon (Katz, 1986; Wolf & Goodglass, 1986). Although it remains to be determined whether one single deficit underlies all of these difficulties, and what effect reading experience has on these tasks, there is no question that in at least some poor readers, the phonological difficulties extend beyond awareness (Brady, 1986; Wagner & Torgesen, 1987).

In light of these more extensive phonological deficits, it is somewhat problematic that those phonological awareness tasks which correlate most strongly with reading beyond Grade 2 require considerable phoneme manipulation and therefore a heavy memory component (see Yopp, 1988). Such tasks also frequently involve lexical access and production of a verbal response as well. For example, the Auditory Analysis Test (AAT, Rosner & Simon, 1971), which is a strong predictor of reading success from kindergarten to adulthood, requires the child to apply all of these skills. In the AAT (e.g., "can you say smile without the /s/"), the child must repeat an incoming word, hold it in memory, isolate and remove the designated element, and reconstruct and accurately produce the new phonological sequence (also a word) without this segment (see Yopp, 1988, for a discussion). While a confounding of these factors may approximate the actual demands of reading, they obscure understanding how the various phonological deficits of reading are related to each other (Wagner & Torgesen, 1987).

Evidence that stressing working memory may mask the true abilities of children with reading disability derives from work on syntax. Although poor readers often lag behind good readers when asked to complete ungrammatical sentences, to explain ambiguities, or to detect, correct, or explain semantically and/or syntactically anomalous sentences (e.g., Bowey, 1986; Ryan & Ledger, 1984; Siegel & Ryan, 1984), it has been proposed that poor these limitations may derive from inadequate phonological processing (Crain & Shankweiler, 1988; Shankweiler & Crain, 1986). According to this view, although knowledge of syntactic structures may be intact, limitations in phonological processing constrict the operation of working memory, leading to poor performance on sentence-level tasks.

In a direct test of this possibility, Fowler (1988) asked children in regular classes to make judgments of grammaticality and to correct violations of grammaticality on the same set of sentences; memory and syntactic factors were systematically manipulated. Each task was metacognitive inasmuch as it required children to focus on form rather than content (Galambos & Hakuta, 1988; Gleitman, Gleitman, & Shipley, 1972), but only the correction task was expected to stress working memory. The results were consistent with this hypothesis. Although performance in both tasks was significantly affected by syntactic variables,

only in the correction task was performance associated with reading ability or affected by memory manipulations. Thus, syntactic knowledge was associated with reading disability only when its expression involved heavy memory demands, verbal response requirements and manipulation of sentential elements. (See Bentin, Deutsch, & Liberman, 1990, for a replication of these results with a similar population).

One must ask at what point a child's difficulties with a phoneme awareness measure may be wholly accounted for by the heavy and multiple phonological processing demands imposed by the task. Inasmuch as one can free phonological analysis tasks from extraneous memory and production requirements, will the reading disabled child continue to display a deficit in phoneme awareness? Again, for the purpose of the present study, the goal was to design a phoneme awareness task that would ask the child to isolate or identify phoneme segments, but that would not require manipulation or a verbal resonse.

Spelling strategies in phoneme awareness. A third interpretative concern relates to the hypothesis that successful performance on phoneme awareness tasks may result from rather than predict success in acquiring written language (Ehri, 1989; Morais, Alegria, & Content, 1987). Although prediction studies finding an association of early phonological awareness and later reading skill render a strong version of this hypothesis unlikely (e.g., Bradley & Bryant, 1983; Mann & Liberman, 1984; Perfetti, Beck, Bell, & Hughes, 1987; Share, Jorm, Maclean, & Matthews, 1984), it is a critical concern when examining the continuing association between reading disability and phonological awareness. Consider, for example, a study finding that adult readers have a considerable advantage over non-readers when asked to "say *pat* backwards" (Byrne & Ledez, 1983). As argued by Tunmer (1989), this task is handled most efficiently by using an orthographic strategy; non-readers without access to letter representations would be at a considerable disadvantage. Similarly, when a child is asked to "say please without the /l/" (Rosner & Simon, 1971), chil en who car read should be more inclined to work with the more tangible units of spelling wherever possible.

That subjects do indeed employ spelling strategies, where available, is demonstrated in several studies by Ehri and her colleagues. For example, when asked to count the number of sounds in *rich* and *pitch*, fourth graders report pitch to have one more segment (Ehri & Wilce, 1980). Similarly, invented spellings of young children indicate that post-vocalic nasals (e.g. *bump, think*) are considered a part of the vowel until spelling conventions indicate otherwise (Ehri, 1984).

Consequently, the advantage of the better readers on many phonological awareness tasks may derive from their ability to read or spell, rather than from an independent skill underlying reading, spelling and phonological awareness. Thus, it is important to have a phonological awareness measure that controls for spelling knowledge.

PURPOSE

The present study was designed to test the hypothesis that phonological awareness continues to be specifically and significantly associated with reading ability beyond the early stages of instruction. A new task was developed to measure phonological awareness which, paired with a nonverbal control task, would assess and control for the contribution of metacognitive ability and spelling strategies, while minimizing memory demands. To determine whether general metacognitive abilities contribute to individual differences in reading and phonological awareness, a phonological analog was created to parallel an existing measure of visual analytic ability, the Embedded Figures Test (EFT, Satz, Taylor, Friel, & Fletcher, 1978).[1] In the Embedded Figures Test, subjects are asked to identify which of three complex designs contains a specified shape. The parallel phonogical measure created for this study, the Embedded Phonemes Test (EPT), required subjects to identify which of three words, indicated both by pictures and oral presentation, contains a specified phoneme. To address further concerns that many tasks ostensibly measuring phoneme awareness make multiple demands on phonological processing, this task was designed to minimize extraneous requirements. No manipulation of segments was involved and the use of pictures reduced memory load and eliminated the need of a verbal response. The final consideration in designing the phoneme awareness task concerned the suggestion that superior performance on phoneme awareness tasks is a result of orthographic strategies, rather than consideration of the phonological representation. To assess this possibility, we compared performance on items in which consideration of the word's spelling would aid in identifying the embedded phoneme with peformance on those items in which knowledge of the word's spelling would not be of assistance.

This pair of measures was given to 48 second to fourth grade children, along with measures of word recognition, nonsense word reading and receptive vocabulary (Peabody Picture Vocabulary Test-Revised, PPVT-R, Dunn & Dunn, 1981). Subjects were required to have standard scores above 80 on the PPVT-R and were selected to vary widely in reading ability. Although all 48 subjects received the same battery of measures, different analyses on distinct subsets of the larger subject pool focussed on separate issues. For purposes of exposition, these different treatments are introduced as separate experiments below.

EXPERIMENT 1

The first experiment focussed on the contribution of phoneme awareness skills and general metacognitive skills to reading ability in ten pairs of school-aged children selected from the sample of 48 such that each pair was matched for age and vocabulary level, but differed on both reading measures. In this comparison, phoneme awareness, but not metacognitive skill, was predicted to be associated with reading. With regard to the spelling manipulation, it was predicted good readers would retain an advantage even when orthographic knowledge could not be invoked (Spelling Foil). However, consistent with Ehri (1989), the advantage was expected to be greater on those items in which spelling aids phoneme identification (Spelling Aid Condition). In short, this first analysis was expected to provide further, stronger, support to the large body of research suggesting that phoneme awareness is specifically associated with reading in schoolchildren.

Method

Subjects

To compare extreme reading groups matched on age and vocabulary score, ten pairs of more and less skilled readers were selected from the larger sample and individually matched on age and PPVT-R standard score. Reading group assignment required a consistent score on word recognition, word attack, and teacher evaluation. (See Table 1 for descriptive statistics).

Materials

Vocabulary. The *Peabody Picture Vocabulary Test-Revised* (Dunn & Dunn, 1981) was used to assess vocabulary; the subject selects which of four pictures corresponds to an orally presented label. The PPVT-R provides both a vocabulary age equivalent and a standard score which correlates highly with IQ scores on omnibus cognitive measures.

Table 1. *Characteristics of skilled and less skilled readers matched on age and vocabulary level.*

	Reader group			
	Less skilled readers (n=10)		More skilled readers (n=10)	
	M	SD	M	SD
(Age (years, months)	8,10	(0,8)	8,8	(0,8)
PPVT-R standard score	102.7	(10.4)	102.8	(9.8)
Word recognition (WJ13)				
raw score	28.5	(6.2)	36.7	(2.1)
standard score	102.6	(9.9)	122.4	(6.1)
grade equivalent	3.0		6.1	
Word attack (WJ14)				
raw score	7.8	(4.4)	19.4	(3.4)
standard score	94.6	(11.6)	118.2	(7.2)
grade equivalent	2.8		12.9	

Reading. The reading measures included two subtests from the *Woodcock-Johnson Psychoeducational Battery* (Woodcock & Johnson, 1977): a test of word recognition (WJ13: Letter-Word Identification), and a nonsense word decoding task (WJ14: Word Attack).

Non-linguistic analysis measure. The *Embedded Figures Test* (EFT) developed by Satz et al. (1978) was used to control for general metacognitive factors and other task variables common to phonological and non-verbal measures. The EFT measures the ability of the subject to recognize a simple component shape embedded within a more complex figure. As depicted in Figure 1, each target item is presented at the top of the page with three choices provided below in a row. Across all trials, the correct responses are evenly distributed among the three positions. The test includes 24 items graded in difficulty.

Experimental phonological awareness measure. The *Embedded Phoneme Test* (EPT) was developed to assess phonological awareness using a format parallel to the EFT. In the EPT, subjects were first presented with a familiar pictured item (*"Listen to the first sound in the word MAIL"*) and were then asked to pick which of the three words pictured below had that sound embedded within it (e.g., *Can you hear that sound anywhere in the word SWIMMING, in the word SNOWING, or in the word SWINGING*). (See Figure 2).

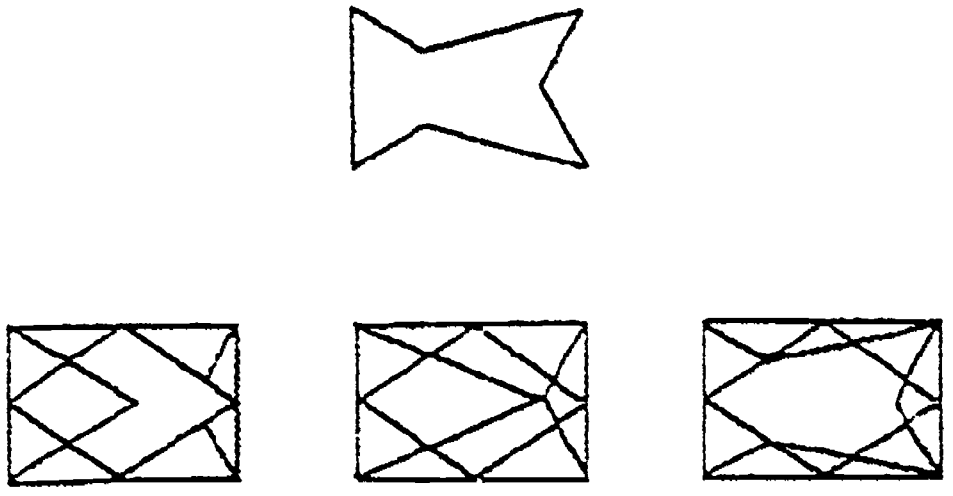


Figure 1. **Sample item from the Embedded Figures Test.**

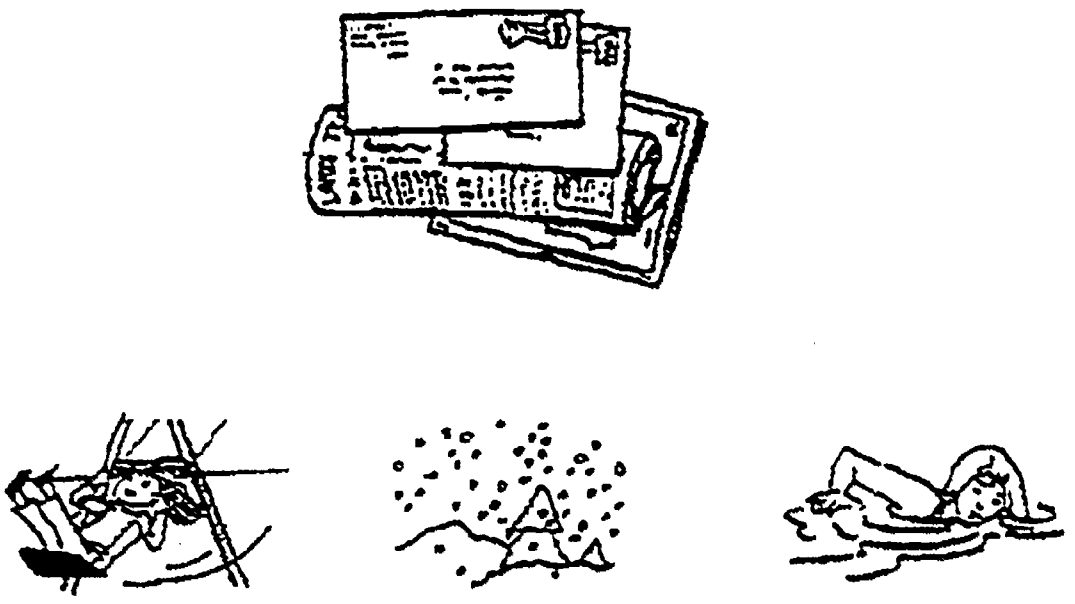


Figure 2. **Sample item from the Embedded Phonemes Test.**

The EPT was loosely modeled after a kindergarten measure requiring children to listen to a target word and to decide which of three words following had the same initial consonant (Stanovich, Cunningham, & Cramer, 1984a). In that study, kindergarten children performed at 73% correct on this task and performance was predictive of later reading skill. Although items from Stanovich (1984) were included for training and to establish a baseline measure of performance, the EPT measure differed from the earlier measure in two major respects. First, in the EPT, pictures accompanied the oral presentation. This served both to reduce memory demands and to parallel the non-verbal control. Secondly, in the EPT, the target segment was not always found in initial position but could occur anywhere within a word, serving to make the task more challenging for the school-aged child.

To directly assess the role of spelling strategies in performing this phoneme awareness task, two spelling conditions were included in addition to the three baseline items, for a total of 27 items. In approximately half of the cases (Spelling Aid Condition, 11 trials), consideration of the word's spelling would aid in identifying an embedded phoneme (e.g., *doll/bald-globe-block).* For the remaining items (Spelling Foil Condition, 13 trials), a spelling strategy alone could not lead to a correct response. In the Spelling Foil items the first letter of the target word either occurred in more than one of the choices (hence spelling was ambiguous, as in *ape/plant-pail-plaid*) or the first letter did not occur in any of the choices (it simply did not help, as in *zip/price-wasp-peas*)). In no case, however, was there a direct conflict between a spelling and sound response. (Refer to the Appendix for complete list of items).

The segments tested sampled evenly across stops, fricatives, nasal/liquids and vowels. These segments were embedded in monosyllabic (*ape/pail*), bisyllabic (*mail/swimming*) and multisyllabic (*cat/skeleton*) words. Target segments were assigned to either the first half or the latter half of the test word (*leg/glass* vs. *chair/watch*). Similarly, half the consonants were embedded in a cluster (*run/astronaut*); half were not (*ship/dalmatian*).

Distractors were selected such that each included at least one segment closely related to the target phoneme, differing from the target in only one feature. Thus, in the example shown in Figure 2, the [*n*] in *snowing* and the [*ng*] in *swinging* are phonetically similar to the target phoneme [*m*]. As in the EFT, all three choices were highly similar in overall configuration (e.g., number of syllables, morphemic structure). Further, to minimize erroneous strategies, the medial and final segments in the target item were consistently supplied or deleted in the three choices. Thus, in *zip/price-wasp-peas*, the final segment [p] of *zip* was present in all choices; the medial segment [I] was present in none of the choices. Words and pictures were carefully screened to avoid an association on the basis of semantic grouping (e.g. *fork /food* would not be acceptable).

In sum, the phonemes task was presented in a format nearly identical to that of the nonverbal figures task. In both, subjects were first shown a target picture (a geometric shape in the first case; a pictured object with a one-syllable label in the EPT). They were then asked to indicate by pointing which of the other three pictures contained the targeted shape or phoneme.

Procedure

All measures were administered to the subjects on an individual basis in two visits, each lasting 20-25 minutes. In the first, the vocabulary and reading measures were given following standard procedures. In the second, several weeks later, both the EFT and the EPT were presented, with order counterbalanced across children.

The metacognitive measures were introduced as "detective games." In the case of the EFT, they were told, *"Your job will be to find a shape hidden within a design. See this figure here? Can you find one just like it in any one of these figures?"* For the EPT, subjects were told, *"Now you are going to play another detective game, but this time, your job is to find sounds hidden in words."* Two items from Stanovich et al. (1984) kindergarten measure served to introduce the task. Feedback was provided on these items. (*"Listen to the first sound in the word FACE. Can you find that sound in the word PIG, in the word FORK, or in the word TOP?"*) Children were then tested on three similar items to determine a baseline level of phonological awareness. At this point, the experimenter introduced longer words, in which the target phoneme could be embedded anywhere. *"Listen to the first sound in FAN. Can you find that sound anywhere in the word CAMERA, DINOSAUR or BUTTERFLY?"* Subjects were explicitly encouraged to pay attention to the sounds of words rather than to their spelling and feedback was provided until the child understood the task. Although response latencies were recorded for both measures, they failed to play a significant role in performance on either task at this age level and are not entered in the analyses presented here.

Results

As can be seen in Table 2, mean performance on both the EPT and the EFT was well above chance level (33%).[2] Individual scores ranged from 44 to 96% correct in the Phonemes task; and from 50 to 95% correct on the Figures. Children performed almost perfectly (97% overall) on the baseline items drawn from Stanovich et al.'s (1984) kindergarten battery, indicating both that the children understood the task and could identify the initial phoneme. Thus, the difficult component of the EPT is in detecting the initial phoneme in one of the three choices.

Table 2. *Percentage correct on metacognitive measures as a function of reading ability.*

Metacognitive measures	Reading group: Less skilled readers (n=10)		More skilled readers (n=10)	
	M	SD	M	SD
Embedded Figures Test	75.8	10.4	78.8	10.7
Embedded Phonemes Test	57.0	7.7	69.6	15.1
Spelling Aid	57.3	8.6	71.8	22.0
Spelling Foil	46.9	13.8	60.0	18.1

Note: Overall score on the EPT is somewhat inflated due to inclusion of baseline measures; these are not included in the subtests.

As predicted, skilled readers performed significantly better than less skilled readers on the EPT, $t(18) = 2.35$, $p < .05$. In contrast, the two

groups did not differ on the nonverbal analogue, the EFT, t (18) = 0.54, $p > .05$. The correlation between the two analysis measures was near zero, $r(20) = .01$, indicating that despite common task requirements, each is tapping quite a different ability. To determine the contribution of spelling strategies to subjects' performance, the percentage correct per reading group was calculated separately for the Spelling Aid and Spelling Foil conditions. In a repeated measures ANOVA, there was a significant effect of both reading group, $F(1,18) = 5.24$, $p < .05$, and spelling condition $F(1,18) = 7.06$, $p < 0.02$, with no interaction between them, $F(1,1) = .03$, $p > .85$.

Group differences were comparable in both the Spelling Foil and the Spelling Aid conditions; both were marginally significant in post-hoc Scheffe analyses, $F(1,18) = 3.3$ and 3.8, $p < .10$. Although the finding that both groups performed better in the Spelling Aid condition is consistent with Ehri (1989), spelling knowledge alone cannot account for the skilled reader's advantage on this phoneme awareness task.[3]

An even more rigorous test of the hypothesis that individual differences in phoneme awareness contribute to reading ability/experience involves the use of reading level controls (Bryant & Goswami, 1986). In such a design, older poor readers are matched to younger good readers on reading level and IQ; a continued deficit in phoneme awareness despite comparable reading levels, suggests that phonological awareness may play a causative role in reading ability. Just such a design was employed in the next set of analyses.

EXPERIMENT 2

This experiment was designed to address concerns that differences in performance on phoneme awareness tasks may derive from rather than contribute to a child's reading level. To this end, pairs of children were selected such that the poor readers were at least one year older and one year ahead in school than the younger good reader; the two groups were matched on reading level and PPVT-R standard score at the time of testing. Using this research method, it was predicted here that a deficit in phonological awareness as assessed here should continue to characterize the older poor reader (Bradley & Bryant, 1978). Poor readers were expected, however, to have an advantage on nonverbal tasks unrelated to reading as a function of their greater maturity and vocabulary age.

Method

Subjects

To compare good and poor readers when reading ability is matched, 16 pairs of older poor readers and younger good readers were individually matched on their both their word recognition raw scores and their PPVT-R standard scores (see Table 3 for descriptive statistics). To further emphasize differences in reading experience, the younger child in each pair was a grade or more behind the older child. In making these matches; where differences on the matching variables were observed between the two groups, it was the poor readers who were given the advantage.

Table 3. *Reading level comparisons.*

	Reading group		
	More skilled readers n=16 M (SD)	Less skilled readers n=16 M(SD)	T(30), p-value
Matching variables			
Word recognition - raw score	31.9 (4.7)	33.1 (3.9)	0.78, $p < .44$
Word attack - raw score	12.8 (5.1)	11.6 (4.2)	0.72, $p < .48$
Vocabulary - standard score	99.4 (10.6)	101.6 (13.4)	-0.51, $p < .61$
Distinguishing variables			
Age (years, months)	8,2 (0,4)	9,5 (0,5)	-9.59, $p < .0001$
Grade	2.4 (0.5)	3.5 (0.5)	-5.84, $p < .0001$
Word recognition - st'd score	116.8 (9.7)	104.5 (10.4)	3.46, $p < .002$
Word attack - st'd score	109.9 (8.6)	98.9 (9.5)	3.40, $p < .002$
Metacognitive tasks (percentage correct)			
Embedded Figures Test	75.4 (7.5)	72.5 (12.1)	0.73, $p < .47$
Embedded Phonemes Test	63.3 (10.7)	53.3 (5.9)	3.30, $p < .002$
Spelling Aid	65.3 (15.6)	55.7 (11.4)	1.99, $p < .06$
Spelling Foil	52.9 (16.6)	42.8 (8.4)	2.17, $p < .04$

Results

Despite significant advantages in chronological age, vocabulary age, years of schooling, and current grade level; and equivalent or higher scores on word recognition, older poor readers performed significantly more poorly than younger good readers in locating phonemes embedded in words t (30) = 3.31, $p < .003$. This suggests that individual differences in phoneme awareness play a causal role in determining reading ability. Contrary to expectations, but further strengthening the specific nature of the phoneme awareness deficit, the EFT failed to differentiate the two groups t (30) = 0.73, $p < .47$. A breakdown of scores according to spelling condition indicates that the significant difference on the EPT was upheld in both the Spelling Aid and the Spelling Foil condition, though the good readers had the greater advantage when spelling could not be brought to bear.

One might argue that the greater phoneme awareness of the good readers in this sample is directly related to their nonsignificant advantage on the nonword decoding test. To assess this possibility, a further comparison was made including only the 11 pairs of subjects in which the older poor reader was equal to or better than the younger good reader on the reading levels achieved on both word recognition and word attack subtests. Even under these more stringent conditions, the good readers maintained their overall EPT advantage (young M = 17.1, SD = 2.8; older M = 14.3, SD = 1.5), t (20) = 2.97, $p < .01$). They also maintained their advantage in the Spelling Foil condition (young M = .53, SD = .14 ; older M = .41, SD = .08), t (20) = 2.52, $p < .03$. Where spelling could provide assistance, however, it appears that the older poor readers capitalized on the knowledge they had acquired; this is reflected in the failure to find a difference between the two groups in the Spelling Aid condition (younger M = .64, SD = .14 ; older M = .57, SD = .12), t (20) = 1.30, $p < .20$.

To summarize, the analyses presented thus far indicate that subjects who are unequivocally "skilled" at both decoding and word recognition possess greater phoneme awareness than do subjects who are unequivocally "less skilled" on these same measures, largely independent of age, vocabulary knowledge and grade. This holds true even when both groups have attained comparable word recognition reading levels. However, the results also hint that progress in word recognition may proceed somewhat independently of both phoneme awareness and word attack skills. This hint is followed up in subsequent analyses involving a larger sample. including "average" readers whose performance is more variable and whose difficulties and strengths are not necessarily specific to reading. Using multiple regression analyses, it becomes possible both to replicate the specific association between reading and phoneme awareness observed in the previous analyses as well as to more analytically assess the association between phoneme awareness and each of the reading measures.

EXPERIMENT 3

In the current set of analyses, the association between phoneme awareness and reading ability was examined in the entire set of 48 subjects, selected to vary broadly in age, vocabulary level and reading ability. In an effort to extend the prior finding of a specific association between phoneme awareness and reading ability to a more heterogeneous sample, multiple regression analyses were employed to control for the possible contributions of age, vocabulary, and metacognitive ability as assessed by the EFT. Of particular interest was the opportunity to separately assess the association between phoneme awareness and the two reading measures (word recognition and word attack), once these other variables had been controlled for; the inclusion of "average" readers enhanced the possibility of greater dissociation.

In this sample of readers, simple correlations were expected to reveal associations between reading and all other variables, and between EPT and EFT. Once these associations were controlled for, it was predicted that EPT performance would continue to correspond with both reading measures. It was further predicted that word attack would explain further variance in performance on the EPT, even after controlling for differences in word recognition knowledge and the other variables.

Subjects

The subjects included 48 second, third and fourth graders between the ages of 7 1/2 and 10 years. Children were excluded if they obtained a standard score below 80 on the Peabody Picture Vocabulary Test-Revised (PPVT-R, Dunn & Dunn, 1981); or if English was not the primary language spoken in the home.[4] To obtain a cross-section of children, care was taken to find an equal distribution of Low, Mid and High readers with a spread of vocabulary levels at two age groups; equal numbers of boys of girls were included in each of the categories. (Refer to Table 4 for subject description).

Table 4. *Characteristics of all subjects participating in the study.*

	Age and reading ability levels							
	YOUNGER			*OLDER*			*TOTAL*	
	Low	Mid	High	Low	Mid	High	*M*	*SD*
n	8	8	8	8	8	8		
Age in months	98.4	99.8	99.0	112.0	113.3	112.0	105.6	7.7
PPVT-R standard score	99.6	100.6	104.6	94.8	102.4	113.5	102.5	12.2
WJ13 raw score	22.8	32.8	36.0	30.8	34.1	38.3	32.4	5.5
standard score	98.3	116.4	124.6	98.8	107.5	120.5	111	12.3
WJ14 raw score	4.6	11.8	17.8	9.0	13.2	20.3	12.8	5.9
standard score	91.5	107.8	117.5	94.4	102.5	116.3	105	12.6

Results

The correlation coefficients presented in Table 5 highlight the different pattern of associations characterizing the two metacognitive measures. Whereas general metacognitive skill, as assessed by the EFT, was correlated only with PPVT-R, phoneme awareness was specifically correlated only with reading. While the specificity of the association between phoneme and reading is consistent with earlier analyses, in this case, the two reading measures were allowed to diverge in subject selection. Although the two reading measures were highly correlated with each other, it is clear that it is decoding that most crucially depends on phoneme awareness; EPT was not correlated with word recognition ability. The association between the reading measures and EPT was further explored in multiple regression analyses.

Table 5. *Intercorrelations among experimental and predictor variables.*

	(n=48)				
	PPVT	WJ13	WJ14	EPT	EFT
1. Age	-.08	.29*	.15	-.04	.06
2. PPVT-R standard score	--	.32*	.28*	.06	.27+
3. Word recognition-WJ13		--	.86**	.15	.17
4. Word attack-WJ14			--	.37**	.21
5. Embedded Phoneme Test				--	.22
6. Embedded Figures Test					--

*$p < .05$
**$p < .01$
+$p < .06$

First, to determine how the reading measures were related to EPT performance, a best subsets analysis was performed comparing all possible combinations of the five predictor variables (age, word recognition, word attack, PPVT-R, and EFT). The optimal model to explain performance on the EPT included both reading measures: $r^2 = .240$, $F_{2,45} = 7.10$, $p < .01$; the inclusion of all predictors explained no further variance ($r^2 = .26$, $F_{5,42} = 2.96$, $p < .05$). Although the best model for explaining EPT performance included both reading measures, the two measures exerted quite different effects. The word attack measure was the major predictor ($r = .37$, $p < .01$); entered first, it had a positive standard coefficient of 0.91. Somewhat surprising, however, is the fact that in this sample the word recognition measure, when entered into the model after decoding, makes a significant, but *negative* contribution to EPT performance (standard coefficient = -0.63). Closer inspection of individual performance patterns indicates that children whose word attack skills were low relative to their word recognition scores performed even less well on our phoneme test than children whose word attack abilities were consistent with their word recognition vocabulary; this effect was over and above the general effect of word recognition and may suggests differences in reading strategies.[5]

The strength of the association between phoneme awareness and decoding skill is further indicated in a multiple regression analysis involving word attack as the dependent variable. Although word recognition, vocabulary, EFT, and age together explain 75% of the variance in performance on the word attack measure, phoneme awareness contributes an additional 5%

of the variance after all these variables have been entered, $F(1,42) = 11.05, p < .01$.

These results suggest that individual differences in performance on phoneme awareness tasks exist apart from general intellectual and metacognitive factors, and cannot be readily explained away even by experience. Rather, even in children making apparent progress in acquiring a word recognition vocabulary, phoneme awareness remains crucial to the decoding of novel words.

EXPERIMENT 4

The goal of the final set of analyses was to determine whether conclusions drawn with these novel phoneme awareness measure could fairly be extended to phoneme awareness measures in general. At the same time, it is important to assess the validity of the the EPT by comparing it to a widely used test of phonological analysis of proven reliability and validity. Because the original intent of this study was to create a nonverbal analog of an existing phonological awareness measure, the first 23 subjects tested were also administered the *Lindamood Auditory Conceptualization Test* (LAC, Lindamood & Lindamood, 1971). This provides the opportunity to determine whether the new phonological awareness involve the same abilities as the more established measure, the LAC.

Because the LAC was not specifically designed to minimize memory requirements, spelling strategies, or cognitive demands also involved in reading, it was predicted that the first order correlations between reading and the LAC would be greater than those between reading and EPT. For the same reason, it was predicted that performance on the LAC would correspond both to the PPVT-R score and EFT. On the other hand, it was also predicted that the two phoneme awareness measures, LAC and EPT, would be significantly correlated, despite the differences in task requirements (Stanovich et al., 1984). Consistent with the prior analyses involving the EPPT, it was predicted that the LAC would continue to correlate with reading performance even after controlling for PPVT and EFT.

Method

Subjects

Of the 48 children discussed in Experiment 3, a randomly selected subset (n = 23) were given both phonological measures. Descriptive measures on these children show them to be representative of the group as a whole (mean age 8 yrs, 9.1 mos., *SD* 8.0 mos.; PPVT-R mean 102.9, *SD* 13.6; WJ13 mean 32.7; *SD* 5.4, WJ13SS mean 110.5; *SD* 11.7, WJ14 mean 13.4; *SD* 6.1, WJ14SS mean 105.3; *SD* 13.4).

Materials

The materials are the same as in Experiments 1 through 3, but for the addition of the *Lindamood Auditory Conceptualization Test* (LAC, Lindamood & Lindamood, 1971), a widely used test of phonological analysis of proven reliability and validity. In this study, it was used for comparison with the EPT. In the LAC, the child is asked to manipulate colored blocks corresponding to the phonemic segments in nonsense words. Using this technique, the LAC measures the ability to isolate and compare sounds, and to discern their number and order within nonsense syllables. In the first part of the test, subjects are asked to place colored blocks in response to isolated speech sounds (e.g., to place two blocks of the same color to represent /z/ /z/, but to use two different colors to represent /z/ /m/). The second part of the test also requires the subject to represent sounds with colored blocks, but now syllables are introduced containing two to four segments. Items are presented in a pre-determined order such that only one change is required per item. For instance, the experimenter would present three different colored blocks and say, "this is /vop/, now show me /vops/." The next item would proceed from that: "this is /vops/, now show me /vups/." The score obtained is a combination of weighted raw scores from the two parts of the test. The maximum possible score is 100. The LAC correlates highly (.66 to .81) with combined reading and spelling scores on the Wide Range Achievement Test in children in grades K to 12 (Calfee, Lindamood & Lindamood, 1973); it was also highly significant in distinguishing good/poor reader groups in the third grade and in adult education classes (Pratt & Brady, 1988).

Procedure

The LAC was administered following standard procedures. It was presented in a separate 12 minute session within the same week as the PPVT-R and the reading tests.

Results

The group of subjects included in these analyses obtained scores on the EPT and EFT virtually identical to those achieved by the larger group from which they were drawn (see Experiment 3 for those statistics). The mean score on the LAC was 69.74 (SD 19.58), with a range in performance from 33 to 100. This score would place the

children between the second and third grade, according to the LAC norms. Since the mean grade level of our subjects was 3.4, they were, on average, performing below grade level on this test. Correlations among vocabulary, reading, and analysis measures reveal associations similar to those observed in other subsets of subjects, but for an even higher overall correlation between EPT and the reading measures. PPVT-R and EFT continued to be nonsignificant correlates of EPT.Consistent with the findings of Calfee et al. (1973) and of Pratt & Brady (1988), the LAC scores of our sample also correlated significantly with reading measures, r_{23} = .47 to .57; $p < .05$. Also consistent with expectations, the LAC also correlated significantly with the EPT, r_{23} = .53; $p < .01$. (See Table 6 for correlations).

Table 6. *A comparison of two phoneme awareness measures.*

Correlations with reading, vocabulary, age and EFT (n=23)		
	Phoneme awareness measure	
Predictor variable	EPT	LAC
Word recognition	.42*	.47*
Word attack	.61**	.48*
PPVT-R standard score	.20	.61**
Age	.15	-.12
Embedded Figures Test	.39	.41+

+$p < .06$
* $p < .05$
** $p < .01$

The LAC diverged from the EPT in regard to their degree of association with the PPVT-R and the two reading measures. Stepwise regression analyses to predict LAC showed PPVT-R to be the most important predictor of LAC performance ($F_{1,21}$=12.33, $p < .05$) accounting for 37% of the variance. The second best predictor was the standard score on word Recognition; these two measures together accounted for 45% of the variance on the LAC ($F_{2,20}$= 9.13, $p < .01$). Age, EFT, and WJ14 explained no further significant variance. On the other hand, the EPT score, included after both PPVT-R and word recognition, further contributed an additional 12% of the variance on the LAC, the second phoneme awareness measure contributed ($F_{1,21}$ = 5.0; $p <$ .05). In sum, LAC peformance appears to depend heavily, though not solely, on general intellectual factors.

In contrast to the LAC performance, stepwise regression analyses to predict EPT performance in this subgroup, selected variables. Consistent with results reported for the entire group, the only significant predictor was word attack ($F_{1,21}$=12.20, $p < .05$, r^2=.37). EFT was a marginal predictor of EPT in this subgroup ($F_{1,21}$=4.05, p = .06), with WJ14 and EFT together explaining 45% of the variance ($F_{2,20}$=9.02, $p < .05$). Performance on the LAC did not contribute further explanatory power above and beyond that accounted for by the decoding measures. Non-significant predictors included word recognition, PPVT-R, age and sex.

In sum, although the EPT measure correlated fairly highly with the LAC, the two were by no means equivalent in how they related to decoding and word recognition. Whereas the EPT correlated best with a pure measure of word decoding (WJ14), the LAC correlated better with WJ13 - a standard measure of word recognition which taps experience, vocabulary and decoding skill. Whereas PPVT-R standard scores did not correlate significantly with EPT performance, it was the most significant predictor of performance on the LAC, suggesting that general intelligence and task demands appear to play an important role in performance on the LAC.[6]

GENERAL DISCUSSION

This study was directed at identifying and assessing the role of three factors that may account for the association between reading and phonological awareness once reading instruction has begun. To this end, a pair of measures was designed to simultaneously control for non-linguistic task demands and general analytic skill, while minimizing requirements of verbal production and working memory. In addition, the measures were designed to assess the contribution of reading experience and spelling strategies in explaining individual differences. Performance on this pair of measures by schoolchildren is consistent with and extends the findings of previous studies. The results indicate that reading ability, and particularly decoding skill, continues to be significantly associated with phonological awareness even when a number of other factors have been taken into account.

Metacognitive factors. First, the association between phonological awareness and reading ability appears to be largely independent of general metacognitive skill and non-linguistic task factors. Although phonological and nonlinguistic analysis

tasks were presented in a nearly identical format, only the phonological measure successfully discriminated skilled from less skilled readers matched on age and vocabulary level. The lack of correlation between phonological awareness and and nonlinguistic awareness or between phonological awareness and general verbal ability suggests that a failure to understand and cope with extraneous task factors cannot explain poor performance on the EPT. Further support for the independence of phoneme awareness from general cognitive factors.

The present study goes beyond previous studies in also examining the pattern of abilities in a relatively unselected population of children whose reading profiles varied. Once again, the newly developed phonological awareness measure failed to correlate with general verbal ability or the nonlinguistic measure. Although this is not the case for all phonological awareness measures, as indicated in the comparison with the Lindamood task (LAC), the point to be drawn is that the relationship between phoneme awareness measures and reading does not simply derive from general cognitive factors.

The evidence that the particular nonlinguistic measure selected was useful as a pre-school predictor of reading success in other studies, but failed to correlate with reading in the present sample, is consistent with the suggestion that metacognitive factors may serve a catalytic, rather than an ongoing role in reading acquisition. In sharp contrast, the particular abilities required in phonological awareness tasks continue to be highly associated with reading well into the school years.

Extraneous memory demands. The second factor hypothesized to play a central role in explaining the association between phonological awareness and reading is working memory, defined broadly to include entering, maintaining, and retrieving items in a phonological store. In this study, such factors were not so much controlled for as minimized. Thus, the task did not require the subject to produce a verbal response or to reverse or manipulate phoneme segments. Further, the use of pictures allowed subjects to refresh their memory as needed under no time constraints, rather than requiring the subject to store and compare four different words in memory. With this format, each of the three alternative could be considered individually. That the phonological awareness task remained correlated with decoding skill rules out some of the less interesting explanations for the strong association between reading and phonological awareness; one need not artificially stress working memory, metacognitive factors or the production system to obtain an association.

On the other hand, it can not be concluded from these results that memory factors do not contribute to phonological awareness. Even in this simple task, the subject must maintain a segment in mind, and simultaneously scan one word at a time until a matching phoneme is encountered. Certainly, a dramatically impoverished verbal store will compromise performance even on this task (Dreyer, 1989). The present results instead serve to sharply restrict the potentially critical significance of working memory factors to their most intrinsic role in isolating and identifying individual phonemes at a single word level.

Reading experience and spelling strategies. The third factor, the role of reading level in affecting performance on phonological awareness measures, was addressed in two ways. First, although both skilled and less skilled readers were more accurate in their performance when spelling and phonological representations uniqely specify the same result (Spelling Aid), the continued advantage of skilled readers when spelling alone failed to provide cues makes it quite clear that it is not spelling alone that is conferring an advantage on the skilled readers. They can, at a better rate than the less skilled readers, go beyond their spelling knowledge to access and accurately select from among the phonological representations of the various choices presented. They are better able to isolate and identify the relevant sound units.

That phoneme awareness is not a simple function of reading and/or spelling expertise is further demonstrated by a comparison of older poor readers with younger good readers at equivalent reading-levels, where the poor readers had the advantage of age, vocabulary knowledge, and grade level. The finding that older poor readers lagged behind younger controls in both spelling conditions suggests that phoneme awareness continues to be an important determinant of reading aptitude during the school years.

Finally, the results of this study are consistent with the claim by Yopp (1988), that each phonological awareness measure brings its own task requirements. Although both of the phonological awareness measures compared here were associated with reading and with each other, they differed importantly in other respects. Whereas the LAC made heavy processing demands, the EPT appeared to be a measure of phonological awareness that is less confounded by

general intelligence as assessed by a vocabulary measure. But even after controlling for these more general factors in the LAC, a word recognition measure explained an additional 10% of the variance in performance. The results of this study concur with others in showing that across very distinct phonological awareness measures, and after a number of task variables are controlled for, a core ability remains which is robustly associated with reading (Stanovich et al., 1984; Yopp, 1988).

In sum, this paper supports the view that phoneme awareness, specifically involving the isolation and identification of phonemic segments, is a continuing area of difficulty for schoolchildren with reading disability, deserving of our attention and concern.

REFERENCES

Ball, E., & Blachman, B. (1988). Phonemic segmentation training: Effect on reading readiness. *Annals of Dyslexia, 38*, 208-225.

Bentin, S., Deutsch, A., & Liberman, I. (1990). Syntactic competence and reading ability in children. *Journal of Experimental Child Psychology, 49(1)*, 147-172.

Bowey, J. A. (1986). Syntactic awareness and verbal peformance from pre-school to fifth grade. *Journal of Psycholinguistic Research, 15*, 285-308.

Bradley, L., & Bryant, P. (1983). Categorizing sounds and learning to read—a causal connection. *Nature*, 301, 419-421.

Bradley, L., & Bryant, P. (1978). Difficulties in auditory organization as a possible cause of reading backwardness. *Nature, 271*, 746-747.

Brady, S. (in press). The role of working memory in reading disability. In S. Brady & D. Shankweiuler (Eds.), *Phonological processes in literacy: A tribute to Isabelle Y. Liberman*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Brady, S. (1986). Short-term memory, phonological processing and reading ability. *Annals of Dyslexia, 36*, 138-153.

Brady, S. A., Shankweiler, D., & Mann, V. A. (1983). Speech perception and memory coding in relation to reading ability. *Journal of Experimental Child Psychology, 35*, 345-367.

Bryant, P., & Goswami, U. (1986). Strengths and weaknesses of the reading level design: A comment on Backman, Mamen and Ferguson. *Psychological Bulletin, 100*, 101-103.

Byrne, B., & Ledez, J. (1983). Phonological awareness in reading-disabled adults. *Australian Journal of Psychology*, 35, 185-197.

Calfee, R.C., Lindamood, P., & Lindamood, C. (1973). Acoustic-phonetic skills and reading—Kindergarten through twelfth grade. *Journal of Educational Psychology, 64*, 293-298.

Crain, S., & Shankweiler, D. (1988). Syntactic complexity and reading acquisition. In A. Davidson & G. Green, (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dreyer, L. (1989). *Phonological processing in reading: Phonological memory as a component of decoding ability*. Doctoral dissertation, Columbia University.

Dunn, L., & Dunn, L. (1981). *Peabody Picture Vocabulary Test—Revised*. Circle Pines, MN: American Guidance Service.

Ehri, L. (1984). How orthography alters spoken language competencies in children learning to read and spell. In J. Downing & R. Valtin (Eds.) *Language awareness and learning to read*. New York: Springer-Verlag.

Ehri, L. (1989). The development of spelling skill and its role in reading acquisition and reading disability. *Journal of Learning Disabilities, 22*, 356-365.

Ehri, L., & Wilce, L. (1980). The influence of orthography on readers' conceptualization of the phonemic structure of words. *Applied Psycholinguistics, 1*, 317-385.

Fowler, A. (1988). Grammaticality judgments and reading skill in grade 2. *Annals of Dyslexia, 38*, 73-84.

Galambos, S., & Hakuta, K. (1988). Subject-specific and task-specific characteristics of metalinguistic awareness in bilingual children. *Applied Psycholinguistics, 9*, 141-162.

Gleitman, L., Gleitman, H., & Shipley, E. (1972). The emergence of the child as a grammarian. *Cognition, 2*, 137-164.

Gleitman, L., & Rozin, P. (1977). The structure and acquisition of reading: Relation between orthography and the structured language. In A. S. Reber & D. L. Scarborough (Eds.), *Toward a Psychology of Reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jorm, A. (1983). Specific reading retardation and working memory: A review. *British Journal of Psychology, 74*, 311-342.

Katz, R. B. (1986). Phonological deficiencies in children with reading disability: Evidence from an object naming task. *Cognition, 22*, 225-257.

Liberman, A. (1989). Reading is hard just because listening is easy. In C. van Euler (Ed.) *Wenner-Gren International Symposium Series: Brain and Reading*. Hampshire, England: MacMillan.

Liberman, I. Y. (1971). Basic research in speech and lateralization of language: Some implications for reading disability. *Bulletin of the Orton Society, 21*, 71-87.

Liberman, I. Y., & Shankweiler, D. P. (1985). Phonology and the problems of learning to read and write. *Remedial and Special Education, 6*, 8-17.

Lindamood C. H., & Lindamood, P. (1971). *Lindamood Auditory Conceptualization Test*. Boston: Teaching Resources.

Lundberg, I. (1978). Aspects of linguistic awareness related to reading. In A. Sinclair, R. J. Jarvella & W. J. M. Levelt (Eds.), *The child's conception of language*. Berlin: Springer-Verlag.

Lundberg, I., Frost, J., & Peterson, O. (1988). Effects of an extensive program for stimulating phonological awareness in pre-school 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.

Mann, V. A. (1984). Longitudinal prediction and prevention of reading difficulty. *Annals of Dyslexia, 34*, 117-137.

Mann, V. A. (1986). Phonological awareness: The role of reading experience. *Cognition, 24*, 65-92.

Mann, V. A., & Liberman, I. Y. (1984). Phonological awareness and verbal short-term memory: Can they presage early reading problems? *Journal of Learning Disabilities, 17*, 592-599.

Mann, V. A., Liberman, I. Y., & Shankweiler, D. (1980). Children's memory for sentences and word strings in relation reading ability. *Memory & Cognition, 8*, 329-335.

Mann, V. A., Tobin, P., & Wilson, R. (1988). Measuring the causes and consequences of phonological awareness through the invented spellings of kindergarten children. *Merrill-Palmer Quarterly, 33*, 365-391.

Mattingly, I. (1972). Reading, the linguistic process, and linguistic awareness. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye*. Cambridge, MA: MIT Press.

Morais, J., Alegria, J., & Content, A. (1987). Segmental analysis and literacy. *Cahiers de Psychologie Cognitive, 7*, 415-438.

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.

Palley, S. (1986). *Speech perception in dyslexic children.* Unpublished doctoral dissertation, The City University of New York.

Perfetti, C., Beck, I., Bell, L., & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal: A longitudinal study of 1st grade children. *Merrill-Palmer Quarterly, 33*, 283-319.

Pratt, A., & Brady, S. (1988). The relationship of phonological awareness to reading disability in children and adults. *Journal of Educational Psychology, 80*, 319-323.

Rapala, M. M., & Brady, S. (1990). Reading ability and short-term memory: The role of phonological processing. *Reading & Writing, 2*, 1-25.

Rosner, J., & Simon, D. P. (1971). The auditory analysis test: An initial report. *Journal of Learning Disabilities, 4*, 384-392.

Rozin, P. (1975). The evolution of intelligence and access to the cognitive unconscious. In J. Sprague & A. N. Epstein (Eds.), *Progress in psychobiology and physiological psychology (Vol. 6)*. New York: Academic Press, 1975.

Ryan, E. B., & Ledger, G. W. (1984). Learning to attend to sentence structure: Links between metalinguistic development and reading. In J. Downing & R. Valtin (Eds.), *Language awareness and learning to read*. New York: Springer-Verlag.

Satz, P., Taylor, H., Friel, J., & Fletcher, J. (1978). Some developmental and predictive precursors of reading disabilities: A six year follow-up. In A. L. Benton & D. Pearl (Eds.) *Dyslexia: An appraisal of current knowledge*. New York: Oxford U. Press.

Siegel, L., & Ryan, E.B. (1984). Reading disability as a language disorder. *Remedial and Special Education, 5*, 28-33.

Shankweiler, D., & Crain, S. (1986). Language mechanisms and reading disorder: A modular approach. *Cognition, 24*, 139-168.

Share, D., Jorm, A., MacLean, R., & Matthews, R. (1984). Sources of individual differences in reading acquisition. *Journal of Educational Psychology, 76*, 1309-1324.

Smith, L., & Kemler, D. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology, 24*, 279-298.

Stanovich, K. E. (1988). The right and wrong places to look for the cognitive locus of reading disability. *Annals of Dyslexia, 38*, 154-177.

Stanovich, K. E. (1982). Individual differences in the cognitive processes of reading: I. Word coding. *Journal of Learning Disabilities, 15*, 449-572.

Stanovich, K. E., Cunningham, A. E., & Cramer, B. B. (1984a). Assessing phonological awareness in kindergarten children: Issues of task comparability. *Journal of Experimental Child Psychology, 38*, 175-190.

Stanovich, K., Cunningham, A., & Feeman, D. (1984b). Intelligence, cognitive skills, and early reading progress. *Reading Research Quarterly, 19*, 120-139.

Torgesen, J. (1985). Memory processes in reading disabled children. *Journal of Learning Disabilities, 18*, 350-357.

Treiman, R. (1985). Spelling of stop consonants after /s/ by children and adults. *Applied Psycholinguistics, 6*, 261-282.

Treiman, R., & Baron, J. (1981). Segmental analysis ability: Development and relation to reading ability. In G. E. MacKinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice (Vol. 3)*. New York: Academic Press.

Tunmer, W. E. (1988). Metalinguistic abilities and beginning reading. *Reading Research Quarterly, 23*, 134-158.

Tunmer, W. E., Herriman, M., & Nesdale, A. (1988). Metalinguistic abilities and beginning reading. *Reading Research Quarterly, 23*, 134-158.

Tunmer, W. E., & Nesdale, A. R. (1985). Phonemic segmentation skill and beginning reading. *Journal of Educational Psychology, 77*, 417-427.

Vogel, S. (1974). Syntactic abilities in normal and dyslexic children. *Journal of Learning Disabilities, 7*, 47-53.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing in the acquisition of reading skills. *Psychological Bulletin, 101*, 192-212.

Wolf, M., & Goodglass, H. (1986). Dyslexia, dysnomia and lexical retrieval: A longitudinal investigation. *Brain and Language, 28*, 159-168.

Woodcock, R. W., & Johnson, M. B. (1977). *Woodcock-Johnson Psychoeducational Battery*, Boston: Teaching Resources.

Yopp, H. K. (1988). The validity and reliability of phonemic awareness tests. *Reading Research Quarterly, 23*, 159-177.

FOOTNOTES

Journal of Experimental Child Psychology, submitted.

†Also Department of Human Development, Bryn Mawr College.

[1]We first attempted to use the *Lindamood Auditory Conceptualization Test* (LAC), with a musical control task (M-LAC), following Pratt (1984). Extensive work to adapt the M-LAC procedure to our needs was unsuccessful; Pratt's procedure paralleled only the first half of the LAC procedure and the lack of difference in her groups on this measure may have been due to ceiling effects. It was the lack of a suitable nonverbal control that motivated our move toward a new measure of phoneme awareness.

[2]Scores are presented in terms of percentage correct to facilitate comparison across different measures and conditions, which vary in the number of items included.

[3]In the current version of the EPT, the goal of the Spelling Foil condition was only to assess whether the advantage of the poor readers would be maintained when spelling crutches were absent. However, because the two conditions were not perfectly matched in all respects, the opposite conclusion cannot be drawn, that is, that the difference between performance on the Spelling Aid and Spelling Foil conditions derives from orthographic factors alone. This question is currently being investigated in a follow-up study in which an expanded version of the EPT includes a completely balanced design within the two spelling conditions.

[4]The pair of measures introduced in the present study were developed to be used as part of an extensive battery of language and nonlanguage measures in a large scale dyslexia subtyping study being conducted by Haskins Laboratories and Yale University Medical School. For future comparability with that study, subjects' ages and exclusionary criteria were determined by those established for the subtyping study.

[5]The independence of WJ13 and WJ14 is illustrated quite clearly among the eight children between 81/2 and 91/2 years of age who achieved "high" scores (37 to 40) on the WJ13 Word Identification Subtest. Four of these children made zero or one error on the WJ14 Word Attack test; these "decoders" achieved a score of between 74.1 and 96.3% correct on the EPT. The other four misread between four and ten items on WJ14; they scored between 44.4 and 66.7% correct on the EPT.

[6]These results differ from those of Pratt and Brady (1988) who report that vocabulary was not a significant factor in LAC performance among IQ-matched extreme reader groups in the second grade. Pratt and Brady (1988) did, however, find vocabulary to be associated with both reading and LAC performance in adults attending remedial education classes.

APPENDIX

EMBEDDED PHONEME TEST

Baseline items

PEAR	PEN	TILE	MASK
SOAP	KING	DIME	SALT
MILK	DATE	MOON	BAG

Monosyllabic

DOLL	BALD	GLOBE	BLOCK	*PAN	PHONE	SPONGE	SPHERE
CHAIR	BUSH	WATCH	DUST	*ZIP	PRICE	WASP	PEAS
LEG	GLASS	FROG	GRAPES	*WIG	WORM	WRITE	WRENCH
UP	KISS	MICE	BUS	*APE	PLANT	PAIL	PLAID

Bisyllabic

*TIE	CASTLE	FEATHER	PASTRY	*GAME	PIGEON	PICNIC	PENGUIN
*SOCK	TREASURE	WHISTLE	POISON	*JUICE	MEASURE	PICTURE	SOLDIER
MAIL	SWIMMING	SNOWING	SWINGING	*YOLK	MUSIC	LIQUID	NICKEL
AX	WINDOW	WAGON	WEDDING	*ICE	SUNSHINE	SANDWICH	SAILBOAT

Multisyllabic

BOX	VOLCANO	OCTOPUS	VEGETABLES	*CAT	SKELETON	SPAGHETTI	STRAWBERRY
VAN	UMBRELLA	ENVELOPE	GRANDFATHER	RUN	AMBULANCE	ASTRONAUT	ELEPHANT
NUTS	LAWNMOWER	HAMBURGER	FAMILY	*SHIP	GRASSHOPPER	EXPLOSION	DALMATIAN
INK	POCKETBOOK	BILLYGOAT	LAWNMOWER	*EAT	FISHINGROD	SPIDERWEB	JELLYFISH

*Spelling foil

Haskins Laboratories Status Report on Speech Research
1990, SR-103/104, 153-164

Short-term Serial Recall Performance by Good and Poor Readers of Chinese

Nianqi Ren† and Ignatius G. Mattingly†

Chinese second-grade students who had been classified either as "good readers" or as "poor readers" were subjects in a visual serial recall experiment in which the items in the series to be remembered were Chinese characters. Three series types were used: orthographically similar, phonologically similar, and nonsimilar. The same students were subjects in a parallel auditory serial recall experiment in which the items to be remembered were spoken Chinese words, and the series were phonologically similar or nonsimilar. The results of these experiments were broadly parallel to the results of experiments with good and poor readers of English: Good readers performed better than poor readers and, in the visual experiment, were relatively more affected by phonologically similar series than the poor readers. It is concluded that, whether the writing system is alphabetic or nonalphabetic, the phonological mechanism used in short-term recall of visually-presented verbal material is the same mechanism that is used for reading. This mechanism is, in fact, the linguistic module that controls speaking and listening. Representations of phonological segmental structure are automatically computed by the module, even though they may have no apparent relevance for a particular linguistic task, such as the reading of Chinese characters.

A correlation has been clearly demonstrated between the ability of young children to read an alphabetic orthography and their performance in tasks requiring serial recall of alphabetic material: good readers recall more than poor readers (Shankweiler, Liberman, Mark, Fowler, & Fischer, 1979). Here we investigate whether a similar correlation can be found for young readers of a nonalphabetic orthography, that of Chinese.

It is generally accepted that material to be remembered for more than a few milliseconds is coded phonologically, if such a coding is possible. Such phonological coding has been demonstrated in serial recall experiments not only for spoken utterances (Wickelgren, 1965, 1966), but also for pictures of nameable objects (Conrad, 1972), letters (Conrad, 1964; Conrad & Hull, 1964), alphabetically-written words (Kintsch & Buschke, 1969), Chinese logograms (Tzeng, Hung, & Wang, 1977), Japanese logograms (Erickson, Mattingly, & Turvey, 1977), and American Sign Language signs with obvious English equivalents (Hanson, 1982). When subjects try to recall a series of such items, the errors they make are most reasonably interpreted as phonological confusions. Moreover, if a series to be recalled consists of phonologically similar items, more errors will be observed for this series than for a nonsimilar series. Apparently, the mechanism subjects use to remember verbal material is phonological, even in a condition that might seem to favor use of a non-phonological mechanism if one were available. This phonological mechanism may be, as Baddeley and Hitch (1974) propose, a buffer in "working memory." Or it may be, as we will argue, that "working memory" or "verbal short-term memory" is simply the rehearsal of verbal material with the aid of the linguistic mechanism or module (Fodor,

This work was supported by NICHD Grant HD 01994 to Haskins Laboratories. We are very grateful to Huixin Hu for her help in running the subjects and preparing the figures, to Chuanliang Cui for arranging for projection equipment, to Baomin Ye and Yi Xu for help and advice on Chinese characters, to Leonard Katz for perceptive comments on an earlier draft of this paper, to the students at the Kongjiang Er Cun Elementary School, Shanghai, who took part in the experiments, and to their teachers who so generously cooperated with us.

1983) that supports speaking and listening, and that necessarily produces phonological (indeed, linguistic) representations.

But, while orthographic items may be phonologically encoded for memorial purposes, it does not necessarily follow that reading requires phonological encoding. A reader is ordinarily trying, not to remember a text verbatim, but to understand it. It is thus of great interest that, in fact, a correlation has been demonstrated between the ability of young children to read in an alphabetic orthography and their serial recall of alphabetic material (Shankweiler et al., 1979). These investigators asked good and poor second-grade readers to recall visually-presented series of letters and auditorially-presented series of letter-names. The letter-names in some series were phonologically similar, i.e., rhymed (e.g., B, C, D, G, P) and in other series were non-similar (e.g., H, K, L, Q, R). It was found that in both modes of presentation, good readers recall more than poor readers, and are more affected by phonological similarity than poor readers. Similar results have been found for spoken words and sentences (Mann, Liberman, & Shankweiler, 1980).

Evidently, alphabetic reading and serial recall are relying on the same phonological mechanism. This finding is surely an important clue to the nature of the reading process, but we would be better able to interpret it if we knew whether the correlation is specific to alphabetic reading or obtains for the reading of nonalphabetic orthographies as well. If no correlation were found for nonalphabetic reading, we would conclude that because of its segmental character, alphabetic orthography makes some special demand on phonological processes also used in working memory, while for nonalphabetic orthographies, in which the symbols correspond to larger linguistic units, there is no obvious reason for phonological segments to have any role in reading unless the text is to be remembered. On the other hand, if a correlation *were* found for nonalphabetic orthography, we would be led to conclude that, regardless of the particular character of the orthography, phonological processes are intrinsic to reading, either because reading requires working memory, as Baddeley (1979) has argued, or because reading, like the remembering of verbal material, necessarily relies on the language module.

Some recent work with Japanese second-graders encourages the expectation of a correlation for nonalphabetic orthographies (Mann, 1985). In this study, a group of good readers and a group of poor readers were asked to remember both auditory and visual stimuli in a recurring recognition paradigm (Kimura, 1963). The auditory material consisted of Japanese nonsense syllables; the visual material, of abstract designs, photographs of male Japanese faces, hiragana symbols that in the Japanese writing system represent phonological moras, and kanji characters that in this writing system represent morphemes. The good readers performed better than the poor readers on the phonologically codeable material—the nonsense syllables, the hiragana and the kanji—but not on the abstract designs or faces. However, this study did not vary phonological similarity systematically, and it used a paradigm that minimized the effects of rehearsal.

The present investigation is concerned with the reading of the other major nonalphabetic orthography in modern use, Chinese. Spoken Chinese is, in fact, a group of dialects that, although having a common historical origin, are now quite different from one another, and, not necessarily even mutually intelligible. The differences, however, are mainly phonological; the syntax and the basic stock of monosyllabic morphemes inherited from Classical Chinese is common to all the dialects. Traditionally, the same writing system could be used for these different dialects because its characters stand for these common morphemes. Thus 石 stands for the morpheme that is pronounced [ʂi] in Mandarin and [zəʔ] in Shanghainese, and means 'stone' in both. A Chinese morpheme may function either as a monomorphemic word or as constituent of a polymorphemic word, but this distinction is not indicated in the writing. There are spaces between characters but no specific word-boundary markers.

The character 石, and many others, are unitary signs, but over 90% of the characters used in modern Chinese writing are "phonetic compounds." A phonetic compound has two parts, the "signific" and the "phonetic," each of which can, in general, appear also as a separate, free-standing character. The signific is usually at the left or the top of the compound character; the phonetic at the right or the bottom. In principle, the signific is supposed to indicate the semantic category of the morpheme that the compound character stands for; the phonetic, its pronunciation. For instance, the character 材, cái (in the pinyin Romanization, with tone-marking added), 'lumber,' consists of a signific,

 that, as a separate character, stands for mù 'wood'; and a phonetic  that, as a separate character, stands for cái 'talented person.' But because of changes in both language and writing system over a long period, neither the signific nor the phonetic now necessarily provide very reliable information. For instance,  táng, 'hall' consists of the signific  tǔ, 'earth' and the phonetic  shàng, 'esteem.'

During their first two years of school, children in the People's Republic of China, regardless of their native dialect, are required to learn Putong Hua ("the common speech"), that is, Mandarin, the idealization of the Beijing dialect that is the official language of the P.R.C. They also learn pinyin, the official Romanization system for Mandarin. And most remarkably, they learn to recognize, write, pronounce in Mandarin, and use in sentences some 1400 Chinese characters. (In later years of elementary schooling, they learn many more). Thus, except for native speakers of the Beijing dialect, Chinese children learn to read in their second language.

The experiments on Chinese we report here are comparable to the Shankweiler et al. (1979) serial recall experiments. Like those researchers, we investigated the ability of good and poor readers to recall linguistic material presented visually and auditorially. But the items to be recalled in their experiments were letters and spoken letter-names, while in our experiments the items were Chinese characters and spoken monomorphemic Mandarin words. There were also certain methodological differences. They compared simultaneous and sequential visual presentation of the items in a series; we used only sequential presentation. In the visual presentation, we included an orthographically similar condition; they did not. They compared performance with immediate and considerably delayed subject response; we simply used a moderately delayed response. They required a written response to the auditory as well as to the visual presentation; we required a spoken response to the auditory presentation. Perhaps most important, they required their subjects to report the entire series after each trial, while we used a Waugh-Norman procedure (Waugh & Norman, 1965), so that the subjects reported only a single item.

EXPERIMENT 1: VISUAL PRESENTATION

Subjects

The subjects used in both the visual and the auditory experiments were native speakers of Shanghainese completing the second grade in several different classes in an elementary school in Shanghai. The principle for selecting "good readers" and "poor readers" was that the two groups should differ in reading ability, but not in general intelligence. Thus, the recommendation of classroom teachers and scores on the final examination in reading at the end of second grade were used to define the two groups, but students who had done poorly on the final examination in mathematics were excluded. This procedure yielded 40 candidate good readers and 40 candidate poor readers. Then the Draw-a-Man IQ test (Goodenough, 1926) was given to 78 of these students (two were not available), and those whose IQ scores fell outside the range 90-120 were eliminated, leaving 25 good readers and 25 poor readers. Two of the poor readers failed to attend, so that 25 good readers and 23 poor readers actually took part in the experiments. The good readers ranged in age from 90 to 104 months; the average was 98. Their IQs ranged from 96 to 118; the average was 107. The poor readers ranged in age from 90 to 107 months; the average was 99. Their IQs ranged from 92 to 120; the average was 103.

Materials

All the characters used in the experiment were drawn from the inventory of characters that Chinese children are expected to know by the end of second grade. Fifteen series of six characters each were prepared; they are given in the Appendix. This series length had been found, in pilot testing, to yield a useful range in number of recall errors for both the best-performing and poorest-performing subjects. The series were of three types. In an "orthographically similar" (OS) series, the characters all had the same signific, but the monomorphemic words for which they stood had no particular phonological resemblance to one another, either in Mandarin or in Shanghainese, nor (except perhaps from a historical standpoint) were they semantically related. In a "phonologically similar" (PS) series, all the words for which the characters stood rhymed and had the same tone, whether read as Mandarin or as Shanghainese, but they were semantically unrelated, and the characters had no

significs or phonetics in common. In a "nonsimilar" (NS) series, neither the characters nor the words for which they stood had any common properties. There were five series of each type. Altogether, 90 different characters were used, none more than once. A separate 2" × 2" slide was made from a drawing of each character.

Procedure

The characters were shown by a slide projector, in black and white, on a large screen at a comfortable distance from the subject. The rate and duration of presentation were controlled by a timer attached to the projector.

A Waugh-Norman paradigm was employed. On each trial, a six-character series was presented sequentially. Each character was visible for approximately 1.5 seconds, and there was an interstimulus interval of .2 seconds. After a delay of four seconds, one of the first five characters in the series was presented again, as a "probe," and the subject was required to write down the character that had followed the probe in the series. An advantage of this procedure, compared with one in which the subject must write down the entire series, is that there is less opportunity for differences in handwriting ability to confound differences in recall. This was an important consideration in the case of the present subjects. Although they were able to recognize the characters used, they were not yet necessarily very adept at writing them down quickly.

Within each series type, a word at a different serial position served as the probe in each of the five series. The ordering of trials was randomized with respect to series type and probe position.

The subjects were tested individually. Before the actual test trials, the subject was given at least three practice trials, and the test did not begin until it was clear that the subject understood the task.

After the main experiment had been carried out, each subject was shown each of the 90 characters and asked to pronounce it. The average error rate for the 48 subjects was 2.5%. Nineteen subjects performed this task perfectly, and none had an error rate greater than 8%. Errors in the main experiment may thus be confidently interpreted as predominantly errors of recall rather than errors of recognition.

Results and Discussion

The results are shown in Figure 1 and Figure 2 (left). In Figure 1, each panel shows the percentage of correct responses at each serial position for one of the three types of trial, with reading ability as the parameter. In Figure 2, the percentages of correct responses for good readers and poor readers and for each trial type are collapsed across serial position.

An analysis of variance was performed on reading ability (good or poor), trial type (OS, PS, NS), and serial position (2 through 6) of the character probed for.

As would be expected from the results of many other experiments in serial recall of verbal material, performance generally declined across earlier serial positions (2 to 4) and improved across later positions (4 to 6), and there is a significant main effect of serial position [$F(4, 184) = 20.72$, $p < .001$].

Overall performance was best on the OS trials, slightly poorer on the NS trials, and considerably poorer on the PS trials, and there is a significant main effect of trial type [$F(2, 92) = 13.50$, $p < .001$]. A contrast analysis showed that there was no significant difference between performance on NS and on OS trials [$F(1, 46) < 1$], but that performance on PS trials was significantly poorer than performance on either NS trials [$F(1, 46) = 17.33$, $p < .001$] or OS trials [$F(1, 46) = 24.13$, $p < .001$]. Performance was especially poor for medial serial positions on PS trials, and there is a significant interaction between serial position and trial type [$F(8, 368) = 3.10$, $p < .005$].

As shown in Figure 2 (left), the overall performance of the good readers was much better than that of the poor readers, and there is a significant main effect of reading ability [$F(1, 46) = 11.95$, $p < .005$]. The effects of serial position on good readers and poor readers were similar, and there is no significant interaction between reading ability and serial position $F(4, 184) < 1$].

The effects of trial type on good readers and poor readers were different. While the good readers performed better than the poor readers on all three trial types, the difference in performance was significant for NS trials [$F(1,46) = 12.54$, $p < .001$] and for OS trials [$F(1,46) = 66.33$, $p < .05$], but not for PS trials [$F(1,46) < 1$]. There is an interaction between reading ability and trial type that falls just short of significance [$F(2, 92) = 3.00$, $p = .055$]. If just the NS and PS types are considered, this interaction is clearly significant [$F(1,46) = 5.68$, $p < .05$]: good readers were more affected by phonological similarity than were poor readers. In fact, good readers performed significantly better on NS trials than on PS trials [$F(1,24) = 18.53$, $p < .001$], but poor readers did not [$F(1,22) = 1.96$, $p > .05$].

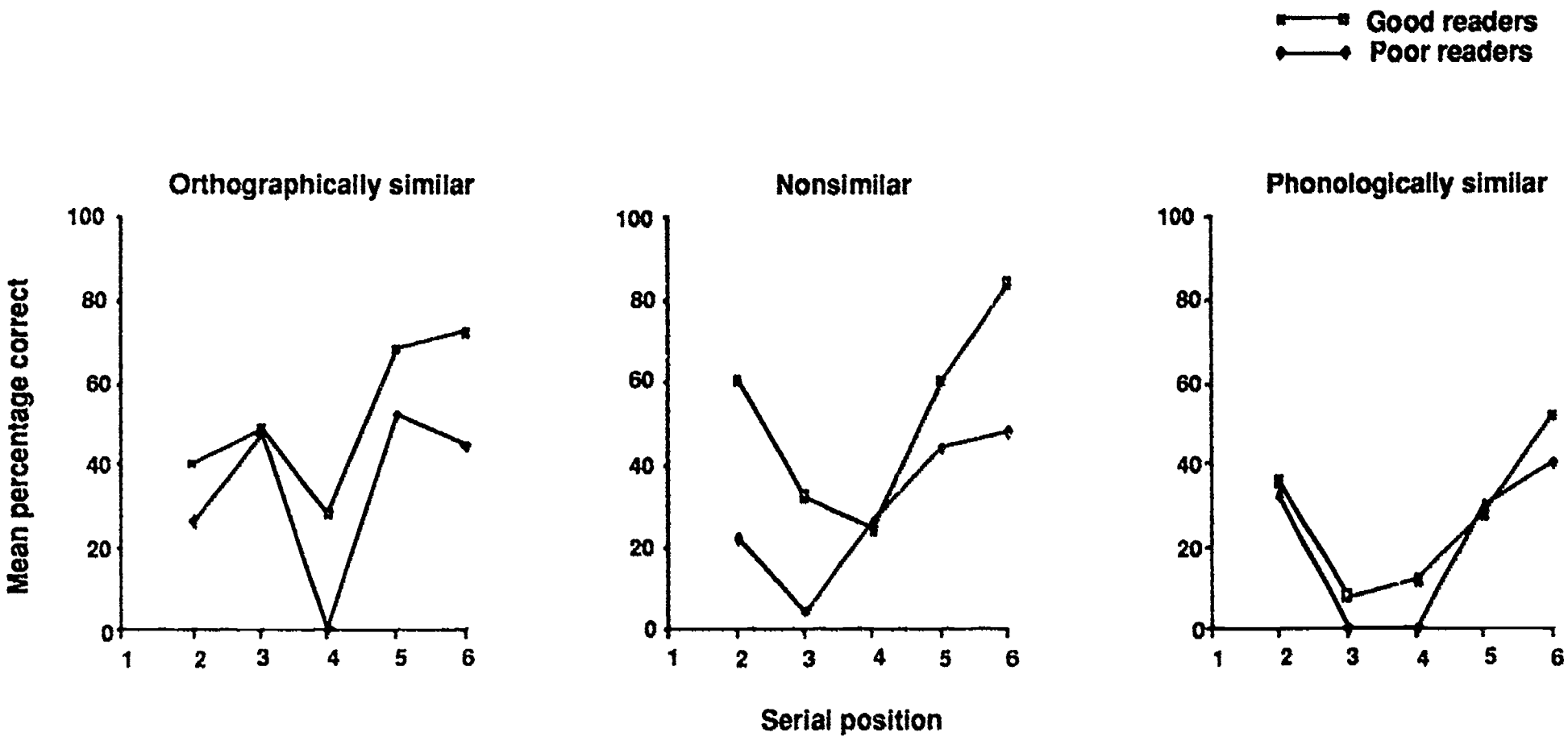


Figure 1. Percentage of correct recall of visually-presented Chinese characters by good readers and by poor readers, at each serial position, for OS, NS and PS trial types.

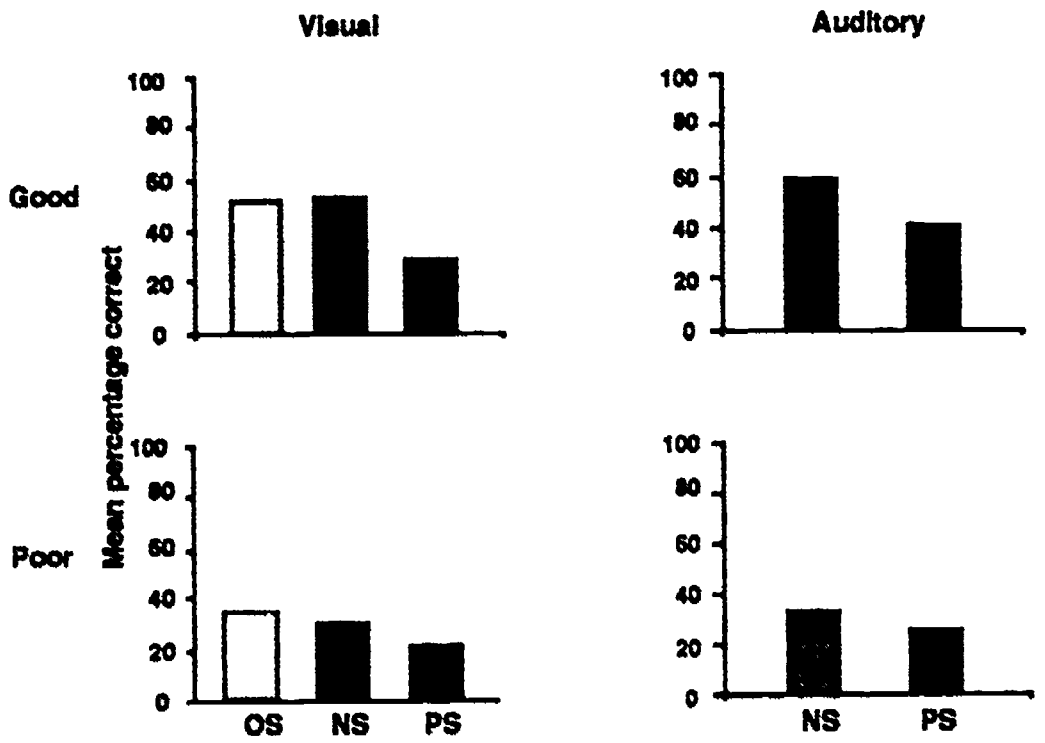


Figure 2. **Percentage of correct recall of visually-presented Chinese characters, and of the corresponding auditorially presented Chinese words, by good readers and by poor readers, for OS, NS, and PS trial types, combined across serial position (2 to 6).**

Neither group performed significantly differently on the OS trials than on the NS trials [good readers: $F(1,24) < 1$; poor readers: $F(1, 22) < 1$]. It might therefore be expected that a comparison of the PS and OS trials would parallel the comparison between PS and NS trials. But this proves not to be the case, because, as can be seen from Figure 2, the poor readers actually performed slightly *better* on the OS trials than would be expected from their performance on the NS trials. As a result, poor readers performed significantly better on OS trials than on PS trials [$F(1,22) = 7.36$, $p. < .05$], just as did good readers [$F(1,24) = 17.55$, $p < .001$]; and if just OS and PS types are considered, there is no significant interaction between reading ability and trial type [$F(1,46) = 2.891$, $p > .05$].

The results of Experiment 1 parallel the findings of Shankweiler et al. for recall of letters by good and poor readers of English. The ability to recall a string of characters correlates with reading ability in Chinese, just as the ability to recall a string of letters correlates with reading ability in English. Phonological similarity affects the performance of good readers more than it does the performance of poor readers. Orthographic similarity, not explored by Shankweiler et al. for English writing, but certainly a plausible potential source of confusion in the case of Chinese writing, in fact has no effect.

EXPERIMENT 2: AUDITORY PRESENTATION

Materials

The materials were the series of spoken monomorphemic Mandarin words corresponding to the five series of phonologically similar characters and the five series of nonsimilar characters in Experiment 1. These series were tape-recorded by the first author, a native speaker of Shanghainese who is fluent in Mandarin. The format in which the material was recorded paralleled the procedure of Experiment 1, but the order of trials was determined by a different randomization, and in each trial, a word at a different serial position served as the probe. A metronome was used to control the rate at which the words in a series were spoken. There was an interval of 2 seconds between the onsets of successive words in a series, and the onset of the probe word occurred 4 seconds after the onset of the last word of a series.

Procedure

The procedure was parallel to that of Experiment 1. The subject heard each series and the following probe word over a loudspeaker, and responded by speaking a word. This response was tape-recorded.

Results and Discussion

The results for the auditory presentation are shown in Figure 2 (right) and Figure 3.

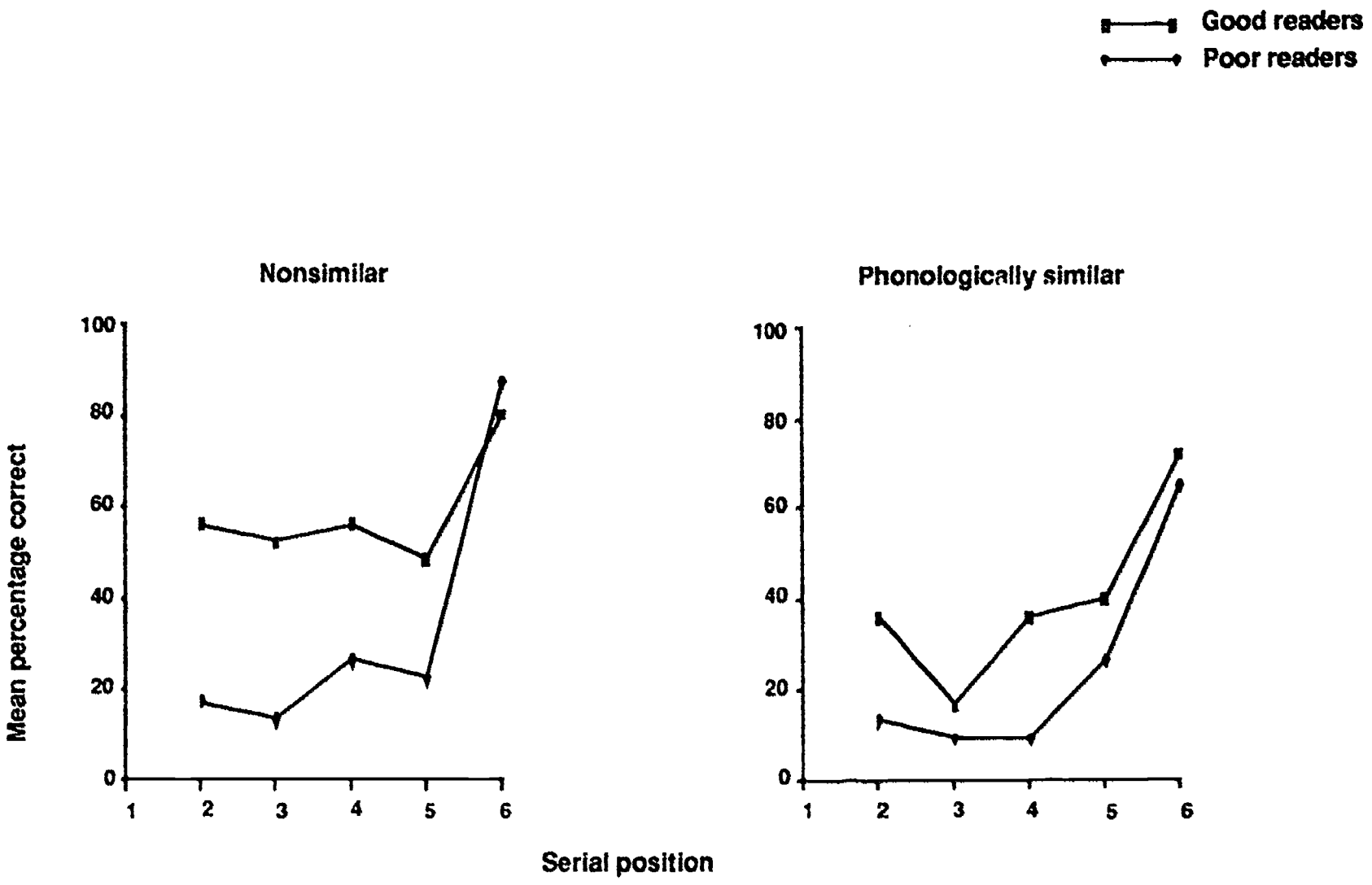


Figure 3. **Percentage of correct recall of auditorially-presented Chinese words by good readers and by poor readers, at each serial position, for NS and PS trial types.**

An analysis of variance was performed; the factors were the same as for the analysis in Experiment 1, except that only two trial types had to be considered. Recall is much better for position 6 than for earlier positions, and there is a significant main effect of serial position [$F(4, 43) = 16.63$, $p < .001$]. Performance on NS trials is significantly better than on PS trials [$F(1, 46) = 13.40$, $p < .001$]. (The graphs of the responses in the NS condition show an unexpected dip at position 5; this happened because, for this position in this particular sequence, we had inadvertently selected a word that was phonologically similar to the preceding word that served as the probe.) Good readers perform significantly better than poor readers [$F(1, 46) = 13.21$, $p < .001$]. But it should be noted that this is not true for position 6, at which the good readers and poor readers perform equally well. It may be that a spoken response for later positions in auditory presentation relies on echoic memory as well as, or rather than, phonological memory (Crowder & Morton, 1969). If so, differential effects of phonological similarity would be less obvious than with written responses to visually presented stimuli. At any rate, and in contrast with the results for visual presentation, good readers and poor readers are about equally disadvantaged by phonological similarity, and the interaction between reading ability and trial type is not significant. Nor is there any significant interaction for any other combination of factors.

Auditory and visual performance for each serial position are compared in Figure 4, combining across reading ability and NS and PS trial types. Another analysis of variance was made for the responses by all subjects to PS trials and NS trials in both experiments. In addition to the factors considered in the previous analyses, the within-subjects factor modality of presentation (visual or auditory) was added. Performance with auditory presentation is better than with visual presentation, and there is, accordingly, a significant main effect of modality [$F(1, 46) = 6.45$, $p < .05$].

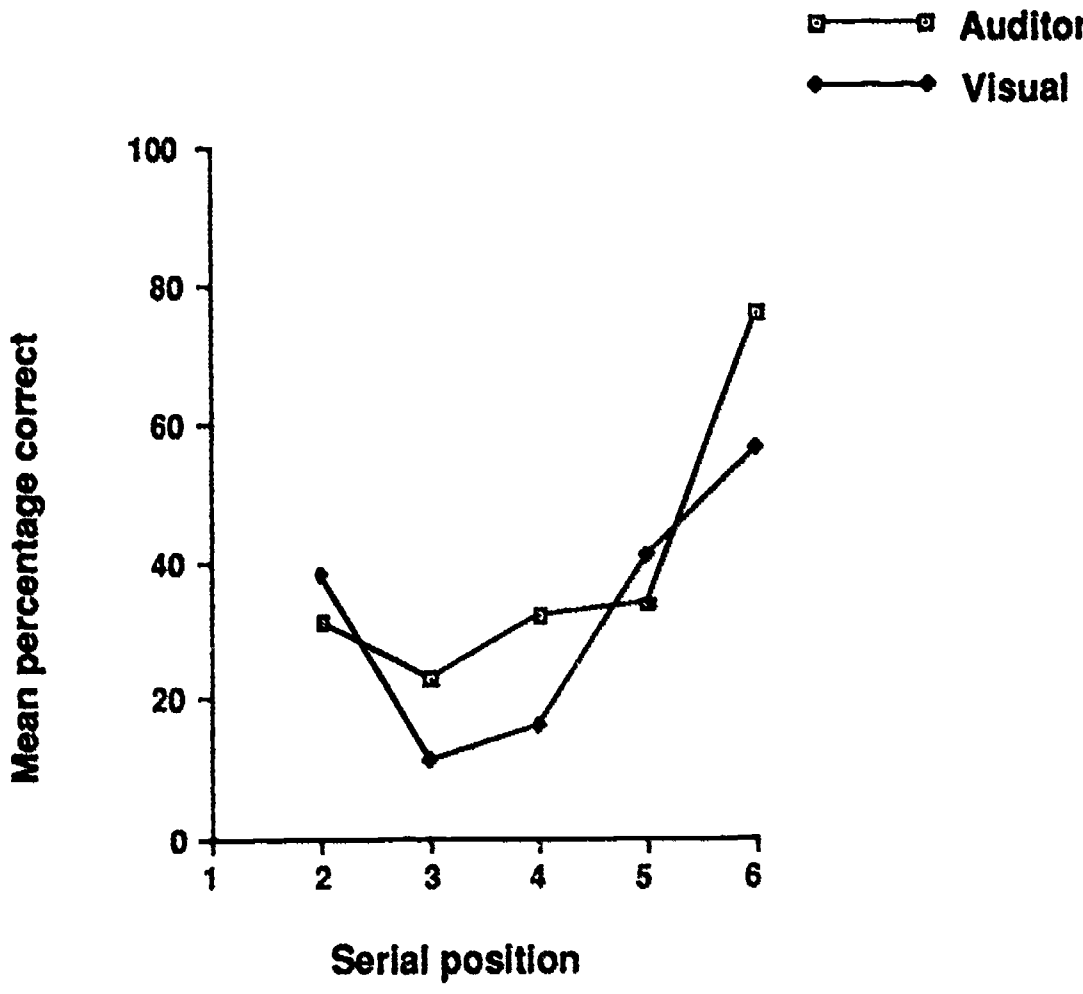


Figure 4. Percentage of correct recall of visually-presented Chinese characters and of the corresponding auditorially-presented Chinese words, at each serial position, combined across reading ability (good and poor) and trial type (NS and PS).

Serial position functions differ across modality, with performance in medial positions being better, relative to performance in early positions, in the auditory modality. Thus there is a significant interaction between modality and serial position [$F(4, 43) = 3.99$, $p < .01$]. Good readers perform well at the last position in both modes, whereas the performance of poor readers at this position is poor with visual presentation but, as we have just noted, equal to that of the good readers with auditory presentation. Hence there is a significant interaction between modality, serial position and reading ability [$F(4, 43) = 2.82$, $p < .05$]. There were no other interactions with modality, even though, given a significant interaction between trial type and reading ability with visual presentation, but not with auditory presentation, a three-way interaction among trial type, reading ability, and modality might have been expected.

In the auditory presentation, good and poor Chinese readers differed in their performance, just as Shankweiler et al.'s good and poor English readers did. Phonological similarity, as would be expected, reduced performance. But, probably for methodological reasons, a differential effect of phonological similarity on good readers and on poor readers was not demonstrated, in contrast with the finding of Shankweiler et al. Auditory performance was better than visual performance, but given necessary differences in experimental procedures, no special importance can be assigned to this result.

General Discussion

In serial-recall experiments in which the items to be remembered were Chinese characters and the corresponding spoken Mandarin words, and the subjects were young readers of Chinese, we have obtained results essentially parallel to those obtained by Shankweiler et al. (1979), in experiments in which the items were Roman letters and their spoken English names, and the subjects were young readers of English. We found that the ability of Chinese subjects to read Chinese, like the ability of English subjects to read English, correlated with the efficiency of the mechanism used for serial recall of verbal material, whether spoken or written. As has long been recognized, this mechanism is, in some important sense, phonological.

Our findings suggest several conclusions. First, this phonological mechanism is clearly needed for the reading of Chinese and presumably of other nonalphabetic orthographies, and the mechanism is of such importance that its relative efficiency distinguishes good readers from poor readers. This is what might have been expected, indeed, given earlier serial recall experiments with nonalphabetic orthographies, and given also experiments for both English and Chinese demonstrating that phonetic distractor tasks slow down detection of sentence anomaly (Kleiman, 1975; Tzeng et al., 1977). But it would still have been possible to maintain, before the present results, that reading was not the same process as short-term recall or sentence-anomaly detection, and, at least in the case of a nonalphabetic orthography, need not involve a phonological mechanism.

Second, there may not be very many different ways to read; perhaps only one. Given the apparent variety in the orthographies that have been and continue to be used by different cultures, it might have been supposed that there were many different possible cognitive strategies for reading and writing. But closer inspection of these orthographies reveals that there are really just two basic types, one of which employs syllabic units and the other, phonemic units (Gelb, 1963). Now, the present experiment indicates that even this significant structural difference is not crucial, for the reading of the two orthographic types relies on the same mechanism.

Third, the nature and cognitive status of this phonological mechanism needs to be reconsidered. According to a widely-held view, the mechanism is "working memory," used for cognitive problem-solving tasks, including the parsing and understanding of sentences. Working memory includes a phonological buffer in which words are stored while awaiting higher-level processing, as in sentence understanding, or simply when short-term verbatim retention is required, as when a name or a number is temporarily remembered (Baddeley & Hitch, 1974). On this thoroughly "horizontal" account, a memorial mechanism that happens to be phonological supports various higher-level cognitive processes, of which sentence-processing is merely one.

It has never been quite clear whether the contents of the phonological buffer in working memory were supposed to derive immediately from the spoken or written input, before lexical access, or from phonological specifications in lexical entries, after lexical access. On the former assumption, it is difficult to explain the results of short-term serial recall experiments with nonalphabetic writing; surely, the lexical entries for Chinese words would have to be accessed to get the segments into the phonological buffer. On the

latter assumption, the difficulty is that what would appear to be required for post-lexical parsing is the syntactic and semantic information stored in the lexicon for words, rather than the phonological information. The results of the present experiment further embarrass the "working memory" account. The problem is that the mechanism that appears to be used by readers of Chinese, as well as by recallers of Chinese and readers and recallers of English, deals in phonological segments, and is therefore inhibited by similarity of such segments in a serial recall task. Why should such mechanism be used at all for reading the Chinese orthography, in which the units are morphosyllabic, not segmental? If we insist on the working-memory account, we have to say that readers of Chinese will not be very efficient unless they are able to access and temporarily store information that is not directly specified in the input and would appear to be quite irrelevant to sentence-parsing and sentence understanding.

A way around this difficulty is to be found in a "vertical" account, according to which the phonological mechanism is not as a general-purpose "working memory" system, but a language module, in the sense of Fodor (1983). The language module controls the primary processes of speaking and listening and is exploited, we would claim, in the secondary processes of reading and writing. The language module provides cognition with a representation of the linguistic structure of an utterance, including its segmental phonological structure; it also represents in some way the meaning of the utterance. The operation of the module, in the presence of appropriate stimulation, is automatic and compulsory; it always outputs these representations, even if they are of no use on a given occasion. Thus, whether the module is used for the reading of Chinese or of English, its output will include phonological structure; variations in orthographic structure do not affect the character of the output. Moreover, because the representation of phonological structure is part of the output, there is no need to attribute its existence to the requirements of sentence parsing. The module doubtless employs various intermediate representations in its computations, but there is no reason to identify its output with any of them.

From this point of view, "short-term memory" and "working memory" are merely somewhat misleading names for further ways in which cognition can exploit the ability of the language module to compute representations. Verbal material is temporarily remembered by rehearsal, that is, by the repeated computation of a fresh linguistic representation from a previously computed, now decaying one. Propositions needed for problem-solving are maintained in the mind by rehearsing sentences that assert them.

A modular account of reading, to be sure, has problems of its own. It has to be explained how a system that is presumed to be biologically specialized for speaking and listening to speech can be effectively accessed using arbitrary and conventional signs in another modality. (For some speculation on this question, see Mattingly, 1991).

Finally, we would claim that it is perhaps for learning to read, rather than for the actual process of reading, that differences in orthographic structure are most significant. Mastering a sufficient inventory of Chinese characters requires years of memorization; it does not, however, require the segmental phonological awareness that alphabetic writing both demands and fosters in the reader (Mattingly, 1972). It is sometimes suggested that phonological awareness has some direct connection with the phonological mechanism discussed earlier, as if limitations of awareness might be explained by, or might themselves account for, limitations of this mechanism (Shankweiler & Crain, 1986). The present experiments, however, provide no encouragement for this "unitary" view, for they show that the phonological mechanism, that is, the language module, is necessary for the reading of an orthography for which segmental awareness is not necessary. It appears to be one thing for the module to produce linguistic representations efficiently; another, for the reader to become aware of the particular aspect of these representations that alphabets exploit: their segmental character.

REFERENCES

Baddeley, A. D. (1979). Working memory and reading. In P. A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), *Processing of visible language* (pp. 355-370). New York: Plenum.

Baddeley, A. D., & Hitch, G. B. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and activation* (Vol. 4, pp. 47-90). New York: Academic Press.

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology, 55*, 75-84.

Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British Journal of Psychology, 55*, 429-432.

Conrad, R. (1972). Speech and reading. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationships between speech and reading* (pp. 205-240). Cambridge, MA: MIT Press.

Crowder, R. G., & Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception & Psychophysics, 5*, 365-373.

Erickson D., Mattingly I. G., & Turvey, M. T. (1977). Phonetic activity in reading: An experiment with kanji. *Language and Speech, 20*, 384-403.

Fodor J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Gelb, I. J. (1963). *A study of writing* (2nd ed.). Chicago: University of Chicago Press.

Goodenough, F. L. (1926). *Measurement of intelligence by drawings*. New York: World Book Co.

Hanson, V. L. (1982). Short-term recall by deaf signers of American Sign Language: implications of encoding strategy for order recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 572-583.

Kimura, D. (1963). Right temporal lobe damage. *Archives of Neurology, 8*, 264-271.

Kintsch, W., & Buschke, H. (1969). Homophones and synonyms in short-term memory. *Journal of Experimental Psychology, 80*, 403-407.

Kleiman, G. M. (1975). Speech recoding in reading. *Journal of Verbal Learning and Verbal Behavior, 14*, 323-329.

Mann, V. A. (1985). A cross-linguistic perspective on the relation between temporary memory skills and early reading ability. *RASE Remedial and Special Education, 6(6)*, 37-42.

Mann, V. A., Liberman, I. Y., & Shankweiler, D. (1980). Children's memory for sentences and word strings in relation to reading ability. *Memory & Cognition, 8*, 329-335.

Mattingly, I. G. (1972). Reading, the linguistic process, and linguistic awareness. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationships between speech and reading* (pp. 133-147). Cambridge, MA: MIT Press.

Mattingly, I. G. (1991). Reading and the biological function of linguistic representations. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 339-346). Hillsdale, NJ: Lawrence Erlbaum.

Shankweiler D., Liberman, I. Y., Mark L. S., Fowler C. A., & Fischer, F. W. (1979). The speech code and learning to read. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 531-545.

Shankweiler, D., & Crain, S. (1986). Language mechanisms and reading disorder: a modular approach. *Cognition, 24*, 139-168.

Tzeng, O. J. L., Hung, D. L., & Wang, W. S.-Y. (1977). Speech recoding and reading Chinese characters. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 621-630.

Waugh, N. C., & Norman, D. H. (1965). Primary memory. *Psychological Review, 72*, 89-104.

Wickelgren, W. A. (1965). Distinctive features and errors in short-term memory for English vowels. *Journal of the Acoustical Society of America, 38*, 583-588.

Wickelgren, W. A. (1966). Distinctive features and errors in short-term memory for English consonants. *Journal of the Acoustical Society of America, 39*, 388-398.

FOOTNOTE

[†]Also University of Connecticut, Storrs.

APPENDIX: CHARACTER SERIES

A. Orthographically similar

富 宫 客 室 宝 安
fù gōng kè shì bǎo ān

样 村 林 材 枝 校
yàng cūn lín cái zhī xiào

伟 他 住 什 作 件
wěi tā zhù shén zuò jiàn

拼 技 报 投 拔 拨
pīn jì bào tóu bá bō

运 这 迈 远 过 进
yùn zhè mài yuǎn guò jìn

B. Nonsimilar

闹 院 更 徒 抗 查
nào yuàn gèng tú kàng chá

桌 墙 民 黑 同 忆
zhuō qiáng mín hēi tóng yì

千 上 盖 玩 澄 有
qiān shàng gài wán chéng yǒu

碰 会 斤 郭 菊 赛
pèng huì jīn guō jú sài

枯 里 虫 漂 分 马
kū lǐ chóng piāo fēn mǎ

C. Phonologically similar

向 让 傍 创 亮 望
xiàng ràng bàng chuàng liàng wàng

空 松 中 聪 东 工
kòng sōng zhōng cōng dōng gōng

貌 道 笑 教 料 绕
mào dào xiào jiào liào rào

床 航 防 王 狼 忙
chuáng háng fáng wáng láng máng

首 狗 走 斗 丑 口
shǒu gǒu zǒu dǒu chǒu kǒu

Haskins Laboratories Status Report on Speech Research
1990, SR-103/104, 165-172

Recall of Order Information by Deaf Signers: Phonetic Coding in Temporal Order Recall*

Vicki L. Hanson†

To examine the claim that phonetic coding plays a special role in temporal order recall, deaf and hearing college students were tested on their recall of temporal and spatial order information at two delay intervals. The deaf subjects were all native signers of American Sign Language. The results indicated that both the deaf and hearing subjects used phonetic coding in short-term temporal recall, and visual coding in spatial recall. There was no evidence of manual or visual coding by either the hearing or the deaf subjects in the temporal order recall task. The use of phonetic coding for temporal recall is consistent with the hypothesis that recall of temporal order information is facilitated by a phonetic code.

There is a strong tendency for normally-hearing adults to recode printed letters into a phonetic code in tasks of short-term serial recall of linguistic stimuli. These adults persist in using this form of memory representation even in some situations in which doing so has detrimental effects on their recall ability. For example, detrimental effects due to phonetic coding have been obtained when items rhyme (Conrad, 1962; Conrad & Hull, 1964; Healy, 1974) and when concurrent competing articulation is required (Conrad, 1972; Healy, 1977; Murray, 1967, 1968). Questions have arisen as to the reason(s) for the use of this code and whether other codes can effectively substitute for phonetic coding in short-term recall.

Studies with deaf subjects have provided one useful means of differentiating between some explanations for this phenomenon and of delineating the role of phonetic coding in short-term memory. One hypothesis tested with deaf subjects was the proposal that the use of a phonetic code reflects primary language experience (Shand, 1982; Shand & Klima, 1981). Contrary to the predictions of this hypothesis, the evidence indicates that deaf signers, for whom sign language is primary, do not always use sign coding. In particular, evidence that deaf signers recode printed words into a manual code in serial recall is lacking, while other evidence indicates that in some cases deaf native signers will use phonetic coding in short-term recall (Hanson & Lichtenstein, 1990). Both of these findings are inconsistent with the primary language hypothesis.

Another prominent hypothesis for the use of phonetic coding in serial recall is that it reflects the sequential character of speech (Baddeley, 1979; Crowder, 1978; Healy, 1975; Penney, 1985, 1989). According to this hypothesis, there are properties of a phonetic code that make it particularly well-suited for temporal recall. That is, the auditory/vocal aspects of speech, being temporally arrayed, promote recall of temporal information.

If a phonetic code is well suited for maintaining temporal order information, then deaf individuals, who would be expected to have difficulty in using a phonetic code, ought to have difficulty in maintaining temporal order information. Evidence

This research was supported by Grant NS-18010 from the National Institute of Neurological and Communicative Disorders and Stroke to the author and by Grant HD-01994 from the National Institute of Child Health and Human Development to Haskins Laboratories. For providing help at Gallaudet University, I am grateful to Drs. Michael Karchmer and Patrick Cox. The work of Nancy Fishbein, Deborah Kuglitsch, Eliza Goodell, and Dan Weiss in testing the subjects is gratefully acknowledged.

consistent with this claim has been repeatedly obtained (for a review, see Cumming & Rodda, 1985). In tasks requiring the serial recall of linguistic stimuli (whether printed words, digits, signs, fingerspelling, or pictures), deaf subjects consistently have been found to recall fewer items than hearing subjects, even when confounds with spatial order recall are eliminated (e.g., Bellugi Klima, & Siple, 1975: Blair, 1957; Hanson, 1982; Krakow & Hanson, 1985; McDaniel, 1980; Pintner & Paterson, 1917; Wallace & Corballis, 1973; Withrow, 1968). However, in the temporal recall of nonsense stimuli, deaf subjects have not been found to recall fewer items than hearing subjects (McDaniel, 1980; Olsson & Furth, 1966).

It appears to be only in the recall of linguistic stimuli that deaf subjects are at a disadvantage. Deaf subjects are at no disadvantage, compared with hearing subjects, in spatial recall of stimuli, regardless of whether the stimuli are linguistic or nonsense (Carey & Blake, 1974; Das, 1983). Indeed, there has even been some evidence that deaf individuals are at an advantage in spatial recall (Blair, 1957).

In a series of experiments, Healy (1975, 1977, 1978, 1982) convincingly demonstrated that hearing subjects use a phonetic code for the short-term retention of temporal, but not spatial, order information. Using procedures that isolated temporal and spatial information, Healy found evidence of phonetic confusions (e.g., B for P and F for S) in the recall of temporal order information. These phonetic confusions were limited to short retention intervals, suggesting rapid decay of the phonetic component to short-term temporal recall. Healy found that recall accuracy, correspondingly, dropped significantly at the longer intervals during temporal order recall.

The question addressed in the present experiment is whether deaf signers similarly use phonetic coding specifically for the short-term retention of temporal information. Such a finding would present a strong argument for the importance of phonetic coding in temporal recall. For prelingually, profoundly deaf individuals, developing the use of a speech code is a formidable task. Deaf individuals who use sign have potentially another memory code—a manual code—more readily available. Thus, if a manual code can provide an effective medium for retaining temporal order information, then deaf signers would be expected to use it. In the case of the present study, this manual code would be one based on the handshapes of the American manual alphabet. In this alphabet, there is a handshape for each letter, and words can be spelled out, letter by letter, on the hand. Shown in Figure 1 are some examples of letter handshapes from the American manual alphabet.

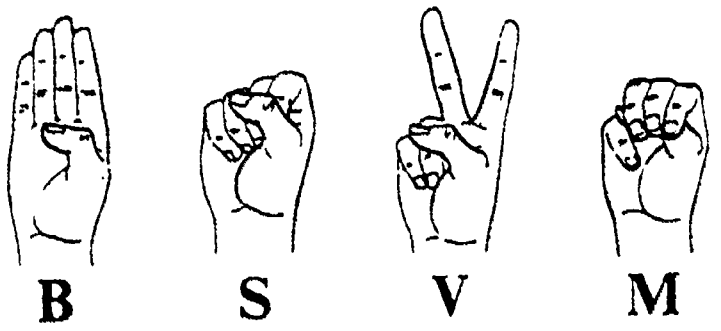


Figure 1. The handshapes B S V M of the American manual alphabet.

Studies have shown that some deaf signers use phonetic coding in short-term recall of printed material (for a review, see Hanson, 1989, in press). The present study represents an attempt to clarify the conditions under which this phonetic coding is used; specifically, recall of temporal order information was isolated from recall of spatial information at two short-term retention intervals. Another question investigated was whether or not, in contrast to hearing subjects, deaf subjects make use of manual or visual coding alternatives in the recall of temporal information. Finally, the use of phonetic, visual, and manual codes by deaf signers in spatial recall was also examined.

In the present study, the procedures of Healy (1975) were used to isolate temporal and spatial recall. The general procedure was that on each trial, a series of four letters were shown to the subjects, one letter at a time. After all four letters were shown, there was an interference task, in which the subjects were required to name each digit in a string of digits. In the present study, two interference intervals were used—a short one (3 digits) and a longer one (15 digits). Following the interference task, subjects were asked to write down the four letters from that trial.

Letter sets designed specifically to test for phonetic, manual, and visual confusions were employed. The identity of the four letters was always known to the subjects, thus eliminating a

confound with recall of item information. The subjects were a group of deaf college students and a control group of hearing college students. These subjects participated in both recall tasks, thus allowing a within-subjects comparison of temporal and spatial order recall.

Method

Subjects

The deaf subjects were 8 students at Gallaudet University who were paid for their participation. All were congenitally and profoundly deaf, with a hearing loss of 90 dB or greater, better ear average. In addition, all had two deaf parents and were native signers of American Sign Language (ASL). Their median reading proficiency was grade 9.2 (range 3.3 - 12.9+), according to the comprehension subtest of the *Gates MacGinitie Reading Tests* (1978, Level F, Form 2) administered to each subject. These subjects were thus excellent deaf readers when rated against national surveys of the reading proficiency of deaf students (Karchmer, Milone, & Wolk, 1979). To obtain a measure of speech production ability for individual subjects, speech intelligibility measures were obtained from school records. This measure rates the speech production ability of individuals on a scale of 1 - 5, in which 1 is readily intelligible to listeners and 5 is unintelligible. According to school records, the speech intelligibility ratings of these 8 subjects were as follows: One had speech that was rated a "3," 2 had speech that was rated a "4," 3 had speech that was rated a 5," and 2 had speech ratings that were listed as "NONE." Thus, with one exception (the subject with the rating of "3"), the subjects here had poorly intelligible speech as judged by listeners.

The hearing subjects were 8 undergraduates from the University of Connecticut who were paid for their participation. All had normal hearing.

Stimuli

The stimuli for both the temporal and spatial order recall tasks were the upper-case letters B, S, V, and M. The handshapes corresponding to these letters are shown in Figure 1. This character set provided a test of phonetic similarity with the two letters B-V (Wolford & Hollingsworth, 1974). These two letters are manually (Richards & Hanson, 1985) and visually (Wolford & Hollingsworth, 1974) distinct. This letter set also provided a test of manual similarity with the two letters M-S, which are similar in the American manual alphabet (Richards & Hanson, 1985), but are phonetically and visually distinct (Wolford & Hollingsworth, 1974). Finally, this stimulus set provided two tests of visual similarity with the letter pairs V-M and S-B (Wolford & Hollingsworth, 1974) both of which are phonetically (Wolford & Hollingsworth, 1974) and manually (Richards & Hanson, 1985) distinct.

The same four consonants appeared on each trial in both the temporal and the spatial order recall tasks. In each task, a test sequence of 48 trials was generated. The 24 permutations of these letters each appeared once at the short retention interval (3 digits) and once at the long retention interval (15 digits). The order of trials was randomized, with the constraint that in every block of eight trials there were four trials at the short interval and four trials at the long interval. The digits presented during the retention interval were the digits 1 - 9. No digit occurred more than once in succession.

Procedure

The subjects were individually tested in the two tasks on successive days. Half of the subjects received the temporal order recall task first; half received the spatial order recall task first. The reading comprehension subtest of the *Gates-MacGinitie Reading Tests* was administered to the deaf subjects on the 2nd day of testing.

Stimulus presentation in both the temporal recall and spatial recall condition was controlled by a microcomputer. A trial began with four horizontally-arrayed boxes shown on the monitor with computer graphics. Each box was 5 in. high × 2 in. wide. The four letters of a trial then appeared one at a time in these four boxes. The letters were 1⅜ in. × ¾ in. Each letter was presented for 1000 ms, followed by a 1,000 ms ISI. The digits were presented simultaneously, following the fourth ISI, appearing as a string of digits. The presentation duration averaged 400 ms per digit, such that the 3 digits of the short interference interval were presented for 1,200 ms and the 15 digits of the long interval were presented for 6,000 ms. Following the offset of the digits, a message appeared on the computer screen instructing the subjects to "write your answer now." After a 16-sec interval during which subjects wrote their responses, the next trial began.

Instructions were signed for the deaf subjects by a deaf experimenter, a native signer of ASL, and they were spoken for the hearing subjects by a hearing experimenter. The subjects were instructed that on every trial they would see four letters, one at a time, and that these letters would be followed by a series of single-digit numbers—either 3 or 15 digits. Deaf subjects were told to simultaneously sign and mouth (pronounce

without voicing) each letter and digit. Hearing subjects were told to pronounce aloud each letter and digit. The subjects were told that following offset of the digits, they were to write the four letters in the answer booklets provided. They did not have to fill in the four boxes sequentially. Each page in the answer booklets had 12 rows of four boxes drawn on it.

For the temporal order recall task, subjects were told that they would see the letters B, S, V, and M on every trial, and that the B would always appear in the left-most box, the S in the next box, the V in the next box, and the M in the right-most box. They were told to write the letters in the boxes to show the temporal order in which the letters appeared—that is, to write the first letter they saw in the first (i.e., left-most) box on their answer sheets, the second letter they saw in the second box, and so forth.

For the spatial order recall task, the subjects were told that they would see the same letters on every trial—B, S, V, and M. They were also told that the B would always appear first, the S second, the V third, and the M fourth. The spatial location of each of these four letters would vary from trial to trial. The subjects were told to write the letters in the boxes to show the left-to-right spatial order in which the letters appeared.

The subjects received four practice trials in each condition before beginning the test trials. The practice trials were taken from the same letter and digit sets as were used in the experimental trials.

Results

Responses were scored as incorrect if the correct letter did not appear in the correct serial position. Table 1 gives the percentage correct responses for deaf and hearing subjects in the temporal and spatial order recall conditions at the two retention intervals. An arcsine transformation was applied to this accuracy data, and an analysis of variance was performed for the between-subjects factor of group (deaf, hearing) and the within-subjects factors of recall condition (temporal, spatial) and retention interval (3, 15 digits). The analysis indicated a main effect of retention interval, $F(1,14) = 69.38$, $MS_e = .0196$, $p < .001$, as well as an interaction of recall condition × retention interval, $F(1,14) = 11.11$, $MS_e = .0135$, $p < .005$. This interaction reflected a larger decline in accuracy as the retention interval increased in the temporal order recall condition than in the spatial order recall condition. No significant interactions involving subject group emerged (all $ps > .10$).

Table 1. *The percentage of correct responses in the temporal order and spatial order recall tasks.*

	Temporal Order		Spatial Order	
Interval	Hearing	Deaf	Hearing	Deaf
3 digits	96.1	97.4	95.6	92.3
15 digits	80.2	79.2	90.5	83.2

The conditional probabilities of making confusions related to the phonetic, manual, or visual similarity of the stimulus letters were next computed for each subject, then an analysis of variance was performed on these probabilities. The conditional probability of making a phonetic error was the percentage of incorrect responses of V for B or B for V; these two letters were phonetically similar, but manually and visually distinct. That is, for errors that were made on the letters V and B, this was the percentage of responses with the phonetically similar letter. The conditional probability of making a manual error was the percentage of incorrect responses of S for M or M for S; these two letters are similar manually, but are phonetically and visually distinct. Finally, the conditional probability of visual errors was the percentage of incorrectly responding V for M, M for V, S for B, and B for S; these letter pairs are similar visually, but they are manually and phonetically distinct. The conditional probability of these visual confusions was computed as the percentage of errors on these four letters that were visually similar. The conditional probabilities for the two subject groups for each error type are shown in Table 2 for both temporal and spatial order recall.

Table 2. *Conditional probabilities of each error type for deaf and hearing subjects in the temporal order and spatial order recall conditions.*

	RECALL CONDITION					
	Temporal Order			Spatial Order		
Interval	Hearing	Deaf	*M*	Hearing	Deaf	*M*
3 digits						
Phonetic	60.7	68.8	64.7*	14.3	4.3	9.3
Manual	8.3	34.4	21.4	14.3	13.5	13.9
Visual	21.9	10.4	16.1	70.5	51.6	61.1*
15 digits						
Phonetic	57.1	35.6	46.3	8.7	19.6	14.2
Manual	47.4	41.3	44.4	24.2	23.0	23.6
Visual	21.6	31.3	26.5	61.5	56.3	58.9*

*Greater occurrence than other error types at that interval, $p < .01$.

The analysis of these conditional probabilities on the factors of subject group × error type (phonetic, manual, visual) × delay interval (3 digits, 15 digits) × recall condition (temporal, spatial) revealed an interaction of error type × recall condition, $F(2,28) = 18.90$, $MS_e = 1,477.97$, $p < .001$. This interaction indicated that different error types predominated for temporal and spatial order recall. To determine the source of the interaction of error type × recall condition, the temporal and spatial order data were analyzed separately.

In the temporal order condition, there was a significant main effect of error type, $F(2,28)=7.28$, $MS_e = 1,332.04$, $p < .005$, that was qualified by an interaction of error type × delay interval that approached significance, $F(2,28) = 2.87$, $MS_e = 1,258.13$, $p < .08$. Separate post hoc analyses of the short and long delay intervals revealed significantly more phonetic confusions than either manual or visual confusions at the short interval (Newman-Keuls, $p < .05$), but no difference in the frequency of the three types of confusions at the long interval (Newman-Keuls, $p > .05$). There were no significant effects involving subject group in this analysis of the temporal order condition (all ps > .10).

In the spatial order condition, there was also a main effect of error type, $F(2,28)=15.39$, $MS_e = 1,413.83$, $p < .001$, this time reflecting more visual confusions then either phonetic or manual confusions (Newman-Keuls, $p < .05$). The proportions of phonetic and manual confusions were not significantly different from each other (Newman-Keuls, $p > .05$). In this spatial order condition, the interaction of error type × delay interval was not significant, ($F < 1$), indicating a similar pattern of errors at the two delay intervals. As with the temporal order condition, there were no significant effects involving subject group (all $ps > .10$).

DISCUSSION

Consistent with Healy (1975, 1982), evidence was found in the present study for the use of a phonetic code in temporal order recall among the hearing subjects, as indicated by the predominance of phonetic confusions in temporal recall at the short interval. This phonetic code appeared to decay relatively quickly, with the consequence that temporal order recall ability declined significantly after only a few seconds of interpolated activity.

Deaf children receive training in lipreading and speaking. They also have everyday exposure to watching other people speak. Through such experiences, they may pick up information about the phonetic structure of language. Of interest here is the finding that the deaf subjects used a phonetic code, as indicated by the predominance of phonetic confusions in temporal recall at the short interval. As with the hearing subjects, this phonetic code appeared to decay relatively quickly. Associated with this decay of the phonetic code, the accuracy of the deaf subjects declined significantly at the long interval.

The pattern of recall confusions indicated different coding in the spatial than in the temporal order recall task for subjects in both groups. There was no indication of phonetic coding in spatial order recall. Rather, the error pattern showed a predominance of visual confusions in letter recall at both delay intervals. Thus, there was evidence for the use of a relatively long-lasting visual code mediating spatial recall for subjects in both groups.

As noted previously, deaf subjects typically perform more poorly on temporal recall of linguistic stimuli than hearing subjects do. Yet, in the present study, recall accuracies of the deaf and hearing subjects were comparable. There are two factors, either or both of which may have contributed to this unexpected result. The first is the use of phonetic coding among the deaf subjects. Previous research has indicated that deaf subjects who do not show evidence of the use of phonetic coding tend to recall fewer items than hearing subjects who do use phonetic coding (e.g., Conrad & Rush, 1965; Wallace & Corballis, 1973). The accuracy of deaf subjects may also be less even when phonetic coding is used, although there is evidence that as deaf subjects' use of phonetic coding increases, temporal recall accuracy improves (Conrad, 1979; Hanson, 1982; Hanson & Lichtenstein, 1990). In the present study, where no differences in the use of phonetic coding between hearing and deaf subjects were obtained, the deaf subjects' temporal recall accuracy was comparable to that of the hearing subjects. The second factor that may have contributed to the comparable accuracy of the two groups is the overall high level of accuracy of both deaf and hearing subjects. This level of accuracy is higher than is generally obtained with this paradigm (see Healy, 1975, 1982), most likely due to the use of longer stimulus presentations in this study than in others. These longer intervals likely allowed for more phonetic rehearsal during stimulus presentation. In particular, ceiling effects of hearing subjects may have masked any potential group differences.

Anecdotally, it is worth reporting that one difference in the performance of the hearing and the deaf subjects in the present study was that the temporal order recall task appeared more "natural" to the hearing subjects, while the spatial order task seemed more "natural" to the deaf subjects. That is, when giving instructions to the subjects, the hearing subjects assumed a temporal recall task. It required elaboration for them to understand what to do in the spatial order task. In contrast, many deaf subjects assumed a spatial recall task, and needed elaboration of the temporal order instructions. This observation is consistent with experimental evidence of differences in temporal and spatial order preference for hearing and deaf subjects (O'Connor & Hermelin, 1972, 1973). Despite the preference on the part of the deaf subjects for spatial order recall, their accuracy was not significantly better in this condition than in the temporal order recall condition (see also Das, 1983). Similarly for the hearing subjects, their preference for temporal order recall did not translate into greater accuracy in this condition than in the spatial order recall condition, although, as noted previously, there may have been ceiling effects obscuring possible differences here.

The deaf subjects' use of a phonetic code in temporal recall is consistent with the claim that recall of temporal order information for linguistic stimuli is facilitated by the use of a phonetic code. For deaf subjects, the acquisition and use of a phonetic code is extremely difficult. Thus, they would be expected to use visual or manual codes, if such codes were effective. Yet, in the present study, evidence was not obtained that the deaf subjects, all native signers of ASL, relied on these alternatives for temporal order recall.

The finding that the deaf subjects used a phonetic code is especially impressive, given the fact that most of them had speech that was rated as only poorly intelligible. These intelligibility ratings, however, are based on listeners' ratings of intelligibility, and may not reflect the extent to which an individual deaf subject can effectively use a phonetic code to mediate temporal recall. The fact that speech production ability does not reflect ability to use a phonetic code is dramatically demonstrated by research with some brain damaged patients who have lost the ability to produce speech. These patients can still retain the ability to use a phonetic code (e.g., Baddeley & Wilson, 1985; Bishop & Robson, 1989; Martin, 1987).

There is reason to believe that the results obtained here might not have been obtained if less skilled deaf readers had been tested. Within the deaf population, the evidence indicates that phonetic coding is used primarily by good readers, whether beginning readers (Hanson, Liberman, & Shankweiler, 1984), high school students (Conrad, 1979; McDermott, 1984), or college students (Lichtenstein, 1985). The subjects of the present study were college students, and most were excellent readers when rated against national norms for the reading levels of deaf adults (Karchmer et al., 1979). The importance of the present study, therefore, is not the suggestion that phonetic coding in temporal recall will be characteristic of all deaf subjects, but rather the finding that, for those deaf subjects who *are* able to use a phonetic code, that this code is specifically used for temporal recall.

There were no significant effects of manual coding for either temporal or spatial recall. In the short interval condition, where temporal order recall was best, the incidence of manual confusions by the deaf subjects was only half the incidence of phonetic confusions. Were the deaf subjects using manual coding that the experimental procedures failed to detect? This possibility cannot be completely ruled out solely by failure to find an effect of manual coding. For example, it is possible that the letters used to test sign confusability were not similar enough to produce confusions. Although the letters M and S have been judged to be manually similar by deaf subjects such as those of the present study (Richards & Hanson, 1985), it is possible that this pair was not *as* similar as the phonetically or visually similar pairs. Indeed, such a ranking of similarity across stimuli is not possible. It should be noted, however, that the letters M and S were part of a stimulus set previously found to produce performance decrements in written letter recall among deaf children (Hanson et al., 1984).

Some studies, done with self-report (Lichtenstein, 1985) and observation of overt rehearsal (e.g., Liben & Drury, 1977; Locke & Locke, 1971), have suggested that deaf subjects may use manual rehearsal in the temporal recall of letters and words. However, such reports and observations do not, necessarily, indicate the use of an internal code based on signs. That is, rather than being an indicator of internal coding, overt manual rehearsal may, at least in some cases, be serving as a supplemental storage mechanism. This overt use of sign may provide some recall of

information in addition to that supplied by the internal code. For example, in various studies over the years, the author has noticed that deaf subjects will use memory "tricks" (unless specifically directed not to do so), such as manually recording some stimulus letters on their hands, nd keeping the fingers in position for these letters throughout the stimulus sequence while memorizing other letters in the sequence. Similar observations were reported by Locke and Locke (1971). In these cases, the manual signal appears to serve not as an internal code, but rather as a visible reminder of the stimuli.

This discussion should not be taken to mean that a manual code based either on signs or on handshapes of the manual alphabet cannot serve as a short-term memory code. There is clear evidence in the literature for manual coding in short-term memory studies. If we look closely at these studies, however, we notice that a pattern begins to emerge. For signed stimuli, evidence of sign intrusions or decrements related to the formational similarity of signs has been reported in temporal recall (Bellugi et al., 1975; Hamilton & Holzman, 1989; Hanson, 1982; Krakow & Hanson, 1985; Shand, 1982). Recall accuracy for a sequence of signs in these studies, however, tends to be poorer than hearing subjects' recall accuracy for a sequence of words. Moreover, correlations between the use of sign coding and memory span have not been demonstrated (Kyle, 1980). In contrast to these studies with signed stimuli, there has been no clear evidence of manual coding obtained in studies that have examined the temporal order recall of printed letters or words (Hanson & Lichtenstein, 1990). Evidence of manual coding of print has generally been obtained only under conditions in which temporal order recall is not required. For example, evidence of sign coding has been obtained in paired associate and free recall tasks, facilitating, in these cases, the learning of items that have formationally similar signs (Conlin & Paivio, 1976; Moulton & Beasley, 1975; Odom, Blanton, & McIntyre, 1970; Putnam, Iscoe, & Young, 1962).

In conclusion, the finding that the deaf college students in the present study used a phonetic code specifically in temporal recall, despite their difficulty in using speech and despite their having a manual code code available to them as an alternative short-term memory code, is consistent with the claim that a phonetic code is particularly well-suited for recall of temporal order information (Baddeley, 1979; Crowder, 1978; Healy, 1975; Penney, 1985, 1989). This evidence also adds to a growing body of literature indicating that deaf subjects have available to them a variety of short-term memory coding options, the use of which varies as a function of specific subject characteristics (e.g., reading proficiency), stimulus characteristics (e.g., signed vs. print stimuli), and task characteristics (e.g., temporal vs. spatial order recall).

REFERENCES

Baddeley, A. D. (1979). Working memory and reading. In P. A. Kolers, M. Wrolstad, & H. Bouma (Eds.), *Processing of visible language* (Vol. 1, pp. 355-370). New York: Plenum.

Baddeley, A., & Wilson, B. (1985). Phonological coding and short-term memory in patients without speech. *Journal of Memory and Language, 24*, 490-502.

Bellugi, U., Klima, E. S., & Siple, P. (1975). Remembering in signs. *Cognition, 3*, 93-125.

Bishop, D. V. M., & Robson, J. (1989). Unimpaired short-term memory and rhyme judgement in congenitally speechless individuals: Implications for the notion of "articulatory coding." *Quarterly Journal of Experimental Psychology, 41*, 123-140.

Blair, F. X. (1957). A study of the visual memory of deaf and hearing children. *American Annals of the Deaf, 102*, 254-263.

Carey, P., & Blake, J. (1974). Visual short-term memory in the hearing and the deaf. *Canadian Journal of Psychology, 28*, 1 - 14.

Conlin, D., & Paivio, A. (1975). The associative learning of the deaf: The effects of word imagery and signability. *Memory & Cognition, 3*, 335-340.

Conrad, R. (1962). An association between memory errors and errors due to acoustic masking of speech. *Nature, 193*, 1314-1315.

Conrad, R. (1972). Speech and reading. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationships between speech and reading* (pp. 205-240) Cambridge, MA: MIT Press,.

Conrad, R. (1979). *The deaf schoolchild*. London: Harper & Row.

Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British Journal of Psychology, 55*, 429-432.

Conrad, R., & Rush, M. L. (1965). On the nature of short-term memory encoding by the deaf. *Journal of Speech and Hearing Disorders, 30*, 336-343.

Crowder, R. G. (1978). Language and memory. In J. F. Kavanagh & W. Strange (Eds.), *Speech and language in the laboratory, school, and clinic* (pp. 331-376). Cambridge, MA: MIT Press.

Cumming, C. E., & Rodda, M. (1985). The effects of auditory deprivation on successive processing. *Canadian Journal of Behavior Science, 17*, 232-245.

Das, J. P. (1983). Memory for spatial and temporal order in deaf children. *American Annals of the Deaf, 128*, 894-899.

Gates-MacGinitie Reading Tests (2nd ed.). (1978). Boston: Houghton Mifflin.

Hamilton, H., & Holzman, T. G. (1989). Linguistic encoding in short-term memory as a function of stimulus type. *Memory & Cognition, 17*, 541-550.

Hanson, V. L. (1982). Short-term recall by deaf signers of American Sign Language: Implications of encoding strategy for order recall. *Journal of Experimental Psychology: Learning, Memory and Cognition, 8*, 572-583.

Hanson, V. L. (1989). Phonology and reading: Evidence from profoundly deaf readers. In D. Shankweiler & I. Y. Liberman (Eds.), *Phonology and reading disability: Solving the reading puzzle* (pp. 69 - 89). Ann Arbor: University of Michigan Press.

Hanson, V. L. (in press). Phonological processing without sound. In S. Brady & D. Shankweiler (Eds.), *Phonological processes in literacy*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hanson, V. L., Liberman, I. Y, & Shankweiler, D. (1984). Linguistic coding by deaf children in relation to beginning reading success. *Journal of Experimental Child Psychology, 37*, 378-393.

Hanson, V. L., & Lichtenstein, E. H. (1990). Short-term memory coding by deaf signers: The primary language coding hypothesis reconsidered. *Cognitive Psychology, 22*, 211-224.

Healy, A. F. (1974). Separating item from order information in short-term memory. *Journal of Verbal Learning and Verbal Memory, 13*, 644-655.

Healy, A. F. (1975). Coding of temporal-spatial patterns in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14*, 481-495.

Healy, A. F. (1977). Pattern coding of spatial order information in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 16*, 419-437.

Healy, A. F. (1978). A Markov model for the short-term retention of spatial location information. *Journal of Verbal Learning and Verbal Behavior, 17*, 295-308.

Healy, A. F. (1982). Short-term memory for order information. In G. H. Bower (Ed.), *The psychology of learning and motivation:* (Vol. 16, pp. 191-238). New York: Academic Press.

Karchmer, M. A., Milone, M. N., Jr., & Wolk, S. (1979). Educational significance of hearing loss at three levels of severity. *American Annals of the Deaf, 124*, 97-109.

Krakow, R. A., & Hanson, V. L. (1985). Deaf signers and serial recall in the visual modality: Memory for signs, fingerspelling, and print. *Memory & Cognition, 13*, 265-272.

Kyle, J. G. (1980). Sign coding in short term memory in the deaf. In B. Bergman & I. Ahlgren (Eds.), *Proceedings of the first international symposium on sign language research*. Stockholm: The Swedish National Association of the Deaf.

Liben, L. S., & Drury, A. M. (1977). Short-term memory encoding strategies of the deaf. *Journal of Experimental Child Psychology, 24*, 60-73.

Lichtenstein, E. (1985). Deaf working memory processes and English language skills. In D. Martin (Ed.), *Cognition, education, and deafness* (pp. 111-114). Washington, DC: Gallaudet College Press.

Locke, J. L., & Locke, V. L. (1971). Deaf children's phonetic, visual, and dactylic coding in a grapheme recall task. *Journal of Experimental Psychology, 89*, 142-146.

McDaniel, E. D. (1980). Visual memory in the deaf. *American Annals of the Deaf, 125*, 17-20

McDermott, M. J. (1984). *The role of linguistic processing in the silent reading act: Recoding strategies in good and poor deaf readers*. Unpublished doctoral dissertation, Brown University.

Martin, R. C. (1987). Articulatory and phonological deficits in short-term memory and their relation to syntactic processing. *Brain and Language, 32*, 159-192.

Moulton, R. D., & Beasley, D. S. (1975). Verbal coding strategies used by hearing-impaired individuals. *Journal of Speech and Hearing Research, 18*, 559-570.

Murray, D. J. (1967). The role of speech responses in short-term memory. *Canadian Journal of Psychology, 21*, 263-276.

Murray, D. J. (1968). Articulation and acoustic confusability in short-term memory. *Journal of Experimental Psychology, 78*, 679-689.

O'Connor, N., & Hermelin, B. (1972). Seeing and hearing and space and time. *Perception & Psychophysics, 11*, 46-48.

O'Connor, N., & Hermelin, B. (1973). The spatial and temporal organization of short-term memory. *Quarterly Journal of Experimental Psychology, 25*, 335-342.

Odom, P. B., Blanton, R. L., & McIntyre, C. K. (1970). Coding medium and word recall by deaf and hearing subjects. *Journal of Speech & Hearing Research, 13*, 54-58.

Olsson, J. E., & Furth, H. G. (1966). Visual memory-span in the deaf. *American Journal of Psychology, 79*, 480-484.

Penney, C. G. (1985). Elimination of the suffix effect on preterminal list items with unpredictable list length: Evidence for a dual model of suffix effects. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 11*, 229-247.

Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory & Cognition, 17*, 398-422.

Pintner, R., & Paterson, D. (1917). A comparison of deaf and hearing children's visual memory for digits. *Journal of Experimental Psychology, 2*, 76-88.

Putnam, V., Iscoe, I., & Young, R. K. (1962). Verbal learning in the deaf. *Journal of Comparative and Physiological Psychology, 55*, 843-846.

Richards, J. T., & Hanson, V. L. (1985). Visual and production similarity of the handshapes of the American manual alphabet. *Perception & Psychophysics, 38*, 311-319.

Shand, M. A. (1982). Sign-based short-term coding of American Sign Language signs and printed English words by congenitally deaf signers. *Cognitive Psychology, 14*, 1-12.

Shand, M. A., & Klima, E. S. (1981). Nonauditory suffix effects in congenitally deaf signers of American Sign Language. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 464-474.

Wallace, G., & Corballis, M. C. (1973). Short-term memory and coding strategies in the deaf. *Journal of Experimental Psychology, 99*, 334-348.

Withrow, F. B. (1968). Immediate memory span of deaf and normally hearing children. *Exceptional Children, 35*, 33-41.

Wolford, G., & Hollingsworth, S. (1974). Evidence that short-term memory is not the limiting factor in the tachistoscopic full-report procedure. *Memory & Cognition, 2*, 796-800.

FOOTNOTES

Memory & Cognition, 18(6), 604-610 (1990).

†Also IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, New York

Haskins Laboratories Status Report on Speech Research
1990, SR-103/104, 173-182

The Processing of Inflected Words*

Leonard Katz,† Karl Rexer,†† and Georgije Lukatela†††

Is an inflected word identified by first decomposing it into stem plus suffix or, instead, is it recognized as a whole? Several lexical decision experiments studied the recognition of inflected words in English (a language with few inflections) and Serbo-Croatian (a heavily inflected language). If recognition depended on decomposition, preceding the inflection with a brief exposure of the stem (<100 ms) should have primed the lexical entry for the stem and, therefore, facilitated recognition of the whole inflected word that followed. It did not. It was also found that the speed of recognizing an inflected word was more strongly associated with the frequency of the whole inflected form than with the frequency of its stem. The results suggested that in word recognition, lexical contact is first made with the whole word form. Nevertheless, morphological decomposition may still occur in subsequent processing.

From a linguistic point of view, the study of derivational and inflectional morphology is largely a matter of discovering the ways that morphological variants relate to more basic, primary, forms. For example, the English adjective *national* is derived from the noun *nation* by adding the suffix *-al*. The past tense verb *walked* is an inflected variant of the stem *walk*; adding the suffix *-ed* to a regular verb stem changes it to the past participle. To the linguist, there is no question that the relation between a variant and its primary form can be described in terms of compositional processes such as these; the variant is composed of the stem plus a morpheme which syntactically modifies the original meaning of the stem.

In contrast, a serious question remains for the psychologist who asks whether the linguistic evidence of componentiality reflects the processing these forms undergo when they are perceived. For the psychologist, the question is: Does the process of understanding a derived or inflected word involve the morphological components of the word as distinct entities or, instead, are these words understood by reference to the whole word form, without any utilization of their components?

A refinement of the question is quickly necessary. When we hear a variant for the first time, it is undeniable that we will apply our tacit knowledge of morphology to an analysis of the separate components in order to interpret the meaning. When you read the novel sentence, "He Nixoned out of it," you understand that the name of an American ex-president is being used as a past tense verb form, even though you may never have seen this form before. There is no other way to explain how we can understand *Nixoned* except by appealing to a process that analyzes it into its components. But a question still remains for variants that are heard more frequently and, therefore, are familiar. For example, it is far less clear that we make a componential analysis on walked when we hear the sentence, "He walked out of it." In contrast to *Nixoned*, we have heard and read the form *walked* many times; because of this, it may be understood via a process that avoids the more complex route by which novel forms must be processed. Instead, we may process the familiar form as an unanalyzed whole.

This paper will present some new experimental evidence that addresses the processing question. We will limit the discussion to the processing of inflection partly in order to make the discussion manageable but mainly because inflectional

This research was supported by National Institute of Child Health and Human Development grants HD-01994 to Haskins Laboratories and HD-08495 to Belgrade University. We were assisted by Mira Peter in several phases of the research. We gratefully acknowledge the criticisms of our colleagues Laurie Feldman and Ram Frost.

morphemes are more likely than derivational morphemes to be candidates for componential processing, as we shall show. Thus, inflection provides us with a generous test of the conjecture that stem and morpheme are habitually processed independently; if separability can not be demonstrated for inflection, it is unlikely that it exists for derivation either. But before considering the evidence from laboratory experiments, I would like to consider, briefly, evidence from what we might call "natural" experiments: those that produced the languages of the world.

Linguistic Evidence

It can be shown that the inflectional morphemes in many languages were, in earlier forms of the language, separate words: distinct lexical items. As the language changed, those items that were associated with stems of a particular word class became attached to (and sometimes fused with) the stems they were associated with. For example, Greenberg (1978) shows that gender markers on nouns were sometimes, at an earlier stage in the history of a language, separate demonstrative pronouns. Over time the pronouns migrated to the noun they described, mutated phonologically, and fused with the noun in the form of an affix. The processing implication of this is that because the inflection was once not only a distinct concept but also a distinct lexical item, it may be only weakly fused to the stem and, therefore, may be easily detached from the inflected stem during processing. This implication is, of course, far from compelling but, nevertheless, is suggestive.

The argument that inflection is only weakly attached to the stem is bolstered by another universal linguistic fact. When a derivational and inflectional morpheme occur together in a word, the inflection is always farther from the stem than the derivation (Greenberg, 1966). The greater fusion of the derivational morpheme with the stem is consistent with the idea that it is more idiosyncratically associated with that particular stem; it can be applied less consistently to *other* items in that stem's word class. Moreover, it changes the meaning of the stem more than does the inflection (Bybee, 1985). Consider the following example. The plural of the word *doll* is the inflected word *dolls*. The diminutive (derivational) form of *doll* is *dolly*. When the two are combined, the form is *dollies*. No language would produce the reverse order *doll-s-y* to convey the same meaning.[1] Thus, there is a gradient of fusion; even if derivational morphemes turn out to be fused inseparably to the stem, inflectional morphemes may be less strongly attached.

Experimental Evidence

One way to pose the question of processing separability more precisely is to ask whether the mental lexicon contains a separate entry for each morphological variant or, instead, contains only one single form, the stem. If only one exists, then the processing system would have to separate the stem within any variant from the rest of the form so that the stem could be recognized lexically. The remaining morpheme (the inflectional or derivational morpheme) would have to be analyzed by some additional process, which might be characterized as a syntactic process. On the other hand, the lexicon might contain the stem (but only if it is a complete word, as is the usual case in English) and all variants. In this second possibility, the stem and its variants would each have the status of a separate word. We used two experimental paradigms to investigate this question. In the first, the rationale was to facilitate the recognition of an inflected word by presenting it in a way that was consistent with the subject's hypothesized processing structure. If we assume that the processing system prefers to recognize the stem and inflection as separate components, recognition of the inflected word should be best when the components themselves are presented to the subject as distinct entities. Thus, we structured the stimulus presentation so that the word was already divided into its components when it was perceived by the processing system; the stimulus word was preanalyzed for the subject. We looked for improvement in the speed of word recognition for these divided stimuli.

The second paradigm arose adopting from the opposite assumption: if the lexicon contains both the stem and each of the whole word inflected variants, then the speed of word recognition for any given word form should be related to its frequency of occurrence—the frequency of that specific whole word form. Alternatively, if it is only the stem that is in lexicon, then recognition speed ought to be related to the frequency of the stem because it is the stem component that is common to them all and is the basis for lexical activation.

Experiment series 1: Dividing the inflected word into stem and suffix

Assume, for the moment, that the lexicon contains only stems. If so, then the processing system should recognize stems more quickly than it recognizes inflected forms because inflected forms must first be decomposed, i.e., analyzed into stem and inflection, before the stem will match, and therefore be able to activate, its lexical

representation. Therefore, if the inflected word that is presented to the subject is first divided by the experimenter into its stem and inflection, the processing system should have easy access to the stem and will recognize the inflected word quickly. In order to accomplish this, we divided a stimulus word temporally, creating a stimulus onset asynchrony between stem and inflection.

Experiments in English

One hundred regular English verbs were selected. All were in the past tense and all ended in *ed*. They were selected so as to cover a wide range of frequency of occurrence. An equal number of pseudowords was selected. Their stems were generated by altering one or two letters of real English verbs (not verbs used in this experiment) and all were "inflected" with *-ed* to create pseudo-past tense forms. Thus, the pseudowords had pseudostems but real inflections.

On each trial, subjects were presented with a fixation point in the middle of a computer screen (500 ms), followed first by a verb stem and then by the addition of a suffix inflection "ed" to the stem. The letters were black on a white background. Stimulus onset asynchronies (SOA) between the onset of the stem and the subsequent addition of the inflection were measured in "ticks," the non-interlaced refresh rate of the computer monitor: one tick equalled 1/60 of a second. Five SOAs were used: zero (i.e., no delay between stem and inflection), and 1 to 4 ticks. In each stimulus list, an equal number of words and pseudowords received each of the 5 SOAs. Five lists were created such that each stimulus received all 5 SOAs across the lists.

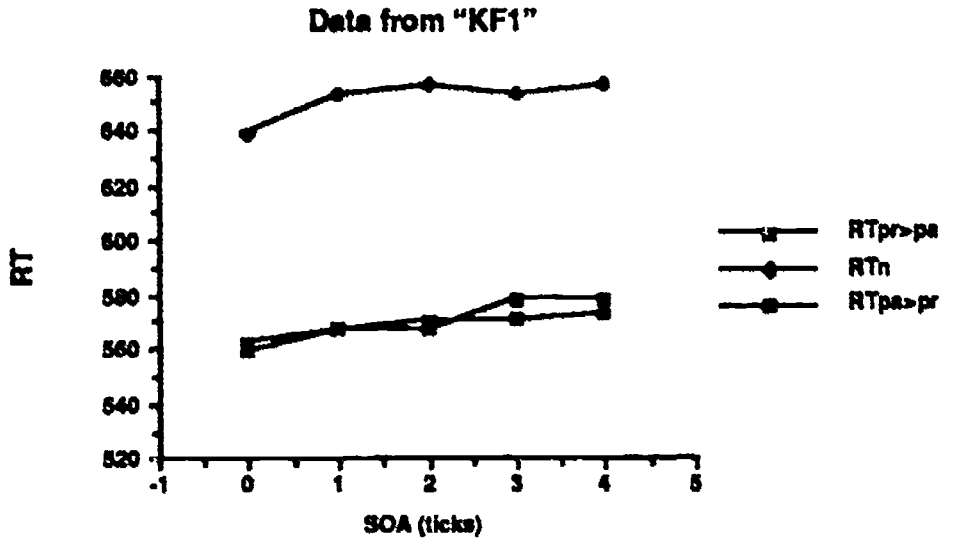
Error rates for individual subjects were not allowed to exceed 10% on either words or pseudowords. Six subjects were run on each of the five lists for a total of 30. A small number of other subjects was discarded for exceeding the error criterion. Subjects received 40 practice trials at the end of which they were given feedback on errors and response speed. A given subject saw only one of the five lists.

Recall the prediction: If only stems exist in the lexicon, then stimuli should be recognized more slowly at an SOA of zero than at one or more of the other SOAs. At zero SOA, the processing system must find the stem, which is necessary for lexical access according to the hypothesis. But the stem is embedded in the whole word so that some processing energy and (presumably) time must be allocated to the extraction of the stem. On the other hand, at one or more of the other SOAs, the system might process the stem until lexical activation occurs and, without missing a beat, then allocate processing to the inflection. Therefore, we should find the fastest recognition times at one of the nonzero SOAs.

Figure 1 presents a graph of the results; RT is plotted against number of ticks. Clearly, the fastest RTs occurred at zero ticks, when the stem and inflection came on simultaneously. Thus, past tense verb forms were recognized fastest when they were presented undivided; in contrast, when they were presented so as to give the stem processing precedence over the suffix, recognition was slower.

The experiment was replicated with a different set of regular verbs. It was thought that those verbs whose past tense is more frequent may not be decomposed but verbs whose present tense is more frequent might show the expected effect in which zero SOA is slower than one of the others. In the second experiment, half the verbs were more frequent in their present tense and half were more frequent in their past tense form. Despite this change however, the SOA effect was similar to the first experiment, for both types of verbs, as the graph in the bottom half of Figure 1 shows.

Other possible artifacts were also explored. Suppose that subjects simply ignored the suffix. In principle, a stimulus word could have been recognized by processing the stem alone. The *-ed* suffix was, in fact, redundant: All stimuli, stems and pseudostems alike, had the same suffix. No information was really carried by the inflection. Note that this accounting does not adequately explain why zero SOA should be fastest; it seems reasonable that other SOAs, in which the stem was initially exposed without the "extraneous" inflection should still be easier to process. Nevertheless, perhaps the subsequent appearance of the suffix following the stem took up processing capacity of some other kind (attentional, perhaps), thus slowing the recognition RT. In order to test this explanation, we changed the composition of the pseudowords in the last experiment so as to require subjects to attend to the suffix (leaving the words as they were). Of the 100 pseudowords, 20 were replaced with words that had real verb stems but pseudosuffixes. The suffixes *-eg*, *-el*, *-ev*, and *-en* were added to real stems to produce pseudowords like "walkeg" and "playel." The remaining 80 pseudoverbs contained nonstems and the real *-ed* suffix (e.g., *refeamed*). Thus, a correct response could not be made on the basis of identifying the stem alone; instead, subjects had to attend to the entire word.



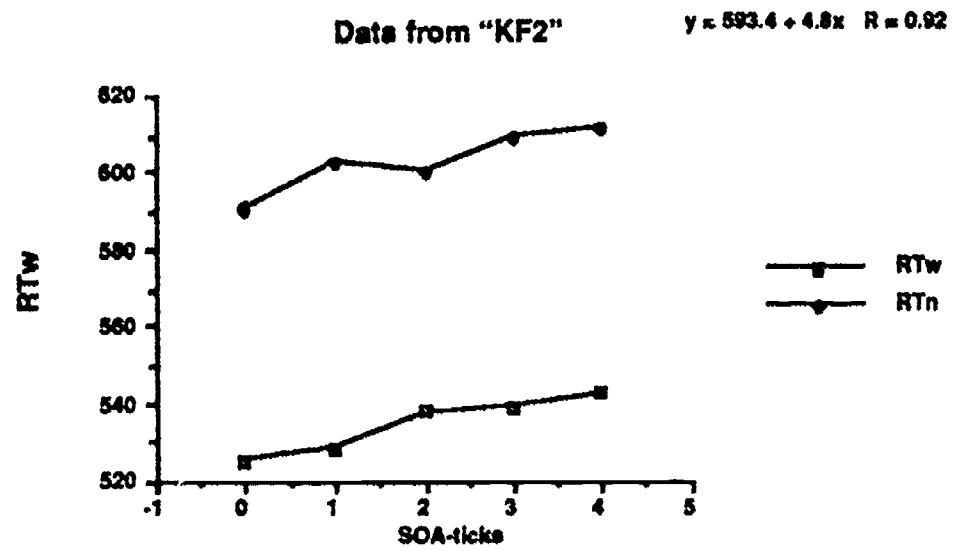


Figure 1. **Stimulus onset asynchrony between stem and inflection for regular past tense verbs.**

Despite the changes, the results of the third experiment were, in part, indistinguishable from the previous two; recognition of words was fastest at zero SOA and rose slowly as SOA increased further. Clearly, failure to attend to the inflection did not account for the previous result with words. However, the result for pseudowords was different; a significant curvilinear function for RT was obtained. The fastest pseudoword rejection was at 2 ticks (33 ms). For the 5 SOAs from zero to 4 ticks, respectively, RTs in milliseconds were 653, 642, 639, 646, 650.

Despite the last pseudoword result, the major results suggest that when we break apart the past tense verb into word stem and *ed,* we seem to be breaking apart a word that is ordinarily recognized as a whole, despite the fact that, linguistically, the division is on a morphological boundary. This indicates that we ought to get the same result if we break apart a word that contains no morphological units, a word that is an indivisible whole, morphologically, such as the word *select*? What would happen if we divided such a word into two arbitrary pieces in a manner that was analogous to the inflected word—the difference being that the new word, unlike the inflected word, would not be divided into two morphemes (stem and inflection). What would happen to recognition latency of the word *select* if we manipulated the SOA between the initial letter string *sele* and the final letter string *ct*? Such an experiment had been run three years earlier (Macaruso & Katz, 1986) and the outcome had been the same: zero SOA was fastest and RTs increased slowly with increasing SOA. Although inflected verbs had not been included and, therefore, no quantitative comparison of slopes is possible, qualitatively, *select* behaves like *walked;* the implication is that the former is as unitized as the latter.

Finally, we ran a variation in which the *-ed* was presented first, before the stem. The rationale for this procedure was the notion that the processing system might normally process the inflection *before* recognizing the stem and, therefore, one of the nonzero SOAs would be optimal for stem recognition. Some researchers (e.g., Forster, 1976) have proposed a prelexical affix stripping mechanism that is consistent with this rationale. In our suffix-first procedure, the *ed* came on the screen zero to 4 ticks before the stem. The spatial positions of the two morphemes remained as in normal print; only the temporal order was reversed. The results were, again, quite indistinguishable from the conditions in which the stem appeared first: recognition under zero SOA was faster than under any other delay. The recognition system preferred the past tense verb as a complete unit.

Experiments in Serbo-Croatian

Several lexical decision SOA experiments were run in Serbo-Croatian. In the latter, which is a highly inflected language, one might expect to get more positive evidence of decomposition than in English. Because most word stems have many inflectional variants, and many of these variants are in frequent use, it would seem that the processing system could benefit from the great reductions in storage that would result from storing only stems in lexicon. First, a series of experiments was run in which the target stimulus was a future tense verb. In Serbo-Croatian, the future is regularly formed by attaching a person-number suffix inflection to the verb stem. For example, the verb *raditi* (to work) forms the future first person singular by taking the verb stem *radi-* and adding the inflection *-cu* to form *radicu* (I shall work). The experiments consisted of presenting the stem zero to four ticks before the appearance of the inflection, as in the English experiments. In spite of our expectations that Serbo-Croatian would show a decomposition effect, the results were similar to the results we reported for English; responses were fastest for zero SOA and increased monotonically from zero to four ticks.

A second series of lexical decision experiments was run using inflected nouns and adjectives. In Serbo-Croatian, the case inflection of an adjective or noun occurs as a suffix. The inflection is very informative syntactically; it indicates number, gender, and one of seven cases. Kostić (in press) presents a more complete description of the Serbo-Croatian cases. The design of all experiments was similar. Only a single SOA was used and there were three main conditions: (1) The stem and inflection appeared simultaneously, (2) the stem preceded the appearance of the inflection, or (3) the inflection preceded the stem. Both parts of the word remained on the screen together until the subject responded (up to 2 sec). Some of the pseudoword stimuli had nonword stems and some had legal stems but were paired with illegal inflections (the latter were comparable to the English pseudowords, discussed above, like *walkeg*). The inflection or pseudoinflection was always two letters long. Average latencies ranged between 600 and 800 ms and error rates were below 6%.

In the first experiment, SOA was fixed at 2 ticks (33 ms) for all conditions. In order to manipulate the amount of processing time given to each stem, half the stems were one syllable in length and half were two syllables. The inflection was always one syllable long. For one syllable stems, reaction time to simultaneous presentation was slightly slower (679 ms) than reaction time to stem - first (665), although not significantly so. For the inflection - first condition, RT was significantly slower (693). Thus, the results were in the expected direction (stem - first was faster) but were not statistically significant. For two syllable stems, the results again showed no statistical difference between simultaneous (661) and stem - first (664) and a significant difference between these conditions and inflection - first (693).

In subsequent experiments, the SOA was changed to 3 ticks (50 ms). For one syllable stems, the simultaneous, stem - first, and inflection - first conditions gave latencies of 685, 690, and 711. But for two syllable stems, the stem - first condition became significantly slower than the simultaneous condition (673 vs. 690) although the inflection - first condition remained significantly slowest of all (733 ms). Additional experiments supported this pattern: stem - first was roughly equivalent to simultaneous appearance (either slightly faster or slightly slower, although not significantly so) as long as the stem was not too long (i.e., one syllable). This suggested that the early part of the word recognition process involved the first part of the word (the leftmost letters). Was this simply due to the fact that the first letters of a word are more informative than the final letters? Or was it due to the particular morphological separation between the first and last letters? In other words, should the results be attributed to the particular lexical and syntactic statuses of, respectively, the stem and inflection or not?

To answer this question, we ran a final lexical decision experiment. The word stimuli were one

syllable masculine gender nouns that were presented in the nominative case; such words are citation forms whose stems do not require any inflection: the entire form is a stem. Other cases in the declension append an inflectional suffix to the nominative form. (Linguists describe this form as having a "null inflection"). Thus, no suffixed words were presented to the subject. As before, the stimuli were presented (1) all letters simultaneously, (2) the first letters preceding the last two, after an SOA of 3 ticks, or (3) the last two letters preceding the first by 3 ticks. The critical difference from the previous experiments was that the first letters in condition (2) were not the stem of the whole word and the two letters in condition (3) were never a legal inflection. Thus, the division of the word was entirely arbitrary, like the arbitrary division made in the English experiment described above in which words like *select* were divided into *sele* and *ct*. The results contrasted with the previous Serbo-Croatian experiments in which the words were divided morphologically. Simultaneous presentation produced a recognition time of 673 ms. When the first letters preceded, RT was significantly slower, 712 ms, and when the last two letters preceded, RT was 729 ms. Thus, when the first few letters did not correspond to a lexical entry (i.e., did not correspond to a stem), recognition time appeared delayed compared to the earlier experiments in which the stem was presented intact.

The Serbo-Croatian experiments taken together suggest, then, that whether the stimulus is a whole inflected word or a stem, subjects have about equal access to it. This was the case both for stems that were complete words in themselves and stems that were not. Thus, the Serbo-Croatian results suggest something a little different than the English results. For Serbo-Croatian, the stem may have a lexical representation distinct from the whole inflected form. Nevertheless, there was no evidence that recognition of the inflected form proceeds via decomposition.

Experiment series 2: Using a word's frequency of occurrence

After our inability to find any substantial evidence for the hypothesis that the stem alone is involved in accessing inflected forms, we decided to take a different tack and explore the consequences of the opposite hypothesis: that the different inflectional variants are instantiated separately in lexicon. What kind of results should we expect if inflected verbs are, in fact, represented as a separate form from the stem in the lexicon? Assuming that *walk* and *walked*, for example, are both in lexicon, the lexical decision time for each form ought to be a function of that form's specific frequency of occurrence. For example, the recognition time for *walked* should be a function of the frequency of *walked* and should not be related to the frequency of *walk*. In contrast, if inflected verbs are decomposed into stem and inflection during the process of recognition, then recognition time should be controlled by the time it takes to contact the stem in lexicon; therefore, lexical decision time should be related to the frequency of the stem. Thus, the relevant question is: what predicts RT to a word form better—the frequency of occurrence of that specific form or the frequency of the word's stem?

We abandoned the SOA procedure of the previous set of experiments and, instead, adopted a more standard lexical decision paradigm: On each trial a single word or pseudoword was presented undivided. Our main tool was word frequency of occurrence, according to the Kučera-Francis corpus (Kučera & Francis, 1967). We selected one hundred regular verbs covering a wide range of frequencies. Subjects saw lists of verbs and pseudoverbs in which present and inflected forms were mixed randomly from trial to trial. Each subject saw all 100 verbs and 100 pseudoverbs twice, once in each of two lists separated by an unrelated secondary task of 10 minutes duration. If a word had been presented in the present tense in the subject's first list, it was inflected in the subject's second list (and vice versa).

Two experiments were run. In the first experiment, half the verbs and pseudoverbs were in present tense form (e.g., *walk)* and half were in past participle form (e.g., *walked*). In the second experiment, half the verbs and pseudoverbs were again in present tense form but the other half were in present participle form (e.g, *walking*). The present participle is typically much less frequent than the present tense form and, so, it provides a rather strong test of the hypothesis that form-specific frequency drives recognition time.

There are two methods that can be used reasonably to index the frequency of a word's stem. One method counts the total occurrence of the stem accumulated over every inflected and derived form of the word (but only those derived forms in which the stem remains unaltered) plus, of course, the occurrence of the pure stem by itself. Reaction time should be a function of this total frequency according to a model in which the stem is extracted from a morphologically complex word and used for lexical access; recognition means

matching the stem that was extracted from the stimulus with its identical lexical representation. But for a model that does not make this assumption, i.e., the model in which an inflected form has a separate lexical status from the stem, separate frequencies of the occurrence (of the pure stem by itself and of the inflected form by itself) are appropriate. We present the data for two experiments, analyzing both experiments first according to the frequency of the specific form and then according to the total stem frequency.

For each stimulus, response times were averaged over subjects and were then analyzed by multiple regression. RT was regressed on log Kučera-Francis frequency (F) for the present tense and log F for the past tense. In the top half of Figure 2 are the results of the first experiment. The histograms present the partial regression coefficients, i.e., the b weights, for present tense and past tense log frequencies (F) as predictors of RT in the regression equation: $RT = a + b_1F(present) + b_2F(past)$. The b weights are either negative (i.e., RT decreases as frequency increases) or are statistically zero. Results for trials when the target was a present tense verb form (e.g., *walk*) are on the left and past tense verb forms (e.g., *walked*) are on the right. It is clear from the analysis that when a target verb was in the present, the frequency of its present tense form predicted RT better than the frequency of its past. In contrast, when a verb was in the past tense, it was the past tense frequency that was the better predictor. In fact, the inappropriate frequencies were not statistically significant in these data although the appropriate frequencies are.

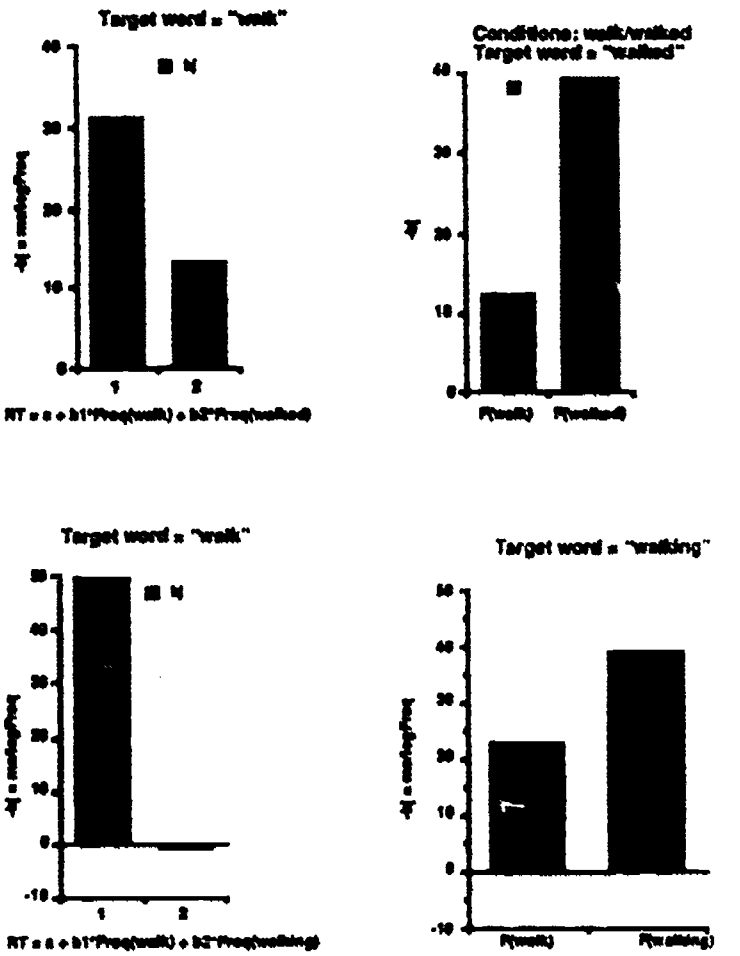


Figure 2. Partial regression weights (b_i) for log frequency of verb stem form or log frequency of the inflected form as predictors of lexical decision time.

A second experiment was run in which the present tense form was mixed with trials of the present participle form of the verb (e.g., *walking*). In the bottom half of Figure 2, we see again that RT is best predicted by the frequency of the actual form that was presented. The one change from the previous results is that, when the present participle is presented, RT is also significantly—although secondarily—related to the present tense form's frequency.

Similar results are obtained when we regress RT on an index of stem frequency which counts, cumulatively, the total occurrence of every inflected and derived form. Figure 3 presents new regressions on the same responses presented in Figure 2 but now the partial regression coefficients represent the log total frequency of the stem in addition to the log frequency of the actual form that was presented. In the top half of Figure 3 are the results of the first experiment.

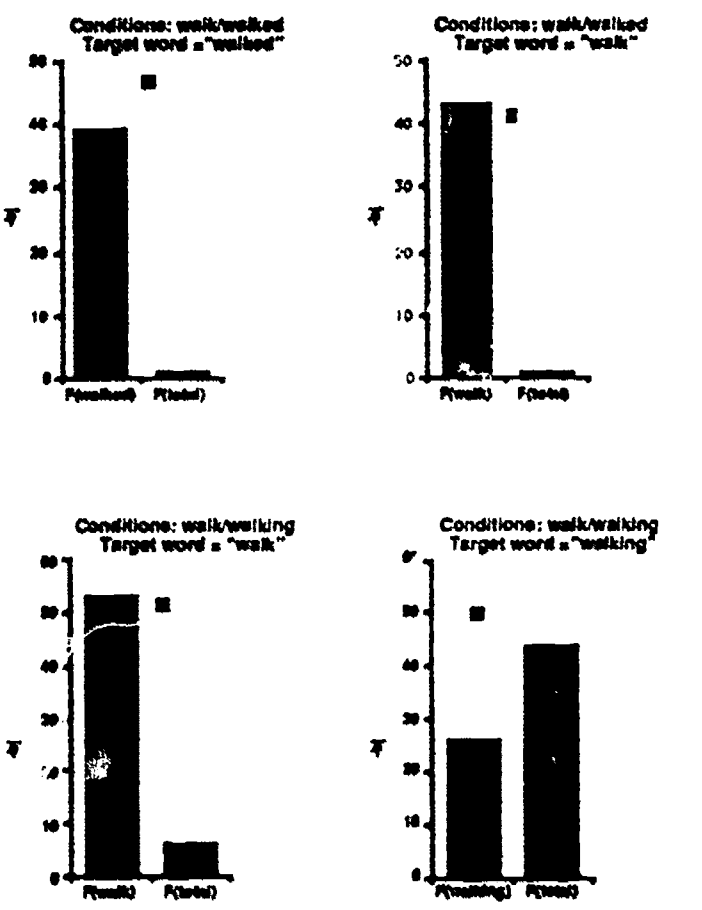


Figure 3. Partial regression weights (b_i) for log frequency of verb stem form or log of total stem frequency (all forms) as predictors of lexical decision time.

Trials when the target was a present tense verb form are on the left and past tense trials are on the right. It is clear from the analysis that when a target verb was in the present, the frequency of its present tense form predicted RT better than total frequency. Only the coefficient for present frequency is significant. Similarly, when a verb was in the past tense, it was the past tense frequency that was the significant predictor, not total frequency.

In the bottom half of Figure 3 are the corresponding results for the second experiment, in which the stimuli consisted of the present tense and the present participle. When the stimulus was in the present tense, present tense frequency was, again, a better predictor than total frequency. However, the results look somewhat different when the stimulus was the present participle; total frequency is the stronger predictor here, although both coefficients are significant ($p < .001$). This latter significance is the only outcome that suggests that total frequency can have an effect on recognition time over and above the stimulus form's actual frequency. We have no explanation for this latter result except that it may be related to the large difference in actual and total frequencies.

Discussion

In the introduction of this paper we discussed the universal linguistic facts suggesting that inflections may be only weakly attached to stems. In spite of this evidence, the experiments presented here have not been able to find clear evidence in favor of one implication of this suggestion, viz., that the lexicon contains only the stems of inflected words. Instead, our results suggest that inflected forms are recognized as wholes and that the morphological information contained in them is not referenced during that process. These results echo the structural explanation given by Lukatela, Gligorijević, Kostić and Turvey (1980) for the organization of Serbo-Croatian inflected nouns. For Lukatela et al., each inflected case had an independent representation in the lexicon; all these representations of a given lexeme were, however, linked.

Our results are also consistent with a model for lexical access of letter strings proposed by Caramazza and his associates, the Augmented Addressed Morphology model or AAM (Caramazza, Laudanna, & Romani, 1988). They suggest that the lexicon contains both a whole word representation and a morphologically decomposed representation. Of these two forms, the one that determines the lexical decision response itself depends on which one is activated beyond threshold first: According to Caramazza et al., this form will be the whole word form when the word is not novel to the subject.

An additional consistency between the AAM model and the present data is found in the pseudoword results. Because pseudowords are novel stimuli for the subject, they ought to be morphologically decomposed before processing. Therefore, in the SOA experiments, we would expect a pseudoword in which the stem and inflection are presented simultaneously to be rejected more slowly than a presentation in which the pseudostem is presented first. This result did obtain for the SOA experiment in which the suffix stimulus was sometimes an illegal inflection such as "eg" appended to a real word stem. According to the model, a componential analysis of the stimulus will determine the response if no entry for the whole word is found in lexicon. Because an inflected pseudoword has no entry in lexicon, its morphological components will be analyzed, an operation that will increase its recognition time. This disadvantage can be countered, evidently, if the pseudoword is presented with its morphemes already divided temporally by a stimulus onset asynchrony.

However, we still need to explain why the same process did not occur also in the two prior experiments in which the pseudowords carried only legal inflections. In these two experiments, the pseudowords behaved like the familiar real words. Perhaps the reason for the difference is that, in the two prior experiments, subjects did not attempt to process the inflection at all because it was redundant. The word—nonword decision could be made on the basis of the stem alone, which was always either a real stem (and, therefore, a real word) or not. Such a strategy could not work in the third study where real stems were sometimes combined with illegal suffixes (e.g., "walkeg"). Note, however, that even if subjects did adopt a different strategy in the third experiment, the same result was obtained for real words as in the two previous experiments.

In spite of this evidence against decomposition, there are still plenty of reasons to believe that linkages among inflectional relatives exist. We interpret our findings to suggest that inflectional analysis does not take place prior to the initial lexical activation of a word (each inflected variant seems to have a separate lexical status). Analysis may follow or, instead, activation of a target word's relatives may occur subsequent to

activation. Consistent with this idea are the data from paradigms in which priming occurs over a long delay between the prime and its relative, e.g., the repetition priming paradigm and the paradigm in which the prime and the target are embedded in different lists. Using the former, our colleagues at Haskins Laboratories, Laurie Feldman and Carol Fowler and their associate Shirley Napps have shown morphological processing effects that are distinct from episodic and semantic/associative effects (e.g., Fowler, Napps, & Feldman, 1985; Napps, 1989).[2] For example, in contrast to semantic/ associative priming effects which occur only when the prime and target are contiguous items and are temporally close, morphological effects can be found at long lags between prime and target. Thus, there is reason to suspect that the mechanisms that underly semantic priming and morphological priming follow different time courses. We may speculate that the morphological process, which has the slower decay time, may also have the slower onset as well. Murrell and Morton (1974) showed morphological priming effects when the prime was a variant in a list that preceded the presentation of the target list. In this case, as well, there is no compelling reason to believe that what was activated by the prime was the stem that is common to all relatives. Instead, the representation that was contacted in lexical access could have been the specific form, i.e., the variant itself; the facilitatory effects may have been a result of a consequent activation of the target's relatives via structural linkages between those relatives.

In addition, several studies using the lexical decision paradigm have been interpreted as consistent with a post-access origin for morphological priming. Macaruso (1988) used a lexical decision paradigm with cross-modal priming that precluded priming artifacts based on orthographic relations between prime and target. He found both variant—specific effects and evidence of linkage between relatives. Sanchez-Casas, Garcia-Albea, and Bradley (in press) demonstrate a morphemic effect using two syllable bimorphemic words with a presentation SOA similar, in part, to ours. Notably, however, they could find an effect only when the SOA between morphemes was increased to 200 ms, much longer than our own maximum of 67 ms. This suggests that their effects were generated relatively late in the processing of the word. Finally, Nagy, Anderson, Schommer, Scott, and Stallman (1989) used a design in which only word and nonword stems were presented to the subject; this effectively precluded the subjects from adopting a strategy based on morphological decomposition because they never saw any morphologically complex forms. In analyses similar to those we presented here, lexical decision latency to the stem was regressed on several variables, including stem frequency, total frequency of inflectional relatives, total derivational frequency, and others. Overwhelmingly, the primary predictor was stem frequency itself, which accounted for 43% of the variance. The frequency of the stem's inflectional relatives was not significant (although it appears to have been confounded with a word's age of acquisition, which was significant). Derivational frequency was significant, but accounted for only an additional 5%, by itself and in interaction with an index of the word's part of speech. Nevertheless, in several analyses of subsets of the data, significant and slightly more substantial effects of a stem's morphological relatives were found. This mix of results about the effects of a word's relatives lead them to the conclusion that "... morphologically related words are grouped together under the same entry in the internal lexicon, or perhaps in linked main entries." That is, all variants exist in the lexicon and either the stem is the point of lexical access or each variant is accessed directly. In either case, there probably exists a linked network among all the members of the family such that, when one is accessed, all are activated.

Thus, we do not abandon the notion that morphological information is activated during the processing of words. Moreover, it seems reasonable to suspect that the processing of the two kinds of morphology, derivation and inflection, may differ between themselves. The former may not be analyzed apart from the stem or root: at least, not as consistently as inflection may be. Derivation is likely to be fused more closely to the stem or root, as we have indicated in the introduction to this article. Related to the difference between inflections and derivations is the relative consistency of the position of inflection in relation to the stem; this may be important in determining whether the component is processed separately from the stem or not. Clearly, when the morpheme follows the stem, it is possible for the processing system to operate on-line in a way that takes advantage of knowledge provided by the stem (i.e., lexical information about the kinds of arguments and morphemes appropriate to that stem) thereby facilitating the perception and integration of the morphemic information with the stem information. Inflection tends to be a suffix process

in languages that have inflection. Even in languages that are not considered inflectional languages, like Chinese, there are often some inflection-like morphemes (e.g., a past tense marker) and these are typically suffix morphemes. The implication of this is that there is something natural about processing an inflection after the stem. To bolster the claim that suffix morphology is more natural, note that there are no languages that are exclusively prefixing languages (Greenberg, 1966). Prefixing is used only in addition to (but never instead of) suffix morphology.

In conclusion, it is suggested that the lexicon's representation of the whole word form (undecomposed) is the first point of lexical contact in the recognition of familiar forms, including morphologically complex forms. Whether or not a word undergoes subsequent morphological analysis remains to be explored. However, it seems that we should now focus our research efforts on the likelihood that the activation of morphological information occurs later in processing, after initial lexical activation.

REFERENCES

Bybee, J. H. (1985). *Morphology*. Amsterdam: John Benjamins.

Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition, 28*, 297-332.

Feldman L. (in press). The contribution of morphology to word recognition. *Journal of Psychological Research* .

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. C. T. Walker (Eds.), *New approaches to language mechanisms*. Amsterdam: North Holland.

Fowler, C., Napps, S. , & Feldman, L. (1985). Lexical entries are shared by regular and irregular, morphemically related words. *Memory & Cognition, 13*, 241-255.

Greenberg, J. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), *Universals of human language*. Cambridge: MIT Press.

Greenberg, J. (1978). Gender markers. In J. Greenberg (Ed.), *Universals of human language,3*. Cambridge: MIT Press.

Kostić, A. (in press). A new approach to isolated word recognition. *Journal of Psychological Research* .

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Lukatela, G., Gligorijević, B., Kostić, A., & Turvey, M. T. (1980). Representation of inflected nouns in the internal lexicon. *Memory & Cognition, 8*, 415-423.

Macaruso, P., & Katz, L. (1986). Decomposition of verb stem and inflection. Paper presented to Psychonomic Society, New Orleans.

Macaruso, P. (1988). *Lexical organization of inflected words*. Doctoral dissertation, University of Connecticut.

Murrell, G. A. , & Morton, J. (1974). Word recognition and morphemic structure. *Journal of Experimental Psychology, 102*, 963-968.

Nagy, A., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24(3)*, 262-282.

Napps, S. (1989). Morphemic relationships in the lexicon: Are they distinct from semantic and formal relationship? *Memory & Cognition, 17*, 729-739.

Sanchez-Casas, R. M., Garcia-Albea, J. E., & Bradley, D. (in press). On access representation: The temporal separation technique. *Journal of Psychological Research* .

FOOTNOTES

*Journal of Psychological Research (in press).

†Also University of Connecticut, Storrs.

††University of Connecticut, Storrs.

†††Also University of Belgrade.

[1]I know of only a few exceptions; these can be accounted for by assuming that the inflected form has, itself, become lexicalized: that it is, in effect, a stem. For example, in German we have *das Kind*, and *die Kinder*, but *die Kinderschen* (not *Kind-schen-er*).

[2]However, Feldman (in press) also presents evidence of a failure to find decomposition effects.

Haskins Laboratories Status Report on Speech Research
1990, SR-103/104, 183-202

Steady-state and Perturbed Rhythmical Movements: A Dynamical Analysis*

Bruce A. Kay,† Elliot L. Saltzman, and J. A. Scott Kelso††

The purpose of this study was to derive the qualitative dynamical properties of a simple type of voluntary rhythmical activity. To this end, rhythmic finger movements were examined in the steady-state and when momentarily perturbed by a torque pulse. It was found that: (a) movement frequency, amplitude, and peak velocity were stable under perturbation, signalling the presence of an *attractor*; and (b) the *dimensionality* of that attractor, as measured by the correlation integral, was approximately equal to that of the simplest limit-cycle oscillators. Also, (c) the *strength* of the attractor was constant with increasing movement frequency; and (d) the *Fourier spectra* of the steady-state trials showed an alternating harmonic pattern. These results are consistent with a previously-derived nonlinear oscillator model. However, (e) the oscillation was *phase-advanced* by perturbation overall, and a consistent phase-dependent phase-shift pattern occurred. These phase-response results are inconsistent with our previous limit-cycle model. The overall phase-advance also shows that any central pattern generator responsible for generating the rhythm must be non-trivially modulated by the limb being controlled.

The origin and form of biological rhythms have been the objects of intense inquiry for years, and many studies have focused on the neurophysiological bases of the oscillatory mechanisms (so-called central pattern generators, CPGs) that underlie such rhythms. For example, circadian rhythms (e.g., Pittendrigh & Daan, 1976) and rhythmic motor acts such as locomotion (Grillner & Zangger, 1979), have been studied in these terms. Much research into CPGs has been aimed toward answering the question: What is the actual form of a CPG in terms of neural structure and interactions among the component neurons in particular behavioral and physiological situations (e.g., Carpenter & Grossberg, 1983; Lennard, 1985; Selverston, 1980)? However, because biological rhythms are so ubiquitous and occur over such a large variety of particular physical (e.g., neural or biochemical) structures, it is also important to look for macroscopic commonalities across these instantiations. This latter type of approach focuses on the more global, generic properties of the rhythmic behaviors themselves and complements more microscopic analyses.

In this article, we report on the macroscopic dynamical properties of a simple, one-joint voluntary rhythmic movement. In a previous study (Kay, Kelso, Saltzman, & Schöner, 1987), we found that similar movements displayed an invariant relationship between two basic parameters—frequency and amplitude—that characterize rhythm. We derived a simple dynamical model that could account for this relation. In that study, only steady-state rhythms were investigated, and only these basic kinematic parameters were studied. The purpose of the present article is to provide as complete and detailed a dynamical description of a rhythmic task as possible, both during steady-state rhythms

An earlier version of this paper was submitted in partial fulfillment of the doctoral degree program of the first author at the University of Connecticut, Storrs, CT. The research reported herein was supported by Contract No. N00014-83-K-0083 from the U.S. Office of Naval Research and NINCDS Grant NS-13617. Preparation of this manuscript was provided by Sloan Foundation grant number 87-2-16 to the first author.

Thanks to Arthur Winfree for comments on an earlier draft, Paul Lennard for helpful discussion, and Esther Thelen and Gary Riccio and two anonymous reviewers for their welcome critiques.

and when the moving limb is momentarily perturbed. In developing a broader data base relevant to this class of movements, we also discuss more detailed, yet still abstract, models of how these movements might be generated and controlled.

Qualitative Dynamics

The approach adopted in our work for conceptualizing and modeling complex biological behavior is that of qualitative dynamics, an outgrowth of dynamical systems theory (e.g., Abraham & Shaw, 1982; Thompson & Stewart, 1986). Qualitative dynamics takes as data the evolution of a system's observable characteristics/descriptors, and describes that evolution in terms of a set of equations of motion. The first step, prior to writing down any equations at all, is to classify the system's behavior qualitatively into generic categories or behavioral forms. As the data base is expanded, more refined and more quantitative statements are derived, including candidate dynamical equations and precise values for the equations' parameters. For example, the movement of a single limb to a target is described qualitatively as a point attractor. A point attractor is a stable equilibrium point that attracts all trajectories from arbitrary initial conditions. Point attractor dynamics have the property that a transient perturbation applied during movement does not deter achievement of the equilibrium point: Transient perturbation is equivalent dynamically to a resetting of initial conditions of the system's state variables (e.g., position and velocity). A simple example of point attractor dynamics is the damped linear mass-spring, describable by a second-order linear differential equation:

$$m\ddot{x} + \alpha\dot{x} + k(x-x_0) = 0 \qquad (1)$$

where m is the mass, α the linear damping coefficient, k the linear stiffness coefficient, and x_0 is the equilibrium position. In rhythmical tasks the appropriate qualitative dynamical description is a cycle. If the cycle is stable in certain respects, it is termed a periodic or limit-cycle attractor. A limit-cycle attractor, like a point attractor, attracts trajectories from arbitrary initial conditions, but a stable oscillation of fixed amplitude and frequency is attained. For the simplest oscillatory models, a suitable set of dynamical coordinates are position and velocity. These coordinates define a *phase space*, and in this space (here, the phase plane), the limit-cycle is a closed loop or orbit. In addition to stable frequency and amplitude, the limit-cycle's orbit can display characteristic scaling relation among its kinematic observables. For example, the well-known van der Pol oscillator (van der Pol, 1926), displays a constant amplitude across a wide range of oscillation frequencies, and is described by the following second-order nonlinear differential equation:

$$m\ddot{x} + \alpha\dot{x} + \gamma(x-x_0)^2\dot{x} + k(x-x_0) = 0 \qquad (2)$$

where γ is the coefficient of the nonlinear van der Pol damping term. In the closely-related Rayleigh oscillator (Jordan & Smith, 1977), frequency and amplitude vary according to the inverse function. In both of these oscillators, the limit-cycle's stability is due to the presence of nonlinear damping, or escapement, functions (Andronov & Chaikin, 1949; Kugler & Turvey, 1987; Kugler, Turvey, Schmidt, & Rosenblum, in press; Minorsky, 1974).

In providing a qualitative dynamical account of rhythmical movements, we have investigated the following properties related to limit-cycle dynamics:

Presence of an Attractor

The first major property is that the behavior being modeled can be characterized as a limit-cycle *attractor*. Is an attractor actually present? This question can be answered by comparing kinematic variables such as frequency and amplitude before and after delivery of a transient mechanical perturbation.

Strength of the Attractor

Given the presence of an attractor, an important property is its *strength* of attraction. Trajectories perturbed away from limit-cycles return more rapidly to strong attractors than to weak ones. Also, different oscillators have different functions relating attractor strength and various kinematic observables, such as frequency. For example, the strength of the van der Pol and other similar oscillators is constant across frequency and is constant regardless of where the oscillation is perturbed. The limit-cycle strength can be estimated by measuring the time taken for the system to return to the limit-cycle following perturbation: Shorter returns to the limit-cycle are associated with stronger attractors.

Phase Response Characteristics

An important descriptor of any rhythmic process is its *phase response* to perturbation (e.g., Stein, 1976; Winfree, 1980; Yamanishi, Kawato, & Suzuki, 1979). The question is whether the perturbation has the effect of shifting the rhythm in time with respect to an unperturbed, control

rhythm. The amount of the shift is usually normalized to the period of the oscillation and is thus termed the phase shift. Phase shift patterns can be used to distinguish candidate oscillator models. For example, the sinusoidally-forced linear damped mass-spring,

$$m\ddot{x} + \alpha\dot{x} + k(x-x_0) = F\cos(\omega t) \qquad (3)$$

cannot be phase-shifted by a mechanical perturbation to the mass. Following the transient, the mass resumes its pre-perturbation phase relation with the driver, which is unaffected by such perturbations. On the other hand, the van der Pol oscillator has no external driving term and can be phase shifted. Furthermore, it has a characteristic pattern of phase shift, depending on where in the cycle the motion is perturbed and also on the magnitude of the perturbation.

Dimensionality of the Attractor

A fourth property of a limit-cycle is its dimensionality. The limit-cycles that the Rayleigh and van der Pol oscillators produce are one-dimensional because they form a simple closed curve in phase space: The motions are purely periodic. Oscillating biological systems are not ideal mathematical systems, however. They are never exactly periodic: When plotted on the phase plane, a rhythmic movement trajectory appears as a *band* around some average closed curve (see Figure 1a).

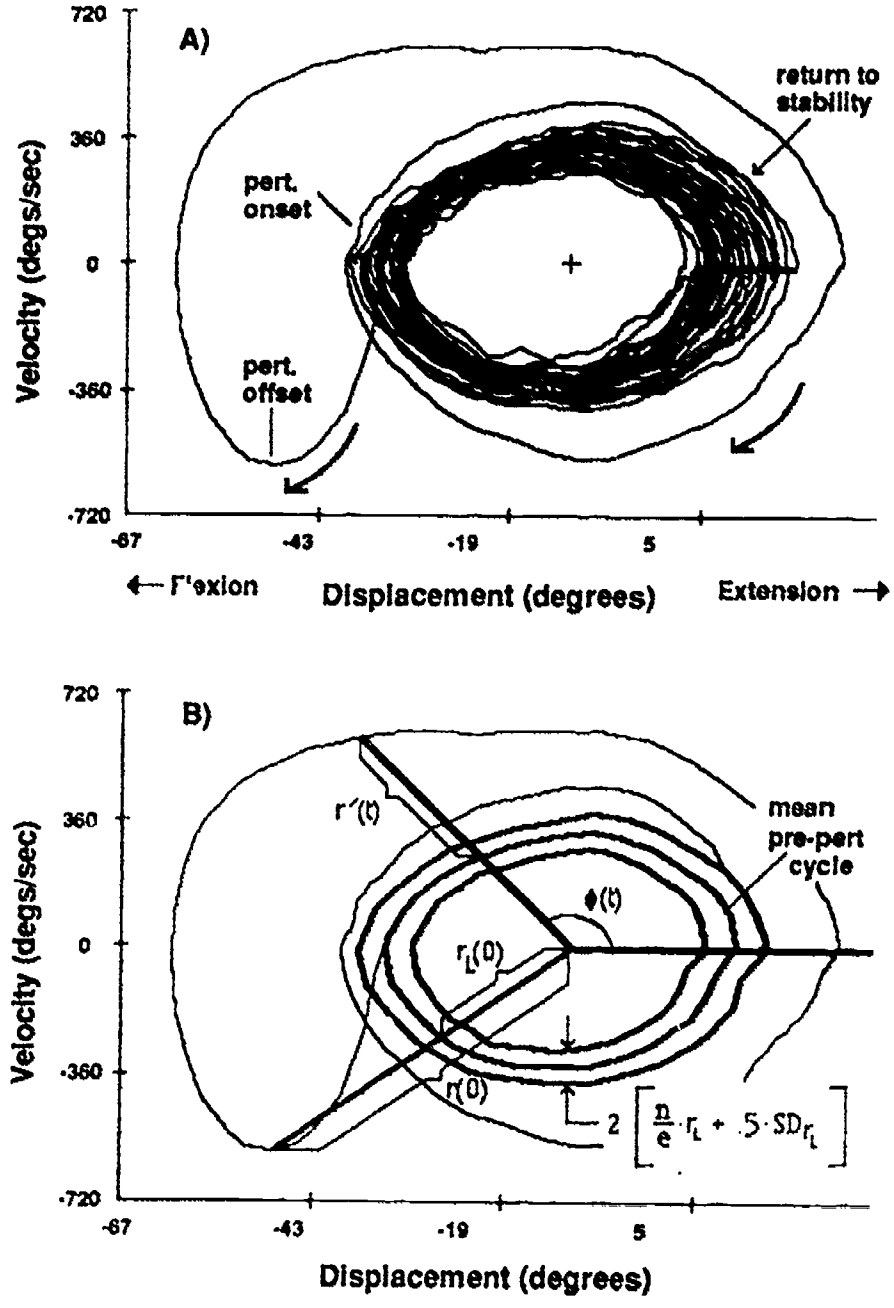


Figure 1. (a) Typical perturbation trial plotted on the phase plane. The central cross indicates the location halfway between average pre-perturbation peak extension (more positive displacement values) and peak flexion (more negative). *(b)* Notational conventions used in analyzing the response (see text).

Is the variability we see due to a purely random, noisy process, or is another deterministic process present? If the variability is due to noise, then it may be modeled by the addition of a stochastic forcing function to the main oscillator equation; for example, the stochastic function may be interpretable as random fluctuations in the recruitment of motor units. On the other hand, the band of variability may be the result of an additional deterministic process; for example, a secondary oscillation having a frequency that is incommensurate with the larger-scale oscillation may be present. Such a situation may reflect an interaction of central and peripheral oscillatory processes. In that case, the limit-cycle is two-dimensional—one dimension for each oscillatory process—and is an example of a very different class of dynamical behavior, *quasi-periodicity*. A third way in which bands on the phase plane may be produced is by a deterministic *chaotic* process (Thompson & Stewart, 1986), which exhibits trajectories that have fractional (or fractal; Mandelbrot, 1983) dimension. If the band of variability were to be produced by either of the deterministic methods, second-order dynamics would be inadequate. That is, second-order dynamics cannot generate two-dimensional limit-cycle or chaotic attractors; the order of the dynamics must be third or greater (Thompson & Stewart, 1986). Such models are different at a fundamental level of qualitative dynamics from lower-order models with added noise (which can be thought of as an additional infinite-dimensional process).

Thus, in order to distinguish among these three possibilities, one must assess the dimensionality of the attractor. Standard waveform analysis techniques cannot be relied upon to perform this task. The Fast Fourier Transform (FFT), for example, may not detect the presence of a second incommensurate frequency if it is very close to the dominant frequency, and the spectrum of a chaotic system may be indistinguishable from that of stochastic noise (Bergé, Pomeau, & Vidal, 1984). In the present article, we use a computation that allows us to estimate the dimensionality of our movement trajectories directly (Grassberger & Procaccia, 1983; see Kay, 1988, for a tutorial on dimensionality analysis in the context of motor skills research).

Fourier Spectra of the Movement Trajectories

Although standard Fourier series methods do not afford a determination of the dimensionality of the attractors underlying the movement trajectories observed, they can provide key information regarding the structure of the dynamics producing such behaviors. All of the oscillators mentioned so far have characteristic spectra: The spectrum of the driven linear mass-spring has only one Fourier component, the fundamental, after decay of the initial transient. The van der Pol and Rayleigh oscillators' spectra contain a fundamental frequency and its odd harmonics, that is, the frequencies that are odd multiples of the fundamental. Their spectra contain much less energy in the even harmonics, and this is true for a very broad class of oscillators that contain certain symmetries in the structure of the dynamical equations. For example, the Rayleigh oscillator contains damping terms that are all in odd powers of the velocity ($\dot{x}$ and $\dot{x}^3$), and the damping function is symmetric about the origin $(x,\dot{x}) = (x_0,0)$. Such a symmetry is mathematically termed odd symmetry. If even-powered damping terms are added, this symmetry is broken, and the spectrum contains even harmonics as well.

Preview of Methodology

In generating a broad descriptive base for rhythmic movements, we adopted two basic experimental protocols, involving perturbation trials and steady-state trials. We used transient mechanical perturbations to test for the presence of an attractor, to measure the attractor strength, and to measure the system's phase response characteristics. In these trials, (a) the subject rhythmically cycled the index finger for several cycles, (b) the experimenter delivered a single torque-pulse perturbation, and (c) the subject continued cycling for several more cycles. Because of the great number of trials required for the phase response analysis, these trials were kept to a relatively short duration, 15 s to 25 s. In order to assess the dimensionality and Fourier spectrum of movement trajectories, the same subjects were required to perform steady-state (uninterrupted) trials of 50 s duration.

Method

Subjects

The subjects were 4 right-handed male volunteers; Subjects 2 and 4 were trained musicians (cellist and percussionist, respectively), whereas subjects 1 and 3 had no musical training. Each subject participated in two experimental sessions, each session consisting of 3 hr of actual data collection.

Apparatus

The apparatus was a modification of one described in detail on previous occasions (Kelso, Holt, Rubin, & Kugler, 1981; see Figure 2). It consisted of a freely rotating finger manipulandum that allowed flexion and extension about the first joint (metacarpophalangeal) of the index finger in the horizontal plane. Two transducers were attached to the apparatus's vertical rotation shaft: a precision DC potentiometer (which measured angular position of the finger) and a tachometer (which measured angular velocity). In addition, a DC torque motor was attached to the top of the shaft. The output signals of the transducers and the control voltage applied to the torque motor were recorded with a 16-track FM tape recorder for later digitization. The tachometer signal was also used in combination with threshold-detection and delay circuitry to provide a trigger for delivering perturbations.

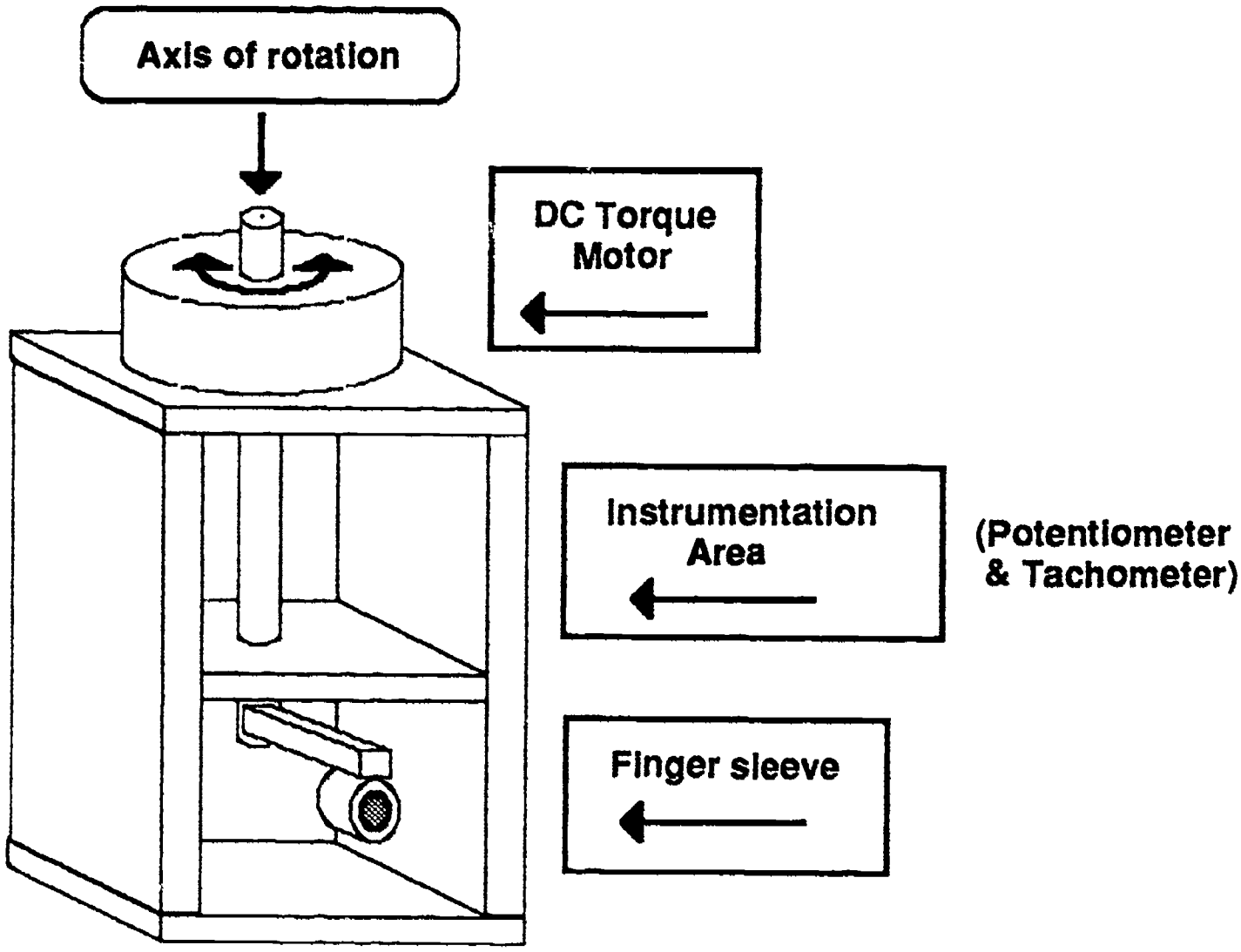


Figure 2. Schematic figure of the finger positioning apparatus (instrumentation details omitted).

Procedure

Overall Experimental Protocol

Subjects were placed in a dentist's chair, and their right (dominant) forearm was placed in a pre-formed splint, which was rigidly attached to the finger-positioning device in such a way that the flexion-extension axis of the index finger's first joint was directly in line with the positioner's vertical axis. Vision of the finger was not excluded.

For all trials, we instructed subjects to cyclically flex and extend their index finger "at a comfortable rate." They were not instructed explicitly about the amplitude of movement; for example, they were not told to move the finger maximally. For the steady-state trials, subjects were required to cycle continuously for a period of 1 min; 50 s from the middle portion of these trials were later analyzed.

Perturbation trials were conducted as follows: First, the subject started cycling and, after a few cycles, would indicate to the experimenter if he was comfortable. The experimenter waited several more cycles, and then set the perturbation delivery process into motion (see next section). After the delivery of a perturbation, the experimenter collected several more cycles of data. A trial lasted from 15 to 25 s, depending on the subject's preferred tempo. Subjects were instructed not to actively resist the perturbation, but rather to try to return to a steady rhythm similar to that produced before the perturbation as quickly and as easily as possible.

Five blocks of trials, each consisting of 1 steady-state trial followed by 32 perturbation trials (a total of 165 trials), were collected in each session; each block lasted approximately 30 min.

Delivery of Perturbations

On every perturbation trial, a torque-pulse perturbation of one of two magnitudes (estimated at approximately 30 and 60 in-oz of torque) and directions (flexion, extension) was delivered by the DC torque motor. The pulse was 50 ms in duration, and of constant amplitude.

In order to sample sufficiently the various portions of a movement cycle, we attempted to insert the perturbation in eight different angular sections on the phase plane, spaced at 45° intervals (see Figure 3). After initiation by the experimenter, external circuitry detected positive and negative peaks in the velocity signal, introduced one of four delays (zero, one-eighth, one-fourth, and three-eighths of a cycle period, plus a constant 30 ms), and triggered the appropriate torque pulse. For example, perturbations were delivered at approximately 0° by detecting a positive velocity peak and introducing a one-quarter period delay. The cycle period used to calculate each perturbation-trial block's phase delays was measured for the last few cycles of the immediately preceding steady-state trial. Thus, the set of delays that were used throughout a perturbation-trial block were specific to that block. All four combinations of perturbation magnitude and direction were delivered at each phase angle within each block.

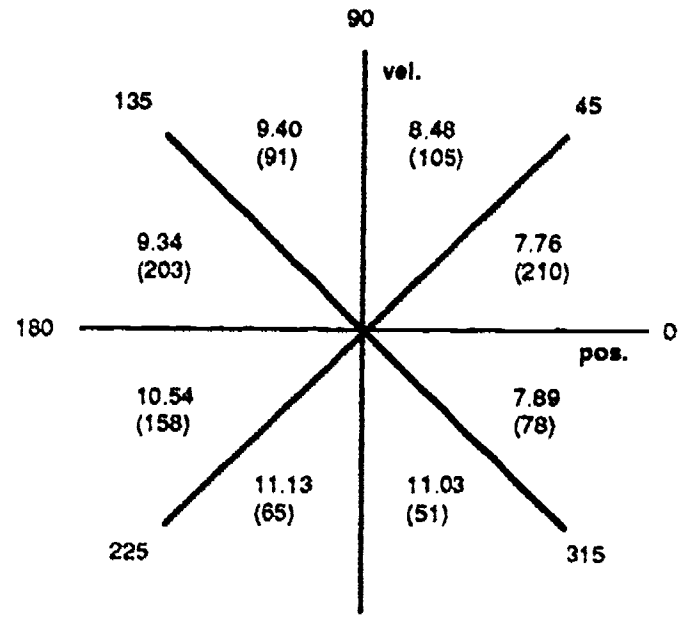


Figure 3. **Phase-plane sectors used in perturbing the finger. Numbers in parentheses are the total number of trials collected in each sector. Numbers outside parentheses are mean attractor strengths (σ) for the sectors.**

Combining the two magnitudes, two directions, and eight phases of perturbation gave a total of 32 perturbation conditions. One trial of each of these conditions was collected in each block; order of presentation was randomized within each block. However, perturbation was not always delivered at the desired phase angle, because of false triggering (especially for the slower movements, which exhibited somewhat noisy velocity signals) and variations in cycle duration within and across trials. Also, a small number of perturbation trials (16 of 1,280) were lost because of experimenter error. Nevertheless, a wide scatter of phase locations was sampled (see Figure 3).

Data Reduction and Dependent Measures

Signal Processing

After the experimental sessions, the recorded transducer and torque control signals were played

back and digitized at 200 samples/s, with 12-bit resolution. For the perturbation trials, the digitized position data were smoothed with a 35 ms (5 sample) triangular window. In two of the eight sessions (Session 1 for Subjects 2 & 3), we experienced mechanical problems with the tachometer (although a good scatter of perturbations on the phase plane was still delivered), and so for these sessions we computed angular velocity from the smoothed position data by using a two-point central difference algorithm. Both transduced and derived velocity data were smoothed with a 55 ms (9 sample) triangular window (see Kay, Munhall, V.-Bateson, & Kelso, 1985, for smoothing and differentiation details). The digitized torque control signal was used to locate the onsets and offsets of the torque pulses. All of the following analyses were performed on the digitized signals. An example of all three digitized signals is shown in Figure 4.

Perturbation Trial Measures

Kinematic Measures: Frequency, Amplitude, and Peak Velocity. For the perturbation trials, we measured frequency, amplitude, and peak velocity of movement cycles both before perturbation and after return to stable behavior following perturbation (see below for the return-to-stability criteria). A cycle was defined as the occurrence of two (successive) peak extension events, which, along with peak flexions, were identified by a peak-picking algorithm applied to the position data. Peak velocities were measured using the same peak-picker on the velocity data, for velocities of extension and flexion movements (positive and negative peaks, respectively). Cycle frequency (in Hz) was defined as the inverse of the time (in seconds) between two peak extensions, and cycle amplitude (peak-to-peak, in degrees) was defined as the average of the extension-flexion, flexion-extension half-cycle excursions. After obtaining these measures for each cycle, means and standard deviations across all steady-state cycles (pre- and post-perturbation) for each perturbation trial were obtained. The Results section reports these within-trial summary data, because of the large number of cycles collected (approximately 20,000).

Estimate of relaxation time and attractor strength. We displayed all perturbation trials on the position-velocity phase plane in order to estimate the point at which the movement trajectory returned to stable rhythmic behavior (see Figure 1b). Two criteria had to be met simultaneously in order for the post-perturbation rhythm to be termed *stable*: (a) The trajectory on the phase plane had to return within a band around the average pre-perturbation cycle and remain either inside of or reasonably near this band; and (b) the frequency of cycling had to settle to a stable value, whether or not that frequency was equal to the pre-perturbation frequency.

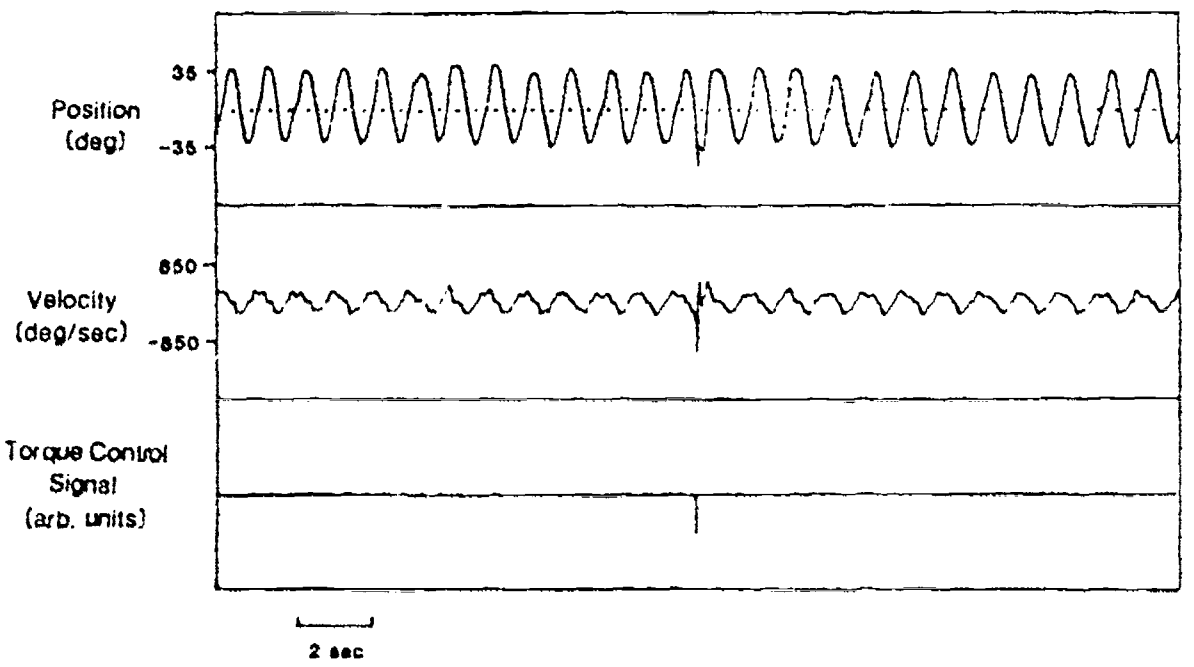


Figure 4. Time-series plot of a typical perturbation trial. Top to bottom: angular displacement (extension = positive, flexion = negative), angular velocity, and the torque control voltage signal.

The band of return was computed as follows: An average amplitude of the pre-perturbation oscillation on the phase plane, $r_L(\theta)$, was computed for each trial in each of 32 phase-plane sectors. The band of return was set to [1 ± n/e] times $r_L(\theta)$, plus one-half of its standard deviation. In this calculation , e is the base of the natural logarithm (e = 2.717...), and the number n was chosen so that the trajectory was perturbed outside the band (n was set to 0.5 for most trials, and to 0.25 for trials in which the perturbation had a smaller effect). The elapsed time between the offset of the perturbation and the time at which both of these criteria were first met was termed the relaxation time, T_{rel}. We found that frequency distributions of T_{rel} were positively skewed, typical of measurements constrained to be greater than zero. To perform statistical analyses, we used the log transform of the raw data; the reported results are the anti-log transforms of the means so obtained.

The T_{rel} in theory can depend on how far the trajectory is perturbed from the limit-cycle, and therefore gives only an indirect indication of the attractor's strength. In order to measure more directly a limit-cycle attractor's strength, the actual form of the return process should be known a priori. Lacking that knowledge, we assumed that trajectories away from the limit-cycle take the form of spirals winding back to the limit-cycle, which is in qualitative accord with the appearance of our trajectories on the phase plane. That is, we assumed that the return process can be approximated by relaxation in amplitude only (excluding, e.g., phase angle), which can be modeled by a first-order linear system displaying exponential decay. The return process was deemed to begin at the time of perturbation offset and end at T_{rel}. For each sample of the return process, the following expression for σ was computed (Figure 1b; see Appendix for details of σ's derivation):

$$\sigma = (1/t)\cdot\ln\left|\,(r(0) - r_L(0))/(r(t) - r_L(t))\,\right| \qquad (4)$$

where t is the elapsed time (in seconds) after the perturbation offset; r(t) is the displacement from the center of oscillation of the return trajectory at time t; $r_L(t)$ is the average displacement of the pre-perturbation limit-cycle in the same phase sector as the return trajectory at time t; and r(0) and $r_L(0)$ are the same quantities but measured at the moment of perturbation offset. The sample values of σ were averaged to provide a value for the entire return process. Like T_{rel}, the frequency distributions of σ were positively skewed; we applied the log transform to σ in order to perform statistical tests.

Phase response analysis. The phase of the new (post-return) rhythm in relation to the old (pre-perturbation) rhythm was determined as follows (see also Winfree, 1980; Yamanishi et al., 1979): At peak extension events, the phase of the oscillation was defined as zero, and for all other points between peak extensions, phase = (t/T), where t is the time (in seconds) from the most recent peak extension to another event of interest and T is the average pre-perturbation cycle period (in seconds). Thus, phase is not defined according to the phase angle on the phase plane. Rather, it is defined as a measure of relative temporal location within the cycle, normalized to units of cycle periods, 0 ≤ phase < 1. The time at perturbation offset, t_p, was normalized as before by T to define the trial's old phase, φ. The elapsed time from the offset of perturbation to the first peak extension event following return to stable oscillation, t_θ, was measured and normalized to define the cophase of the new rhythm, θ. The temporal shift, Δt, was defined as $\Delta t = (t_p + t_\theta) \pmod T$, and normalized to define the rhythm's phase shift, $\Delta\phi = \Delta t/T = (\phi + \theta) \pmod 1$ and the new phase $\phi' = (\phi + \Delta\phi) \pmod 1$ (see Figure 5).

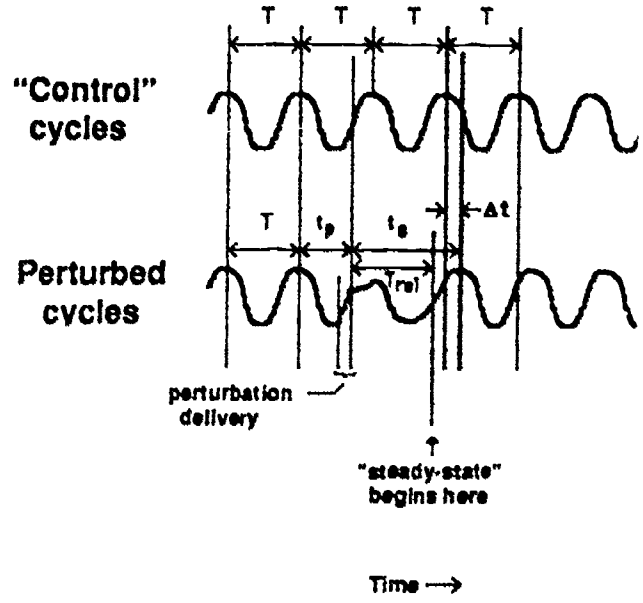


Figure 5. Conventions used in the phase response analysis. Old phase $\phi = t_p/T$; cophase $\theta = t_\theta/T$; phase shift $\Delta\phi = \Delta t/T$.

Thus, Δφ is the amount the post-perturbation rhythm has been phase-shifted in relation to the pre-perturbation rhythm, and φ′ is the phase at which the perturbation would have been turned off in a new cycle if it were delivered at the same old phase in a control cycle (a cycle formed by the continuation of the average pre-perturbation rhythm). The computation is appropriate only if the pre- and post-perturbation cycle periods agree

exactly; however, because these numbers almost never exactly agreed, we accepted post-perturbation deviations of ±5% from the pre-perturbation period (Yamanishi et al., 1979).

To assess the pattern of phase shift that the rhythm exhibits, we constructed a phase transition curve (PTC), plotting new phase ϕ' vs. old phase ϕ. Because this function is biperiodic, in the sense that ϕ' is a periodic function (with period 1) of ϕ and vice versa, a biperiodic function was used to curve-fit the data; standard linear regression techniques are inappropriate here because the arithmetic is actually being performed on the unit circle (Yamanishi et al., 1979). A least-squares fit was performed using the following function:

$$\phi'_{pred} = C_0 + a\phi + \sum_{k=1}^{2} (B_k \sin(2\pi k\phi) + C_k \cos(2\pi k\phi)) \quad (5)$$

where a is the overall linear slope of the function; C_0 is the intercept, which indicates the phase shift averaged across all old phases; and B_k and C_k are the weights on the sine and cosine terms. The error, E, to be minimized was defined as a function of the difference, d, between the measured ϕ' data and the estimated ϕ'_{pred}. Specifically, $d = (\phi' - \phi'_{pred})$ and

$$E = \begin{cases} d & \text{if } -.5 \le d \le .5 \\ d - 1 & \text{if } d > .5 \\ d + 1 & \text{if } d < -.5 \end{cases} \quad (6)$$

Thus, $-.5 \le E \le .5$, preserving the distance relation between points on the unit circle. For example, with $\phi' = 0.9$ and $\phi'_{pred} = 0.1$, d = 0.8; however, these points are close together on the unit circle and E = -0.2 reflects this.

Steady-State Trial Measures

Dimensionality. We assessed the *correlation dimension* of the angular position time-series for the steady-state trials by computing the spatial correlation integral, C(L) (Grassberger & Procaccia, 1983). Representing the data points of each steady-state trial as x(i) (i = 1,...,N; N = the number of data points), C was defined for each possible difference in angular position, L, between sample values as:

$$C(L) = \lim_{N\to\infty} (1/N^2) \cdot \{\#\text{ pairs } (x(i),x(j)) : \text{where} |x(i) - x(j)| < L\} \quad (7)$$

where i,j = 1,...,N; $i \neq j$; and $|\, x(i) - x(j)\,|$ is the Euclidean distance between angular position values. In the limit, as N approaches infinity,

$$C(L) = aL^{\nu} \quad (8)$$

where a is an arbitrary constant and ν is the correlation dimension of the time series (Grassberger & Procaccia, 1983). This relation must hold over a finite range of distance values for the computation to be valid; that is, in a log-log plot, C and L should be linearly related over some values of L, where the slope of the line over that range is interpreted to be the correlation dimension.

In addition, C(L) was computed for a range of *embedding dimensions*. These computations entailed creating multi-dimensional vectors from time-shifted copies of the original single-dimensional time series as follows:

$$X(i) = (\, x(i), x(i+T), x(i+2T), ..., x(i+(k-1)T)\,) \quad (9)$$

where k is a positive integer indexing the embedding dimension, and T is a fixed number of samples of time-delay. The standard higher-dimensional Euclidean metrics were used in defining L for embedding dimensions k > 1. Embedding dimensions greater than one were used because the dimension of the underlying dynamics are a priori unknown, and the computation is valid only when k is greater than or equal to the dimension of the dynamics (one criterion for k is that it must be at least 2ν + 1; Holzfuss & Mayer-Kress, 1986). For each embedding dimension, Equations 7 and 8 were used to compute C(L) and ν.

For this data set, L, the inter-sample distance, ranged over 4096 (2^{12}) discrete values for trials with maximal signal range; the average range was approximately three-fourths of this range, or 3072 discrete values. The L was further discretized into 256 distance bins for the actual correlation integral computation, so that C(L) would have a sufficient number of data points for each bin. In addition, the 10,000-point data records were down-sampled 4:1 to reduce the computational burden. The effective N for each trial was thus 2,500—a relatively small number for this kind of computation (see Abraham et al., 1986, for use of the algorithm with small data sets). We computed C(L) for embedding dimensions 1 to 10, to ensure that k was much greater than 2n + 1. An inter-dimension lag (T) corresponding to 1/4 of the mean cycle period for each trial was used (Abraham et al., 1986; see Fraser & Swinney, 1986, for another criterion for picking lags). Across trials, this ranged from 6 to 32 samples.

A typical plot of C(L) versus L, in logarithmic coordinates, is shown in Figure 6a. Plotting the local slope of this function versus L, a length interval over which the function approximated a straight line was visually determined for

each embedding dimension (i.e., where the slope was roughly constant; see Figure 6b).

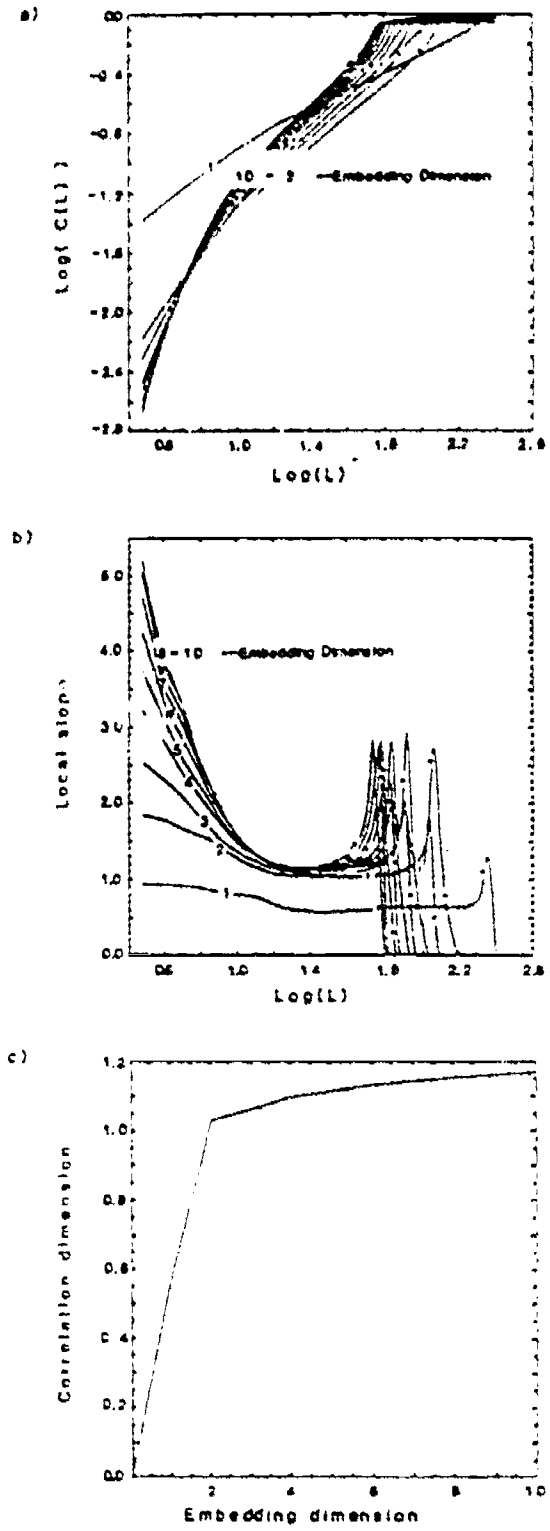


Figure 6. Typical dimensionality plots. *(a)* Logarithmic (base 2) plot of the spatial correlation integral (C; arbitrary units) versus distance (L; arbitrary units). *(b)* Local slope of the function in (A) versus log(L). *(c)* Computed correlation dimension (v) versus embedding dimension (k).

This scaling interval was about 25 distance bins wide in most cases. On this interval, the segment having the least-squared error for a linear fit of log(C(L)) versus log(L) was found. The best-fit segment lengths were constrained to be greater than 5 distance bins and less than 25 distance bins in length (Kay, 1988). We then plotted the slope of the best-fit segment for each embedding dimension, interpreted as ν, as a function of embedding dimension (see Figure 6c for a typical plot) to see whether it converged to a stable value. Although this method of computing the correlation dimension does not afford an estimation of error, the cross-trial standard deviations of ν serve as a rough indication of error (see Holzfuss & Mayer-Kress, 1986, for direct error analysis methods).

Spectral analysis. The digital Fourier transform (DFT) was applied to the unfiltered position data from the steady-state trials. The 1024-point DFT on the data revealed no significant spectral components (viz., above the noise floor) above 20 Hz. Subsequently, trials were down-sampled, to 5:1, to produce an effective Nyquist frequency of 20 Hz, and then a 1024-point DFT was performed on the middle portion of the resultant 2000-sample files (Figure 7).

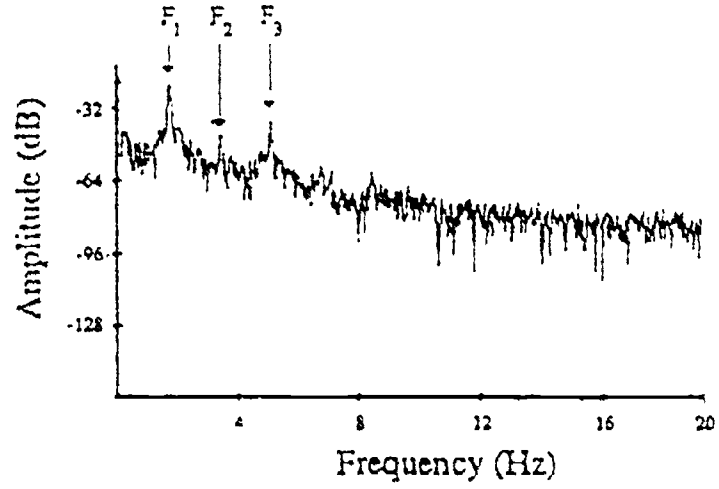


Figure 7. DFT plot for a typical steady state trial, from the downsampled time-series (5:1), Nyquist frequency = 20 Hz.

Spectral peaks were determined visually on a high-resolution graphics screen, and their amplitudes calibrated to degrees. Because of the amount of data that could reasonably be collected in the perturbation trials, DFTs will not be reported on the steady-state portions of those trials; the spectral resolution was very poor in these much smaller data sets.

Results

Perturbation-Trial Data

In 300 of the trials in which the perturbation was delivered in the flexion direction, the finger hit the flexion stop-post (this never happened for extension-pulse trials). We eliminated these trials from the phase response and relaxation time analyses, because these computations rely on the precise nature of the system's transient behavior, which may have been obscured by this effect. The resultant N for these measures was 962 trials. In 19 trials, the initial displacement of the trajectory from the average pre-perturbation cycle ($r(0) - r_L(0)$) was less than all subsequently measured displacements; that is, the trajectory was perturbed too near the average pre-perturbation cycle to perform a measurement of σ. The N for this measure, then, was 943. All trials were retained, however, for the following analyses of the other kinematic observables (frequency, amplitude, and peak velocity), since only the steady-state portions of the trials are used for these analyses.

Stability of kinematic observables. Tables 1 and 2 present the average pre- and post-perturbation values for within-trial frequency, amplitude, and peak velocity, averaged across all trials within each experimental session.

Table 1. *Pre- and post-perturbation frequency and amplitude. (mean values; standard deviations are in parentheses.)*

Subject	Session	Frequency (in deg.)		Amplitude (in deg.)	
		Pre	Post	Pre	Post
1	1	.702	.704	42.42	43.23
		(.173)	(.182)	(8.16)	(9.27)
	2	.379	.363	46.79	41.74
		(.042)	(.078)	(4.67)	(14.76)
2	1	.692	.687	73.84	74.91
		(.109)	(.116)	(3.48)	(3.85)
	2	.982	.983	68.75	68.22
		(.030)	(.035)	(2.97)	(3.37)
3	1	1.982	1.972	42.75	40.88
		(.830)	(.826)	(5.84)	(5.56)
	2	.830	.825	54.97	54.81
		(.053)	(.053)	(9.27)	(9.87)
4	1	.527	.528	55.67	55.06
		(.088)	(.084)	(4.05)	(3.67)
	2	.522	.519	56.41	56.40
		(.036)	(.035)	(5.57)	(5.47)
Mean		.828	.824	55.04	54.28
		(.483)	(.483)	(12.21)	(13.99)

Table 2. *Pre- and post-perturbation peak velocity, in degs/s. (mean values; standard deviations are in parentheses.)*

Subject	Session	Extension		Flexion	
		Pre	Post	Pre	Post
1	1	124.1	137.2	137.3	150.3
		(29.3)	(33.3)	(30.3)	(39.6)
	2	79.0	79.4	88.2	87.8
		(19.4)	(24.0)	(19.3)	(26.9)
2	1	155.5	159.1	178.7	181.9
		(29.2)	(31.3)	(38.2)	(40.0)
	2	217.8	214.2	226.6	234.6
		(17.9)	(16.8)	(36.8)	(40.0)
3	1	359.7	337.6	372.5	360.0
		(38.9)	(38.8)	(34.6)	(35.0)
	2	263.5	253.4	251.0	240.4
		(93.8)	(84.9)	(83.8)	(72.1)
4	1	178.3	181.8	139.9	145.5
		(36.6)	(36.9)	(35.0)	(38.6)
	2	192.9	197.2	163.0	163.3
		(27.0)	(27.4)	(25.0)	(27.2)
Mean		196.8	195.4	194.8	195.6
		(92.0)	(83.8)	(93.0)	(88.0)

In 2×2×2×2 repeated-measures analyses of variance (ANOVAs), with session, perturbation direction (toward extension or flexion), perturbation magnitude (small or large), and time of measurement (pre-, or post-perturbation) as variables, no main effects or interactions were found for either frequency or amplitude. For peak velocity, a 2×2×2×2×2 ANOVA was performed, with the additional factor of cycle half (extension or flexion), and the only significant effects found were the Session × Time of Measurement two-way interaction ($F[1,3] = 10.66$, $p < .05$), which was embedded within the four-way Session × Direction × Magnitude × Time of Measurement interaction ($F[1,3] = 11.54$, $p < .05$), neither of which are of interest here. Overall, pre- and post-perturbation values for all three kinematic observables were quite similar. In addition, there was no systematic difference between peak velocity going into extension and peak velocity going into flexion.

In summary, the stability of the kinematic observables of frequency, amplitude, and peak velocity indicate that a periodic attractor is indeed present.

Relaxation time and attractor strength. Table 3 lists mean relaxation time T_{rel} and attractor strength σ collapsed across all trials within each experimental session. A 2×2×2 ANOVA on T_{rel}

with session, perturbation direction, and perturbation magnitude as variables revealed only a main effect of perturbation magnitude ($F[1,3] = 192.71$, $p < .001$), with T_{rel} being longer for the stronger perturbations, averaging 551 ms as opposed to 395 ms for the weaker pulses. Also, the range of T_{rel} values observed was not large—the range of the session means was only 204 ms, whereas the range of mean cycle periods was 2,160 ms.

Table 3. *Relaxation time and attractor strength. (mean values).*

Subject	Session	T_{rel} (ms)	σ (1/s)
1	1	548	8.99
	2	418	10.13
2	1	488	9.15
	2	486	10.73
3	1	499	8.53
	2	344	9.89
4	1	437	8.62
	2	440	9.39
Mean		458	9.41

The same ANOVA on σ revealed significant main effects of session ($F[1,3] = 24.70$, $p < .05$) and perturbation magnitude ($F[1,3] = 48.18$, $p < .01$). The σ was smaller (indicating a weaker attractor) in the first session than the second (8.82 and 10.05 respectively), and was larger for the weak torque pulses than for the strong ones (10.10 and 8.58 respectively). The three-way Session × Direction × Magnitude interaction was also significant ($F[1,3] = 36.91$, $p < .01$).

We performed one-way ANOVAs on T_{rel} and σ to determine how uniform each was with respect to the phase angle of perturbation onset. Observed onset angles were binned into eight sectors, 0°-45°, 45°-90°, ..., 315°-360°. Phase angle had no effect on T_{rel} ($F[7,21] = 1.09$, $p > .1$), but significantly affected σ ($F[7,21] = 3.09$, $p < .05$). As can be seen in Figure 3, σ increases with increasing phase angle, starting from about 315° and going counterclockwise on the phase plane.

In order to assess how T_{rel} and σ scale with movement frequency, we performed correlations across all subjects' data, because the range of observed frequencies within each subject's data was too small to allow such an analysis. The linear correlation of log(T_{rel}) and log(frequency) was not significant, $r(N = 961) = .05$, $p > .1$. The linear correlation of log(σ) and log(frequency) was also nonsignificant, $r(N = 943) = -.050$, $p > .1$.

In summary, neither relaxation time nor attractor strength were correlated with movement frequency, that is, both were effectively constant across the range of frequencies we observed. The range of T_{rel} values was rather small, and σ was not uniform on the phase plane.

Phase response. Only trials in which the mean post-perturbation frequency differed from the mean pre-perturbation frequency by less than or equal to ±5% were used in this analysis. 823 of the 962 no-hit trials met this criterion. A typical phase transition curve (PTC; old phase ϕ vs. new phase ϕ') is shown in Figure 8, along with the data's best periodic fit. For each of the subjects' PTCs, the data tended to scatter around a line parallel to the $\phi' = \phi$ line.[1]

Accordingly, the linear slope parameter a in Equation 5 was fixed to 1.0, and the remainder of the parameters were determined by least-squares fitting. Three constrained multiple regressions were performed. First, the intercept was forced to zero, and the coefficients on the sine and cosine terms were obtained, as well as the overall R^2. The second regression fixed the sine and cosine terms to zero, leaving the intercept free to vary. The third fit was performed with all parameters (except slope) free to vary. Table 4 lists the intercepts found in the third fit and the R^2s for all three fits.

Table 4. *Coefficients of the Best Biperiodic Fits for Each Session.*

Subject	Session	Intercept	R^2 fit 1	R^2 fit 2	R^2 fit 3
1	1	.0631	.869	.882	.892
	2	.0488	.895	.902	.923
2	1	.1094	.662	.785	.812
	2	.1030	.629	.612	.666
3	1	.1217	.619	.768	.776
	2	.0263	.848	.833	.855
4	1	.0595	.882	.895	.916
	2	.0681	.804	.793	.856

A paired t-test comparing the R^2s of the first and third fits indicates that the intercept added significantly to the regression of new phase on old phase ($t(7) = 2.948$, $p < .05$). The intercept was always in the interval (.0, .5), indicating that the post-perturbation rhythm phase-led the control rhythm; that is, the phase response averaged

across all old phases in all cases was a phase advance—no phase delays were observed.[2] The mean phase shift was .0750, significantly greater than zero (one-tailed $t(7) = 6.43$, $p < .01$). Incremental *F*s (Kerlinger & Pedhazur, 1973) for each Subject × Session combination were all significant at the .05 level or better (range: $F(1,106) =$ ` to $F(1,123) = 86.45$).

Most .e flexion-pulse trials that were omitted .n this analysis (due to subjects hitting the stop) were delivered during the flexion portion of the cycle, and so these phases were somewhat underrepresented in the phase response analysis. Thus, the possibility exists that the fits obtained from the overall data set were biased. Accordingly, separate analyses (fits 1 and 3 described earlier) were carried out on the extension- and flexion-pulse data. In neither case did the paired t-test comparing the R^2s of fits 1 and 3 reach significance ($t(7) = 1.46$ and 1.37, for extension and flexion respectively), probably because the Ns in these sub-analyses were so low. However, for the *extension-pulse* data, in which few data points were eliminated, 6 of 8 of the incremental *F*s were significant (range: $F(1,61) = 2.56$ to $F(1,70) = 60.84$), and the corresponding overall phase shifts were phase advances. The two nonsignificant intercepts were also in the phase advance direction. For the *flexion-pulse* data, only 3 of 8 incremental *F*s reached significance (range: $F(1,27) = .471$ to $F(1,29) = 61.22$). Again, these were phase advances, and 4 of the remaining 5 intercepts were in the direction of phase advance. The only phase delay observed in the entire data set was nonsignificant. Phase advance did not depend on direction of perturbation ($t(7) = 0.86$, $p > .4$), averaging .0710 and .0596 for extension and flexion pulses respectively, both of which were significantly greater than zero (one-tailed $t(7) = 5.99$, $p < .01$, and 3.14, $p < .05$, respectively).

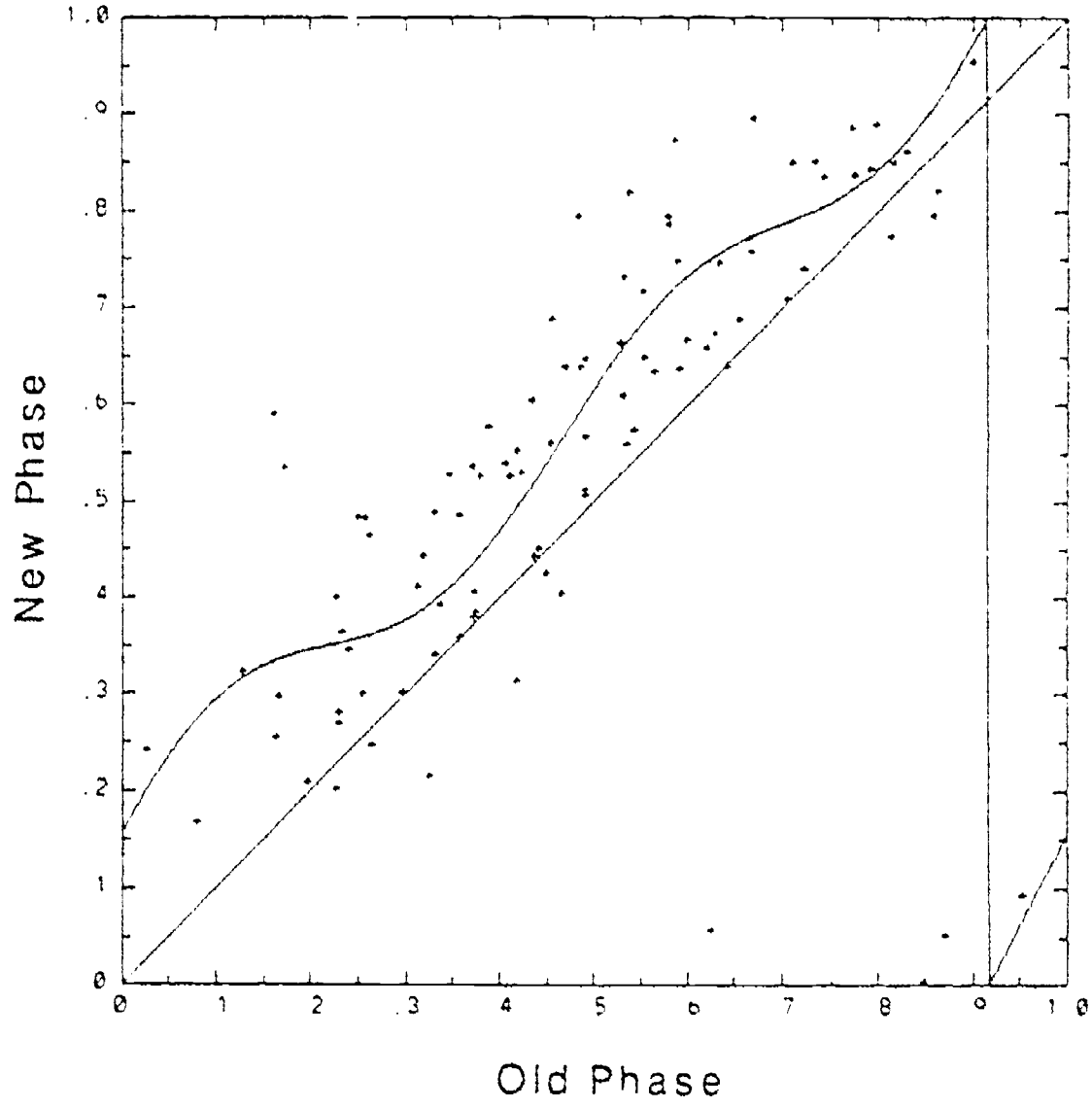


Figure 8. **Phase transition curve data, and biperiodic fit, for Subject 2's Session 1 results.**

Similar results were obtained when the two perturbation-strength conditions were separately analyzed. A paired t-test of the R^2s comparing fits 1 and 3 was significant for the weak condition ($t(7) = 2.58$, $p < .05$), but was not significant for the strong condition ($t(7) = 1.875$, $p > .1$). Six out of eight of the incremental Fs reached significance in both weak and strong conditions (weak range: $F(1,55) = .315$ to $F(1,52) = 58.28$; strong range: $F(1,33) = 1.598$ to $F(1,61) = 62.13$). The magnitude of the phase advance did not depend on perturbation strength ($t(7) = 0.43$, $p > .6$), with cross-subject means of .0675 and .0589 for the weak and strong perturbations respectively, both of which were significantly greater than zero (one-tailed $t(7) = 4.73$, $p < .01$, and 2.95, $p < .05$, respectively).

In addition to the overall phase advance, there was a consistent phase-dependent *pattern* of phase shift, as revealed by the contributions of the sine and cosine terms. A paired t-test comparing the R^2s of fits 2 and 3 was significant ($t(7) = 4.024$, $p < .01$). In 6 of the 8 sessions the incremental Fs were significant (range: $F(4,106) = 2.448$ to $F(4,86) = 9.554$). Collapsing the data across these six sessions, there were two old phases at which the phase advance reached local maxima, at $\phi = .22$ and $\phi = .67$. These correspond to about halfway through the flexion and extension half-cycles of movement, respectively. Furthermore, there were local minima at approximately $\phi = .4$ and $\phi = .96$, corresponding roughly to peak extension and peak flexion (see Figure 9).

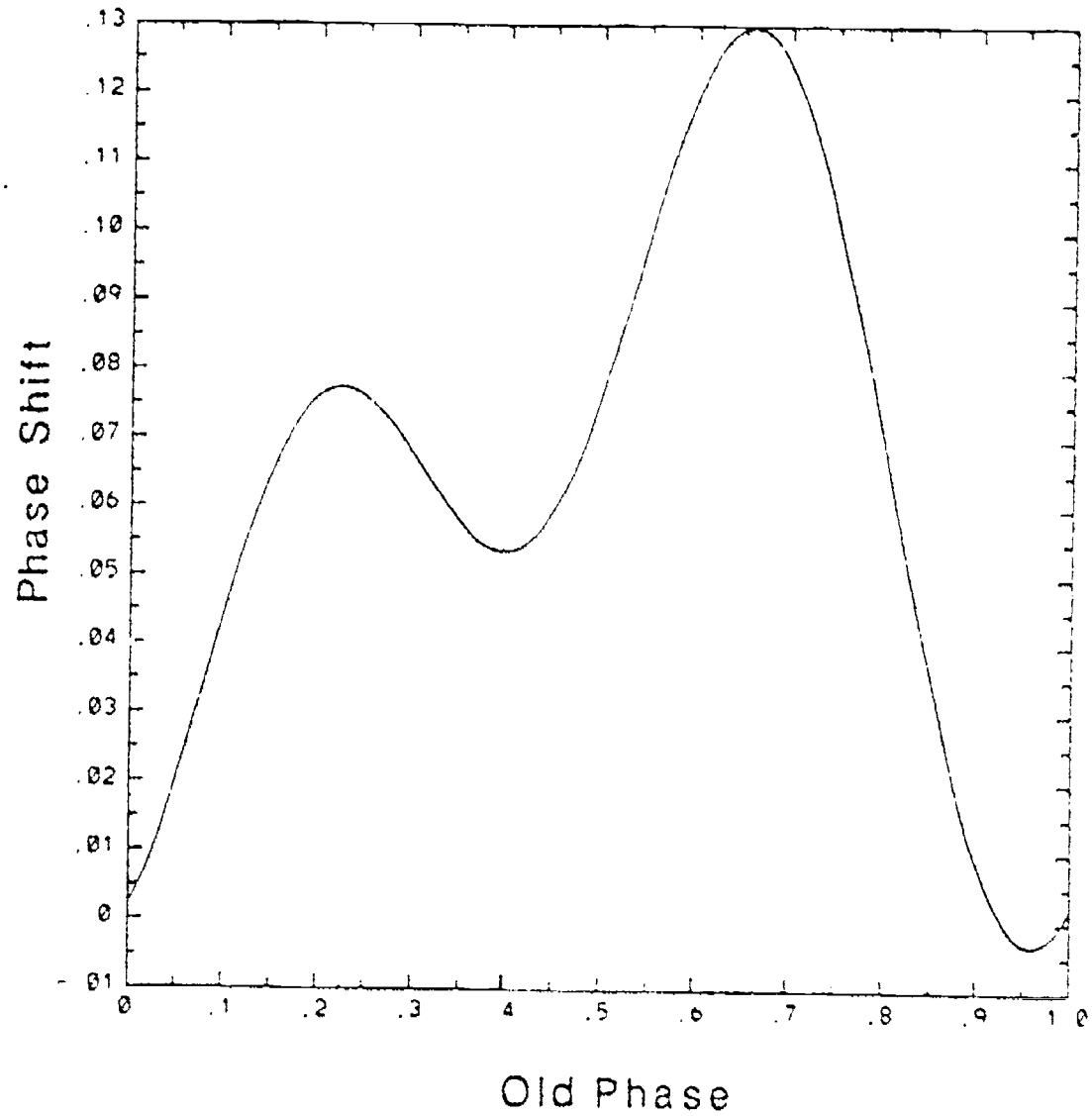


Figure 9. **Phase shift ($\Delta\phi$) as a function of old phase (ϕ), for the biperiodic fit, across all sessions that showed a significant effect for sinusoidal components.**

The Ns for the extension- and flexion-pulse and weak- and strong-pulse data subsets were too small to allow separate analyses of phase-dependent patterns within these conditions, because of the large number of parameters in these fits.

In all of the results here, no differences were found between the non-musicians (Subjects 1 & 3) and musicians (Subjects 2 & 4) (cf. Yamanishi et al., 1979).

In summary, we found that the rhythm was consistently phase-shifted overall (i.e., phase-advanced across all sampled old phases) and that there was a consistent phase-dependent pattern of phase shift. Also, phase response did not depend on direction or magnitude of perturbation.

Steady-State Trial Data

Dimensionality. For all steady-state trials, the computed correlation dimension ν did not converge to a stable value with increasing embedding dimension, but rather continued to increase slightly, a pattern frequently observed in such analyses (Mayer-Kress, 1986; see Figure 6c). For a few trials, the computation was repeated for embedding dimensions k = 1 to 20, and the same lack of convergence occurred. Grassberger and Procaccia's (1983) algorithm is known to introduce systematic errors in the estimates of ν for high values of embedding dimensions (Mayer-Kress, 1986), but the embedding dimension used for the final determination of ν should be as large as possible, to ensure that it is larger than the dimension of the underlying dynamics. Mayer-Kress and Layne (1986) have recommended that the correlation dimension be measured from the results for k = 7, for low dimensional attractors. Table 5 reports the within-session means and standard deviations of the computed correlation dimension ν and the scaling interval endpoints over which the computation was performed, for k = 7.

We performed 2×5 ANOVAs, with session and trial as variables, were performed on ν and the interval endpoints; no effects or interactions were found. The observed ν-values were similar to the value found for a sine wave with added quasi-random noise (1.20 ± .1, Abraham et al., 1986).[3] Note that the scaling intervals include the amplitudes of the fundamentals found in the spectral analysis (see Table 5). For amplitudes below the scaling interval, ν scaled linearly with embedding dimension; that is, it was approximately equal to k for all k, indicating infinite-dimensional (e.g., stochastic) behavior for these short lengths. For amplitudes above the scaling interval, the correlation integral C(L) saturated, indicating zero-dimensional behavior at the largest scale (i.e., at the largest length scales, all behavior looks like a point attractor; Kay, 1988).

Table 5. *Correlation dimension (ν), scaling interval, and spectral amplitudes of the steady-state trials. (mean values; standard deviations are in parentheses.)*

Subject	Session	ν	Scaling Interval (in degrees) From	To	F_1 (degs)	F_2 (degs)	F_3 (degs)	F_4 (degs)
1	1	1.206	4.63	5.82	4.74	.166	.328[a]	.052
		(.071)	(.52)	(.63)	(.75)	(.105)	(.117)	(.049)
	2	1.200	4.92	6.04	5.16	.188	.222	.054
		(.036)	(.73)	(.86)	(.60)	(.181)	(.046)	(.074)
2	1	1.174	7.39	9.72	8.70	.528	.232	.052
		(.022)	(.46)	(.62)	(.71)	(.180)	(.283)	(.116)
	2	1.182	7.51	9.24	8.14	.376	.112	.000
		(.036)	(.57)	(.36)	(.79)	(.242)	(.157)	(.000)
3	1	1.199	5.34	7.20	5.56	.154	.382[a]	.014
		(.045)	(.50)	(.68)	(.88)	(.095)	(.054)	(.020)
	2	1.103	6.36	8.37	7.58	.066	1.088[a]	.000
		(.128)	(1.21)	(1.08)	(1.39)	(.148)	(.680)	(.000)
4	1	1.104	5.28	7.24	6.76	.294	.772[a]	.114
		(.047)	(.32)	(.49)	(1.02)	(.419)	(.280)	(.110)
	2	1.149	6.68	9.41	7.52	.498	.940[a]	.224
		(.024)	(.82)	(.66)	(1.00)	(.237)	(.055)	(.135)
Mean		1.165	6.01	7.88	6.77	.271	.510[b]	.064
		(.068)	(1.12)	(1.53)	(1.44)	(.156)	(.337)	(.073)

[a] $F_3 > F_2$, $p < .05$. [b] $F_3 > F_2$, $p < .01$.

In summary, the dimensionality results indicate that the attractor is a low dimensional one, consistent with a single oscillatory process at the medium length scales coexisting with a stochastic process at the smaller length scales.

Spectral analysis. Table 5 lists the amplitudes of the first four spectral peaks found in the dimension trials, averaged within session, from the fundamental F_1 to the fourth partial of the fundamental, F_4. All partials found were harmonically related to the fundamental (within the resolution of the DFT). The data exhibit relatively weak harmonic content, the sum of the amplitudes of F_2, F_3, and F_4 on average being only 12% of the amplitude of F_1. However, the third partial was greater in amplitude than the second (paired-trial $t(39) = 2.724, p < .01$).

In summary, the odd harmonics appear to predominate in the movements' spectra, consistent with the simple nonlinear oscillators discussed in the Introduction.

DISCUSSION

To a great extent, these rhythmical movements display the dynamical behavior of very simple limit-cycles. The stability of the kinematics in the face of perturbation indicates that an attractor is present. The dimensionality results show that this attractor is of low dimension, not inconsistent with a one-dimensional limit-cycle at the intermediate amplitude scales and stochastic noise at the smaller amplitude scales.

Many of the present results are consistent with a previously derived hybrid model (Kay et al., 1987). The hybrid oscillator is a combination of the van der Pol and Rayleigh oscillators:

$$m\ddot{x} + \alpha\dot{x} + \beta\dot{x}^3 + \gamma(x-x_0)^2\dot{x} + k(x-x_0) = 0 \qquad (10)$$

where β is the coefficient of the Rayleigh nonlinear damping term (cf. Equation 2). Kay et al. found that this oscillator best modeled the covariation of frequency and amplitude of wrist movements in the unperturbed case. In the present data set, the observed relaxation time is a function of the magnitude of perturbation only and was constant across movement frequency, as predicted by the hybrid model. The attractor strength is also constant across movement frequency, also as predicted by the hybrid model. A further agreement with this model is the spectra of the steady-state trials. The harmonic structure of the hybrid (as well as the van der Pol and Rayleigh) oscillator contains only every other harmonic for a wide range of attractor strength. Although our movement data exhibited both even and odd harmonics, the observed pattern—with the even harmonic F_2 relatively attenuated compared with the odd harmonic F_3—was similar in form to the harmonic structure of these simple nonlinear oscillators.

However, there are two discrepancies between the model and the present data, both of which were revealed by perturbation. First, the movement's strength of attraction is non-uniform on the phase plane, unlike the hybrid model. Thus, the behavior of the movement when it is not on the limit-cycle is different from that of the model.

Second, the observed pattern of phase response differs from the pattern predicted by the hybrid or similar oscillators. Although these oscillators can be phase-shifted by mechanical perturbation at particular old phases, they do not exhibit an overall phase-shift when averaged across all possible old phases. The model of mechanical perturbation to the hybrid oscillator, for example, is as follows: When the perturbation is off, the equation of motion is Equation 10. When the perturbation is on, a torque pulse is added as a forcing function:

$$m\ddot{x} + a\dot{x} + \beta\dot{x}^3 + \gamma(x-x_0)^2\dot{x} + k(x-x_0) = \Gamma \qquad (11)$$

In effect, the perturbation serves only to reset the initial conditions, that is, the mechanical state of the system, to effectively new starting values. For small Γ, the post-perturbation initial conditions are only slightly different from the pre-perturbation locus, and the phase shift produced for any particular old phase (ϕ) is small, diverging only slightly from the $\phi = \phi'$ line. Consider one direction of perturbation: Suppose Γ is positive, and, in a model of our rhythmic movements, corresponds to extension pulses. In the oscillator's "extension" half-cycle, Γ has the effect of speeding up the motion, because it is in the direction of motion and so assists it in getting to peak extension. Thus, in this half-cycle the hybrid oscillator is phase-*advanced* by an extension pulse. In the flexion half-cycle, an extension pulse has the effect of slowing down the motion, because it is in the direction opposite to the motion, holding it back from peak flexion. In this half-cycle, the oscillator is phase-*delayed* by an extension pulse. With uniform sampling of perturbation phase over the entire cycle, these effects balance exactly, and no average phase shift results whatsoever (see Figure 10).

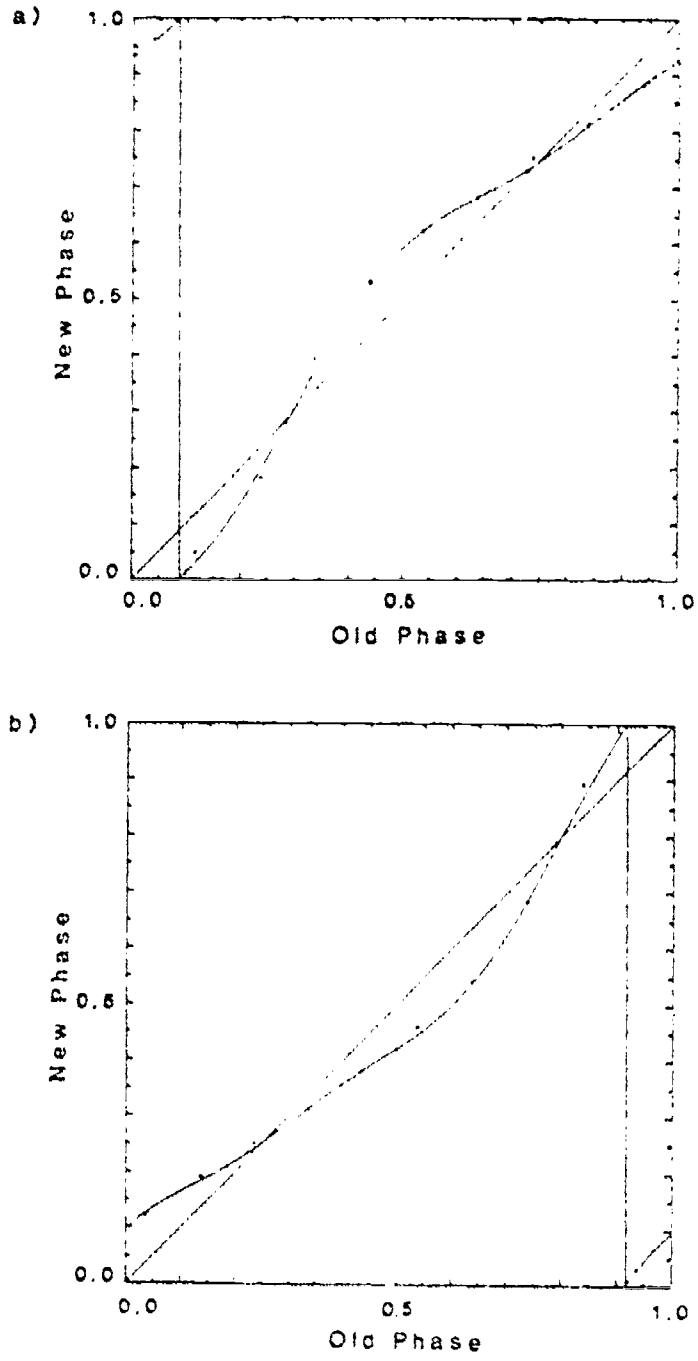


Figure 10. **Phase transition curve (ϕ vs. ϕ') of the simulated hybrid oscillator, perturbed with (a) weak extension and (b) weak flexion torque pulses.**

Thus, when the PTC for the hybrid is fitted across all possible perturbation phases, the intercept is always nonsignificant for small perturbations, in contrast to the significant positive intercepts found in the present data.[4] In addition, the phase-dependent pattern of phase shift for the hybrid oscillator is quite different from the observed data: instead of two peaks at roughly halfway through the extension and flexion half-cycles, this oscillator's PTC exhibits a peak during the extension half-cycle and a trough during the flexion half-cycle, for extension pulses. The pattern of advances and delays is reversed for flexion pulses ($\Gamma < 0$), but, again, the overall phase shift is zero and only one peak occurs in the PTC.

It seems likely, therefore, that second-order, time-invariant dynamical systems may be inadequate for capturing the patterns of phase-response of our movement data. One possibility is to modify the assumption of time-invariance: perhaps, in addition to transiently affecting the mechanical *state* of the system (which it surely does), the perturbation has an additional transient effect on the system's *parameters*.

The overall phase advance we observed indicates that, on average, the movement temporarily speeds up in response to the perturbation. This would occur if the stiffness of the movement system temporarily increases, either during the perturbation itself or for a short period of time afterward, or both. Stiffening up the hybrid oscillator leads to the following equation of motion for the duration of the perturbation:

$$m\ddot{x} + \alpha\dot{x} + \beta\dot{x}^3 + \gamma(x-x_0)^2\dot{x} + k(\Gamma)(x-x_0) = \Gamma \qquad (12)$$

where k is now a function of the perturbation. In order to match the experimental results, this stiffening function should be independent of both magnitude and direction of the torque pulse (Γ), and so may be written as

$$k = \begin{cases} k_0 \text{ when } \Gamma = 0 \\ k_0 + \Delta k \text{ when } \Gamma \neq 0 \end{cases} \qquad (13)$$

where k_0 is the stiffness during steady-state portions, and Δk is some positive number. This model's PTC has an overall phase advance, but it has the same phase-dependent pattern of phase shift (one maximum and one minimum) as the hybrid in Equation 11. In order to match the data's phase-dependent pattern, we must choose another stiffness function, and it may have to be a radically different one. It is also possible, of course, that the hybrid is simply the wrong candidate dynamical system, and other systems of different structure or higher order are required to model our data.

One example is composed of two oscillatory components: a central nervous system oscillator driving a peripheral limb segment with its own oscillatory biomechanical dynamics. Counting each oscillatory component as a second-order process, this situation would be described by a fourth-order system of differential equations. Our results suggest that if this model is to be taken seriously, the central oscillator is not independent of the limb's dynamics (Grillner, 1981; Saltzman

& Kelso, 1987; Turvey, Rosenblum, Schmidt, & Kugler, 1986; Wing, 1980). If there is a central timekeeper, it is affected by perturbations delivered to the limb being controlled. In other words, *the coupling between the central timer and the peripheral musculo-skeletal oscillator is fundamentally bi-directional,* not uni-directional. Explicit central pattern generator models for this activity must, therefore, include feedback from the peripheral, controlled system.[5]

In summary, we have obtained a fairly complete characterization of the dynamical behavior of a simple rhythmical voluntary movement. Many of these characteristics agree quite closely with simple one-dimensional limit-cycle dynamics. However, certain aspects of the present data set do not, and work remains to be done to understand these differences. The present data set constrains the form that any model of this type of behavior might assume; some classes of dynamics (e.g., some forms of second-order dynamics with time-invariant parameterizations) have been excluded as possible models. It remains to be seen whether these results generalize to different classes of rhythmic movement—for example, involuntary tremor oscillations or multi-joint and multi-limb tasks—or whether, in those instances, qualitatively different dynamics prevail.

REFERENCES

Abraham, N. B., Albano, A. M., Das, B., de Guzman, G., Yong, S., Gioggia, R.S., Puccioni, G. P., & Tredicce, J. R. (1986). Calculating the dimension of attractors from small data sets. *Physics Letters, 114A*, 217-221.

Abraham, R. H., & Shaw, C. D. (1982). *Dynamics-The geometry of behavior*. Santa Cruz, CA: Ariel Press.

Andronov, A., & Chaikin, C. E. (1949). *Theory of oscillations*. Princeton, NJ: Princeton University Press.

Bergé, P., Pomeau, Y., & Vidal, C. (1984). *Order within chaos: Towards a deterministic approach to turbulence*. New York: Wiley.

Carpenter, G. A., & Grossberg, S. (1983). A neural theory of circadian rhythms: The gated pacemaker. *Biological Cybernetics, 48*, 35-59.

Fraser, A. M., & Swinney, H. L. (1986). Independent coordinates for strange attractors from mutual information. *Physical Review A, 33*, 1134-1140.

Grassberger, P., & Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica, 9D*, 189-208.

Grillner, S. (1981). Control of locomotion in bipeds, tetrapods, and fish. In J. M. Brookhart & V. B. Mountcastle (Eds.), *Handbook of physiology, Section 1: The nervous system, Vol. II: Motor control, Part 1* (pp. 1179-1236). Bethesda, MD: American Physiological Society.

Grillner, S., & Zangger, P. (1979). On the central generation of locomotion in the low spinal cat. *Experimental Brain Research, 34*, 241-262.

Holzfuss, J., & Mayer-Kress, G. (1986). An approach to error-estimation in the application of dimension algorithms. In G. Mayer-Kress (Ed.), *Dimensions and entropies in chaotic systems: Quantification of complex behavior* (pp. 114-122). New York: Springer-Verlag.

Jordan, D. W., & Smith, P. (197). *Nonlinear ordinary differential equations*. Oxford, England: Clarendon.

Kay, B. A. (1988). The dimensionality of movement trajectories and the degrees of freedom problem: A tutorial. *Human Movement Science, 7*, 343-364.

Kay, B. A., Kelso, J. A. S., Saltzman, E. L., & Schöner, G. (1987). The space-time behavior of single and bimanual rhythmical movements: Data and model. *Journal of Experimental Psychology: Human Perception & Performance, 13*, 178-192.

Kay, B. A., Munhall, K. G., V.-Bateson, E., & Kelso, J. A. S. (1985). A note on processing kinematic data: sampling, filtering, and differentiation. *Haskins Laboratories Status Report on Speech Research, SR-81*, 291-303.

Kelso, J. A. S., Holt, K. G., Rubin, P., & Kugler, P. N. (1981). Patterns of human interlimb coordination emerge from the properties of non-linear limit-cycle oscillatory processes: Theory and data. *Journal of Motor Behavior, 13*, 226-261.

Kerlinger, F. N., & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York: Holt, Rinehart, and Winston.

Kugler, P. N., & Turvey, M. T. (1987). *Information, natural law, and the self-assembly of rhythmic movement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kugler, P. N., Turvey, M. T., Schmidt, R. C., & Rosenblum, L. D. (1990). Investigating a nonconservative invariant of motion in coordinated rhythmic movements. *Ecological Psychology, 2*, 151-189.

Lennard, P. (1985). Afferent perturbations during "monopodal" swimming movements in the turtle: Phase-dependent cutaneous modulation and proprioceptive resetting of the locomotor rhythm. *Journal of Neuroscience, 5*, 1434-1445.

Mandelbrot, B. B. (1983). *The fractal geometry of nature*. San Francisco: Freeman.

Mayer-Kress, G. (Ed.) (1986). *Dimensions and entropies in chaotic systems: Quantification of complex behavior*. New York: Springer.

Mayer-Kress, G., & Layne, S. (1987). Dimensionality of the human electroencephalogram. In S. H. Koslow, A. J. Mandell, & M. F. Schlesinger (Eds.), *Proceedings of the New York Academy of Sciences Conference on Perspectives in Biological Dynamics & Theoretical Medicine* (pp. 62-89). New York: New York Academy of Sciences.

Minorsky, N. (1974). *Nonlinear oscillations*. New York: Krieger.

Pittendrigh, C. S., & Daan, S. (1976). A functional analysis of circadian pacemakers in nocturnal rodents. IV. Entrainment: Pacemaker as clock. *Journal of Comparative Physiology, 106*, 291-336.

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

Selverston, A.I. (1980). Are central pattern generators understandable? *Behavioral and Brain Sciences, 3*, 535-571.

Stein, P. S. G. (1976). Mechanisms of interlimb phase control. In R. N. Herman, S. Grillner, P. S. G. Stein, & D. G. Stuart (Eds.), *Neural control of locomotion* (pp. 465-487). New York: Plenum Press.

Thompson, J. M. T., & Stewart, H. B. (1986). *Nonlinear dynamics and chaos: Geometrical methods for engineers and scientists*. New York: Wiley.

Turvey, M. T., Rosenblum, L. A., Schmidt, R. C., & Kugler, P. N. (1986). Fluctuations and phase in coordinated rhythmic movements. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 564-583.

van der Pol, B. (1926). On relaxation-oscillations. *Philosophical Magazine, 7, 2*, 978-992.

Winfree, A. T. (1980). *The geometry of biological time*. New York: Springer.

Wing, A. M. (1980). The long and short of timing in response sequences. In G. E. Stelmach and J. Requin (Eds.), *Tutorials in motor behavior* (pp. 469-486).. *Amsterdam: North-Holland*.

Yamanishi, J., Kawato, M., & Suzuki, R. (1979). Studies on human finger tapping neural networks by phase transition curves. *Biological Cybernetics, 33*, 199-208.

FOOTNOTES

Journal of Experimental Psychology: Human Perception and Performance, in press.

†Also Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology; currently at Department of Psychology, Brown University, Providence.

††Center for Complex Systems, Florida-Atlantic University, Boca Raton.

[1]Winfree (1980) has termed this type of PTC Type 1 phase resetting, because the average slope of the PTC is 1. Another type of PTC occurs when the average slope is 0. In this case, the rhythm is strongly reset, in the sense that all information about the phase of perturbation is lost. Type 0 behavior did not occur in our experiment.

[2]Phases in this interval are conventionally termed advances, whereas in the interval (0.5, 1.0) they are termed delays (see Winfree, 1980, p. 145, Figure 14).

[3]We also performed the computation on simulated data consisting of a digitally-generated sine wave with added quasi-random noise; v converged to a value of 1.26. with some variability in the next significant digit.

[4]For larger perturbations, the hybrid, van der Pol, and Rayleigh oscillators exhibit Type 0 phase resetting.

[5]It is interesting to note that Lennard (1985) reached a similar conclusion in his study of perturbed monopodal swimming movements in the turtle. In perturbing the turtle triceps muscle nerve, he obtained a consistent phase shift, also a phase advance, but for only two phases of perturbation. Intriguingly, these phases (at ϕ = .25 and ϕ = .85) correspond roughly to the points of maximum phase shift found in our experiment.

APPENDIX

Derivation of a Measure of Limit-Cycle Attractor Strength

Assume that the return to the limit-cycle following perturbation is a linear relaxation process in amplitude only; that is, the return trajectory spirals back to the limit-cycle, with the radial displacement from the limit-cycle decaying exponentially back. Let

$$r'(t) = r(t) - r_L(t) \tag{A1}$$

where r(t) is the radial displacement of the return trajectory at time t from the center of oscillation and $r_L(t)$ is the radial displacement of the limit-cycle from the center of oscillation at the same phase angle of the return trajectory at time t. The assumption of linear relaxation can now be explicitly written as

$$\frac{dr'}{dt} = -\sigma r' \tag{A2}$$

Integration gives the solution:

$$r'(t) = r'(0)e^{-\sigma t} \tag{A3}$$

where $r'(0)$ is the displacement from the limit-cycle at t = 0. Solving for σ gives

$$\sigma = -\frac{1}{t} \ln \left| \frac{r'(t)}{r'(0)} \right| \tag{A4}$$

or

$$\sigma = \frac{1}{t} \ln \left| \frac{r(0) - r_L(0)}{r(t) - r_L(t)} \right| , \tag{A5}$$

where ln is the natural logarithm function. The σ can be measured any time after the perturbation is turned off, but for practical purposes, some end to the return process must be defined. Under the decay assumption, the computation is valid for trajectories starting both outside and inside the limit-cycle. Also, the return process takes an infinite amount of time, but a real return would be buried in the variation of the observed behavior after some finite time. Furthermore, the computation is valid only when the numerator in the argument of the logarithmic function is larger than the denominator.

Note that the σ in this article corresponds to α/m in Equation A9 in Kay et al. (1987). That is, it represents the relaxation constant of the linearized return process of the van der Pol oscillator, normalized to unit mass. The return processes of the van der Pol, Rayleigh, and hybrid oscillators are all nonlinear (Jordan & Smith, 1977). They are more nonlinear given larger nonlinear damping terms and greater distances from the limit cycle. We have chosen to assume linearity in order to simplify the extraction of our attractor strength measure.

Appendix

SR #	Report Date	DTIC #	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 252-90
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/AS	
SR-97/98	January-July 1989	PB 90-121161/AS	ED321317
SR-99/100	July-December 1989	PB 90-226143/AS	ED321318
SR-101/102	January-June 1990	PB 91-138479	
SR-103/104	July-December 1990		

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

END

U.S. Dept. of Education

Office of Educational
Research and Improvement (OERI)

ERIC

Date Filmed
September 24, 1991