DOCUMENT RESUME

ED 128 870 CS 501 502

TITLE	Status Report on Speech Research: A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for Its Investigation, and Practical Applications, July 1 - September 30, 1976.
INSTITUTION	Haskins Labs., New Haven, Conn.
REPORT NO	SR-47-(1976)
PUB DATE	76
NOTE	172p.
EDRS PRICE	MF-$0.83 HC-$8.69 Plus Postage.
DESCRIPTORS	Acoustics; *Articulation (Speech); Educational Research; Higher Education; Music; Nonverbal Communication; *Oral Communication; *Research; Research Methodology; *Speech; *Speech Skills; Theories

ABSTRACT

This report, covering the period in 1976 from July 1 through September 30, is one of a regular series on the status and progress of speech research. Manuscript topics are: stop-consonant recognition--release bursts and formant transitions as functionally equivalent, context-dependent cues; modes of perceiving; discrimination of intensity differences carried on formant transitions varying in extent and duration; discrimination functions predicted from categories in speech and music; right-ear advantage for musical stimuli differing in rise time; dichotic competition of sounds--the role of acoustic stimulus structure; distance measures for speech recognition--psychological and instrumental; laryngeal timing in consonant distinctions; phonetic aspects of time and timing; static and dynamic acoustic cues in distinctive tones; the effects of selective adaptation on voicing in Thai and English; perception of nonspeech by infants; categorical perception along an oral-nasal continuum; stop-voicing production--natural outputs and synthesized inputs; and shifts in vowel perception as a function of speaking rate. (JM)

SR-47 (1976)

Status Report on

SPEECH RESEARCH

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications

1 July - 30 September 1976

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

ACKNOWLEDGEMENTS

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

National Institute of Dental Research
Grant DE-01774

National Institute of Child Health and Human Development
Grant HD-01994

Assistant Chief Medical Director for Research and Development, Research Center for Prosthetics, Veterans Administration
Contract V101(134)P-342

Advanced Research Projects Agency, Information Processing Technology Office, under contract with the Office of Naval Research, Information Systems Branch
Contract N00014-76-C-0591

United States Army Electronics Command, Department of Defense
Contract DAAB03-75-C-0419(L 433)

National Institute of Child Health and Human Development
Contract N01-HD-1-2420

National Institutes of Health
General Research Support Grant RR-5596

HASKINS LABORATORIES

Personnel in Speech Research

Alvin M. Liberman,* President and Research Director
Franklin S. Cooper, Associate Research Director
Patrick W. Nye, Associate Research Director
Raymond C. Huey, Treasurer
Alice Dadourian, Secretary

Investigators

Arthur S. Abramson*
Thomas Baer
Peter Bailey[1]
Fredericka Bell-Berti*
Gloria J. Borden*
James E. Cutting*
Ruth S. Day*
Donna Erickson
Frances J. Freeman*
Jane H. Gaitenby
Thomas J. Gay*
Terry Halwes
Katherine S. Harris*
Alice Healy*
Isabelle Y. Liberman*
Leigh Lisker*
Ignatius G. Mattingly*
Paul Mermelstein
Seiji Niimi[2]
Lawrence J. Raphael*
Bruno H. Repp
Philip E. Rubin*
Donald P. Shankweiler*
Linda Shockey
George N. Sholes
Michael Studdert-Kennedy*
Quentin Summerfield[1]
Michael T. Turvey*
Robert Verbrugge*

Technical and Support Staff

Eric L. Andreasson
Elizabeth P. Clark
Donald Hailey
Harriet G. Kass*
Elly Knight*
Sabina D. Koroluk
Roderick M. McGuire
Agnes McKeon
Terry F. Montlick
Nancy R. O'Brien
Loretta J. Reiss
William P. Scully
Richard S. Sharkany
Edward R. Wiley
David Zeichner

Students*

Mark J. Blechner
Steve Braddon
David Dechovitz
Susan Lea Donald
F. William Fischer
Hollis Fitch
Carol A. Fowler
Morey J. Kitzman
Andrea J. Levitt
Roland Mandler
Leonard Mark
Sandra Prindle
Abigail Reilly
Robert Remez
Helen Simon
Emily Tobey
Harold Tzeutschler
James M. Vigorito

---

*Part-time
[1]Visiting from The Queen's University of Belfast, Northern Ireland.
[2]Visiting from University of Tokyo, Japan.

CONTENTS

I. Manuscripts and Extended Reports

II. Publications and Reports

III. Appendix: DDC and ERIC numbers (SR-21/22 - SR-45/46)

I. MANUSCRIPTS AND EXTENDED REPORTS

Stop-Consonant Recognition: Release Bursts and Formant Transitions as Functionally Equivalent, Context-Dependent Cues

M. F. Dorman,* M. Studdert-Kennedy,+ and L. J. Raphael++

ABSTRACT

Three experiments studied the roles of release bursts and formant transitions as acoustic cues to place of articulation in syllable-initial voiced stop consonants. Experiments I and II assessed the weight of these cues by systematically removing them from American English /b,d,g/, spoken before nine different vowels by two speakers. Experiment III assessed the functional invariance of the release burst by transposing it from the nine syllables of speaker 2 across all eight vowels for each class of stop consonant. The results showed that labial and apical bursts were largely invariant in their effect before all vowels; velar bursts before front vowels and velar bursts before central-back vowels were also invariant within their set. However, release bursts carried significant perceptual weight in only one syllable out of 27 for speaker 1, in only 13 syllables out of 27 for speaker 2. For speaker 2 labial and velar bursts carried significant weight primarily before central-back, rounded vowels, apical bursts primarily before high, front, unrounded vowels. Furthermore, burst and transition tended to be reciprocally related: where the perceptual weight of one increased, the weight of the other declined. They were thus shown to be functionally equivalent, context-dependent cues, each contributing to the rapid spectral changes that follow consonantal release. The results were interpreted as pointing to the important role played by the front-cavity resonance in signaling place of articulation.

---

*Also Arizona State University, Tempe.

+Also Queens College and Graduate Center of The City University of New York.

++Also Lehman College and Graduate Center of The City University of New York.

Acknowledgment: We thank Alvin Liberman, Paul Mermelstein and, especially, Gary Kuhn for their careful comments on an early draft of this paper. We thank Agnes McKeon for her expert preparation of the figures. We also thank Suzi Pollack and Tony Levas for help in running subjects and tabulating data. This research was supported in part by NIH HD01994.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

INTRODUCTION

The present paper deals with an aspect of the problem of perceptual constancy--the invariance problem--in speech recognition. At the level of phoneme recognition the problem is manifest in the variety of acoustic signals that may be categorized as the same phoneme. This variability arises from several sources. For a fuller discussion than can be given here, see Studdert-Kennedy (1974).

One source is differences among speakers' vocal-tract dimensions. Since the area function of the vocal tract determines the resonant (formant) frequencies by which a particular phoneme is cued, the formant patterns of signals produced by a child may be quite different from those produced by an adult: in fact, formant frequencies for a given vowel may differ by as much as 30 percent. Moreover, the formant frequencies for a child often approximate those of a different vowel spoken by an adult. Even within a single speaker, several sources contribute to vowel variability. Lindblom (1963), for example, found that formant frequencies may vary by a factor of 2.3:1, depending on whether the vowel is spoken in isolation or in consonantal context. Moreover, in rapid speech the tongue often does not reach the articulatory "targets" achieved in deliberate speech, so that vowel formant frequencies tend to be "reduced."

Phonetic context and rate also alter the acoustic cues for consonants. As an example of the effects of context, syllable-initial /b/ before the vowel /a/ is characterized by an upward spectral change, syllable-final /b/ following /a/, by a downward spectral change. For a second example, the voiced-voiceless distinction in stop consonants (/b/ vs. /p/, /d/ vs. /t/, /g/ vs. /k/) is cued primarily by voice onset time (VOT) (the interval between consonantal release and the onset of phonation) in dissyllables with stress on the second syllable; by the intersyllable interval in dissyllables with an unstressed second syllable; by vowel duration in syllable-final stops when unreleased, and by the spectrum of the release burst in syllable-final stops when released. As an example of the effects of rate, the VOT distributions of voiced and voiceless English stops do not overlap if the stops are spoken in citation form, but may overlap considerably if the stops are spoken in sentence context (Lisker and Abramson, 1967).

In the present paper we are concerned with yet another aspect of the invariance problem--the variation in acoustic cues for a given stop consonant as a function of the following vowel. Many studies have demonstrated that formant transitions are generally sufficient cues for stop-consonant recognition. Since the shape of these transitions varies with the following vowel, accounts of stop-consonant recognition have generally emphasized the role of context-conditioned cues (perhaps relational invariants) within the consonant-vowel syllable. Recently, however, Cole and Scott (1974a, 1974b) have suggested that stop consonants before different vowels may be recognized in terms of a context-independent acoustic cue (or simple invariant), namely, the burst produced at the release of stop-consonant occlusion.

In the following experiments we explore these cues in some detail with natural speech. We assess, first, the extent to which separable components

of the complex of acoustic cues for initial, voiced stop consonants--the release burst, the devoiced, and the voiced formant transitions--are sufficient cues for the perception of place of articulation. We ask, second, whether one of the components--the burst--is an invariant cue for stop-consonant recognition. Finally, we discuss the implications of our results for an account of stop-consonant recognition.

Acoustic Segmentation of Stop-Consonant-Vowel Syllables

Acoustic analysis of /bV, dV, gV/ syllables, reveals five qualitatively distinct segments before a stable vowel formant pattern is reached (cf. Fischer-Jørgensen, 1954, 1972; Halle, Hughes, and Radley, 1957; Fant, 1969): (1) a period of occlusion (usually silent, though occasionally voiced); (2) a transient explosion (usually less than 20 msec) produced by shock excitation of the vocal tract upon release of occlusion; (3) a very brief (0-10 msec) period of frication, as articulators separate and air is blown through a narrow (though widening) constriction, as in the homorganic fricative; (4) a brief period (2-20 msec) of aspiration, within which may be detected noise-excited formant transitions, reflecting shifts in vocal-tract resonances as the main body of the tongue moves toward a position appropriate for the following vowel; (5) voiced formant transitions, reflecting the final stages of tongue movement into the vowel during the first few cycles of laryngeal vibration. Since we are only concerned with stop consonants in the present study, we shall not consider the role of the first segment (occlusion) which serves to distinguish stops from vowels and other consonants. Furthermore, since the explosion and frication, even if separable on an oscillogram or spectrogram, are probably not discriminable by ear, we shall treat them in what follows as a single burst of energy, lasting some 2-30 msec.

The fourth segment (aspirated or devoiced formant transition), although usually distinguishable on an oscillogram with a high resolution time scale, is not always readily apparent on a spectrogram (see Figure 1). Investigators have therefore tended to discount it as an acoustic cue[1] and to concentrate attention on the burst and on the voiced formant transition. The present paper attempts to redress the balance by treating this segment as a separable component of the cue complex.

Bursts and Transitions as Cues for Stop Consonants

Research with synthetic speech has revealed that both bursts and voiced formant transitions may serve as separate cues to place of articulation of initial /b,d,g/. Many studies have shown that transitions of the second and third formants are sufficient cues for the place distinction (for example, Liberman, Cooper, Delattre, and Gerstman, 1954; Delattre, Liberman, and Cooper, 1955), and these are, in fact, the standard cues used in speech synthesis. It is important to note that--since the acoustic shape of formant transitions varies as a function of the following vowel--formant transitions are necessarily context-

---

[1] Voiceless transitions have been given due weight in studies of voiceless stops (Liberman, Delattre, and Cooper, 1958) and fricatives (Harris, 1958).

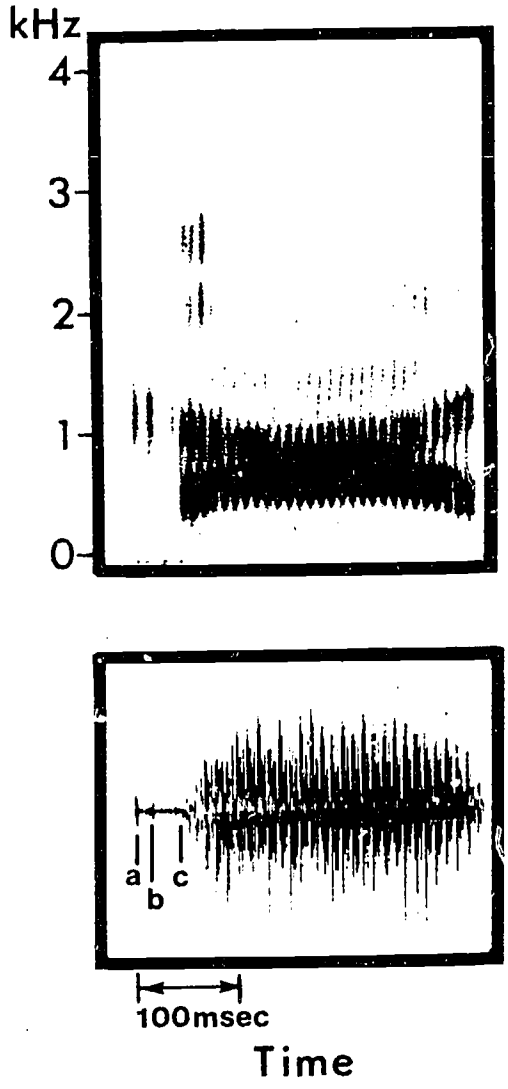


Figure 1: A spectrogram of the syllable /gʌd/, spoken by speaker 2 (top). An oscillogram of the same utterance is shown at the bottom: burst (a-b) and aspiration (b-c) duration and the onset of voicing (c) are indicated by vertical lines.

dependent cues for stop consonants. The same is true of velar bursts. Hoffman (1958) found that while bursts centered at frequencies above 3000 Hz acted as cues for /d/, burst cues for /g/ lay near the second formant of the vowel and were therefore context-dependent (cf. Liberman, Delattre, and Cooper, 1952). Hoffman could find no burst that would serve as a powerful cue for /b/, but this may have reflected, in part, the deficiencies of his synthesizer, rather than of natural speech.

In fact, attention has recently turned to the question of how cues isolated in synthetic speech experiments act and interact in naturally produced speech. With respect to the voiced stop consonants, Cole and Scott (1974b) have

questioned the role of the formant transitions in carrying phonetic information. These authors, following Day (1970) and Liberman, Mattingly, and Turvey (1972), have suggested that a major role of context-dependent formant transitions is to provide information about the temporal order of the segments in the speech signal. Cole and Scott (1974b) go further to suggest that phonetic information is carried primarily by a simple invariant cue, and that for /b,d,g/ the invariant place cue lies in the initial noise energy (burst and aspiration) before the onset of laryngeal vibration.

The latter claim drew apparent support from a recent experiment by Cole and Scott (1974a). Using a tape-splicing procedure to remove formant transitions from /bi,bu,di,du,gi,gu/, thus leaving burst and aspiration followed by steady-state vowel, Cole and Scott found that recognition of the syllables remained essentially unimpaired. Moreover, when the initial energy from /bi/ was transposed to /u/, or the initial energy from /bu/ was transposed to /i/, recognition was again unimpaired. This relation was also reported for /di/ and /du/. However, for the /gi/ to /u/ transposition, 90 percent /b/ responses were reported. The /gu/ to /i/ transposition fared better with 82 percent correct responses. Cole and Scott (1974a:101) concluded that "stop consonants may be recognized before different vowels...in terms of invariant acoustic features."

Implicit in this conclusion is the assumption that bursts are not only invariant, but sufficient cues to place of articulation. For if they are not sufficient, it matters little whether or not they are invariant. However, it has been known for a number of years both from synthesis experiments (Liberman, Delattre, and Cooper, 1952; Hoffman, 1958) and from the acoustic analysis of natural speech (Fischer-Jørgensen, 1954, 1972; Halle, Hughes, and Radley, 1957; Fant, 1969) that, while release burst spectra vary systematically with the following vowel for initial velar stops, they are largely invariant for initial labial and apical stops. The most novel aspect of Cole and Scott's (1974a) conclusion is therefore, that burst cues are <u>sufficient</u> for recognition of stop-consonant place of articulation. Several considerations suggest that this claim may merit more careful consideration.

First, Cole and Scott (1974a) made no attempt to separate the release burst from the context-conditioned voiceless aspiration. If we examine the spectrograms of Figure 2 in Cole and Scott (1974a:104), we see obvious acoustic differences between the transposed portions of syllable pairs. Had listeners been asked to identify the vowels of these transposed portions, they might well have been able to do so, thus demonstrating that the experimenters had transposed not consonants, but whispered consonant-vowel (CV) syllables. In fact, Winitz, Scheib, and Reeds (1972), in an experiment closely related to that of Cole and Scott (1974a), have reported precisely this result for the (admittedly longer) burst and aspiration portions of initial /p,t,k/.

A second reason to question Cole and Scott's conclusion is that they transposed energy for the voiced stops between only two vowels. Since most dialects of English contain approximately 16 distinctive vowel nuclei, transpositions over two vowels represent a rather meager test of their hypothesis. Indeed, Fischer-Jørgensen (1972) has shown for Danish /b,d,g/ that bursts are effective cues for /b/ and /g/ before /i/ and /u/, but not before /a/, while for /d/ a burst is an effective cue before /i/, but not before /a/ or /u/. Thus, there is already evidence from a language other than English that bursts are not adequate cues for the distinction among these stop consonants in certain vowel environments.

A third consideration is that the release bursts, claimed by Cole and Scott (1974a) as sufficient and invariant cues, have not proved to be sufficient for automatic speech recognition. If these cues were indeed sufficient and invariant, it would be a simple enough matter to specify their acoustic values and build the appropriate filters into a speech recognition device. In practice, this has not been done, partly because in natural speech, release bursts are absent from stops in unstressed syllables and from syllable-final stops in all syllables at least as frequently as they are present in the stressed syllables to which Cole and Scott gave their attention.

A final, and perhaps the most important consideration, is that articulatory gestures associated with initial labial, apical, and palatovelar stop consonants before a variety of different vowels give rise to systematic variations in syllabic acoustic structure that make the hypothesis of any single sufficient cue (whether burst or transition) across all environments extremely unlikely. Every researcher who has worked on speech synthesis is familiar with the fact that a "good" rendering of a particular phonetic segment may require different acoustic patterns in different phonetic environments. For example, good initial, voiced apical stops are more readily synthesized with a burst before high, front vowels, but with extensive voiced formant transitions before back vowels. Furthermore, even though isolated cues may serve a valid experimental function, natural speech typically displays a complex of cues with varying acoustic salience and therefore, we may suspect, varying perceptual weight in different environments. It will simplify the description and interpretation of our experimental results, if we here spell out the most important acoustic variations and some possible perceptual consequences. For more detail than we can give here, the reader is referred to Fischer-Jørgensen (1954, 1972), Halle, Hughes, and Radley (1957), Fant (1959, 1960, 1969), Flanagan (1972), Heinz (1974) and Klatt (1975).

Release burst energy. The energy (duration × intensity) in the transient release and its following frication varies as a function of several factors, including the cross-sectional area of the constriction just after release, the resonant cavity in front of the point of release and perhaps, the release gesture itself. Thus, /b/ for which there is essentially no front cavity and for which the release gesture is rapid (Fujimura, 1961; Kuehn, 1973) usually displays a weak transient and virtually no frication, while /g/ for which the cross-sectional area between tongue and palate is relatively large, for which the front cavity is narrowly tuned and for which tongue release is relatively slow, displays the longest burst of the three stops, including, on occasion as Fischer-Jørgensen (1954) noted, a "double" release transient (see Figure 1) [perhaps due to a suction effect (Fant, 1969)]. Burst energy for /d/, with a smaller cross-sectional area between tongue and alveolar ridge and a more broadly tuned front cavity than for /g/, but with a release velocity roughly the same as for /b/, falls midway. We might then predict increasing energy in--and therefore perceptual importance of--the burst as the point of occlusion moves back in the mouth.

Cutting across all three places of articulation however, are possible variations in burst energy due to coarticulation with the following vowel. A major contrast is between front unrounded vowels, such as /i,ɪ,ɛ/ and center-to-back rounded vowels, such as /ɜ,ɔ,u/. For /b/, increased cross-sectional area of the constriction just after release may give rise to a longer and so more effective, release burst before rounded, than before unrounded vowels. For /d/, elongation of the front cavity before rounded vowels is likely to yield lower burst

intensity than before unrounded vowels. For /g/ the effect of front cavity elongation before rounded vowels may be counteracted by increased cross-sectional area of the palato-lingual constriction and narrower front-cavity tuning, than before unrounded vowels. Thus, if we assume that acoustic energy at least partially determines auditory salience and perceptual weight, we might expect the release burst to play a more important role before rounded than before unrounded vowels for /b/ and /g/, but exactly the reverse for /d/.

Release burst spectrum. Spectral sections taken through the release burst of /b/ in nine vocalic environments show a broad curve with peaks over low frequencies, below approximately 2000 Hz (see Figure 2); the low frequency peaks tend to be stronger before rounded than before unrounded vowels. For /d/ the spectral curve is broad and of a relatively high intensity with peaks over higher frequencies, above approximately 2000 Hz (see Figure 2); the peaks tend to shift upward before unrounded vowels and to be somewhat stronger than before rounded vowels. Apart from these minor rounding dependencies, /b/ and /d/ bursts are relatively unaffected by the following vowel. We may note, however, that these bursts do not occupy invariant positions on the frequency scale in relation to their following vowels; the apical burst is spectrally continuous with F_2/F_3 of the high front vowels, but spectrally distinct from F_2 of the back rounded vowels; for the labial burst these relations tend to be reversed. The spectrum of the velar burst, on the other hand, is narrow and of a relatively high intensity with its main peak close to F_3 of a following front vowel, and close to F_2 of the following back vowel, reflecting the changes from the front articulation of /gi/ to the back articulation of /gu/. Thus, while labial and apical bursts are largely invariant on the frequency scale, but variable in relation to following vowel, velar bursts are more or less invariant in relation to the following vowel, but variable on the frequency scale. [For a more comprehensive description of burst spectra in different vocalic environments, see Zue[1] (1976).] The possible perceptual implications of these facts will become clear when we report our results.

Formant-transition range and energy. At least three articulatory factors underlie variations in formant-transition structure. First, are variations in the extent of transitions as a function of place of articulation and following vowel. For bilabials, transitions are longer (and so, presumably more effective cues) before unrounded than before rounded vowels. For apical stops the distance between point of occlusion and vowel-target configuration varies, so that we might expect both devoiced and voiced transitions to be more effective cues to /d/ before back vowels, where transitions are relatively long, than before front vowels, where they are relatively short. Finally, for velars the determining factor is degree of similarity between the velar tongue constriction and that of the following vowel; in general, close vowels (such as /i/) will have relatively little transition, and open vowels (such as /a/), a more marked transition.

A second factor affecting formant-transition structure is the onset of voicing relative to onset of the release burst [i.e., VOT (Lisker and Abramson, 1964)]. An increase in the time taken for consonantal release (i.e., in release burst duration) leads to an increase in the time taken for development of a transglottal pressure drop sufficient to initiate voicing, and so to an increase in VOT. If VOT is increased, transitions into the following vowel may be largely complete at voicing onset, so that the duration of devoiced transitions relative to voiced transitions is increased. Since release burst duration (and so VOT) typically increases from labial to apical to velar points

[1]Acoustic Characteristics of Stop Consonants: A Controlled Study, by Victor Waite Zue, Ph.D. thesis, M.I.T., May, 1976.

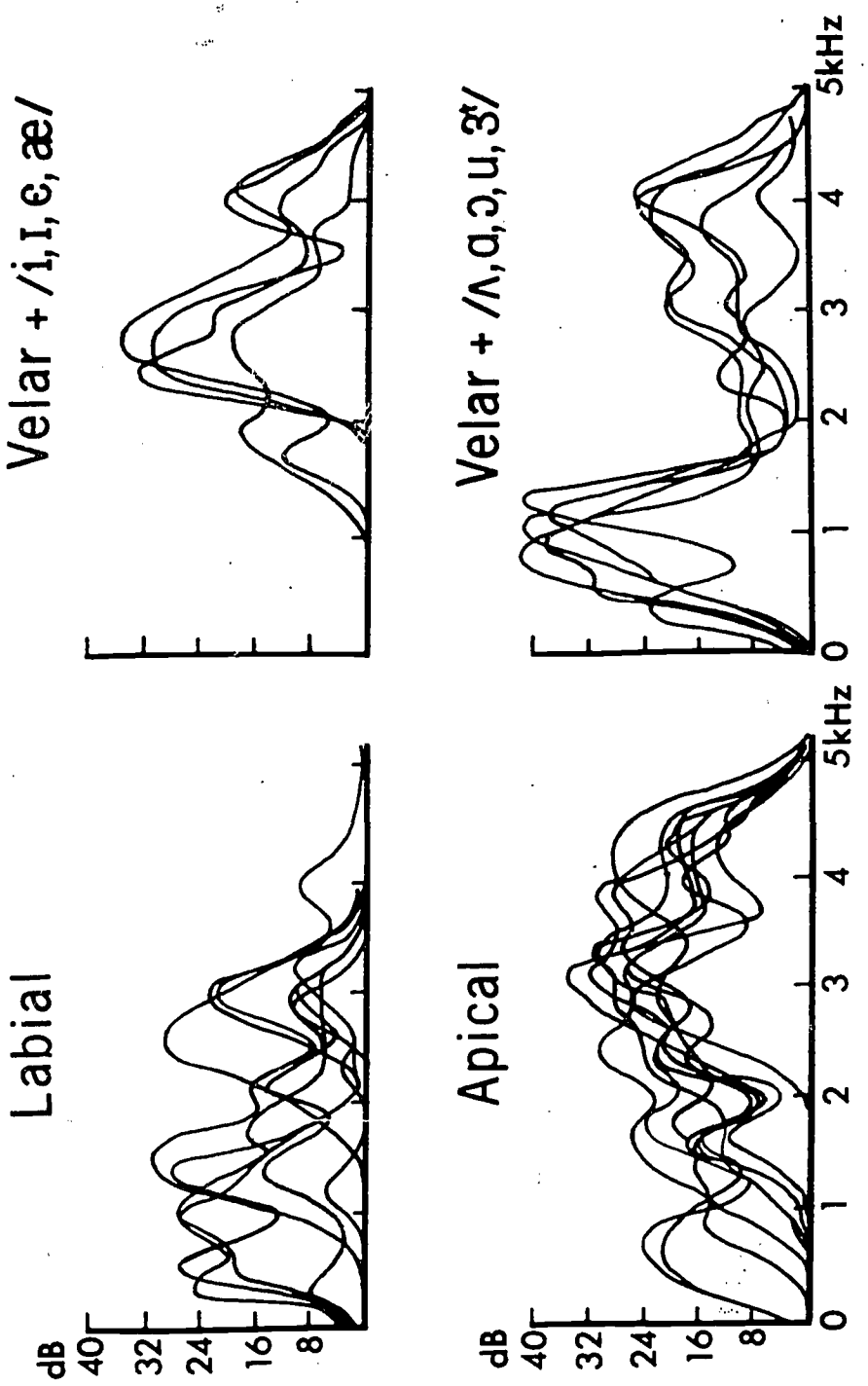


Figure 2: Spectra of the bursts from syllable-initial /b,d,g/ spoken before nine vowels by speaker 2. The velar spectra have been divided into front and center-back vowel series.

of articulation (Lisker and Abramson, 1964:Table 1), we may reasonably predict corresponding increases in the perceptual weight attached to devoiced transitions.

Finally, speakers differ in vocal-tract shape and dimensions, as well as in articulatory habits (Bell-Berti, 1975), and even two phonetically identical utterances of the same speaker are probably never identical acoustically. If we add chance variations in relative effectiveness of bursts and transitions, due to such factors as distance between speaker and listener (or between speaker and microphone), we must conclude that predictions of the perceptual weight attached to the several acoustic cues to place of articulation can be, at best, statistical, and that the likelihood of any single cue being the sole determinant of the percept in all contexts is extremely low.

As will be seen, the results of the following three experiments support this conclusion. Experiment I assesses the role of bursts and formant transitions in the recognition of natural speech by systematically removing them from American English /b,d,g/ spoken before nine different vowels by a single speaker. Experiment II replicated Experiment I with a different speaker. These two experiments are thus concerned with whether the manipulated cues are sufficient for recognition. Experiment III, on the other hand, is concerned with whether the release burst is functionally invariant; it assesses the invariant cue value of the release burst for the second speaker by transposing it from each consonant-vowel-consonant (CVC) syllable across all vowels for each class of stop consonant.

EXPERIMENTS I AND II

Experiment I

Nine CVC syllables were recorded by a male speaker in a carrier phrase, "The little CVC dog," with stress on the CVC. Two tokens of all combinations of initial /b,d,g/, followed by /i,ɪ,ɛ,æ,ʌ,a,ɔ,u,ɜ˞/, with a constant syllable-final /d/ were recorded. In addition, phrases of the type, "The little VC dog" ("The little vowel-consonant dog") were recorded, where V was again one of the nine vowels above and C was again /d/. The phrases were digitized with an effective frequency response of 160-7000 Hz, by means of the Haskins Laboratories pulse code modulation system (Cooper and Mattingly, 1969), and the test syllables were excised and edited. Two parallel sets of 45 experimental signals were then constructed from the oscillograms by the following steps:

1. Each syllable was left in its original form.

2. From each CVC the burst was removed. A burst was defined as an utterance initial, high amplitude (relative to the surrounding signal) component of the signal (see Figure 1). The duration of the burst was determined for each syllable on a high resolution oscillogram; the values were quite consistent across the two tokens of each syllable. Table 1 lists these durations averaged across tokens. The mean burst duration for /b/ was 4.3 msec, for /d/ 6.3 msec, for /g/ 11.7 msec.

3. Each burst was attached to its corresponding VC syllable (for example, the /bid/ burst was attached to /id/), leaving a silent interval between the end of the burst and the first voiced pulse of the vowel, equal in duration to

TABLE 1: Release burst and "aspiration" durations, and voice onset times[1] in milliseconds, for /b,d,g/ followed by nine vowels for two speakers.[2]

	Speaker 1			Speaker 2		
Syllable	Release burst in msec	Aspiration in msec	VOT in msec	Release burst in msec	Aspiration in msec	VOT in msec
/bid/	4	5	9	6	4	10
/bɪd/	5	1	6	9	9	15
/bɛd/	5	5	10	9	4	13
/bæd/	3	2	5	7	8	15
/bʌd/	5	2	7	9	4	13
/bad/	3	1	4	11	4	15
/bɔd/	3	6	9	6	6	12
/bud/	7	4	11	10	14	24
/bɝd/	4	5	9	10	13	23
Mean	4.3	3.4	7.7	8.6	7.3	16.0
/did/	7	1	8	25	12	37
/dɪd/	7	5	12	15	6	21
/dɛd/	6	7	13	12	13	25
/dæd/	8	6	14	13	8	21
/dʌd/	6	6	12	10	5	15
/dad/	5	6	11	7	8	15
/dɔd/	5	7	12	8	7	15
/dud/	6	6	12	5	10	15
/dɝd/	7	7	14	10	15	25
Mean	6.3	5.7	12.0	11.7	9.3	21.0
/gid/	7	12	19	25	10	35
/gɪd/	21	3	24	17	18	35
/gɛd/	7	11	18	22	14	36
/gæd/	12	6	18	29	7	36
/gʌd/	14	9	23	18	13	31
/gad/	11	6	17	21	25	46
/gɔd/	13	7	20	20	15	35
/gud/	8	11	19	20	15	35
/gɝd/	12	11	23	20	21	41
Mean	11.7	8.4	20.1	21.3	15.3	36.7

[1] Voice onset time (VOT) is the sum of release burst, affrication, and aspiration durations.

[2] The values for speaker 1 are the averages of two tokens of each syllable; for speaker 2, of one token of each syllable.

the interval between burst offset and voicing onset in the CVC from which the burst had been removed.

4. For each CVC the entire signal up to the first well-defined voicing pulse was removed. Thus, the burst and devoiced formants (i.e., noise excited resonances) were removed (see Figure 1), and the duration of this segment was measured on an oscillogram of each utterance. Table 1 lists the two-token averages of the devoiced formants ("aspiration"), as well as of the entire segment from burst onset to voice onset (VOT) for each syllable. Mean VOT for /b/ was 7.7 msec, for /d/ 12.0 msec, for /g/ 20.1 msec.

5. Each burst-plus-devoiced formants was then attached to its corresponding VC syllable.

This procedure permitted us to present five different combinations of the cues to place of articulation (burst, devoiced transition, voiced transition) for each syllable: (a) all three together in the original syllable; (b) burst plus vowel; (c) burst and devoiced transitions plus vowel; (d) voiced transitions plus vowel; (e) devoiced and voiced transitions plus vowel.

Three recordings of each of the 45 signals in each set were generated and randomized into two parallel test sequences of 135 items each. One test was administered to 14 Lehman College undergraduates. The stimuli were played at a comfortable level in a sound attenuated room, on a Revox 1122 tape recorder, over an audiometric loudspeaker. The other test was administered to nine students and faculty volunteers from Yale University: the stimuli were played at a comfortable level in a sound attenuated room on an Ampex AG 400 tape recorder over an AR4x loud speaker at Haskins Laboratories.

The listeners were instructed to write the identity of the initial sound of each syllable. The response categories listed on the answer sheets were /b,d,g,p,t,k,?,Ø/.[2] The ? response was for use when the listener thought that the syllable began with a consonant, but could not decide which one. The Ø response was for use when the listener thought that the syllable began with a vowel. Twenty tokens of the stimuli were played to familiarize the listeners with the task. The listeners were then presented with one of the 135-item test sequences.

Experiment II

Exactly the same procedures of stimulus and test construction as those described above were followed for a second speaker, except that he provided only one token of each syllable and therefore only one test. The sentences were read

[2] A relatively open response set provides a sensitive measure of how "stoplike" a signal sounds. In a situation where only /b,d,g/ are permitted as responses, the identifiability of the signals may be overestimated. For example, a signal composed of labial burst and a steady-state vowel such as /ɪ/, sounds like a click followed by /ɪ/. However, if only /b,d,g/ are permitted as responses, then a subject may well feel that, since the click does not sound like a high-frequency alveolar burst, and is not affricated like a velar burst, (s)he should respond /b/. A correct /b/ response would then be made to a signal that does not sound like /b/.

at a very deliberate rate with stress on the initial consonant of the CVC. Table 1 lists the durations in msec of burst, "aspiration" and VOT for each syllable. The durations are very much longer than (almost double) those of speaker 1. However, the pattern of increase in burst and VOT durations, from labial to apical to velar stops, is similar to that of speaker 1.

Eleven Lehman College undergraduates took the test under conditions identical to those of the Lehman College students in Experiment I.

RESULTS

Experiment I (Speaker 1)

The two groups of subjects gave very similar results on the two parallel tests. We have therefore combined their data. Figure 3 displays percentage correct identification of initial consonantal place of articulation as a function of vowel nucleus for the five sets of cue combinations (all cues, burst plus vowel, burst and voiceless transition plus vowel, voiced transition plus vowel, voiced and voiceless transition plus vowel) and the three classes of consonant (labial, apical, velar). Responses were scored for place of articulation only, and voicing errors were disregarded. Each data point is based on 69 responses (23 subjects × 3 repetitions). The vowels have been ordered along the horizontal axis to trace a rough path around the rim of the English vowel loop from /i/ through /a/ to /u/, with /ɜ˞/ appended. The points have been connected by straight lines to facilitate reading of the graphs.

Labial. All the original syllables, except /bud/ (85 percent), were correctly identified more than 90 percent of the time. The burst was relatively ineffective as a cue and performance hovered around chance (20 percent) before all vowels, except /u/ (81 percent) and /ɜ˞/ (51 percent). The voiced transition, on the other hand, served almost as well as the full syllable and performance hovered around 90 percent before all vowels, except /ɔ/ (84 percent), /u/ (63 percent), and /ɜ˞/ (61 percent), the last two vowels being precisely those for which burst performance was at its best. The addition of the devoiced transition, whether to burst or voiced transition, tended to increase performance by a few percentage points, but this cue clearly carried little perceptual weight.

Apical. All the original syllables, except /did/ (87 percent), /dɛd/ (74 percent), and /dæd/ (81 percent) were correctly identified more than 90 percent of the time. The burst was a moderately effective cue before the front vowels, /i/ (57 percent) and /ɪ/ (65 percent), but otherwise carried little weight, and was only marginally aided by addition of the devoiced transition. The full transition (devoiced and voiced portions), on the other hand, was a moderately effective cue (60 percent or higher) before the back and central vowels, but a weak cue before the front vowels. There seems to be a reciprocal relation between burst and transition; if the weight of one is high, the weight of the other is low. Wherever the full transition carried any marked weight, removal of its devoiced portion led to an appreciable drop in performance, particularly before /u/ and /ɜ˞/. In general, neither burst nor transition alone maintained performance at the level of the original syllables.

Velar. All the original syllables except /gid/ (56 percent), /gɪd/ (70 percent), /gɛd/ (61 percent), and /gæd/ (88 percent) were correctly identified

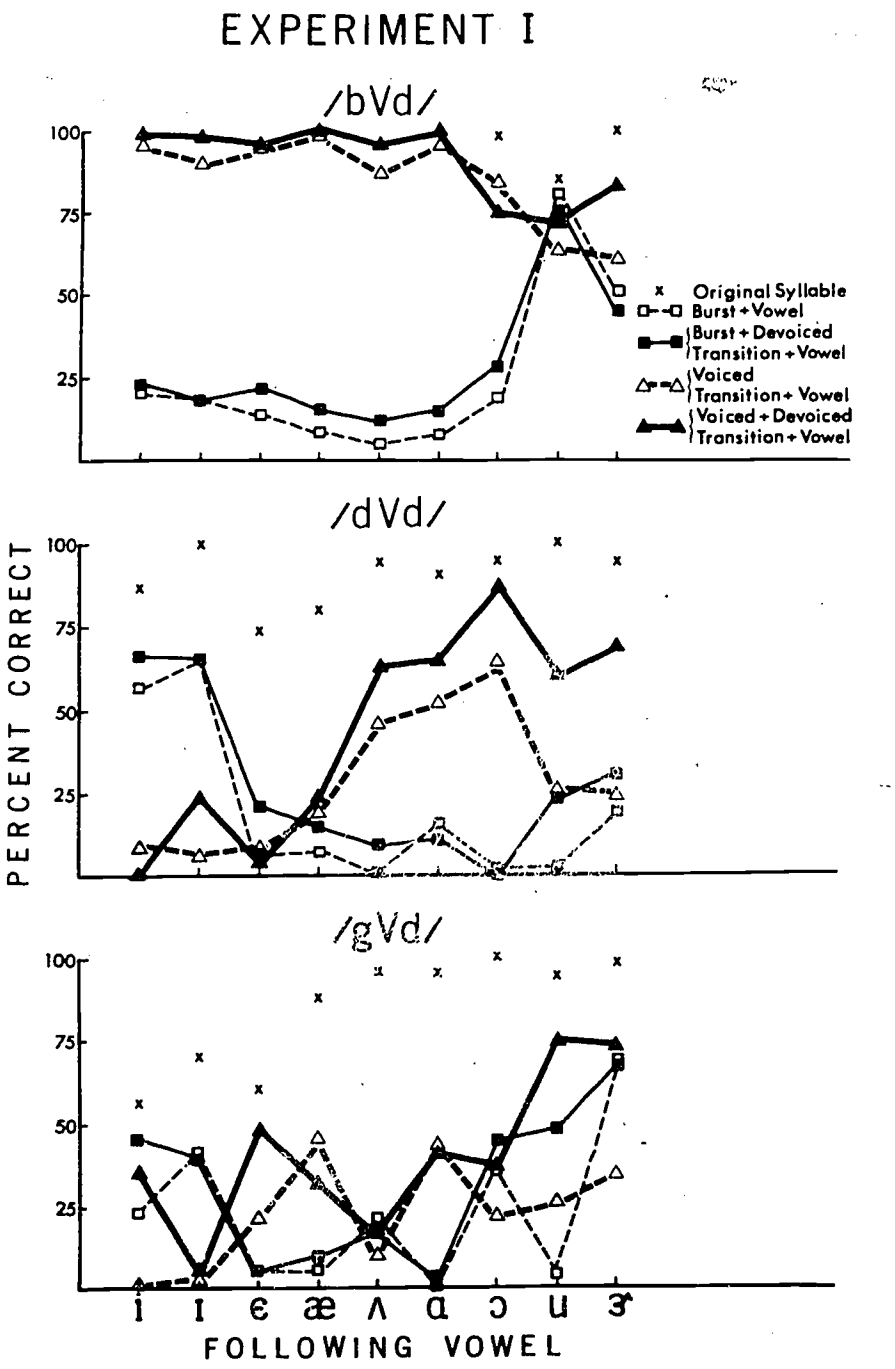


Figure 3: Percent correct recognition of place of articulation for speaker 1 as a function following vowel. The five different combinations of cues to place of articulation are parameters of the curves.

13

more than 90 percent of the time. The burst elicited moderate performances only before /ɛ/ (41 percent) and /ɜ/ (69 percent), and was appreciably aided by addition of the devoiced transition only before /u/. For the full transitions, performance was moderate before /ɛ/ (48 percent), /a/ (41 percent), /u/ (75 percent), and /ɜ/ (73 percent), but weak elsewhere. Just as for /d/, removal of the devoiced portion of the transition had a marked effect before /u/ and /ɜ/. There is again some evidence of a reciprocal relation between burst and transition. Even more obviously than for /d/, no subset of the cues held performance at the level of the original syllables.

Experiment II (Speaker 2)

Figure 4 displays the results for tokens from the second speaker in the same format as Figure 3. For /b/ and /d/, the pattern of results is similar to that of Experiment I, apart from a general increase in level of performance; for /g/ the perceptual weight of the burst is clearly greater than it was for speaker 1. It will be recalled that the duration of the bursts and aspiration segments of speaker 2's utterances was very much greater than (nearly double) that of the corresponding segments of speaker 1 (see Table 1).

Labial. All the original syllables, except /bud/ (85 percent), were correctly identified more than 90 percent of the time. The burst was moderately effective as a cue before all vowels, especially the central to back vowels, /ɔ/ (79 percent), /u/ (79 percent), and /ɜ/ (85 percent), and was as effective as the full syllable for /ɛ/ (97 percent). The full transition served almost as well as the full syllable for all vowels except /a/ (75 percent), /u/ (36 percent), and /ɜ/ (42 percent), the last two again being the vowels for which burst performance was at its best. The perceptual effect of adding the devoiced transition, whether to burst or voiced transition, was generally small, and not reliable.

Apical. All the original syllables were correctly identified more than 90 percent of the time. The burst was a strong cue before /ɪ/ (100 percent) and /ɜ/ (91 percent), moderate before /i/ (72 percent) and /ɛ/ (79 percent), but otherwise carried little weight. Addition of the devoiced transition to the burst had no systematic effect. The full transition was almost as effective as the full syllable for central and back vowels, but was a weak cue before the front vowels. Removal of the devoiced portion of the transition tended to lower performance, especially before /u/. Performances on bursts and transitions were reciprocally related before all vowels except /æ/ and /ɜ/.

Velar. All the original syllables, except /gid/ (64 percent), were correctly identified more than 90 percent of the time. The burst was a moderately effective cue before /ɪ/ (73 percent), /ɛ/ (52 percent), and /ʌ/ (55 percent), almost as effective as the full syllable before /i/, /a/, /ɔ/, /u/, and /ɜ/. Addition of the devoiced transition had no systematic effect. The full transition was a moderately effective cue before /ɪ/ (64 percent) and /ɛ/ (46 percent), a strong cue before /a/ (91 percent), but otherwise carried little or no perceptual weight. Removal of its devoiced portion tended to reduce performance, particularly before /ɪ/ and /a/. Burst and transition again tend to be reciprocally related, particularly before central and back vowels, except /a/.

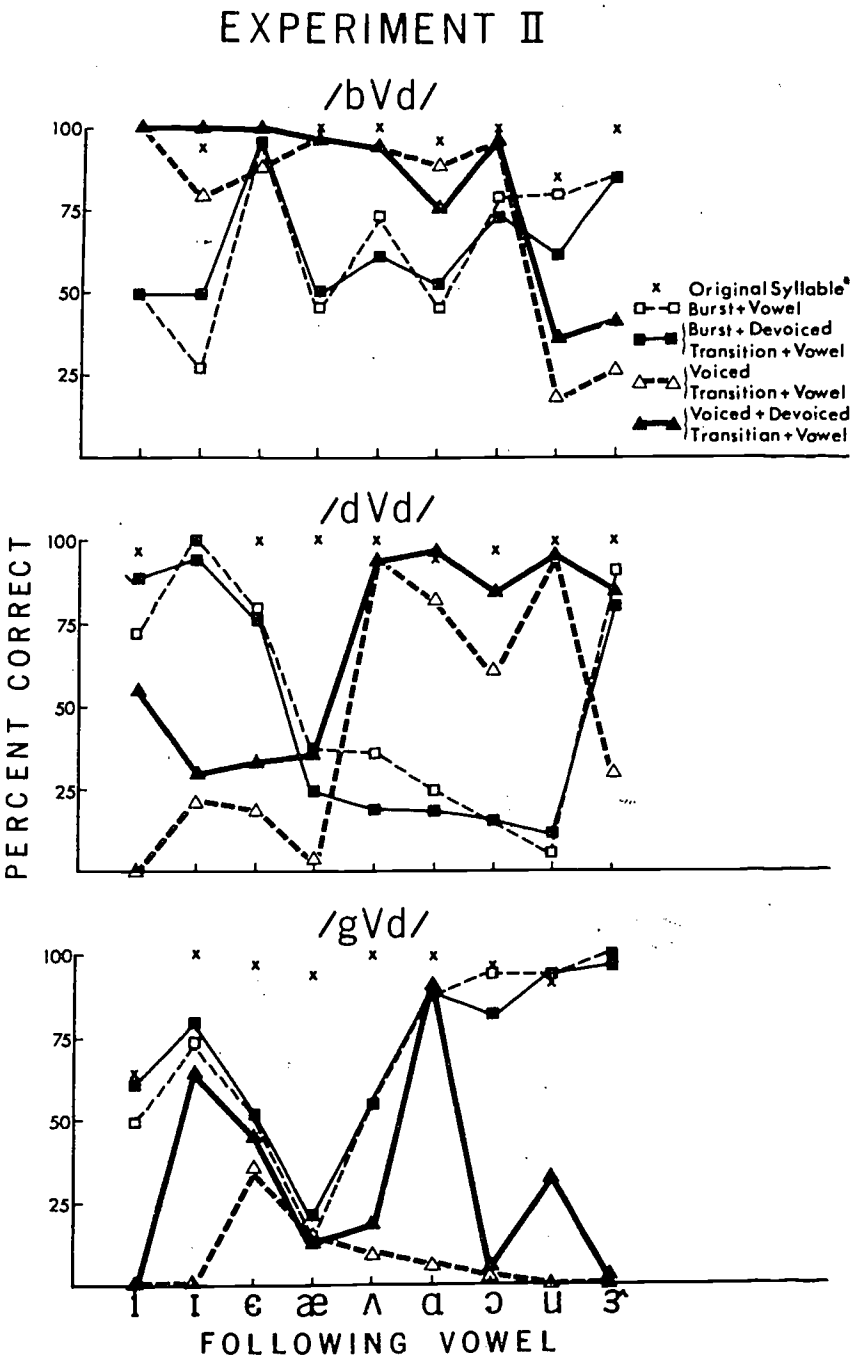


Figure 4: Percent correct recognition of place of articulation for speaker 2 as a function of following vowel. The five different combinations of cues to place of articulation are parameters of the curves.

DISCUSSION

Experiments I and II

The perceptual weight carried by release bursts and formant transitions as cues to place of articulation, varied with consonant, vowel, and speaker. No single cue, or pair of cues, was sufficient for recognition in all contexts. If we take into account variations in acoustic structure, such as those outlined in the introduction, we can make sense of many, though not all, of the results.

Labial. As expected, labial bursts were relatively weak cues. For speaker 2 they were longer in duration and considerably more effective than for speaker 1. Nonetheless, the patterns of performance are quite similar for the two speakers; apart from an anomalous point at /ɛ/ for speaker 2, labial bursts tended to be most effective before rounded vowels. Whether this is due to variations in burst energy or to variations in burst frequency position in relation to the following vowel, will become clearer when we have reported the results of Experiment III. Here we note simply that the rank order correlation between burst duration and percent recognition was not significant for either speaker.

Formant transitions, on the other hand, were almost as effective for both speakers as the full complement of cues, before all nine vowels, except /u/ and /ɝ/. The two exceptions are rounded vowels for which lip constriction necessarily reduces the rise in formant frequency (i.e., the extent of formant transitions) associated with mouth opening.

Apical. As expected, apical bursts tended to be longest and most effective for both speakers, before front vowels. They were weak before all other vowels (/ɝ/ is an exception for speaker 2), and seem to have become systematically weaker as rounding (and so front-cavity length) increased, reducing burst energy (see Table 1). However, burst frequency may also be relevant, and we again defer discussion, noting only the lack of significant correlations between burst duration and performance.

For both speakers (particularly speaker 2), formant transitions were strong cues before central and back vowels /ʌ,a,ɔ,u,ɝ/ where apical transitions are extensive, but weak cues before the front vowels, where transitions are relatively short. Furthermore, as might be predicted from the longer apical than labial VOTs (see Table 1), addition of the devoiced to the voiced transition segments tended to improve recognition of /d/ more than of /b/. However, within the apical series, VOT does not significantly predict the performance gains from addition of the voiceless transitions.

Velar. Speaker differences are most marked for the velar series. The predicted tendency for the burst to be more effective before back, rounded than before front, unrounded vowels was borne out for speaker 2, despite his somewhat longer front than back vowel bursts. However, for speaker 1, the burst was simply a very weak cue before all vowels, except /ɝ/. Again, we note the lack of significant correlation between burst duration and performance, and defer comment on these results.

As expected, the relatively short velar transitions were far less effective cues than were labial and apical transitions for both speakers. At the same time, longer VOTs did tend to increase the effectiveness of devoiced transitions. For speaker 1 the largest performance gains from the addition of devoiced to voiced transitions were for /i/, /ɛ/, /u/, and /ɜ^/, the vowels before which aspiration durations were longest; similarly, for speaker 2 the largest gains were for /ɪ/ and /a/ (see Table 1). However, the rank order correlation between performance and voice onset time was not significant.

Broadly, our results agree with those of Fischer-Jørgensen (1972) for Danish initial-voiced stops in these respects: (1) the burst was a relatively effective cue for /b/ before /u/, but not before /a/; (2) the burst was a relatively effective cue for /d/ before /i/, but not before /a/ or /u/; (3) the burst was a relatively effective cue for /g/ before /i/ and /u/ (speaker 2 only), but not before /a/ (speaker 1 only). Our results disagree with those of Fischer-Jørgensen insofar as: (1) the burst was a relatively ineffective cue for /b/ before /i/; (2) the burst was a relatively ineffective cue for /g/ before /u/ (speaker 1 only); (3) the burst was a relatively effective cue for /g/ before /a/ (speaker 2 only).

Our results do not support the implication of Cole and Scott (1974a) that release bursts alone are sufficient cues to the place of articulation of initial-voiced stop consonants. Nor, contrary to our own expectation, did the addition of devoiced transitions to the bursts reliably improve recognition. If we adopt as an arbitrary (and modest) criterion of significant perceptual weight that recognition performance for release-bursts-plus-vowels should drop by no more than 25 percent below performance for the original syllable, we see that this level was reached for speaker 1 on only one syllable out of 27 (/bud/), for speaker 2 on only 13 syllables out of 27 (/bɛd, bɔd, bud, bɜ^d, did, dɪd, dɛd, dɜ^d, gid, gad, gɔd, gud, gɜ^d/). The role of consonant-vowel (CV) coarticulation in determining burst effectiveness, implicitly denied by Cole and Scott (1974a), is suggested by the preponderance among speaker 2's 13 syllables, of central-back, rounded vowel syllables for /b/ and /g/, of front unrounded vowel syllables for /d/.

EXPERIMENT III

The purpose of this experiment was to test the hypothesis that the initial release burst of /bVd, dVd, gVd/ syllables may be a functionally invariant cue to consonantal place of articulation across a representative set of syllable-nucleus types. The method was to transpose the release burst from each CVC syllable in a series (labial, apical, velar) across all types of VC syllables in that series. For a fair test of the hypothesis we needed tokens from a speaker whose release bursts were known to be at least moderately effective cues in their original syllables. We therefore used the 27 CVC (and 9 VC syllables) recorded by speaker 2 for Experiment II.

Method

The experimental signals were constructed in exactly the same way as the burst-plus-vowel signals of Experiments I and II. The burst was removed from all 27 CVC syllables (for durations see Table 1). Each burst was then attached

to all nine vowel-/d/ syllables (where the vowels were again /i,ɪ,ɛ,æ,ʌ,a, ɔ,u,3^/), leaving a silent interval between burst offset and vowel onset, equal in duration to the devoiced interval for the CVC token being simulated. The result was a set of 81 syllables in each series (labial, apical, velar)--a total of 243.

Three repetitions of each syllable were recorded and randomized into a single test of 729 items. The test was administered to eight Lehman College undergraduates under conditions and instructions identical with those used for the Lehman College students of Experiments I and II.

Results

Figure 5 displays percentage correct identification of initial consonantal place of articulation as a function of following vowel for the nine bursts in each series. Responses were scored for place of articulation only, and voicing errors were disregarded. To facilitate reading, the results for bursts drawn from syllables containing the four front vowels (/i,ɪ,ɛ,æ/) have been grouped in the upper three graphs; the results for bursts drawn from syllables containing the five central and back vowels (/ʌ,a,ɔ,u,3^/) have been grouped in the middle three graphs. The following vowels have been ordered along the horizontal axes to trace a path around the rim of the English vowel loop from /i/ through /a/ to /u/, with /3^/ appended, and points have been connected by straight lines to facilitate reading. For untransposed bursts (i.e., bursts placed before the same vowel as that of the syllable from which they were originally drawn) the data point is circled.

Before considering the three series separately, several general points can be made. First, the highest performance for a given vowel is often not elicited by the burst taken from the original syllable containing that vowel. For example, the burst drawn from the syllable /bad/ elicited a lower performance when attached to /ad/ (the circled point over /a/ in the middle labial graph of Figure 5), than did the bursts drawn from any of the other eight /bVd/ syllables. Similar, if less severe, discrepancies appear for many other syllables.

Second, the highest recognition performance elicited by a particular burst is not always for a syllable containing the same vowel as the syllable from which the burst was drawn. This is most striking in the apical series for which the highest performances elicited by all nine bursts are before /id/ and /ɪd/. Similarly, in the labial series, bursts drawn from all nine syllables, including the front vowel set, elicit their highest performances when attached to back vowel syllables; and in the velar series, bursts from the four central and back vowel syllables, /gad, gɔd, gud, g3^d/, elicit roughly interchangeable performances within their own set.

Both these results suggest a measure of commutability among the bursts of each series. This commutability becomes even more obvious as soon as we notice a third feature of the data, closely related to the first two: the overall form of the performance curves across the vowels is remarkably similar for all bursts within a series, whatever the syllables from which they were drawn. The degree of concordance among the nine curves of each series is a measure of burst commutability or functional invariance. Furthermore, a rather good description of the general curve for each series is provided by simply plotting

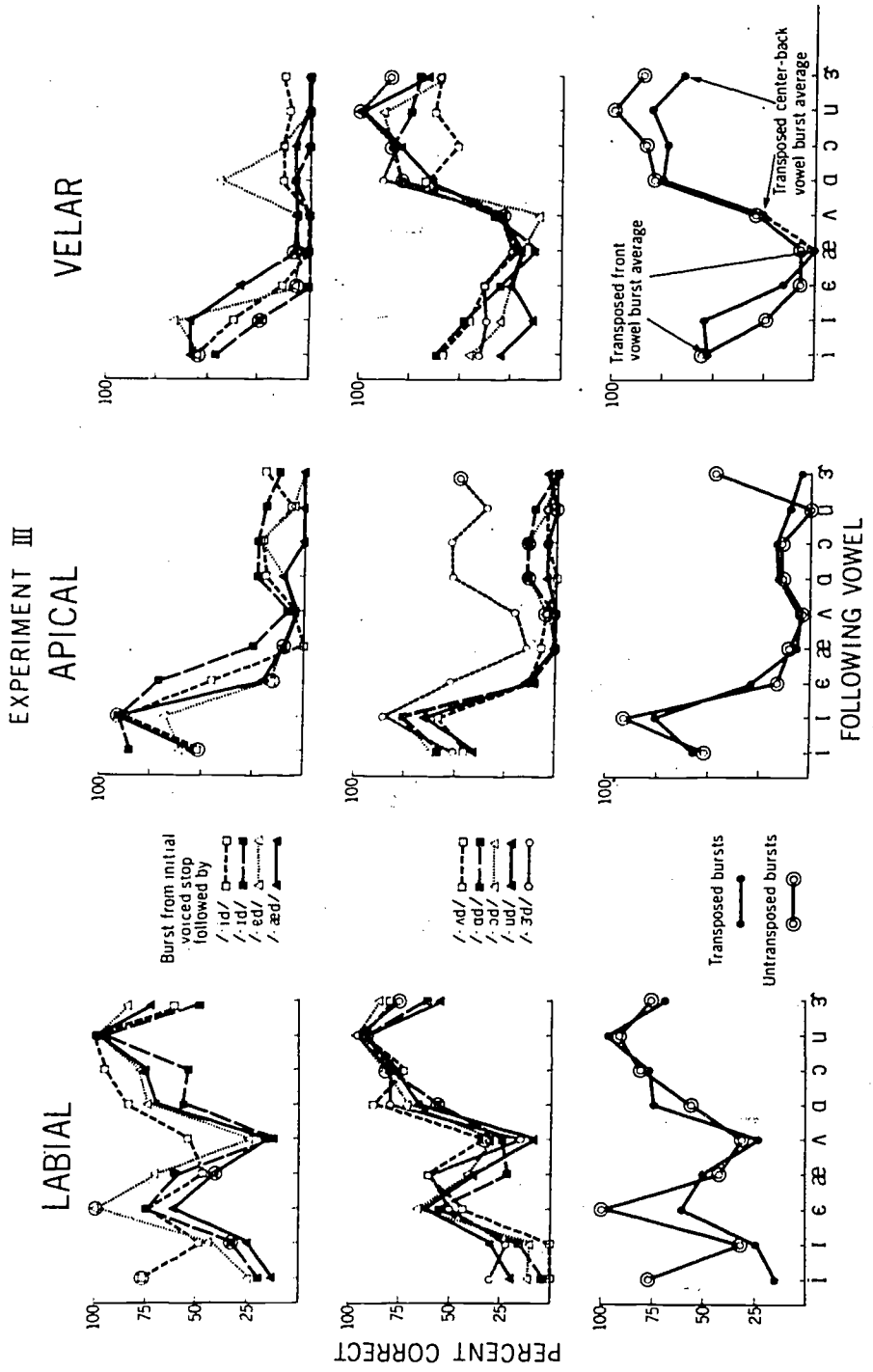


Figure 5: Percent correct recognition of place articulation in burst plus /Vd/ syllables. In the top two rows of figures, each point represents the recognition of syllables composed of a burst taken from one vocalic environment and transposed to each of the other vocalic environments (with the exception of the circled points for the nine untransposed bursts). In the bottom row, average correct recognition scores for syllables in which the bursts were transposed are compared with recognition scores for syllables in which the bursts were attached to the same vowel as that of the syllable from which they were originally taken.

for each vowel the percentage of correct identification of its "untransposed" burst (circled points). These curves are displayed in the lower three graphs of Figure 5, together with a plot of the mean percentage correct for the transposed bursts. The rank order correlation between these curves, that is between performances elicited by transposed and untransposed bursts, is then a second measure of burst commutability or functional invariance.

Labial. All nine labial curves are roughly parallel: bursts from almost every syllable elicit their highest performance before the central-back rounded vowels, /a,ɔ,u,ɝ/, a moderate performance before /ɛ/ (before /æ/ for the /bud/ and /bɝd/ bursts), and relatively weak performances before /i,ɪ,æ,ʌ/ (except the peak for the /bid/ burst before /i/). Kendall's coefficient of concordance (W) among the nine curves is .79 ($p < .0001$). This significant similarity in pattern of burst effectiveness (or sufficiency) demonstrates that the nine bursts are, to a large degree, functionally invariant. However, Spearman's rho between untransposed and mean transposed curves of the bottom labial graph falls short of significance with a value of .53. This failure is clearly due to the peaks for the untransposed /bid/ and /bɛd/ bursts and suggests that release bursts effective in signaling labiality before /i,ɛ/ may be context-dependent.

Apical. All nine apical curves are roughly parallel; bursts from every syllable elicit their highest performances before /ɪ/ and /i/, and apart from fair performances for the /dɪd/ and /dʒʌd/ bursts before /ɛ/, and for the /dʒʌd/ burst before the back vowels and /ɝ/, relatively weak performances elsewhere. Kendall's W among the nine curves is .72 ($p < .0001$). Spearman's rho between the untransposed and the mean transposed curves of the bottom graph (Figure 5) is .60 ($p=.05$), clearly pulled down by the peak for the untransposed /dʒʌd/ burst. The apical bursts like the labial bursts, are to a large degree functionally invariant.

Velar. The curves for the velar bursts fall into two distinct groups--front and central-back vowels. The front vowel bursts elicit moderate performances before /i,ɪ,ɛ/ and, apart from a small peak for the /bɛd/ burst before /a/, weak performances elsewhere. The central-back vowel bursts elicit their highest performances before /a,ɔ,u,ɝ/, weak performances elsewhere, though with a tendency for slightly stronger performances before /i,ɪ/. There is thus a small asymmetry; while front vowel bursts do not concord with back vowel bursts before back vowels, back vowel bursts tend to concord with front vowel bursts before front vowels. As a result, Kendall's W among the nine curves, though significant ($p < .001$), is low (.37). However, if we separate the two groups and compute Kendall's W within them, we find for the front vowels, .69 ($p < .05$), and for the central-back vowels, .66 ($p=.01$). The increased coefficients justify separating the bursts into two groups. Accordingly, the transposed burst curve of the bottom graph (Figure 5) was computed for front vowels and for central-back vowels separately. The result is an excellent fit between transposed and untransposed curves, for which Spearman's rho is .88 ($p < .01$). There is therefore a large degree of functional invariance among the velar front vowel bursts and among the velar central-back vowel bursts.

Discussion

While the release bursts of initial labial, apical, and velar stops display a high degree of functional invariance, they do not display a corollary

degree of sufficiency. In all three experiments, the release burst was seldom sufficient to maintain performance at the level elicited by the original syllable. Vowel-dependent variations in performance are therefore less aptly characterized as variations in "sufficiency" or "cue adequacy", than as variations in the degree to which the burst may be assumed to contribute to the cue complex in natural speech (cf. Stevens, 1975).

In Experiment III, identification of the original syllables was perfect except for /gid/ which the listeners identified with 87 percent accuracy. If we again adopt as an arbitrary criterion of significant perceptual weight that the performance on the untransposed burst-plus-vowel should drop by no more than 25 percent below performance on the original syllables, we arrive at the following set of 14 out of 27 syllables for which the release burst carried weight in judgments of place of articulation in either or both of Experiments II and III: /bid, bɛd, bɔd, bud, bɜ^d, did, dɪd, dɛd, dʒ^d, gid, gad, gɔd, gud, gɜ^d/.

These results bring us into closer agreement with both Fischer-Jørgensen (1972) and Cole and Scott (1974a), since the untransposed burst carried significant weight for /b/ before /i/ in Experiment III. The results also agree very well with those of Liberman, Delattre and Cooper (1952). These authors used the relatively crude Pattern Playback II synthesizer to construct schematic stop bursts before seven two-formant monotone vowels, /i,e,ɛ,a,ɔ,o,u/. Identifications reached 75 percent or higher for /p/ before /i,e,ɛ,ɔ,o,u/, for /t/ before /i,e,ɛ/, for /k/ before /a,ɔ,o,u/. Considering only the vowels common to both experiments, these results agree with our own in finding bursts to carry weight as labial cues before /i,ɛ,ɔ,u/, as apical cues before /i,ɛ/, as velar cues before /a,ɔ,u/. The only discrepancy between the two sets of results is in our finding that a release burst carried weight as a velar cue before /i/. This remarkable agreement between the present natural speech study and an experiment carried out with primitive synthetic speech 25 years ago, suggests that the systematic variations in burst effectiveness common to both experiments reflect a robust perceptual process.

The most obvious source of these variations might seem to lie in release burst energy. Unfortunately, we were not able to make reliable intensity measurements of the release bursts in the present study. However, a scan of the syllables for which release bursts proved adequate and of their durations in Table 1, will reveal no obvious correlation, and as reported above, Spearman's rho between burst duration and performance was not significant for any series. Furthermore, since all schematic bursts synthesized by Liberman, Delattre, and Cooper (1952) were of equal energy, this factor cannot account for their results. Thus, while variations in burst energy may well account for variations in the overall performances elicited by particular bursts or in the recognition of different tokens of a particular stop-vowel syllable (and so for the different levels of performance elicited by the bursts of speakers 1 and 2), they cannot account for systematic variations in burst effectiveness across vowels.

The case is no better when we turn to the absolute spectral properties of release bursts. For example, as remarked in the introduction, spectral sections taken through the apical release burst show a broad high intensity curve over frequencies above about 2000 Hz, largely independent of the following vowel (see Figure 2). We can hardly, therefore, appeal to the absolute spectral

properties of the apical burst to explain the fact that the burst carries appreciable weight before high front vowels such as /i,ɪ/, but essentially no weight before central-back vowels such as /ʌ,a,ɔ,u/.

In fact, the key to the problem may be provided by the work of Kuhn (1975). First, he draws on the acoustic theory of speech production, according to which the resonance of the cavity in front of the point of maximum tongue constriction--that is "the front cavity resonance"--may be associated with any of the first four formants (Fant, 1960:72). He then shows that "the front cavity seems to be associated with what is perhaps the most intense group of formants: with the F_3 group for /i,ɪ,ɛ,æ/, and with the F_2 group for /a,ʌ,ʊ,u/" (Kuhn, 1975:430). Next, he demonstrates that a front cavity frequency estimate can be most readily made for the more constricted vowels and for highly constricted consonants, and that for stop consonants the estimate may be derived from the spectral structure of bursts and transitions. Since the front cavity resonance is a function of front cavity length, and since front cavity length is a function of the place of articulation, an estimate of the resonance is tantamount to an estimate of place of articulation. Finally, a variety of evidence from synthetic speech experiments (e.g., Liberman et al., 1952) suggests that place of articulation is most readily conveyed by stop consonant bursts when their spectral weight lies close to the front cavity resonance of the following vowel. Proximity on the frequency scale may facilitate perceptual integration of the burst with the vowel, so that the listener can track the changing cavity shape characteristic of a particular place of articulation followed by a particular vowel. This hypothesis can account for many of the variations in burst effectiveness observed in Experiments II and III.

Labial. The low frequency labial bursts carried significant weight (by the criterion defined above) before /ɔ,u,ɝ/ in both experiments, and close to significant weight before /ʌ/ in Experiment II, and before /a/ in Experiment III. For all these vowels the front cavity is strongly associated with the second formant and the frequency of that formant lies below 1000 Hz, a region over which the greatest weight of labial burst energy is distributed. The variability in response for /ʌ,a/ may be due to weaker front cavity-to-formant affiliation in less constricted vowels, and the consequent difficulty for the listener in continuous tracking of the changing front cavity resonance in the absence of a formant transition.

The two other vowels before which labial bursts carried significant weight were /i,ɛ/, for which the front cavity is strongly associated with the third formant. However, the untransposed bursts were notably more effective than the transposed and, as remarked above, this suggests a degree of context dependency. The rapid and relatively extensive lip opening before unrounded vowels and the consequent rapid rise in resonant frequencies, may extend the burst frequency range sufficiently high for it to be integrated with F_3 of the following vowel. The ineffectiveness of the burst before /æ/ may again be due to weaker front cavity-to-formant affiliation in a less constricted vowel, and the resulting difficulty for the listener. However, the ineffectiveness of the burst before /ɪ/ is unexplained.

Apical. Apical bursts carried significant weight before /ɪ/ in both experiments and before /i,ɛ,ɝ/ in Experiment II, although performance was very weak for most bursts before /ɛ,ɝ/ in Experiment III. On the assumption that

the high frequency apical burst can be integrated perceptually with the front cavity resonance of F_3 for the high front vowels /i,ɪ/, but less readily, if at all, with the less determinate front cavity resonance of the more open vowels /ɛ,æ/, or with the low frequency front cavity resonance of F_2 for the central-back vowels /ʌ,a,ɔ,u/, these results are very much what we would expect. Nonetheless, there are oddities. For example, it is not clear why the /dɪd/ burst (duration 15 msec) should have been more effective before /i/ than was the untransposed burst from /did/ (duration 25 msec) in Experiment III. Nor is it clear, given the moderate duration of the /dɜ^d/ burst (10 msec), why it should have been a strong cue (91 percent) before the low front cavity resonance of F_2 for /ɜ^/ in Experiment II and a moderately strong cue before /a,ɔ,ɜ^/ in Experiment III.

Velar. Velar bursts carried significant weight before /i,a,ɔ,u,ɜ^/ in both experiments. It will be recalled that the spectral weight of velar bursts tends to lie close to the F_2 frequency of the following vowel. Perceptual integration of the burst with the front cavity resonance of F_3 for the front vowels should therefore be easiest when F_2 and F_3 lie close together as in /i/, precisely as observed. For the central-back vowels /a,ɔ,u,ɜ^/, variation in F_2 frequency, and so of velar burst frequency, is small (roughly from 600 to 1000 Hz). We might therefore expect that velar bursts from all four vowels would be readily commutable and accessible to perceptual integration with the front cavity resonance of F_2. Again, this is precisely what was observed. The systematic decline in performance as F_2 (and so velar burst frequency) decreases from /ɪ/ to /æ/ (see Figures 4 and 5) suggests that the ineffectiveness of velar bursts before /ɪ,ɛ,æ/ may be due to the increasing separation of burst and front cavity resonance (F_3) on the frequency scale. The inadequacy of the burst before /ʌ/ may arise from the relatively weak front cavity-to-formant affiliation for this vowel.

In short, despite several unexplained oddities in the data, our perceptual integration hypothesis provides a remarkably close account of the variations in burst effectiveness in Experiments II and III. At the same time, this account affords insight into the grounds of functional invariance among stop release bursts. Bursts are invariant insofar as they all bear the same relation to any particular following vowel. The relation is that of spectral continuity or discontinuity with the main (or front cavity) resonance of the following vowel. If there is continuity (as in an apical burst followed by /ɪ/, for example), the relation contributes significantly to recognition of consonantal place of articulation; if there is discontinuity (as in an apical burst, followed by /ʌ/, for example), the relation does not contribute significantly to recognition. The invariance is therefore not a simple first-order invariance based on the absolute frequency and/or amplitude of the bursts. Rather, it is a higher order relational invariance based on spectral relations between burst and following vowel.

The general conclusion that the contribution of the burst to the cues for place of articulation depends on the following vowel, is not new. Liberman, Delattre, and Cooper (1952) remarked of their schematic /p/ and /k/ bursts before schematic vowels that: "...the irreducible acoustic stimulus is the sound pattern corresponding to the consonant-vowel syllable" (p. 516). While neither Fant (1959, 1969) nor Stevens (1975) believes that the perceptual process always requires reference to the vowel, both describe the burst in natural speech

as dependent on context for its effect. Stevens (1975) deliberately eschews a description in terms of release bursts and individual formants, since this would imply that these components have independent roles in the cue complex. He emphasizes rather "the overall acoustic spectrum immediately following the release" (p. 311). However, regarding the contribution of the burst to this spectrum he writes:

> "We shall assume that this can be considered as the initiation of the rapid spectrum change at the consonant release, if there is spectral energy in the burst in the vicinity of the major spectral peak for the vowel.... Thus the initial burst of energy in syllables beginning with /g/, and the burst for syllables with a front vowel preceded by /d/ would be considered as part of the rapid spectrum change, since major energy concentrations in these bursts occur in frequency regions where the vowel formant transitions are providing cues for place of articulation of the consonant. The d-burst in a syllable with a back vowel, on the other hand, would not be considered as an integral part of the rapid spectrum change.... The burst at the onset of the consonant /b/ is relatively weak, and may not play a significant role in shaping the rapid spectrum change." (Stevens, 1975:312-313).

The present study suggests that, at least for some speakers and listeners, the contribution of the /g/ burst may not be as strong for open vowels as closed vowels, and that the contribution of the /b/ burst may not always be insignificant. It is precisely to an understanding of such detailed variations that Kuhn (1975) has added by identifying both the burst and "the major spectral peak for the vowel" with the front cavity resonance. In short, Stevens' general description of the conditions under which the burst contributes to the spectral changes following release is consistent both with Kuhn's (1975) front cavity analysis and with our own results. In the following discussion, we attempt to develop some implications of these results for the perceptual process.

GENERAL DISCUSSION

An important feature of the results of Experiments I and II was the tendency toward reciprocal performances on bursts and transitions; where the perceptual weight of one increased, the weight of the other declined. These reciprocal relations follow systematically from the acoustic structure of the syllable. Where transitions are brief (for /b/ before rounded vowels, for /d/ before high front vowels, for /g/ before close vowels), the burst lies near the main formant of the following vowel and contributes significantly to the perceptual outcome; where transitions are extensive (for /b/ before middle, unrounded vowels, for /d/ before central-back vowels), the burst is distinct from the main formant of the following vowel, and contributes little. If we combine this observation with the conclusions of Experiment III, we are led to recognize that, wherever bursts and transitions contribute significantly to the perceptual outcome, they are acoustically and functionally (that is, perceptually) equivalent; both provide a spectrally continuous change from the consonantal release into the following vowel by which the listener can estimate place of articulation. To say that they are equivalent is not, of course, to say that they are alternative. In natural speech, as we have already emphasized, it must be rare that a listener relies on burst alone or on transition alone,

and in Experiments I and II, a single cue was not often sufficient to hold recognition at the level of the original syllable. Bursts and transitions are equivalent and complementary.

Once again, this observation is not new. Over twenty years ago Cooper et al., (1952) remarked that "bursts and transitions complement each other in the sense that when one cue is weak, the other is usually strong." (p. 603). In a similar vein, Fischer-Jørgensen (1954) commented on synthetic speech studies: "The listener does not compare explosion with explosion and transition with transition, but compares artificial syllables comprising either explosion or transition with natural syllables that always contain both" (p. 56). Finally, Fant (1959; 1960:217) has repeatedly emphasized that the qualitatively distinct acoustic segments during the first 10-30 msec after release are probably not auditorily discriminable and "should be regarded as a single stimulus rather than as a set of independent cues" (Fant, 1969:21). And, as we saw above, the acoustic and functional inseparability of burst and transition is implicit in "the rapid spectrum changes" following release that Stevens (1975: 311) describes. In short, the opposition between invariant burst cues and variable transitional cues, imagined by Cole and Scott (1974a, 1974b), is false. Far from being opposed, bursts and transitions are functionally identical.

In conclusion, the results of the present study, and, in particular, the apparent functional equivalence of release bursts and transitions, suggest that the perceptual process may entail continuous tracking of vocal tract resonances. The importance of transitional information for the recognition not only of stop consonants in many contexts, but also of /w,r,l,y/, nasal consonants, fricatives and perhaps even vowels (Lindblom and Studdert-Kennedy, 1967; Shankweiler, Strange and Verbrugge, in press) is attested by an extensive literature (for review, see Liberman et al., 1967; Stevens and House, 1972; Studdert-Kennedy, 1974, 1976; Darwin, in press). We do not doubt that the acoustic invariants for these phonetic segments may eventually be specified; however, we see little grounds for expecting that they will be specified without reference to context.

REFERENCES

Bell-Berti, F. (1975) Control of pharynagel cavity size for English voiced and voiceless stops. J. Acoust. Soc. Am. 57, 456-461.

Cole, R. A. and B. Scott (1974a) The phantom in the phoneme: invariant cues for stop consonants. Percept. Psychophys. 15, 101-107.

Cole, R. A. and B. Scott. (1974b) Toward a theory of speech perception. Psychol. Rev. 81, 348-374.

Cooper, F. S., P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman. (1952) Some experiments on the perception of synthetic speech sounds. J. Acoust. Soc. Am. 24, 597-606.

Cooper, F. S. and I. G. Mattingly. (1969) A computer controlled PCM system for the investigation of dichotic perception. J. Acoust. Soc. Am. 46, 115(A).

Darwin, C. J. (in press) The perception of speech. In Handbook of Perception Vol. 7, ed. by E. C. Carterette and M. P. Friedman. (New York: Academic Press).

Day, R. S. (1970) Temporal-order perception of a reversible phoneme cluster. J. Acoust. Soc. Am. 48, 95(A).

Delattre, P. C., A. M. Liberman, and F. S. Cooper. (1955) Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. 27, 769-773.
Fant, G. (1959) Acoustic description and classification of phonetic units. Ericsson Tech. 1, 1-52.
Fant, G. (1960) Acoustic Theory of Speech Production. ('s-Gravenhage: Mouton and Co.).
Fant, G. (1969) Stops in CV-syllables. Speech Transmission Laboratory Quarterly Progress and Status Report, No. 4, 1-25.
Fischer-Jørgensen, E. (1954) Acoustic analysis of stop consonants. Miscellanea Phonetica 2, 42-49.
Fischer-Jorgensen, E. (1972) Tape cutting experiments with Danish stop consonants in initial position. Annual Report VII (Institute of Phonetics, Univ. of Copenhagen, Copenhagen, Denmark).
Flanagan, J. L. (1972) Speech Analysis Synthesis and Perception. 2nd ed. (New York: Springer Verlag).
Fujimura, O. (1961) Bilabial stop and nasal consonants: a motion picture study and its acoustical implications. J. Speech Hearing Res. 4, 233-247.
Halle, M., G. W. Hughes, and J-P. A. Radley. (1957) Acoustic properties of stop consonants. J. Acoust. Soc. Am. 29, 107-116.
Harris, K. S. (1958) Cues for the discrimination of American English fricatives in spoken syllables. Lang. Speech 7, 1-7.
Hoffman, H. S. (1958) Study of some cues in the perception of the voiced stop consonants. J. Acoust. Soc. Am. 30, 1035-1041.
Klatt, D. (1975) Voice onset time, frication, and aspiration in word-initial consonant clusters. J. Speech Hearing Res. 18, 686-706.
Kuehn, D. P. (1973) A cinefluorographic investigation of articulatory velocities. Unpublished Ph.D. thesis, Iowa University.
Kuhn, G. M. (1975) On the front cavity resonance and its possible role in speech perception. J. Acoust. Soc. Amer. 58, 428-433.
Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1952) The role of selected stimulus variables in the perception of the unvoiced stop consonants. Am. J. Psychol. 65, 497-516.
Liberman, A. M., P. C. Delattre, F. S. Cooper, and L. J. Gerstman. (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monogr. 68, 1-13.
Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1958) Some cues for the distinction between voiced and voiceless stops in initial position. Lang. Speech 1, 153-167.
Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding-Processes in Human Memory, ed. by A. W. Melton and E. Martin. (New York: Halstead Press).
Lindblom, B. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773-1781.
Lindblom, B. and M. Studdert-Kennedy. (1967) On the role of formant transitions in vowel recognition. J. Acoust. Soc. Am. 42, 830-843.
Lisker, L. and A. S. Abramson. (1964) A cross language study of voicing in initial stops: acoustical measurements. Word 20, 384-422.
Lisker, L. and A. S. Abramson. (1967) Some effects of context on voice onset time in English stops. Lang. Speech 10, 1-28.
Stevens, K. N. and A. S. House. (1972) Speech perception. In Foundations of Modern Auditory Theory, Vol. 11, ed. by J. V. Tobias. (New York: Academic Press).

Stevens, K. N. (1975) The potential role of property detectors in the perception of consonants. In Auditory Analysis and Perception of Speech, ed. by G. Fant and M. A. A. Tatham. (New York: Academic Press).

Studdert-Kennedy, M. (1974) The perception of speech. In Current Trends in Linguistics, Vol. 12, ed. by T. Sebeok. (The Hague: Mouton).

Studdert-Kennedy, M. (1976) Speech perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (New York: Academic Press).

Winitz, H., M. E. Scheib, and J. A. Reeds. (1972) Identification of stops and vowels for the burst portion of /p,t,k/ isolation from conversational speech. J. Acoust. Soc. Am. 51, 1309-1317.

Modes of Perceiving: Abstracts, Comments, and Notes*

M. T. Turvey[+] and Sandra Sears Prindle[++]

INTRODUCTION

Intuitively, deliberations on modes of perceiving are intended to flesh out something of the special manner in which humans apprehend their world. In principle, the importance of the enterprise lies in the fact that even an elementary cataloging of modes would significantly fetter the construction of theories of perception and cognition. It goes without saying that in evolving the perceptual styles of humans and animals, nature did not build "general-purpose machines," but rather "special-purpose machines"; and whatever plasticity humans and animals manifest it is a "special-purpose plasticity." Nevertheless, one has the impression that often theory-making proceeds untrammeled by a serious consideration of natural constraints and seems to be oriented toward a general-purpose, context-free perceiver.

While it is the case that deliberating on modes of perceiving is well motivated, unfortunately it is not immediately obvious what it is that one is deliberating. The concept of "mode" is an intuitive object; tacitly we can appreciate the catalytic value of the concept in thinking about matters of perceiving and knowing, but we cannot say precisely and unequivocally what a mode is. Partly in response to this equivocality, our approach to summarizing the volume[1] takes the following form. First, we precis the various papers conveying, ideally, the larger point made by each author. Second, we seek fundamental themes which weave these larger points together in the hope that these

---

* To be published in Psychological Modes of Perceiving and Processing of Information, ed. by H. Pick and E. Saltzman (Hillsdale, N. J.: Lawrence Erlbaum Associates).

+ Also University of Connecticut, Storrs.

++University of Connecticut, Storrs.

---

[1] NOTE: All mention of "the volume" and of the various authors (not included here in the References) refer to the book, Psychological Modes of Perceiving and Processing of Information, ed. by H. Pick and E. Saltzman, Hillsdale, N. J.: Lawrence Erlbaum Associates, in which this paper will appear.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

themes will identify major constraints on the theory of perception. Third, and separately, we gather some of our elementary and rough thoughts on the abstract notion of "style." These we present as notes toward a tenable characterization of the concept of mode in psychology, and in this respect our remarks may be regarded as complementing those of Pick and Saltzman in the initial chapter of Psychological Modes of Perceiving and Processing of Information.

ABSTRACTS AND COMMENTS

A contrast that comes rapidly to mind when one thinks of modes of perceptual processing is that of unconscious and conscious, or as Posner and his colleagues describe it, the contrast of automatic and attentive. Processing of the former kind, we are told by Posner et al., is very much a parallel affair while that of the latter kind is considerably more serial. The significant consequence of attentive processing is that it consumes a portion of the limited resource capacity, thereby curtailing the processing of other concurrent signals, and further, that it induces inertia in the processing apparatus. When there is the intentional selection (attentive) of a particular psychological channel or pathway, it takes effort and time to shift attention to another channel when needed. The costs, therefore, of attentive processing are manifestly plain; among its benefits we may suppose, is a finer grain of analysis.

Inasmuch as the mode of attentive processing can be set by instruction we may ask: To what, precisely, is my processing directed when I am instructed to attend to a given location? It is this question which guides the series of ingenious experiments reported by Posner, Niessen, and Ogden. The conclusion is curious and provocative. Apparently there is little benefit to be gained by knowing ahead of time the external location at which a signal will occur if I do not know the modality which will convey the signal. By inference, attentive processing cannot be directed to a location with the same efficacy that it can be directed to a modality; preference is for knowing the messenger rather than knowing from where the message is coming.

With respect to the inertia induced by the mode of attentive processing, Posner and colleagues (Posner, Niessen, and Kline, 1976) have recently interpreted the peculiar phenomenon of visual capture as being indicative of an inertial assymetry between switching from vision to another modality, and switching from another modality to vision. One is reminded that visual capture refers to the dominating role that vision has in the human conscious experience. When the information for vision and another modality are in conflict, vision is the likely victor. Thus, I will experience my hand as tracing out a curved line when in fact it traces a straight line that has been prismatically distorted for visual consumption (Gibson and Radner, 1937). The relation between visual dominance and the inertial aspects of attentive processing is thus expressed; experiment suggests that vision is not an especially efficient alerting system because the time to switch into vision from another modality significantly exceeds the time to switch between two nonvisual modalities. If the human animal was not in the visual modality at the time of occurrence of an ecologically significant optical signal, it would be, on this account, at a distinct disadvantage. Consequently, one hypothesizes that in response to evolutionary expediency, nature saw fit to bias human conscious experience toward the visual pickup of

information. That the bias is software rather than hardware is suggested by the following observation made when prismatic distortion of vision accompanies haptic exploration: if vision is attended to, the haptic system undergoes an adaptive shift, but if the haptic system is attended to, it is vision that is recalibrated (Kelso, Cook, Olsen, and Epstein, 1976). With other things being equal, it is vision that is attended to by choice.

Herein lies a rationalization of the "primacy of vision" which dovetails with Lee's deliberations, for these also sought to express the supremacy of visual perception. We shall see that while the account of visual primacy derived from Posner's work emphasizes the costs of vision, that the account of Lee's emphasizes the benefits of vision.

The term modality enjoys considerable usage. It is a term befitting the convention of classifying senses according to the qualitatively different conscious experiences. Following this convention, the special sense of vision is a source of visual sensation and the special sense of proprioception is a source of sensation of one's own movements. It has been remarked by Gibson (1966), and echoed enthusiastically by Lee, that it is far more sensible to classify the senses in terms of activities such as looking and listening than in terms of passive conduits transporting qualitatively different sense data. When approached from this perspective, the term "perceptual system" is substituted for the term "senses." And whereas the fundamental role assigned to the senses is that of providing raw materials for the creation of conscious experience, the fundamental role assigned to perceptual systems is that of obtaining information in the service of activity, as Lee so elegantly puts it.

A promissory note of Gibson's (1966) approach is that different perceptual systems can be sensitive to the same information. Here information is defined as information about the environment in a sense of specificity to it; and it is this sense of the term that is intended by Pick and Saltzman. The claim is that the pickup of information of a given type is not necessarily the prerogative of any one perceptual system. It is a claim that is easily glossed over by students of perception but its ramifications are considerable (White, Saunders, Scaddon, Bach-Y-Rita, and Collins, 1970); for those who think in terms of special senses--or special modalities--it is anathema.

Lee reminds us that in the regulation and control of activity three kinds of information are needed: information about surface layout and events; information about relations and changing relations among the limbs; and information about the motion of the body relative to the environment. His argument is that vision supplies all three--it is trimodal--and does so better than the other perceptual systems. Hence, we have the "primacy of vision." Essentially, vision's relation to the other perceptual systems is that of overseer: vision tunes and calibrates those systems which would otherwise be imprecise sources of information relative to the guidance of activity. A dramatic demonstration of vision's role with respect to body-related (proprioceptive) information is provided by Gross, Webb, and Melzack (1974). When asked to plot the position of an arm which rested without moving and out of view (it was hidden by an opaque shield), participants could do so quite accurately if the delay from last seeing the arm was relatively short. However, with the passage of time the position of the resting arm was felt to migrate to one of two positions--flection-adduction or extension-abduction.

It is but a small leap from Lee's paper to Mack's. For the contentions between traditional and Gibsonian perspectives--between indirect and direct realism--that were merely interlineal in Lee's paper are brought to focus in Mack's paper. A departure point is that thorny issue on which Boring and Gibson collided: Can the visual world be apprehended independently of the visual field? Its cognate is perhaps better known: Can perception be indifferent to sensation? This issue takes many forms that are by no means identical. The gist, however, is unmistakable: it is a matter of whether the world can be perceived first hand--directly, or only second hand (by virtue of some surrogate)--indirectly.

Mack distinguishes between proximal and constancy perception. Put bluntly, proximal perception is determined solely by the absolute properties of the retinal image; in contrast, constancy perception is determined by these image properties only partially, or not at all. Obviously the central concept is that of the retinal image and we may, after Gibson (1950), identify two versions of the concept for they are of significance to Mack's remarks. In one version, the image is defined as the anatomical pattern of cells that are excited--this we call the anatomical image; in the other version, the image is defined as the ordinal pattern of excitations indifferent to the location of cells excited--this we call the ordinal image. It was Gibson's (1950) intuition that seeing in terms of the anatomical image and seeing in terms of the ordinal image were two different ways of seeing, two different modes, if you wish.

Generally, when one talks about the retinal image it is the anatomical image one has in mind. Related to this conception is a tendency to talk about the light at an eye in terms of Euclidean geometry and thus to emphasize absolute metrical values. Euclidean geometry was all that was known to the ancients and to the intellectual ancestry who established the conventions and fundamental assumptions of contemporary visual theory. In contrast, the conception of the ordinal image encourages the adoption of projective geometry and its emphasis on abstract relations preserved over projective transformations.

When one describes the retinal image or proximal stimulus in Euclidean terms there is an apparent lack of correspondence between the image and its distal referrent. Consequently, insofar as perception tends to be veridical, it follows that the light at an eye underdetermines perceptual experience. The appropriate perception arises by virtue of processes which supplement the retinal image. Most generally these processes are thought of as memorial or problem solving in nature. The observer in this, the traditional point of view, is much like Sherlock Holmes who must attempt to determine what actually transpired from the limited data or available clues. We refer to this point of view as constructivism, in order to emphasize the central hypothesis that visual perception is built out of a number of ingredients--some of which are provided by the retinal image and some of which are provided by other extra-visual sources (Turvey, 1974, 1975).

Let us now return to Mack's three modes of perception. By all accounts the proximal mode is evident only when the conditions of observation are highly constrained; for example, a two-dimensional nonchanging display exposed briefly against a homogeneous background and viewed from a stationary point of observation. In a phrase, the mode of proximal perception is precipitated by impoverished stimulation.

The subject-relative constancy mode is most obviously an example of constructivism, for the ingredients in the perception recipe include absolute and local anatomical image properties and nonvisual information about eye, head, and body orientation. In subject-relative constancy one must go beyond the light to an eye in order to determine perceptual experience. By our interpretation subject-relative perception uses the anatomical image. And what we would like to believe is that only in rare and artificial circumstances does the anatomical image play a determining role in experience. In short, operating in the proximal and in the subject-relative constancy mode are unnatural recourses for the visual perceptual system.

We are led therefore to the point of view that object-relative constancy perception is representative of the style in which the visual perceptual system maintains contact with the environment. In the object-relative mode, abstract relations in the structured light at an eye provide the optical support for visual perception without supplementing by nonvisual data. Mack informs us that visual perception in the laboratory may sometimes be in error because of the curious bias of the visual system to operate in the object-relative mode when the subject-relative mode is more felicitous for the conditions of observation. But we should not be surprised by this fact. If it is the case that the optical flow pattern at a moving point of observation is structured adjacently and successively in ways that are specific to the observer's movement and to the properties of the environment as Gibson (1966) and Lee argue, then we should suppose that evolution optimized the visual perceptual system of humans and beast to be sensitive to this structure. It is the abstract relational information in the ordinal image understood as the ambient optic array and not the metrical character of the anatomical image, which has constrained the evolution of visual systems. And if that invariant information is specific to the environment, then as the optical support for visual perception it merely has to be detected; it would not have to be supplemented by other sources of knowledge.

Let us summarize to this point. Our quest for the natural style in which humans perceive has realized two dividends. One is that--ceteris paribus--vision preempts conscious experience because it is the most abundant supplier of information about the environment and about one's self; as far as perceptual systems go, it is potentially more costly not to be visually attentive, and considerably more laborious to become so. The other dividend is that, although visual perception may operate in a subject-relative or constructivelike mode, this is not its more natural and preferred style. We pursue the latter proposition in the paper of Shaw and Pittenger.

As remarked earlier, theorizing on matters of visual perception has tended to begin with the retinal image understood as an anatomical arrangement. We can further comment that theorizing has tended to begin with the understanding of the retinal image as a static bidimensional form. The consequence of this attitude is two-fold; first, the analysis of pattern or form perception is taken as propaedeutic to the theory of visual perception; and second, that change, defined as the transformation of an object over time, is said to be inferred from a succession of static retinal images.

The conceptualization of the optical support for visual perception as static and bidimensional has a long tradition. We owe to the 10th century Arab scholar Al Hasan the first comprehensive exegesis of the relation between the image

on the retina and visual perception. Through Berkeley and Von Helmholtz the tradition has been popularly maintained, and it is the source of the fundamental though rarely commented on, suppositions of contemporary visual information processing theory and research (see Niesser, 1967; Haber, 1971). Obviously, if we assume that a two-dimensional static description of the world is the starting point of visual experience, then we have identified the task of perceptual theory; to explain the means by which we arrive at static three-dimensional descriptions (depth perception, object perception) and dynamic three-dimensional descriptions (events). As we have already anticipated, the traditional explanation is that such perceptual experiences are constructed with the assistance of memory.

Suppose, however, that our intuitions about perception are guided not by history and the retinal image, but by the concepts of evolution and ecology. Such being the case, we would recognize that locomotion and the continuous orienting of the perceptual apparatus to the environment are the sine qua non of successful adaptation. We would recognize, in short, that dynamically transforming optic arrays would be the norm and that static frozen optic arrays would be the exception. Furthermore, we would appreciate that an animal would wish to know not simply what kind of object it was looking at but what kind of change the object was undergoing. Perception of the forms of change is of paramount importance to adaptation. In sum, from an evolutionary/ecological perspective we might be led to conjecture that the proper point of departure for a theory of visual perception is kinetic events, and not two-dimensional static forms (Gibson, 1966; Johannson, 1974). This conclusion is cognate with the one that we reached in our discussion of Mack's paper.

An event, Shaw and Pittenger inform us elsewhere (Pittenger and Shaw, 1975), is composed of two things: the object or complex of objects undergoing the change and the change itself. The optical support for the perception of the former (the object) is referred to as the structural invariant, and the optical support for the perception of the latter (the change) is referred to as the transformational invariant. This understanding of the structure of events follows from Gibson's working hypothesis of ecological optics, namely, that for any isolable environmental property there is a corresponding isolable property in the transforming optic array, however complex. By arguing that there are higher-order invariants specific to the styles of change, Shaw and Pittinger express the unorthodox view that the perception of change is direct. They argue, in paraphrase of Gibson's notorious aphorism, that the perception of change is not based on the perception of static forms but, rather, on the detection of formless invariants over time.

Recent examinations of comparatively simple events such as an object moving at constant velocity or accelerating from one position to another, reveal that the perceptions of velocity and acceleration are not based on the prior discriminations of spatial and temporal event (cf. Lappin, Bell, Harm, and Kottas, 1975; Rosenbaum, 1975). Explanations of perceived velocity and acceleration in the constructivist mode would necessitate epistemic mediation, for example, having discriminated at least two spatial positions--taking two retinal snapshots--and having monitored the time elapsed between the two positions, then velocity could be computed by means of a simple formula. The evidence, however, favors the view that velocity and acceleration are not constructed

percepts but directly perceivable attributes of stimulation. This conclusion reflects the larger point that Shaw and Pittenger wish to make, namely, that the nominalistic attitude toward accounts of perceptual experience is fundamentally in error. We can phrase this differently and positively; what Shaw and Pittinger wish to emphasize is the primacy of the abstract.

If this thesis is not already foreign enough for most students of perception to appreciate, it is made all the more so when one considers that in our lifetimes events range from the order of milliseconds to the order of years. How is it possible, we ask, to apprehend slow events without the mediation of memories? What can it possibly mean to detect the transformational invariant of a slow event such as, say, aging? Shaw and Pittenger indicate the direction we might take in search of an answer. More tangibly, they lay bare the absurdity of the conventional story of memory mediation. For if my apprehension of a slow event comes from memory, then I must have some way of collecting the relevant memories, and this implies that I have knowledge of the transformation that relates them to each other. But the transformation that relates the memories to each other is what I have to infer; it cannot be presupposed. Even if we permit a fortuitious gathering of the relevant memories, the memory mediation story fails to work; for now we must attribute to the inferential processes a priori knowledge of transformations in order that we might infer from the nominal data which event transpired.

In the preceding paragraphs we have developed the intuitive notion that visual perceptual theory should be anchored in event perception, that is, in the perception of the transforming optic array. Obviously, within such a framework a static two-dimensional arrangement must be regarded as a type of "frozen" event in which the structured light at an eye has been reduced in its efficiency as a specifier of environmental facts. Belaboring the point somewhat, we may claim that truly static perception is artifactual arising at a relatively late phase in evolution. The perception of paintings, photographs, and the like exemplify the limiting case--and it is just this kind of perception that is examined by Hagen. Her questions are straightforward and they follow naturally from the preceding remarks: Is perceiving pictures much the same as perceiving the ordinary environment, or is there something special going on with pictures? Is there either something special about the information pictures contain or something special that we do with that information? As we might anticipate, Gibson's intuitions on these matters are essentially that the perception of pictures and the perception of the scenes they depict do not differ qualitatively, for the essence of pictures is that the information they convey is structurally equivalent to that of the scenes they depict. In a phrase, picture perception, like event perception, is not epistemically mediated.

Experimentation with a wide range of conditions reveals that when pictures (slides, photographic prints, line drawings) are from the right station point and apparently equate static monocular surface-layout information, the perception of the real scene is always superior to that of the facsimile. This could be because of the perceptual advantages in moving the eye over a real scene rather than over a picture. Alternatively, as Hagen suggests, it could be because, when faced with the task of appreciating the three-dimensional structure specified by the pictorial information, one must suppress the concurrent information specifying that the "frozen" event is actually two dimensional. In either

case, the Gibsonian thesis (Gibson, 1971) that picture perception can be direct like ordinary perception (that is, not epistemically mediated) is not appreciably harmed.

A different conclusion, however, is implied by the "Pirenne paradox." An observer's appreciation of the three-dimensional scene depicted by a two-dimensional picture is significantly enhanced when he or she adopts the wrong station point. This is paradoxical inasmuch as the perspective information provided by a picture is only equivalent to that provided by a real scene at the center of projection for the picture. Pirenne's interpretation of this paradox is clearly in the constructivist mode. Looking at a picture off-center enhances one's awareness of flatness and induces one to use knowledge about the internal components of the picture; by so doing one not only compensates for the perspectival asynchrony, but in addition and more importantly, facilitates the perception of the internal components. The problem with this interpretation as we see it, is that it is not obvious why viewing a picture from an incorrect station point should trigger a compensatory attitude any more than the actual knowledge that one is in the context of picture-viewing. We venture that a more useful approach to the Pirenne paradox lies in noting in what ways a perspective from the wrong station point could be more informative about the internal components than a perspective from the correct station point. Is it that the perspective accompanying an off-center station point specifies the perspective at the on-center station point; in short, that at the wrong station point one has, in some curious fashion, two perspectives on the static object?

All this concern with perception from particular points of view and with the perception of pictures as a possibly particular kind of seeing leads us without too much difficulty, to perceiving--more precisely to visualizing--from no particular point of view. Exemplary of such visualizing is imaging; it has been Paivio's contribution to restore imaging to respectability in academic psychology.

The mechanisms of imaging are part and parcel of a "nonverbal" system which is said by Paivio to mediate both our experience of the environment and our nonverbal actions. This imagery system operates independently of the "verbal" system which supports our linguistic endeavors whether they be performed by ear, eye, or hand. It is the case, as Paivio argues, that the verbal system is dependent on the nonverbal, for while the former communicates what we know about the environment, the latter is the primary source of that knowledge. Nevertheless, the two systems are distinguished by the kinds of objects which comprise their respective memory components. For the imagery system the objects are said to be perceptual analogs, while for the verbal system they are discreet linguistic entities (for example, words).

But how should we characterize the perceptual knowledge that Paivio refers to? On the assumption that the relevant entities are discreet and static images, we might use symbolic logic, formal grammars, machine theory, and the like to characterize them. On this assumption an image could be treated as a symbol, and perceptual knowledge viewed as a symbol manipulating system. Since language can be similarly characterized, the possibility arises that Paivio's imagery and verbal nodes are fed by one and the same symbol manipulating system. This approach is favored by Anderson and Bower (1973) among others, but regarded with skepticism by Paivio.

We have remarked several times in this summary of the volume that the informational support for perceiving and acting consists of abstract invariants defined over time and further, that the kernal units for perceptual theory are kinetic events. If Paivio wishes to maintain that the perceptual knowledge which feeds his imagery system is continuous with perception, then we might wish to propose that perceptual knowledge is most appropriately characterized in terms of events--rather than static images--and cognately, in terms of dynamic abstract invariants.

Our facility with metaphor provides a case in point. If I am requested to remember the sentence: "Rabbits are like children skipping rope down the sidewalk" then an effective prompt at a later date is: "Kangaroos move like a basketball being dribbled" (Verbrugge, 1975). Why should this be so? It stretches the imagination to believe that the equivalence between the two sentences lies in semantic features common to rabbits, children, skipping ropes, kangaroos, and basketballs, or that it could be realized by compounding static images. We may conjecture that the two sentences share a common abstract invariant--periodic up and down motion relative to the ground plane, and it is the detection of this invariant which determines their equivalence.

We alluded above to imaging as perceiving from no particular station point. In a delightful mix of words Verbrugge (1975) remarks that: "Language is more like a piano score--an invitation to create meaning." In his perspective, the listener seeks structure among the virtual objects suggested by a sentence much as he seeks structure in the optic array--except in the linguistic case he does so from no particular station point. The suggestion is that the style in which we perceive language is not qualitatively different from the style in which we perceive or visualize the environment. Our guess is that if Paivio's nonverbal and verbal systems conflate at all, it is not because they use a common propositional format, but because they are both oriented to the abstract invariants which specify events.

Let us pursue the verbal mode a little further. With respect to language perception by ear, there are three aspects of that perception that we might distinguish. We can identify a semantic mode in which we experience the meaning of what we hear, a phonological mode in which we experience what we said distinct from what it means, and an acoustic mode in which we experience certain nonlinguistic aspects of speech (cf. Halwes and Wyre, 1974). Paivio's remarks and our comments in the preceding paragraphs were directed at the semantic mode; the paper by MacNeilage focused on the phonological and the acoustic.

MacNeilage's bone of contention is that perceiving in the phonological mode is qualitatively different from perceiving in the acoustic mode. More precisely, MacNeilage takes issue with the claim that the underlying experiences of language at the phonological level are fundamentally articulatory processes. We may recognize strong and weak versions of this claim. In the strong version, the processes responsible for phonological experience are identical to the neuromotor processes of articulatory coordination involved in speaking, but with the motor commands inhibited at some level prior to inducing mechanical muscular events. In the weak version, phonological experience is constructed from the acoustic data by virtue of knowledge about what human vocal tracts can and cannot do.

The data often cited in support of the motor or articulatory theory of speech perception are no longer as compelling as they might once have been. Thus, one of the cornerstones of the theory, categorical perception, is now known to be indigenous to neither speech nor humans. Nevertheless, there are some curious observations which point to an intimacy between perceiving and producing speech that cannot be dismissed lightheartedly. Among these we might include the tight coupling between hearing and speaking vowels witnessed by the exceptionally rapid shadowing of Chistovich's (1961) subjects, and a recent and provocative discovery compatible with the weaker version of the theory that has been made by Liberman and Dorman (see Liberman, 1975). If two syllables such as /bɛb/ and /dɛ/ are arranged very closely together in time, one of the stop consonants is "masked" so that the listener hears /bɛ/ instead of /bɛb/. However, this perceptual impairment can be readily eliminated by having the two syllables spoken by two different vocal tracts: no matter how temporally proximate is the presentation of the two syllables, as long as they are produced by <u>different</u> vocal tracts they can be heard as separate phonological events. In the perspective of the weaker version of the articulatory theory, this result is interpretable in terms of the listener's tacit knowledge of vocal tracts which specifies that although the rapid transition from one stop consonant to the other is impossible for a single speaker, it can be achieved easily by two speakers.

However, the thrust of MacNeilage's survey is not to be denied; there is relatively little to recommend a motor theory. The hypothesis that speech is perceived by reference to how it is produced is countered by the hypothesis that speech is produced by reference to how it is perceived, that is, the motor theory of perception is nullified by an acoustic theory of production. In view of the latter, we might not wish to regard either phonological perception or production as parasitic on the other, but rather, that perceiving speech and producing speech are related through an abstract structure that is common to both but indiginous to neither (Turvey, 1976). At least for the lowly cricket there is a suggestion that perception and production are manifestations of the same structure: a common gene might mediate the male's song and the female's perception of it (Hoy and Paul, 1973).

Perhaps the larger point to be made with respect to a comparison of perception in the phonological and acoustic modes, is that nonphonological auditory perception has not been treated fairly in theory and research. In studying the auditory perceptual system, insufficient weight has been given to its primary role of detecting environmental sources of mechanical disturbance. Ecologically, the role of audition is to identify the source of sound and the behavior of the identified source (cf. Schubert, 1975). The auditory perceptual system, like its visual counterpart, is oriented to events, but our understanding of auditory perception outside of speech, is based on the perception of sounds that are more nearly abstract than event related.

Consider the common use of artificial sounds in the laboratory; examples are steady-state pure tones or steady-state short bursts of random noise. The most notable feature of the perception of sounds such as these is that they resist reliable identification (Pfafflin and Matthews, 1966; Webster, Woodhead, and Carpenter, 1970). In part, this seems to be owing to the fact that sounds relating to ecological events--the class of sounds to which the auditory perceptual system has been attuned by evolution--involve rapid transients in intensity.

These transients are concomitants of the onsets and offsets of the mechanical disturbances to which the sounds correspond. In the absence of these transients, specification of the identity of the source of the sound is far from ideal (see Saldanha and Corso, 1964; Luce and Clark, 1965, 1967).

Now speech perception is the perception of sound as modulated by articulatory events. But the nonspeech perception with which it is often compared is the perception of sounds that have been stripped of ecological validity. A pure steady-state tone specifies no event whatsoever. The contrast between speech and nonspeech perception or linguistic and nonlinguistic perception is, in our opinion, more often a contrast between event perception and nonevent perception. Such being the case, speculation on how the perception of speech differs from that of nonspeech is premature. Imagine hearing a can or a dish fall to the ground. We can ask with Schubert (1975:102) "Was the can large or small; of heavy or light construction; was it in contact with a hard surface like concrete or an absorbent one like earth or grass? Did the dish shatter or bounce?" Conjecturally, we answer these questions based on the fact that the objects and substances involved, and their interactions, modulate the acoustic array in specific and invariant ways. But what do we know of such invariants and their detection? The answer, unfortunately, is very little. Nevertheless, it is the character of this kind of auditory perception to which the character of speech perception should be compared. There is one modest difference between the two kinds of perception which immediately comes to mind. Differentiating nonspeech environmental events probably takes full advantage of the exteroceptive expertise of vision; in contrast it is roughly apparent that vision's role in the differentiating of speech events is minimal.

At this juncture let us anthologize our review and comments thus far. To the primacy of vision we have now added the primacy of abstract relational information defined over time. The latter is meant to contrast with the more common attitude which asserts the primacy of nominal, punctate, and momentary entities in perception. Furthermore, we have promoted kinetic events as opposed to static retinal images or steady-state sounds as the ecological entities to which evolution has attuned perceptual systems and thus the proper departure point for theorizing. Admittedly this promotion does not reflect the bias of all of the authors of this volume but, ideally, our remarks have been sufficient to support our intuition that the event concept provides a unifying theme.

We consider now the remaining two papers, those of Trevarthan and Halliday. If the papers discussed thus far can be categorized as papers directed to the what and the how of perception, that is, to the issues of what there is to be perceived and how it is perceived, then those of Trevarthan and Halliday may be categorized as papers directed to the who of perception--the epistemic agent or algorist (Shaw and MacIntyre, 1974). As Shaw remarks, the questions of the "what," the "how," and the "who" of perception form a closed set of questions with answers to any one coimplicating answers to the other two. It is fitting, therefore, that the final papers in this volume emphasize the thus far omitted member of the above triad.

Briefly, Trevarthan's major points are these: first, that psychologists are insufficiently sensitive to the implications of anatomy--particularly the somatotopic principal--for perception and action theory; second, that perceptual systems should be considered in the light of mechanisms for action; and finally, that contrary to time-honored claims, infant behavior is intentional. This last point is also the larger point of Halliday's essay.

The organization of the vertebrate midbrain provides an instructive example of both the first and second points. If a map is drawn of the projection from the eyes to the midbrain tectum in the coordinates of the eye, then for animals with frontal-oriented eyes and animals with lateral-oriented eyes, the two maps are quite dissimilar. However, if the maps are drawn in the coordinates of the behavioral field, that is with respect to the asymmetry of the body, we would observe that the two maps are virtually identical. As a general principle, the mapping from eyes to tectum in the coordinates of the behavioral field is relatively invariant; and this mapping of visual loci also maps a topography of points of entry into the action system.

The confluence between seeing and doing was highlighted earlier in Lee's paper and that between hearing and speaking was critically examined in MacNeilage's. A further, though brief comment on the perception-action relation is warranted. The problem of coordination is the problem of controlling the enormous number of degrees of freedom that the biokinematic links--the skeletomuscular hardware--can attain (Bernstein, 1967). In view of the indeterminancy of the peripheral motor apparatus, it is most unlikely that executive processes coordinate movement through the individual control of each degree of freedom. In short, action plans are probably not written in terms of individual muscle contractions. The alternative view(Gelfand, Gurfinkel, Tsetlin, and Shik, 1971) is that action plans are written in terms of muscle linkages, that is, muscle-joint complexes whose activities covary and whose kinematic characteristics are similar. Such linkages may be referred to as coordinative structures (Turvey, 1976). The role of these structures is to reduce the degree of freedom requiring control, for a coordinative structure behaves quasi-autonomously and therefore, from the perspective of an executive procedure it represents but a single degree of freedom. Coordinative structures provide only a partial solution to the problem of degrees of freedom. In the performance of acts the degrees of freedom are regulated with precision, but an action plan is by necessity crudely specified in the language of coordinative structures. We therefore ask: How are movements performed that are precise in their timing, velocity, and displacement? Obviously perception must modulate unfolding action plans, but in order to do so perceptual information must be parsed in ways compatible with the nested components of the evolving act and must be injected into the action system at the right place and at the right time. How this is done is not at all apparent, but we may regard Trevarthan's comments on somatotopic organization and on the contrastive capabilities of focal and ambient vision as preliminary steps in the direction of an answer.

Let us conclude our summary of this volume with the shared insights of Trevarthan and Halliday on the nature of infant behavior. An appropriate backdrop is provided by a brief consideration of Gibson's shift away from perceptual psychophysics. In common with his predecessors, Gibson in his early writings (Gibson, 1950) adopted the causal chain theory of perception; perceptual experience was caused by stimuli. However, as he developed the concept of the optic array it became evident to him that the formulation "stimuli trigger perception" was incorrect and that a more judicious formulation was that "the ambient optic array supports the regulation and coordination of activity." The significance of the reformulation is that it emphasizes exploration and selection with the animal as agent rather than the animal as reactant.

Suppose that we do adopt an agent or algoristic oriented view of the relation between what there is to be perceived and how it is perceived. Do we mean

to hold to this view for all stages of ontogeny? Popular scientific and not-so-scientific opinion would most likely respond "no." For the agentlike qualities of the adult perceiver/actor are said to result from a lengthy apprenticeship; the infant human reacts to stimuli in the ageless story, and only comes to plan and regulate his behavior with respect to information after the slow process of enculturation. The contrary and, perhaps, radical claim of Trevarthan and Halliday is that the infant is inherently purposive. What we witness in Trevarthan's and Halliday's behavioral and protolinguistic analyses of infant line, is the infant as algorist possessing and deploying a stock of fundamental strategies or modes for selectively operating upon the world. The disposition of these strategies rests on the capacity to distinguish between animate and inanimate objects which afford different possibilities of interaction. The infant communicates vocally and gesturally with animate objects, but reaches for and manipulates inanimate objects. We learn from Halliday that the inchoate vocalizations of early childhood are actually basic acts of meaning, intended in part, to procure material ends and to maintain contact with and regulate the behavior of those who enter into the communication scenario. To the claim that the infant is inherently agentlike we add the claim that the infant is inherently social.

NOTES

"Mode" has many synonyms of which "style" and "fashion" are perhaps the most common. We speak about this and that style of dress and we will often pass comment on how fashionable or unfashionable a given style happens to be. Such comment is intended to relate the style in question to the context of contemporary living. It is a matter of whether the style is compatible with some broader context of constraints, although the criteria for adjudicating on this subject are rarely unequivocal.

Fashionableness is a passing quality although there are no fixed time limits on a style's period of grace. Nevertheless, it is fair to claim that the longevity of a style of dress is considerably shorter than that of other styles, such as the style of eating. Other styles are even more perpetual; the style of human locomotion, for example, has undergone relatively little change.

Styles, therefore, may be said to lie on a continuum from persistent to transient, and we additionally propose, from immutable to docile. Consider a further aspect; given several styles of dress, a person cannot be dressed in more than one style at a time. In short, different styles of dress are mutually exclusive. Styles are also said to be stereotypic, invariant ways of doing things. A not uncommon reproach of haute couture by those excluded is that they--the in-crowd--all dress or act alike. The epithet stereotypic must be handled cautiously, for its use is likely to blind one to the important fact that to be in style does not mean that one is a carbon copy of one's comrades in fashion. Rather, one's dress differs perceptibly from that of the others in ways which do not violate the prescribed, although often ineffable, conventions. We may say, therefore, that to be in a style is to be in a certain ballpark of states. We will proceed to define a style as a set of constraints which ensures the realization of an invariant condition over variable instances. Unfortunately, equating style and constraint is not a simple way in which to classify styles. Constraints--and thus, by definition, styles--vary on a scale from

light to severe, with the severity of a constraint measured by the reduction it causes in the number of possible configurations, that is, the extent to which it freezes degrees of freedom.

Consider the relation between style of dress and style of dance. We have remarked already that I cannot be in two styles of dress simultaneously as one excludes the other. Similarly I cannot be dancing in two styles simultaneously. Nevertheless, I can be in a style of dress and dance in a certain style at the same time if one of two conditions exists. First, when my style of dance and my style of dress do not affect one another, as is the case when my style of dress does not restrict my movements, then I am perfectly able to do a certain dance while in a certain style of dress. Second, when my style of dress does restrict my movement in some particular way I can still perform a certain dance if the dance constrains my movements in the same way as my style of dress. For example, one can do the currently popular hustle while wearing platform shoes, since the constraint on bending one's foot is the same for the hustle as for platform shoes. However, an Irish jig and platform shoes are not compatible, since the constraint on bending one's feet imposed by platform shoes is not compatible with dancing the jig.

Speaking more generally, two or more styles are compatible (that is, they can coexist) if (1) they govern different degrees of freedom or (2) they selectively freeze the degrees of freedom which they have in common in the same way.

Returning to our dress-dance metaphor, we intuit that when neither of the above conditions is satisfied, styles behave in a coalitional (free-dominance) fashion. That is to say, styles are not organized in a strictly hierarchical manner. Any one style may take precedence over any other style, depending on the event in which the two styles take part. Thus, I may be intent upon wearing my platform shoes in which case I modify the jig so that I do not bend my feet; or, I may be intent upon doing the jig correctly, so I take my platform shoes off and dance in my bare feet.

Substituting the term mode for that of style, we may summarize as follows: a mode is a set of constraints which guarantees the realization of an invariant condition over variable instances; such sets of constraints may range from temporary to permanent and from flexible to unchangeable; two or more such sets of constraints may operate simultaneously if certain conditions prevail; generally, the organization of modes is coalitional.

In terms of the preceding, we may approach the question of how mode in psychology is to be understood by asking: What constraints are operating when an occasion of perception--a perceptual condition--is labeled as an instance of this or that mode? Ideally, we seek to identify those constraints that are both necessary and sufficient for applying a mode label. As a rough strategy, we can ask initially what constraints are necessary and then inquire whether they are also sufficient.

Reference has been made in this volume to a speech mode and a nonspeech mode, and, in the case of vision, to a focal mode and an ambient mode. It is roughly apparent that the constraints governing the information available to a perceiver (that is, what there is available for the animal to perceive) are necessary for defining a given mode. However, it is also roughly apparent that

those constraints are not sufficient (in and of themselves) for the application of a unique mode label in each particular situation where those constraints occur. Indeed, we argue that while the set of constraints corresponding to the set of answers to the question "what is the animal perceiving?" is necessary for the application of the label for a given mode, it is not sufficient. As a case in point, a musical sequence is easily recognized as such even when the notes of a melody are presented as a speech signal. It has been shown that when the fundamental frequency of a melodic line is inflected on a high quality synthesized syllable (tea), indices for a "nonspeech mode" are obtained even in the presence of overall conditions of stimulation normally associated with a "speech mode" (Darwin, 1969).

Perhaps we should look at the set of constraints governing how information is processed as well as at the constraints on what is processed. In this way, we circumvent the problems caused by attempting to define mode strictly in terms of informational constraints. For example, it has been shown that indices for a nonspeech mode can be obtained with a natural speech stimulus if the perceptual task is a nonlinguistic one (Haggard and Parkinson, 1971). Apparently, a given input is processed in a different way when the nature of the perceiver's task changes. A less equivocal example is provided by the following experiment. When "O" is embedded in a list of digits it can be found more rapidly if the observer is told that he or she is looking for a letter, than if he or she is told that the target is a digit. Conversely, when "O" is a member of a list of letters, latency of search is considerably shorter if one is looking for a digit zero than if one is looking for the letter "oh" (Jonides and Gleitman, 1972).

We see from the above examples that the set of constraints governing how information is processed is by necessity linked with the intent of the perceiver [the epistemic who, as defined by Shaw and McIntyre (1974)], as well as with what information exists in the surrounding medium. An illustration of the coimplicative relations among the what, the how, and the who of perception is provided by the hermit crab's "attitudes" toward a sea anemone. The description of these attitudes is due to von Uexküll (1957). To preface, let us identify the what of perception as the valences (see Gibson, 1966) specified in the ambient optic array structured by the sea anemone; the how of perception as the exploratory and performatory measures taken by the crab in detecting and exploiting the different uses of the sea anemone; and the who of perception as the intents of the crab. In the first case, the hermit crab has been robbed of the actinians which it normally carries on its shell. These actinians serve to protect the crab from its enemy, the cuttlefish. In this case, the crab is described as assuming a "defense tone," and it plants the sea anemone on its shell. In the second case, the shell has been taken from the hermit crab, and the crab attempts often unsuccessfully to crawl into the sea anemone, the crab having assumed a "dwelling tone." Finally, the crab who has been left to starve for some time, assumes a "feeding tone" and proceeds to devour the sea anemone. Thus, if "defense," "dwelling," and "feeding," are mode labels it would seem that answers to each of the what, how, and who questions are necessary for the application of one of the mode labels, and further, that answers to all three questions are sufficient for the application of a unique "mode label" in each particular situation.

In these notes we have attempted, in a most elementary and approximate manner, to sketch the metatheory of modes. To this end we pursued the general

concept of style, teasing from it several principles that we hoped might prove useful to the understanding of the more specific concept of mode in perceptual theory. Of these principles, the most fundamental equates mode with a set of constraints. We were motivated to ask whether, in defining a mode, the information for perception exhausted all the constraints, or whether the information for perception together with the algorithms for its analysis exhausted all the constraints. Our tentative answer to both of these questions is no. Unfortunately, that which appears to provide the full complement of constraints defining a mode is not something that we understand very well at all--namely, the relation among the what, the how, and the who of perception. It is our hunch that an appreciation of the aforementioned relation is the proper departure point for a rigorous analysis of modes of perceiving.

REFERENCES

Anderson, J. and G. H. Bower. (1973) Human Associative Memory. (New York: Academic Press).

Bernstein, N. (1967) The Coordination and Regulation of Movements. (London: Pergamon Press).

Chistovitch, L. A. (1961) Classification of rapidly repeated speech sounds. Soviet Physics and Acoustics 6, 393-398.

Darwin, C. J. (1969) Auditory perception and cerebral dominance. Unpublished Ph.D. thesis, University of Cambridge.

Gelfand, I. N., M. S. Gurfinkel, M. L. Tsetlin, and M. L. Shik. (1971) Some problems in the analysis of movements. In Models of the Structural-Functional Organization of Certain Biological Systems, ed. by I. M. Gelfand, V. S. Gurfinkel, S. V. Fomin, and M. L. Tstelin. (Cambridge, Mass: MIT Press), pp. 329-345.

Gibson, J. J. (1950) The Perception of the Visual World. (Boston: Houghton Mifflin).

Gibson, J. J. (1966) The Senses Considered as Perceptual Systems. (Boston: Houghton Mifflin).

Gibson, J. J. and M. Radner. (1937) Adaptation, after-effect and contrast in the perception of tilted lines I. Quantitative studies. J. Exp. Psychol. 20, 453-467.

Gross, Y., R. Webb, and R. Melzack. (1974) Central and peripheral contributions to localization of body parts: Evidence for the central body schema. Exp. Neurol. 44, 346-362.

Haber, R. N. (1971) Where are the visions in visual perception. In Imagery, ed. by S. Segal. (New York: Academic Press), pp. 36-48.

Haggard, M. P. and A. M. Parkinson. (1971) Stimulus and task factors as determinants of ear advantages. Quart. J. Exp. Psychol. 23, 168-177.

Halwes, T. and B. Wire. (1974) A possible solution to the pattern recognition problem in the speech modality. In Cognition and the Symbolic Processes, ed. by W. Weimar and D. Palermo. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.), pp. 385-388.

Hoy, R. R. and R. C. Paul. (1973) Genetic control of song specificity in crickets. Science 180, 82-83.

Johansson, G. (1974) Projective transformations as determining visual space perception. In Perception: Essays in Honor of J. J. Gibson, ed. by R. B. MacLeod and H. L. Pick. (Ithaca, N. Y.: Cornell University Press), pp. 117-140.

Jonides, J. and H. Gleitman. (1972) A conceptual category effect in visual search: O as a letter or a digit. Percept. Psychophys. 12, 457-460.
Kelso, J. A. S., E. Cook, M. E. Olson, and W. Epstein. (1976) Allocation of attention and the locus of adaptation to displaced vision. J. Exp. Psychol.: Human Perception and Performance 1, 237-245.
Lappin, J. S., H. H. Bell, O. J. Harm, and B. Kottas. (1975) On the relation between time and space in visual discrimination and velocity. J. Exp. Psychol.: Human Perception and Performance 1, 383-394.
Liberman, A. M. (1975) How abstract must a motor theory of speech perception be? Haskins Laboratories Status Report on Speech Research SR-44, 1-16.
Luce, D. A. and M. Clark. (1965) Duration of attack transients on nonpercussive orchestral instruments. J. Audio Engin. Soc. 13, 194-199.
Luce, D. and M. Clark. (1967) Physical correlates of brass instrument tones. J. Acoust. Soc. Am. 42, 1232-1243.
Neisser, U. (1967) Cognitive Psychology. (New York: Appleton-Century-Crofts).
Pfafflin, S. M. and M. V. Matthews. (1966) Detection of auditory signals in reproducible noise. J. Acoust. Soc. Am. 39, 340-345.
Pittenger, J. B. and R. E. Shaw. (1975) Aging faces as viscal-elastic events: Implications for a theory of nonrigid shape perception. J. Exp. Psychol.: Human Perception and Performance 1, 374-382.
Posner, M. I., M. J. Nissen, and R. M. Kline. (1976) Visual dominance: An information-processing account of its origins and significance. Psychol. Rev. 83, 157-170.
Rosenbaum, D. A. (1975) Perception and extrapolation of velocity and acceleration. J. Exp. Psychol.: Human Perception and Performance 1, 395-403.
Saldanha, E. L. and J. F. Corso. (1964) Timbre cues and the identification of musical instruments. J. Acoust. Soc. Am. 36, 2021-2026.
Schubert, E. D. (1975) The role of auditory perception in language processing. In Reading, Perception and Language, ed. by D. D. Duane and M. B. Rawson. (Baltimore, Md.: York Press), pp. 97-130.
Shaw, R. E. and M. MacIntyre. (1974) Algoristic foundations to cognitive psychology. In Cognition and the Symbolic Processes, ed. by W. Weimar and D. Palermo. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.), pp. 305-362.
Turvey, M. T. (1974) Constructive theory, perceptual systems and tacit knowledge. In Cognition and the Symbolic Processes, ed. by W. Weimar and D. Palermo. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.), pp. 165-180.
Turvey, M. T. (1975) Perspectives in vision: Conception or perception. In Reading, Perception and Language, ed. by D. D. Duane and M. B. Rawson. (Baltimore, Md.: York Press), 131-194.
Turvey, M. T. (1976) Preliminaries to a theory of action with reference to vision. In Perceiving, Acting and Knowing: Toward an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).
von Uexkull, J. (1957) A stroll through the worlds of animals and men. In Instinctive Behavior, ed. by C. H. Schiller. (New York: International Universities Press).
Verbrugge, R. R. (1975) Perceiving invariants at the invitation of metaphor. Paper presented at the meeting of the American Psychological Association, Chicago, August.
Webster, J. C., M. M. Woodhead, and A. Carpenter. (1970) A perceptual constancy in complex sound identification. Brit. J. Psychol. 61, 481-489.
White, B. W., F. S. Saunders, L. Scadden, P. Bach-Y-Rita, and C. C. Collins. (1970) Seeing with the skin. Percept. Psychophys. 7, 23-27.

Discrimination of Intensity Differences Carried on Formant Transitions Varying in Extent and Duration*

James E. Cutting[+] and Michael F. Dorman[++]

ABSTRACT

Dorman (1974) found that small intensity differences carried on the initial portions of consonant-vowel syllables were not discriminable. Similar differences carried on steady-state vowels and on isolated formant transitions, however, were readily discriminable. He interpreted the difference between the first and latter conditions as a phonetic effect. Using sine-wave analogs to Dorman's stimuli, Pastore, Ahroon, Wolz, Puleo, and Berger (1975) found similar results. They concluded that the effect is not phonetic, and that it is attributable to simple backward masking. The present studies observed the discriminability of intensity differences carried on formant transitions varying in extent and duration. Results support the conclusion of Pastore et al. (1975) to the extent that the effect is clearly not phonetic. However, these results and others suggest that simple peripheral backward masking is not a likely cause; instead, recognition masking may be involved. Moreover, the finding that phonetic-like processes occur elsewhere in audition does not necessarily impugn the existence of a speech processor; phonemic and phonological processes remain, as yet, unmatched.

Perhaps the most impressive characteristic of speech perception is the efficiency of information reduction (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). The speed and ease of phonemic segmentation is reflected in the rapid transformation of a 40,000 bit/sec acoustic signal into a 40 bit/sec phoneme string (Liberman, Mattingly, and Turvey, 1972), suitable for considerably further savings by conversion into higher-order, meaningful linguistic elements. One empirical manifestation of this process is categorical perception, a

---

*To appear in Perception and Psychophysics.

[+]Also Wesleyan University, Middletown, Conn.

[++]Also Arizona State University, Tempe; Herbert H. Lehman College of The City University of New York; and The Graduate School and University Center of The City University of New York.

Acknowledgment: This research was supported by a National Institute of Child Health and Human Development grant (HD-01994) to Haskins Laboratories and a seed grant from Wesleyan University to the first author. We thank R. E. Pastore for his helpful comments.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

phenomenon in which phonetic properties of a syllable are rapidly extracted and separated from the acoustic waveform. In a discrimination task, acoustically different stop consonants that are labeled the same are typically perceived to be identical. Stops labeled as different, on the other hand, even though they may differ physically by the same amount, are readily perceived to be dissimilar (Liberman, Harris, Hoffman, and Griffith, 1957; Mattingly, Liberman, Syrdal, and Halwes, 1971; Pisoni, 1971, 1973). For example, acoustic information about trajectories of formant transitions--information that contributes directly to the phonemic percept--cannot be retrieved readily from sensory memory.

Dorman (1974) found that phonemically irrelevant acoustic information also cannot be retrieved from sensory memory. He found that intensity differences carried on formant transitions of consonant-vowel (CV) syllables were largely undetectable. However, the same differences were eminently detectable when carried on steady-state vowels or on formant transitions isolated outside the syllable context. It appears that information-reduction mechanisms relevant for speech do not distinguish between phonemically relevant and irrelevant information at this level. This is as it should be. Liberman et al. (1972:323), for example, suggest "that the distinction between speech and nonspeech is not made at some early stage on the basis of general acoustic characteristics," but rather after many speech-relevant processors have been polled for proper speechlike features. In other words, both phonemically relevant and irrelevant auditory signals share some, probably many, early processing stages. This view is supported by the results of a recent study (Pastore et al., 1975) which show that intensity differences carried on frequency ramps before steady-state sine waves are as difficult to discriminate as intensity differences carried on formant transitions of CV syllables.

Dorman's (1974) earlier account of the inability to discriminate intensity differences on formant transitions is incorrect. He noted the similarity between the poor discriminability of intensity differences on formant transitions and poor discriminability of formant frequency within a phoneme category. Both effects were attributed to the uniquely categorical, linguistic processing accorded stop consonants: "After the acoustic cues for stop consonants have been recoded into a phonetic [categorical] representation, all of the acoustic information is stored in a relatively inaccessible short-term auditory memory" (Dorman, 1974:86, italics added). The effect, however, is not necessarily the result of linguistic coding, since categorical perception occurs in several nonlinguistic domains (Cutting and Rosner, 1974; Cutting, in press; Cutting, Rosner, and Foard, in press; Miller, Wier, Pastore, Kelly, and Dooling, in press; see also Locke and Kellar, 1973; Lane, 1965, 1967). Moreover, it does not appear contingent on categorical perception or phonemic processing at all, since the stimuli of Pastore et al. (1975) are likely neither to be perceived categorically (see Pisoni, 1971:Experiment II) nor phonemically (see Cutting, 1974: Experiment III).

Pastore et al. (1975) noted another problem with Dorman's account of his results. They suggested that to change the carrier waveform from a CV syllable to a steady-state vowel syllable, as Dorman did, is to change the task at the same time from one of simple backward-masking detection to one of pedestal detection (see Tanner, 1958; Tanner and Sorkin, 1972). We concur that formal parallels are unmistakable between pedestal detection and the detection of intensity differences carried at the beginning of the vowels. Thus, Dorman's steady-state vowel control does not appear to eliminate simple backward masking

as a cause for poor discriminability of intensity differences carried on CV syllables: pedestal detection experiments appear to be a special kind of masking experiment.

Several important questions about masking arise. First, how might backward masking function in speech perception? For example, if phonemically irrelevant information can be masked at an auditory level, why is it that phonemically relevant information is not masked as well, rendering speech incomprehensible? Second, Pastore et al. (1975) do not suggest a particular relationship between backward-masking detection and pedestal detection tasks. For example, do the two tasks differ in degree or in kind? Should we expect intermediate detectability for speech syllables whose transitions are midway between those of a CV and a steady-state vowel? Or should we expect that all syllables with transitions, regardless of their extent or duration, would inhibit detection of intensity differences since only the steady-state vowel stimulus meets the requisite of having a true pedestal? Experiment I explores the detectability of intensity differences carried on the formant transitions of these intermediate stimuli. The discussion and Experiment II, which follows thereafter, explore the plausibility of simple backward masking versus backward recognition masking as a cause of our results.

EXPERIMENT I

Method

Two arrays of three-formant speech stimuli were generated on the Haskins Laboratories parallel-resonance synthesizer. One array consisted of six items differeing in the extent of formant transitions, with all items identifiable as /ba/ or /a/; the other array consisted of five items differing in duration of formant transitions, with all items identifiable as /ba/ or /bwa/. All stimuli were 300 msec in duration and had a flat pitch contour of 100 Hz. Steady-state /a/ resonances for both arrays centered on 769, 1232, and 2525 Hz for first, second, and third formants, respectively. The six-item /ba/-to-/a/ array contained stimuli whose formant transitions were 60 msec in duration. Transitions decreased in extent by equal increments over this array, in corresponding fashion for all three formants. Stimulus 1 (the prototype /ba/) transitions began at 513, 846, and 2180 Hz for the three formants, respectively; and Stimulus 6 (the steady-state vowel /a/) began with formants of 769, 1232, and 2525 Hz. Intermediate stimuli had intermediate starting frequencies for each formant. The five-item /ba/-to-/bwa/ array contained stimuli whose formant transitions always began at 513, 856, and 2180 Hz, but whose transition durations lasted 40, 60, 80, 100, and 120 msec for Stimuli 1 through 5, respectively. The endpoint stimuli of both arrays are shown schematically in the top panels of Figure 1. Stimuli were digitized and stored on disc file using the pulse code modulation system at Haskins. Further stimulus alteration consisted of decreasing the initial portions of all stimuli by 0, 4, and 8 dB. For the /ba/-to-/a/ array the decreased portion was always 60 msec in duration (like that used by Dorman, 1974), and for the /ba/-to-/bwa/ array it was held to the duration of the formant transitions: 40, 60, 80, 100, or 120 msec. In this manner each of the eleven stimuli was synthesized in three renditions. For an indication of overall amplitude envelope shape of these stimuli see Dorman (1974:Figure 2).

Four diotic stimulus sequences were recorded on audio tape; one identification sequence consisted of random orders of the standard (0 dB) stimuli, 48 and

40 items respectively, for the extent and duration stimuli. Each item in each array appeared eight times. The interval between each item in both sequences was 3 seconds. Listeners wrote down BAH or AH, and BAH or BWAH to identify members of the arrays. Discrimination sequences consisted of 90 and 75 AX trials for the /ba/-to-/a/ and /ba/-to-/bwa/ arrays: (6 and 5 stimuli in the arrays, respectively) × (3 intensities to be discriminated: 0-, 4-, and 8-dB differences between members of the AX pair) × (5 observations per pair). Each discrimination trial began with a 100 msec 1000 Hz warning tone, followed by 500 msec of silence, followed by Stimulus A, another 500 msec silent interval, and Stimulus X. Stimulus A was always the standard stimulus, whereas Stimulus X had formant structures identical to Stimulus A but with its initial portions attenuated by 0, 4, or 8 dB. There was a 3.5 second interval between the offset of Stimulus X and the onset of the warning tone for the subsequent trial. Listeners wrote down S for same if they thought the AX items were identical, and D for different if they were not.

Thirteen Wesleyan University students listened as a group to the four sequences as part of a course project. All were native American English speakers with little experience at listening to synthetic speech. They listened to the audio tapes played on a Crown CX-822 tape recorder, broadcast in a quiet room over an Ampex AA-620 loudspeaker. All listeners sat between 8 and 18 feet from the loudspeaker, which for the standard item delivered approximately 75 dB SPL.

Results

All results are shown in the lower panels of Figure 1. In the left-hand panel, identification functions for /ba/ and /a/ are superimposed on two discrimination functions, those for judgments of 4- and 8-dB differences. Stimuli 1 through 5 were consistently identified as /ba/, and only Stimulus 6 was identified consistently as /a/. The identification "boundary" appears to be located near Stimulus 5, where the two complementary identification functions cross. Discrimination functions (percent correct discrimination of intensity differences at each comparison) show that 8-dB judgments were consistently more successful than the 4-dB judgments [$F(1,144) = 65.1$, $p < .001$]. There was no interaction of intensity with stimulus location along the array; therefore, collapsing across the two intensity differences, there was a significant increase in discriminability as the formant transitions decreased in extent [$F(5,144) = 3.15$, $p < .025$]. Moreover, a trend test (Winer, 1962:132) proved this increase to be linear [$F(1,64) = 49.3$, $p < .001$] with no significant quadratic, cubic, or other higher-order components. The D responses on AA trials (those with 0-dB difference) were scored as false alarms, and the detectability of the intensity differences was then assessed independent of possible response bias. A generally linear increase was obtained: the d' scores for 4-dB judgments were .44, .60, .84, 1.10, 1.08, and 1.15; and those for 8-dB judgments were 1.61, 1.56, 1.89, 2.20, 1.92, and 2.09, respectively, for the six different transition extents.

Results for the /ba/-to-/bwa/ array are shown in the lower right-hand panel of Figure 1. Identification functions are somewhat unimpressive: only Stimulus 1 was consistently identified as /ba/ and, where as Stimuli 3 through 5 were primarily identified as /bwa/, none was so identified with a consistency exceeding 72 percent. The identification "boundary," if one can be said to exist, appears to be near Stimulus 2. The pattern of discrimination results followed very closely that for the previous set of stimuli. Again, 8-dB judgments were superior to 4-dB judgments [$F(1,120) = 58.9$, $p < .001$]; discriminability

FIGURE 1

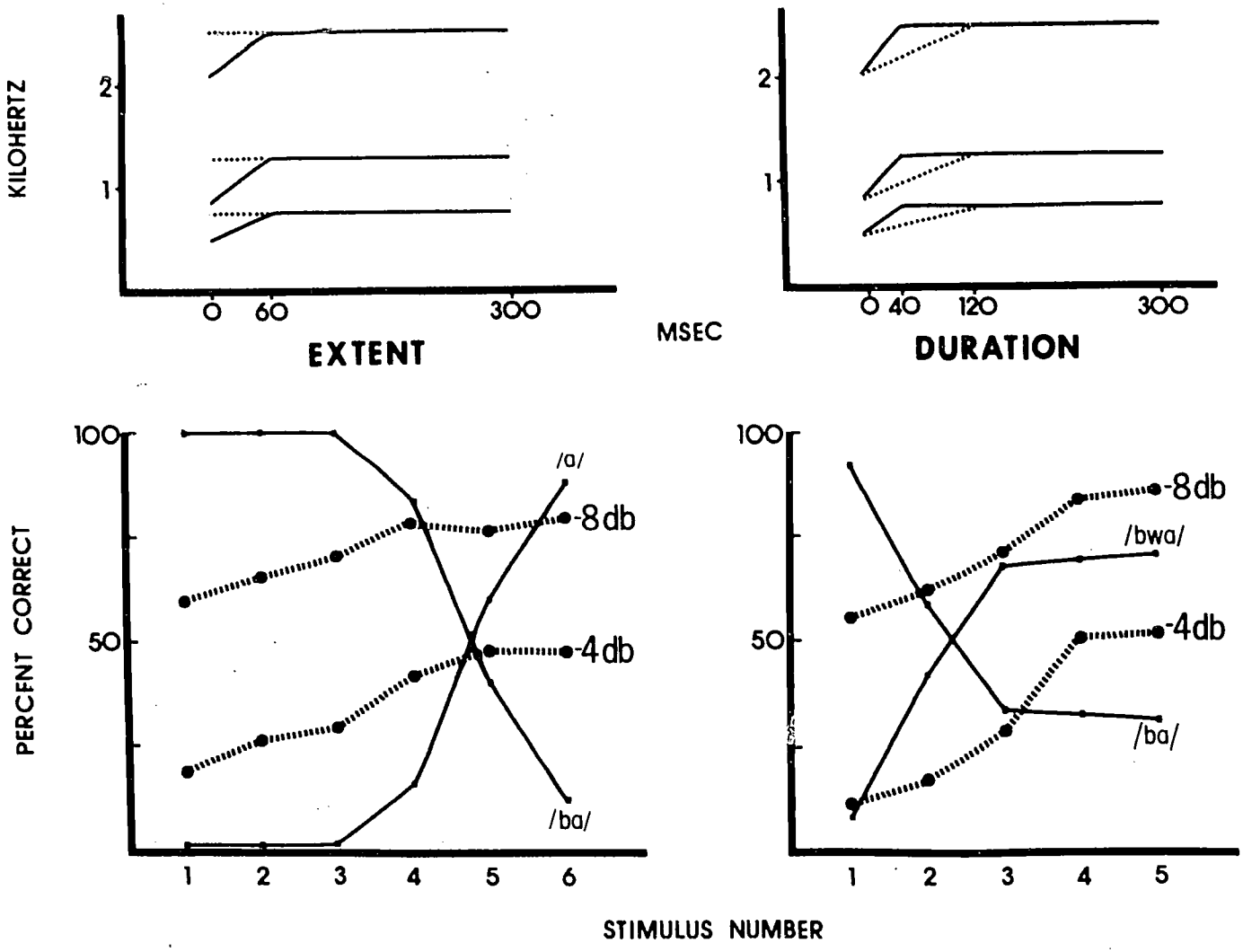


Figure 1: In upper panels, schematic spectrograms indicating the arrays of stimuli whose formant transitions differ in extent or in duration. Actual bandwidth of the formants is 60, 90, and 120 Hz for the first, second, and third formants respectively; very narrow bands are shown only for purposes of clarity in repeating acoustic variation. Respective identification and discrimination results are superimposed on one another in the lower panels.

increased across the stimulus array [$F(4,120) = 7.9$, $p < .001$], and that increase was linear [$F(1,51) = 73.0$, $p < .001$] without significant higher-order components. This linear pattern was repeated in terms of detectability: 4-dB d' scores were .24, .45, .84, 1.26, and 1.28; and 8-dB scores were 1.60, 1.70, 1.96, 2.12, and 2.22, respectively, for the five different transition durations.

Discussion

Two aspects of our results support the primary conclusion of Pastore et al. (1975): the inability to detect intensity differences carried on the formant transitions of stop consonants is a psychoacoustic rather than phonetic effect. First, there is no abrupt increase in detectability of intensity differences as the stimulus arrays change from /ba/ to /a/ for those stimuli differing in extent of transitions, and from /ba/ to /bwa/ for those differing in duration of transitions. If the availability of acoustic information were somehow inhibited by the processing of the highly encoded stop consonant in particular, one would have expected a quantal increase in discriminability in the /ba/-to-/a/ array at about Stimulus 5. Clearly none exists, and thus the effect cannot be directly related to categorical perception. Studdert-Kennedy, Liberman, Harris, and Cooper (1970), among others, would predict discontinuities in the discrimination functions at this point if the phenomenon were related to categorical perception. Second, the increase in discriminability is linear for both arrays. Such linear increases are also at variance with the nonlinear, categoricallike processes associated with phonetic perception.

Our results demonstrate interaction between rate of frequency change and the discrimination of intensity change on formant transitions. That is, for the /ba/-to-/a/ array in particular, the less frequency change that occurs, the more perceptible the intensity differences become. Thus, frequency and intensity appear to be yoked in the percept and contribute in an interactive manner to the traces available to short-term auditory memory. Of course, as Pastore et al. (1975) admit, finding a psychoacoustic basis for the inability to detect such intensity differences here, does not rule out the possibility that a similar outcome could result from processes occurring at other levels. In visual masking, for example, Turvey (1973) demonstrated that when viewers were unable to report a target, the contour information may have been masked peripherally or centrally. At both levels, the effect is similar: viewers are unable to identify the target.

A Second Look at Simple Backward Masking

The secondary conclusion of Pastore et al. (1975), that these results are caused by simple backward masking, is more suspect. While they do not mention these issues, the type of phenomenon they refer to appears to be threshold masking rather than recognition masking (Massaro, 1973, 1975). The locus of the backward masking appears to be peripheral not central, and it appears to result from target-mask integration, not interruption (see Kahneman, 1968; and Turvey, 1973, for arguments with respect to vision). From this view of masking one might not expect to find evidence in any experimental paradigm of the ability to detect 4 to 9 dB intensity differences carried on formant transitions. That is, this information would be buried in background noise considerably prior to the decision making process. There are several reasons to suspect, however, that the intensity information in the Dorman (1974) and present studies is not lost by simple, peripheral target-mask integration.

yllables, as opposed to those on steady-state vowels. Direct comparisons are ifficult: (a) since Dorman used attenuations of 7.5 and 9.0 dB, whereas we sed attenuations of 4 and 8 dB, (b) since Dorman used the carrier stimuli /bæ/ nd /æ/ whereas we used /ba/ and /a/, and (c) since Dorman's listeners heard his timuli through earphones, whereas we played them over a loudspeaker in a revererant room. Nevertheless, a striking trend can be seen when d' scores for his timuli are compared with those for Stimuli 1 and 6 from the /ba/-to-/a/ array.

In the present study, by mixing the CV and V stimuli together with several ntermediate items, the detectability of the intensity differences carried on he CV syllables increased considerably. It decreased, on the other hand, for hose differences carried on steady-state vowels. It would appear, then, that large proportion of the effect is attributable to context, not to masking. hat is, detectability varies according to previous experience and expectations ithin the experiment. The difference in detectability for intensity differnces in CV and V syllable changed from a standard score of more than 3.2 (for orman's 9-dB discriminations) to one of less than .5 (for our 8-dB discriminaions). Such a finding appears to be at variance with the hypothesized effect f simple peripheral masking, and suggests that: (a) the intensity information s available at some level of perceptual analysis and that (b) recognition maskng rather than threshold masking may be involved in the Dorman (1974) and astore et al. (1975) results.

A second avenue of reasoning comes from the many studies of categorical erception of stop consonants, and the fate of within-phoneme-category formant requency information. In ABX (Liberman et al., 1957), odd-ball (Mattingly t al., 1971), and AX (Pisoni, 1971, 1973) paradigms, the discrimination of freuency differences carried on formant transitions has been found to be categorcal--that is, the frequency difference in formant transitions within the same honemic category is discriminated at about chance, while the frequency differnce across categories is discriminated very easily. Despite essentially chance ithin-category preformance, frequency information is netiher masked in auditory rocessing nor lost in the auditory-to-phonetic transformation (see Barclay, 972; Pisoni and Lazarus, 1974). Pisoni and Tash (1974), for example, have nown that "same" reaction times (RTs) to physically different but phonemically identical stop consonants are slower than "same" RTs to physically identical top consonants. Thus, even though the discrimination response implies that the wo signals were perceived identically, and by inference that there was no

the items to be identified properly. Thus, the whole of the stimulus is necessary for perception, and this fact would suggest a holistic mode of processing.

It may be more accurate to account for the present data based on the stimuli's acoustic nature rather than in terms of processing strategy. One particurly important acoustic property of the plucked and bowed stimuli in terms of

distinguishing information left about formant trajectories, the RTs indicate that at some level in the nervous system the information was present. We would expect a similar outcome in an RT analysis with the signals used in the present study. That is, we suspect that the "same" RTs to the physically different (4 dB) signals would be slower than the "same" RTs in the physically identical (0 dB) condition. Experiment II was conducted to test this hypothesis.

EXPERIMENT II

Method

Two stimuli were selected from Experiment I: Stimulus 1 (/ba/) and Stimulus 6 (/a/) from the array with transitions differing in extent. Both were generated in three renditions: the initial 60 msec was attenuated by 0, 4, and 8 dB. One discrimination sequence was assembled exactly as in Experiment I. It contained 120 AX trials: (2 stimuli) × (3 intensities to be discriminated) × (20 observations per item). Listeners pressed, as rapidly as possible, one of two telegraph keys to indicate whether the two items within a trial were the same or different. Reaction times were fed on line into a PDP-11 computer for analysis. They were measured from the onset of the second item to the onset of the keypress.

Four students and staff members at Haskins volunteered for the experiment. All were naive to the purposes of the task. They listened, in groups of two, to stimuli reproduced on an Ampex AG-500 tape recorder and transmitted binaurally through a listening station to Telephonics headphones (TDH-39).

Results and Discussion

The most important reaction time results are shown in Table 2--mean RTs for "same" responses for the 0-dB and 4-dB discriminations. Few "same" responses were made for 8-dB trials, so they are not included. The difference in

TABLE 2: Mean reaction time (and number) of "same" responses to intensity differences carried on the initial 60 msec of CV syllables. Maximum number of trials per cell is 20.

	Intensity difference		
Listener	0 dB	4 dB	
T.B.	659 (19)	942 (9)	$z = 3.02$ $p < .002$, one-tailed
P.B.	609 (16)	694 (11)	$z = 1.19$ $p < .12$
W.F.	612 (17)	818 (8)	$z = 1.92$, $p < .03$
H.S.	669 (18)	1005 (4)	$z = 2.89$ $p < .002$
Mean of means	637	865	

RTs for the two conditions ranged from 85 and 336 msec for the four listeners; the results for three listeners were statistically robust by a Mann-Whitney U test on individual reaction times, while those for the other listener approached significance. (U scores were converted into standard z units, as shown in Table 2.) These results clearly indicate that intensity information not

discriminated on a particular trial is not masked in absolute terms, but is represented in some form throughout the information processing system. The representations for the two stimuli within a trial that differ in the amplitude of their onsets of 4 dB are more difficult to match than are those pairs with the same onset amplitude. These results are congruent with those of Pisoni and Tash (1974) using speech syllables, and with prior results of Emmerich, Gray, Watson, and Tanis (1972) using nonspeech stimuli.

CONCLUDING DISCUSSION

The results of the present studies suggest, first, that the relationship between pedestal-detection and recognition-masking experiments is one of degree rather than kind. There is no discontinuity between the two. Second, the results support the primary conclusion of Pastore et al. (1975): the relative inability to discriminate intensity differences carried on the formant transitions of CV syllables, as compared to those carried on the initial portions of steady-state vowel syllables is an effect that is psychoacoustic rather than phonetic.

Third, our results demonstrate differences between types of masking. Pastore et al. (1975) appear to attribute the inability to discriminate differences carried on formant transitions to simple backward masking. Simple masking, according to Licklider (1951), is the opposite of analysis. Information is simply not processed, and the implication is that masked information is irretrievably buried in background noise. However, results of Experiment II show that phonemically irrelevant acoustic information remains accessible to the listener in some form. This suggests that recognition masking is the phenomenon involved in the Dorman (1974) and Pastore et al. (1975) experiments and Experiment I of the present investigation. Moreover, recognition masking is selective in its effect on auditory versus phonetic memory codes. Fourth, a comparison of the detectability scores from Dorman's (1974) study and those from Experiment I also suggest that this information is not masked absolutely even in recognition terms, but may be used or unused as a function of context in an experimental session.

On the "Speech Processor"

Pastore et al. (1975) suggest that a speech processor is an unneeded construct to account for results in the AX discrimination task. We agree. Nevertheless, whereas our results support this position, we must not ignore the necessity for some such device at some level. The level at which any device is specific to speech is currently a crucial question. Several effects thought to demonstrate the psychological reality of phonetic processing (Wood, 1975:16) have been found to occur in purely auditory domains (Cutting and Rosner, 1974; Blechner, Day, and Cutting, 1976; Pastore, Ahroon, Puleo, Crimmins, Galowner, and Berger, 1976; Cutting, in press; Cutting et al., in press; Miller et al., in press). Thus the mechanism that extracts phonetic information from the speech signal may be the same device that is used elsewhere, for example, in the processing of musiclike sounds. In other words, phoneticlike processing may not be speech-specific processing. Yet these recent findings cut only into the lowest tier of the speech-language hierarchy--that of phonetic processing. The perception of different allophones of the same phoneme as being the same--such as the /p/s in pit, spit, and tip--and the parsing of syllables from a continuous speech stream seem to be processes without nonspeech analogs. Unless (or until)

analogs are found, the notion of a speech processor is not impugned by the existence of phoneticlike processes elsewhere in audition.

REFERENCES

Barclay, J. R. (1972) Noncategorical perception of a voiced stop: A replication. Percept. Psychophys. 11, 269-273.

Blechner, M. J., R. S. Day, and J. E. Cutting. (1976) Processing two dimensions of nonspeech stimuli: The auditory-phonetic distinction reconsidered. J. Exp. Psychol.: Human Perception and Performance 2, 257-266.

Cutting, J. E. (1974) Two left-hemisphere mechanisms in speech perception. Percept. Psychophys. 16, 601-602.

Cutting, J. E. (in press) The magical number two and the natural categories of speech and music. In Tutorial Essays in Psychology, ed. by N. S. Sutherland (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Cutting, J. E. and B. S. Rosner. (1974) Categories and boundaries in speech and music. Percept. Psychophys. 16, 564-570.

Cutting, J. E., B. S. Rosner, and C. F. Foard. (in press) Perceptual categories for musiclike sounds: Implications for theories of speech perception. Quart. J. Exp. Psychol. 23.

Dorman, M. F. (1974) Discrimination of intensity differences in formant transitions in and out of syllable context. Percept. Psychophys. 16, 64-86.

Emmerich, D. S., J. L. Gray, C. S. Watson, and D. C. Tanis. (1972) Response latency, confidence, and ROCs in auditory signal detection. Percept. Psychophys. 11, 65-72.

Kahneman, D. (1968) Method, findings, and theory in studies of visual masking. Psychol. Bull. 70, 404-426.

Lane, H. (1965) Motor theory of speech perception: A critical review. Psychol. Rev. 72, 275-309.

Lane, H. (1967) A behaviorial basis for the polarity principle in linguistics. Language 43, 494-511.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 631-661.

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.

Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin (Washington, D.C.: Winston), pp. 307-334.

Licklider, J. C. R. (1951) Basic correlates of the auditory stimulus. In Handbook of Experimental Psychology, ed. by S. S. Stevens (New York: Wiley), pp. 985-1039.

Locke, S. and L. Kellar. (1973) Categorical perception in a nonlinguistic mode. Cortex 9, 355-369.

Massaro, D. W. (1973) A comparison of forward versus backward recognition masking. J. Exp. Psychol. 100, 434-436.

Massaro, D. W. (1975) Backward recognition masking. J. Acoust. Soc. Am. 58, 1059-1065.

Mattingly, I. G., A. M. Liberman, A. Syrdal, and T. G. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.

Miller, J. D., C. C. Wier, R. E. Pastore, W. M. Kelly, and R. M. Dooling. (in press) Discrimination and labeling of noise-buzz sequences with varying noise lead times: An example of categorical perception. J. Acoust. Soc. Am.

Pastore, R. E., W. A. Ahroon, J. S. Puleo, D. B. Crimmins, L. Golowner, and R. S. Berger. (1976) Processing interaction between dimensions of non-phonetic auditory signals. J. Exp. Psychol.: Human Perception and Performance 2, 267-276.

Pastore, R. E., W. A. Ahroon, J. P. Wolz, J. S. Puleo, and R. S. Berger. (1975) Discrimination of intensity differences on formant-like transitions. Percept. Psychophys. 18, 224-226.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Ph.D. thesis, University of Michigan. [Published in Dissertation Abstracts International, 1972, 32, 6693B (University Microfilms no. 72-14, 964).]

Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253-260.

Pisoni, D. B. and J. H. Lazarus. (1974) Cateogrical and noncategorical modes of speech perception along the voicing continuum. J. Acoust. Soc. Am. 55, 328-333.

Pisoni, D. B. and J. Tash. (1974) Reaction times to comparison within and across phonetic boundaries. Percept. Psychophys. 15, 285-290.

Studdert-Kennedy, M., A. M. Liberman, K. S. Harris, and F. S. Cooper. (1970) Motor theory of speech perception: A reply to Lane's critical review. Psychol. Rev. 77, 234-249.

Tanner, W. P. (1958) What is masking? J. Acoust. Soc. Am. 30, 919-921.

Tanner, W. P. and R. D. Sorkin. (1972) The theory of signal detectability. In Foundations of Modern Auditory Theory, vol. 2, ed. by J. V. Tobias (New York: Academic Press), pp. 63-98.

Turvey, M. T. (1973) On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. Psychol. Rev. 80, 1-52.

Winer, B. J. (1962) Statistical Principles in Experimental Design (New York: McGraw-Hill).

Wood, C. C. (1975) Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analyses. J. Exp. Psychol.: Human Perception and Performance 1, 3-20.

Discrimination Functions Predicted from Categories in Speech and Music*

James E. Cutting[+] and Burton S. Rosner[++]

ABSTRACT

Cutting and Rosner (1974) reported that sawtooth waves varying in rise time and identifiable as either plucked or bowed are perceived categorically according to the strictest criteria. The predicted discrimination functions in that paper were incorrectly calculated. This note gives correct formulae and the predictions that they yield. The original finding is unchanged.

Sawtooth waves differing only in rise time are identifiable as plucked or bowed notes from a stringed instrument. We previously reported (Cutting and Rosner, 1974) that these nonlinguistic sounds are perceived categorically. We also synthesized a continuum of speech sounds by varying only rise time. Listeners identified these sounds as /tʃa/ or /ʃa/ as in CHOP or SHOP, respectively, and perceived them categorically as well.

Our criteria for categorical perception were those suggested by Studdert-Kennedy, Liberman, Harris, and Cooper, (1970): (a) "peaks" of high discriminability between stimuli in restricted regions along the dimension studied; (b) "troughs" of discrimination performance near chance in regions on either side of the peak; and (c) correspondence between discrimination peaks and troughs and the course of identification functions, with peaks occurring at identification boundaries and troughs occurring within each perceptual category. Categorical perception is therefore revealed by a particular combination of results from identification and discrimination tasks. This convergence between identification and discrimination is unusual; a listener generally can discriminate many more stimuli than he or she can identify absolutely (see, for example, Miller, 1956).

---

* To appear in Perception and Psychophysics.

+ Also Wesleyan University, Middletown, Conn.

++ University of Pennsylvania, Philadelphia, Pa.

Acknowledgement: We thank Neil Macmillan for pointing out our error to us and suggesting how it occurred.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

The correspondence between identification and discrimination can be tested quantitatively. Discrimination performance can be predicted from identification data by assuming that discrimination is no better than identification. To the extent that obtained and predicted discrimination scores do not differ significantly, categorical perception has occurred.

Our previous paper described such agreement between obtained and predicted discrimination scores for both the linguistic and musical sounds (Cutting and Rosner, 1974). Unfortunately, the predicted functions were derived through an incorrect formula. This note corrects that error.

To predict discrimination from identification of a two-category continuum, the correct formula for an ABX discrimination task is

$$P(c) = 1/2[1 + (p_1 - p_2)^2] \quad (1)$$

where P(c) is the probability of a correct discrimination, p_1 is the probability of assigning stimulus A to one of the categories, and p_2 is the probability of assigning stimulus B to that same category. The original formula for the three-category case published by investigators at the Haskins Laboratories (Liberman, Harris, Hoffman, and Griffith, 1957) was incorrect as printed; Pollack and Pisoni (1971) give proper formulae for both two- and three-category continua. We will refer to (1) as the Haskins prediction.

Typically, obtained discrimination functions, even for stop consonants, systematically exceed predicted functions by as many as ten percentage points at each comparison along the stimulus array. Thus, the strongest possible relationship between identification and discrimination is not realized (see also Barclay, 1972; Pisoni and Lazarus, 1974; and Pisoni and Tash, 1974). The discrepancy between obtained and predicted discrimination functions is even larger for more "continuously" perceived stimuli such as vowels (Pisoni, 1971, 1973, 1975). By further developing a model that Fujisaki and Kawashima (1970) formulated, Pisoni added a correction factor to prediction formulae such as (1). This factor is based on the asymptotic trough discrimination value; it raises the predicted functions by several percentage points, and it can be interpreted as measuring short-term auditory storage for differences between two stimuli identified alike. For a two-category continuum in an ABX task, the proper Fujisaki-Kawashima prediction formula is

$$P(c) = 1/2[(p_1 - p_2)^2 + p_1(1 - p_2) + p_2(1 - p_1)] + [p_1p_2 + (1 - p_1)(1 - p_2)]T \quad (2)$$

where P(c), p_1, and p_2 are the same as in (1) and T is the asymptotic trough value of the obtained discrimination function. If T = 0.50, (2) reduces to (1). Like the Haskins prediction formula, the Fujisaki-Kawashima formula has suffered the misfortune of appearing incorrectly in print (Pisoni, 1971:44; Pisoni, 1975:13).[1]

---

[1]Page numbers for Pisoni (1971) refer to a version published as a supplement to the Haskins Laboratories Status Report on Speech Research.

Using the correct formulae we have recomputed both the Haskins and the Fujisaki-Kawashima predictions for our data on discrimination of sawtooth waves and of affricate-fricative speech syllables. Predictions were made for each individual listener, then averaged functions were obtained from the individual functions, as Pisoni (1971) suggests.[2] Table 1 shows averaged obtained and predicted discrimination scores.

TABLE 1: Obtained and correctly predicted discrimination values for stimuli differing in rise time. The original predicted functions that appear in Cutting and Rosner (1974) are incorrect.

	Rise time comparison (msec)						
	0-20	10-30	20-40	30-50	40-60	50-70	60-80
Experiment 1							
Sawtooth wave stimuli							
Obtained	61	64	72	78	58	60	59
Haskins predicted	50	51	60	67	56	52	51
Fujisaki-Kawashima predicted	58	59	65	71	61	58	58
Speech stimuli							
Obtained	61	58	59	70	76	61	58
Haskins predicted	51	53	55	62	64	51	50
Fujisaki-Kawashima predicted	58	59	60	67	68	58	58
Experiment 2							
Sawtooth wave stimuli							
Obtained	61	55	66	72	47	50	53
Haskins predicted	50	50	66	73	54	51	50
Sine wave stimuli							
Obtained	54	49	56	68	56	58	53
Haskins predicted	50	51	63	68	54	54	53

The predicted functions in Table 1 are farther below the obtained functions than were those originally published [see Tables 1 and 2 in Cutting and Rosner (1974)]. Nevertheless, the discrepancies between predicted and obtained scores here are not marked. Goodness-of-fit measures calculated from individual-obtained and Haskins-predicted scores revealed no significant differences (see Pisoni, 1971:20), although the observations per comparison may be too few to make small differences statistically reliable. The fit between the data and the correct predictions still supports our prior conclusion that musical stimuli and affricate-fricative consonants differing in rise time are each perceived categorically. Subsequent experiments have provided confirmation; Cutting, Rosner, and Foard (1976) have demonstrated that the musical sounds are perceived as categorically as stop consonants in Pisoni's (1971, 1973) variable-interval AX discrimination task.

In summary, this note presents correct predicted discrimination functions for data previously published (Cutting and Rosner, 1974). The corrections leave the principal conclusion of that study unchanged: nonlinguistic and

[2]The trough value T was not stable for individual listeners; we assumed it to be 0.60 for all listeners for both sets of stimuli represented in Table 1.

linguistic stimuli synthesized with different rise times are perceived categorically. In addition, this note provides correct formulae for predicting discrimination functions. Several previous sources for the formulae are in error.

REFERENCES

Barclay, J. R. (1972) Noncategorical perception of a voiced stop: A replication. Percept. Psychophys. 11, 269-273.

Cutting, J. E. and B. R. Rosner. (1974) Categories and boundaries in speech and music. Percept. Psychophys. 16, 564-570.

Cutting, J. E., B. S. Rosner, and C. F. Foard. (in press) Perceptual categories for musiclike sounds: Implications for theories of speech perception. Quart. J. Exp. Psychol. 28.

Fujisaki, H. and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanisms. Annual Report of the Engineering Research Institute 29 (Tokyo), 207-214.

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.

Miller, G. A. (1956) The magical number seven, plus or minus two, or some limits on our capacity for processing information. Psychol. Rev. 63, 81-96.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Doctoral dissertation, University of Michigan. Dissertation Abstracts International 32 (1972), 6693B (University Microfilms 72-14, 964).

Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253-260.

Pisoni, D. B. (1975) Auditory short-term memory and vowel perception. Mem. Cog. 3, 7-18.

Pisoni, D. B. and J. H. Lazarus. (1974) Categorical and noncategorical modes of speech perception along the voicing continuum. J. Acoust. Soc. Am. 55, 328-333.

Pisoni, D. B. and J. Tash. (1974) Reaction times to comparisons within and across phonetic categories. Percept. Psychophys. 15, 285-290.

Pollack, I. and D. B. Pisoni. (1971) On the comparison between identification and discrimination tests in speech perception. Psychon. Sci. 24, 299-300.

Studdert-Kennedy, M., A. M. Liberman, K. S. Harris, and F. S. Cooper. (1970) Motor theory of speech perception: A reply to Lane's critical review. Psychol. Rev. 77, 234-249.

Right-Ear Advantage for Musical Stimuli Differing in Rise Time

Mark J. Blechner*

ABSTRACT

Nonspeech stimuli differing in rise time, which resemble the sounds of plucked or bowed violin strings, were presented monaurally with contralateral noise, and reaction times for stimulus identification were measured. Reaction times were 12.8 msec faster when the stimulus was presented to the right ear than to the left ear, suggesting left-hemisphere involvement in the processing of these stimuli. This finding, considered along with other studies using the same stimuli, suggests that a single psychological mechanism is involved in the processing of the plucked and bowed sounds and consonant-vowel stimuli. In addition, the data support the theory that the dominant cerebral hemisphere is specialized for the processing of temporal variation.

The distinction between auditory and phonetic processes in the human perception of sounds has been a topic of much debate in recent years. Phonetic processing implies a mode of perception unique to speech stimuli. It is characterized by the fact that there is no one-to-one relationship between the acoustic stimulus and percept, and that perception appears to be modulated by rules of linguistic rather than acoustic organization (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967).

Wood (1975) listed six experimental operations whose results have been thought to converge on the distinction between auditory and phonetic processes.[1] Three of these characteristic data patterns, however, have been found with a particular kind of nonspeech stimulus--sawtooth waves differing in rise time, which resemble the sound of a plucked or bowed violin string. The plucked and bowed sounds, like consonant-vowel (CV) syllables show: categorical perception

*Also Yale University, New Haven, Conn.

Acknowledgment: This research was supported by NIMH Training Grant PHS5T01MH05276-27 to Yale University and by NICHD HD-01994 to Haskins Laboratories. The author thanks James E. Cutting, Michael Studdert-Kennedy, and Andrea G. Levitt for helpful comments on an earlier draft of this paper, and Robert L. Plotz for assistance in running the experiment.

[1] Wood (1975) is cited here because he most clearly summarized the empirical evidence for the distinction between auditory and phonetic processes. However, the notion of special phonetic processing mechanisms was suggested considerably earlier by other researchers. See Studdert-Kennedy (1974) for a review of relevant research and a theoretical exposition of this viewpoint.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

(Cutting and Rosner, 1974), referred to by Wood as the phoneme-boundary effect; boundary shifts due to selective adaptation (Cutting, Rosner, and Foard, 1976); and asymmetric interference with redundancy gain in a speeded classification task (Blechner, Day, and Cutting, 1976). The remaining three experimental results cited by Wood are: right-ear advantage for identifying dichotically presented speech; right-ear advantage for reporting the temporal order of dichotic speech stimuli; and unilateral differences in average evoked potentials during the classification of linguistic and nonlinguistic dimensions. All three of these appear to reflect a single factor, that is, the lateralization of the cerebral hemispheres. It therefore seems quite pressing to determine which, if any, hemisphere is predominantly involved in the perception of plucked and bowed sounds, but so far data on this issue have been indecisive. Cutting, Rosner, and Foard (1975) found that dichotic presentation of the plucked and bowed sounds showed no significant ear advantage, but a null result in a dichotic study need not be considered conclusive, since it could result from the inadequate sensitivity and precision of the measure used.

One way of achieving a decisive finding where null results have predominated, is to use a potentially more sensitive measure, such as reaction time rather than accuracy. Springer (1973) has developed a means of reflecting hemispheric specialization through a reaction-time measure. She presented CV syllables monaurally with contralateral white noise and found a 14-msec advantage for stimuli presented to the right ear.

The purpose of the present study was to detect a potential ear advantage for plucked and bowed sounds using Springer's paradigm, with one modification: Springer had subjects respond only with the right hand, raising the possibility that the observed ear advantage might have been due to intercallosal transfer time rather than to hemispheric processing capacities. In the present study, therefore, both ear of presentation and hand of response were counterbalanced.

METHOD

Stimuli

The stimuli were identical to those used previously by Blechner et al. (1976). They were derived from the sawtooth wave sounds used by Cutting and Rosner (1974), originally generated on the Moog synthesizer at the Presser Electronic Studio at the University of Pennsylvania. The stimuli differed in rise time, reaching maximum intensity in either 10 msec (pluck) or 80 msec (bow). Using the pulse code modulation (PCM) system at Haskins Laboratories, the stimuli were truncated to 800 msec in duration and were stored on disc file in digitized form.

The white noise, which was to be presented contralaterally to the stimuli, was generated by a General Radio random-noise generator (Model 1390-A) and had a bandwidth of 20 kHz. The noise was digitized using the PCM system, truncated to a duration of 1000 msec, and then stored on disc file. The absolute levels of the noise and target stimuli (pluck and bow), as presented to listeners, were 80 and 70 dB SPL, respectively. All sounds were reconverted to analog form at the time of tape recording.

Tapes

All tapes were prepared using the PCM system. A display tape was prepared to introduce the subjects to the stimuli. The two kinds of stimuli (pluck and bow) were played in the same order several times, beginning with three tokens of each item, then two of each, and finally one of each.

Two binaural identification tapes were prepared, each with 32 tokens of the pluck and bow stimuli (16 of each) in random order.

Four dichotic test tapes were recorded. On one channel of each test tape, 60 tokens of the pluck and bow stimuli were recorded in random order with the constraint that every 10 stimuli contained equal numbers of pluck and bow stimuli. Thus, long runs of any one kind of stimulus were prevented. Sixty units of white noise were recorded on the second channel of the tape, with noise onset preceding stimulus onset by 50 msec. In addition, a 50 msec 1000 Hz tone that triggered the reaction time counter was recorded on both channels. The onset of this tone preceded the onset of the noise by 1.55 seconds. An interval of 2 seconds separated the offset of the noise from the onset of the next trigger tone. The intensity of the trigger tone was equivalent to the maximum intensity of the pluck and bow stimuli.

Four dichotic practice tapes were also prepared. These were identical in design with the test tapes but contained only 20 stimuli each.

Subjects and Apparatus

The 16 participants in the experiment included six males and ten females, ranging in age from 18 to 22 years. All were strongly right handed, as indicated by the five most reliable criteria found by Annett (1970). All reported no history of hearing trouble.

The tapes were played on an Ampex AG-500 tape recorder, and the stimuli were presented through calibrated Telephonics headphones (Model TDH39-300Z). Subjects sat in a sound-insulated room and responded with their index finger on either of two telegraph keys mounted on a wooden board. Throughout the experiment, the left key was used for bow responses, while the right key was used for pluck responses. The 50-msec pulse preceding each stimulus triggered a Hewlett-Packard 522B Electronic Counter. When a response on either telegraph key stopped the counter, the reaction time was printed on paper tape by a Hewlett-Packard 560A digital recorder for subsequent analysis. The listener's response choice was recorded manually by the experimenter.

Procedure

Listeners participated individually in a sound-insulated room. At the start of each session, they were informed of the general nature of the experiment and of the particular kinds of sounds that they would be asked to identify. They were told that the difference in rise time would be compared to the difference in sound between a plucked and a bowed violin string.

For preliminary training, subjects listened to the display sequence. They were then instructed on the mode of response, after which they listened to the display sequence twice more, responding to the sounds first with the left hand

and then with the right. Next, they listened to the binaural identification tapes. Eight of the subjects responded to the first tape with the left hand and to the second with the right hand. For the other eight subjects, the order of responding hands was reversed.

Subjects were then told that they would hear the stimulus in one ear, to which they were to pay careful attention, while there would be noise in the other ear, which they should ignore. They were played a few samples from the dichotic tapes to familiarize them with the noise-stimulus combination. They then listened and responded to the four practice tapes, and, after a five minute rest period, to the four test tapes.

For each individual listener, the pluck and bow stimuli were always presented through the same headphone. Ear of presentation was alternated by having the listener reverse the headset. For eight of the participants the stimulus was presented through one of the headphones, while for the other eight it was presented through the opposite headphone.

There were four possible hand-ear configurations. The order of these conditions was determined by a balanced Latin square design, yielding four possible orderings that were administered to four subjects each. The four practice and test tapes, however, were always played in the same order, to prevent any possible confusion between the effects of the random orders and the hand-ear configurations.

Subjects were instructed to respond as quickly and accurately as possible. In the final data analysis, only the last 50 test trials in each block were considered, the first ten functioning as warm-up trials to stabilize performance. The listener, however, was not told that the first ten trials would not count.

RESULTS

All of the subjects were able to identify the pluck and bow stimuli accurately. In the binaural identification trials, no listener made more than 4.7 percent errors.

For the reaction time data of the task with contralateral noise, median reaction time was calculated for each block of test trials for each subject. An analysis of variance was performed on these medians, with order of conditions considered as a between-subject factor, and hand and ear of presentation as within-subject factors.

The mean across subjects of individual medians for right-ear presentation of the stimuli was 662.5 msec, while for left-ear presentation, the mean was 675.3 msec. This 12.8-msec advantage for right-ear presentation was statistically significant, $F(1,12) = 5.69$, $p < .05$. Collapsed over ear of presentation, mean right-hand response was 665.0 msec, while mean left-hand response was 672.8 msec. This 7.1-msec difference, however, was not statistically reliable. All other main effect and interaction terms were not significant.

Accuracy in this experiment was quite high. The mean error rate was 0.9 percent. An analysis of variance on the error data showed no significant main effects or interactions.

DISCUSSION

The Issue of Special Processing for Phonetic Dimensions

The finding of a significant right-ear advantage for the identification of plucked and bowed sounds is very similar to the results for CV syllables, and suggests left-hemisphere involvement in the processing of both kinds of sounds. When the present data are considered along with other studies using plucked and bowed sounds, the parallels between this kind of nonspeech sound and CV syllables are quite compelling. The nonspeech stimuli have yielded all of the basic data patterns cited by Wood (1975) as evidence converging on the distinction between auditory and phonetic processes. The plucked and bowed sounds--like the speech stimuli--show asymmetric interference with redundancy gain in the speeded classification task, categorical boundary effects, selective adaptation of the category boundary, and evidence of left-hemisphere specialization. Considered together, this constellation of results with nonspeech stimuli leads one to question the existence of a special mode of processing for speech stimuli (Liberman, 1970), at least on the phonetic level. One might perhaps argue that identical results with CV syllables and plucked and bowed sounds do not guarantee identical perceptual mechanisms. Nevertheless, at the present time, it seems most parsimonious to account for these results in terms of a single mechanism for processing complex auditory dimensions that cue significant distinctions for a subject, rather than assuming, as Wood (1975) did, that results reflect separate mechanisms for phonetic and higher level auditory processes. It should be emphasized, however, that the conclusion proposed here does not challenge the notion of unique perceptual processes on other levels of linguistic organization.

Relevance to Specific Theories of Hemispheric Specialization

Although the present data have their greatest impact when considered within a set of converging experimental operations, they are relevant also to the specific question of the functions of the two cerebral hemispheres. Kimura (1967) suggested that the left and right hemispheres might be specialized, respectively, for verbal and nonverbal stimuli. This proposition has since been questioned by the discovery of right-ear advantages for nonspeech stimuli (for example; Halperin, Nachshon, and Carmon, 1973). The present study adds another set of data that contradicts the verbal-nonverbal dichotomy of hemispheric specialization.

Bever (1975) has suggested an alternative viewpoint, stressing the importance of different kinds of processing, rather than intrinsic stimulus variables in accounting for lateral asymmetry. He hypothesizes two modes of perception, analytic and holistic, for the left and right hemispheres, respectively. This view purports to account for individual differences in hemispheric specialization for melodies as a function of musical ability (Bever and Chiarello, 1974). However, the analytic-holistic distinction as currently formulated has little predictive value for plucked and bowed sounds. After looking at the data, one might suggest that they require analytic processing. After all, the stimuli differ in rise time, a small difference that might easily be missed if the stimulus were treated more globally. Other evidence, however, contradicts this view. Cutting et al. (1976), for example, have demonstrated that while rise time cues the distinction between plucked and bowed sounds, it is not an entirely sufficient cue. Fully half a second of waveform after stimulus onset is required for

the items to be identified properly. Thus, the whole of the stimulus is necessary for perception, and this fact would suggest a holistic mode of processing.

It may be more accurate to account for the present data based on the stimuli's acoustic nature rather than in terms of processing strategy. One particularly important acoustic property of the plucked and bowed stimuli in terms of hemispheric specialization is their characteristic rapid acoustic variation. Several studies have implicated the resolution of temporal variation as a left-hemisphere mechanism, both in audition (Halperin et al., 1973; Cutting, 1974) and vision (Goldman, Lodge, Hammer, Semmes, and Mishkin, 1968; Carmon and Nachshon, 1971). It may well be that the rate of shift in amplitude which distinguishes the plucked from the bowed sounds, is responsible for the greater left-hemisphere involvement.

REFERENCES

Annett, M. (1970) A classification of hand preference by association analysis. Brit. J. Psychol. 61, 303-321.

Bever, T. G. (1975) Cerebral asymmetries in humans are due to the differentiation of two incompatible processes: Holistic and analytic. Annals of the New York Academy of Sciences 263, 251-262.

Bever, T. G. and R. J. Chiarello. (1974) Cerebral dominance in musicians and nonmusicians. Science 195, 537-539.

Blechner, M. J., R. S. Day, and J. E. Cutting. (1976) Processing two dimensions of nonspeech stimuli: The auditory-phonetic distinction reconsidered. J. Exp. Psychol.: Human Perception and Performance 2, 257-266.

Carmon, A. and I. Nachshon. (1971) Effect of unilateral brain damage on perception of temporal order. Cortex 7, 410-418.

Cutting, J. E. (1974) Two left hemisphere mechanisms in speech perception. Percept. Psychophys. 16, 601-612.

Cutting, J. E. and B. S. Rosner. (1974) Categories and boundaries in speech and music. Percept. Psychophys. 16, 564-570.

Cutting, J. E., B. S. Rosner, and C. F. Foard. (1975) Rise time in nonlinguistic sounds and models of speech perception. Haskins Laboratories Status Report on Speech Research SR-41, 71-94.

Cutting, J. E., B. S. Rosner, and C. F. Foard. (1976) Perceptual categories for musiclike sounds: Implications for theories of speech perception. Quart. J. Exp. Psychol. 28.

Goldman, P. S., A. Lodge, L. R. Hammer, J. Semmes, and M. Mishkin. (1968) Critical flicker frequency after unilateral temporal lobectomy in man. Neuropsychologia 6, 355-363.

Halperin, Y., I. Nachshon, and A. Carmon. (1973) Shift in ear superiority in dichotic listening to temporally patterend nonverbal stimuli. J. Acoust. Soc. Am. 53, 46-50.

Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex 3, 163-178.

Liberman, A. M. (1970) Some characteristics of perception in the speech mode. In Perception and Its Disorders, vol. 48, ed. by D. A. Hainburg and K. H. Pribram (Baltimore, Md.: Williams and Wilkins), pp. 238-254.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Springer, S. P. (1973) Hemispheric specialization for speech opposed by contralateral noise. Percept. Psychophys. 13, 391-393.

Studdert-Kennedy, M. (1974) The perception of speech. In Current Trends in Linguistics, vol. 12, ed. by T. A. Sebeok (The Hague: Mouton), pp. 2349-2386.

Wood, C. C. (1975) Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analyses. J. Exp. Psychol.: Human Perception and Performance 1, 3-20.

Dichotic Competition of Speech Sounds: The Role of Acoustic Stimulus Structure*

Bruno H. Repp+

ABSTRACT

Dichotic consonant-vowel syllables contrasting in two features of the initial stop consonant (voicing and place) were presented for identification in a single-response paradigm without selective attention instructions. The acoustic structure of the syllables was varied within categories on both dimensions [voice onset time (VOT) and formant transitions]. These variations (especially those in VOT) had a clear influence on the pattern of responses (including blends), thus ruling out a simple phonetic feature recombination model. Rather, the auditory properties of the stimuli seem to be preserved at the stage of dichotic interaction. An alternative model (the "prototype model"), which assumes that dichotic integration of information takes place at a "multicategorical" stage intermediate between auditory and phonetic processing, is only moderately supported by the data. Nevertheless, some arguments are presented for maintaining this model as a working hypothesis. A new procedure for estimating the dichotic ear advantage was applied here for the first time, together with the single-response requirement. Most subjects showed unusually large right-ear advantages, which makes the present methodology interesting for the study of hemispheric asymmetry.

INTRODUCTION

Many recent studies of dichotic listening have employed synthetic syllables as stimuli, most often the set /ba/, /da/, /ga/, /pa/, /ta/, /ka/. These syllables offer a number of advantages over other materials. As synthetic syllables, their acoustic properties can be precisely controlled. Phonetically, they are a homogeneous stimulus set that represents all possible combinations of two values

---

*This paper is now in press, Journal of Experimental Psychology: Human Perception and Performance; and portions were presented at the 91st meeting of the Acoustical Society of America, Washington, D.C., April 1976.

+Also University of Connecticut Health Center, Farmington.

Acknowledgment: This research would not have been possible without the generous hospitality of Haskins Laboratories and its director, Alvin Liberman. The author was supported by NIH Grant T22 DE00202 to the University of Connecticut Health Center. I would like to thank Terry Halwes for comments on a draft of this paper.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

of the voicing feature (voiced, voiceless) and three values of the place feature (labial, alveolar, velar). They also yield a reliable right-ear advantage (REA) which often tends to be larger than the REA for other classes of competing speech sounds (Haggard, 1971; Blumstein, 1974; Cutting, 1974).

The Feature Recombination Hypothesis

Detailed studies of the dichotic competition between the six stop consonants have revealed several interesting phenomena, one of which will be of special interest here. When the two competing stimuli differ on both dimensions (voicing and place; for example, /ba/-/ta/), many errors are obtained that combine correct feature values from the two ears, such as /pa/ or /da/ as responses to /ba/-/ta/. These responses have been termed blend errors (Halwes, 1969; Studdert-Kennedy and Shankweiler, 1970). Blend errors are responsible for another finding often called the "feature-sharing advantage" (which actually is a feature-contrast disadvantage): dichotic syllables that differ in both features receive fewer correct responses than syllables that contrast only in a single feature (Halwes, 1969; Studdert-Kennedy and Shankweiler, 1970; Studdert-Kennedy, Shankweiler, and Pisoni, 1972; Pisoni, 1975). These two phenomena--which are basically the same, since blend errors can occur only with double-feature contrasts and therefore lead to higher error rates for these dichotic pairs--have provided the primary support for a feature recombination model of dichotic interaction. In its simplest form, this model assumes that phonetic features are: (1) independently extracted from the auditory information arriving from each hemisphere; (2) stored in a common feature buffer where information about the origin of the feature values is lost; and (3) finally recombined into percepts or responses. In other words, it is assumed that the interaction between dichotic stimuli takes place after the extraction of phonetic features, and that the competing values of a particular feature have equal probabilities of being selected from the feature buffer, independent of other particular features. Although this model has not always been clearly stated in the past, it was implicit in most previous research on dichotic competition (Halwes, 1969; Studdert-Kennedy and Shankweiler, 1970; Studdert-Kennedy, Shankweiler, and Pisoni, 1972; Blumstein, 1974; Pisoni, 1975; Cutting, 1976).

This simple model makes several strong and easily testable predictions, some of which have been examined by Halwes (1969). If all information about the local origin of the feature values is lost, double-feature contrasts should receive an equal number of correct responses and blend errors, and the two possible blend (and correct) responses should also be equally frequent. However, Halwes found correct responses to be twice as frequent as blend errors. This result could be accommodated by assuming that some of the local information is retained, so that feature values that come from the same hemisphere have a better than even chance of being selected together to form a response. However, Halwes also found wide variation in the frequencies of blend errors for different individual stimulus combinations, as well as strong asymmetries in the frequencies of the two possible blend (and correct) responses for individual stimulus pairs. He suggested that unequal salience of different acoustic cues may have played a role, but he did not indicate how this idea could be incorporated in the feature recombination model (which he did not explicitly reject).

In fact, it is possible to maintain the basic structure of the model, if the additional assumption is made that individual phonetic feature values have different strengths or saliencies, which are reflected in unequal probabilities

of selection from the phonetic feature buffer. The question remains: What determines these strengths? One possibility is that they are inherent--that they have a phonetic basis. The other possibility, suggested by Halwes (1969), is that they reflect the acoustic structure of the stimuli. If the latter hypothesis were true, the simple phonetic feature recombination model would have to be rejected, since it rests on the basic assumption that dichotic competition is exclusively phonetic in nature.

In order to test these hypotheses, let us consider another prediction of the model. This prediction is that acoustic stimulus variations within phonetic categories should not affect the frequency of blend errors and, indeed, should leave the whole response pattern unchanged. Since the phonetic features are assumed to be extracted independently before the combination of information from the two hemispheres, acoustic within-category variations can affect only the feature extraction process, but not the subsequent recombination of the features. By definition, within-category variations do not affect the accuracy of phonetic feature extraction (if they do, they are not true within-category variations), so that their effect in dichotic competition should be nil. This null hypothesis, whose maintenance is essential to the survival of the feature recombination model, was the focus of the present study. A rejection of the hypothesis was expected, since an alternative model that predicted specific effects of within-category acoustic variations was available.

The Prototype Model

This alternative model has been proposed by Repp (1976b, in press). It differs from the feature recombination model, as it considers syllables not as bundles of separately extracted phonetic features, but as integral multidimensional entities whose dimensions are inseparable aspects of the whole pattern (cf. Lockhead, 1970, 1972; Garner, 1974; see also the present discussion). The dimensions are assumed to reflect the auditory properties of the stimulus and thus are continuous, not binary. Instead of representing speech sounds as matrices of discrete feature values, they are conceptualized as points in a continuous multidimensional perceptual space. In the same auditory space, a limited number of fixed "prototypes" are located, which represent the listener's "ideal" concepts (his tacit knowledge) of the relevant phoneme or syllable categories. According to this prototype model, a stimulus is identified in three stages: (1) First, auditory processing leads to a mapping of the acoustic information into the multidimensional space. (2) In this perceptual space, the stimulus leads to "activation" of the prototypes in its vicinity, the degree of activation being an inverse and probably nonlinear function of the (Euclidean) distance between stimulus and prototype. This results in a "multicategorical vector" whose elements are the activation values of the prototypes. (3) Finally, a probabilistic decision process selects the prototype with the largest activation value as the response (or percept).

In the prototype model, dichotic interaction is assumed to take place at the level of multicategorical representation, in the form of a weighted averaging of the multicategorical vectors for the two stimuli. A single categorical decision is then made on the basis of this average vector. Thus, the model assumes that the competing information is combined and results in a single percept. This assumption is justified when synthetic syllables with the same fundamental frequency and in the same vocalic context are used because these stimuli strongly tend to fuse in dichotic competition (Halwes, 1969; Repp, 1976b;

Repp and Halwes, in preparation). The nature of the single categorical percept is determined by two factors: ear dominance, represented by the weights in the averaging process, and stimulus dominance, which is determined by the relative distances of the two competing stimuli from the prototypes in the perceptual space. The model predicts that stimuli that are close to a prototype will tend to dominate stimuli that are far from prototypes; this may be called the "category goodness hypothesis" of dichotic competition. Category goodness, that is, the distance from the "correct" prototype, is a function of auditory stimulus characteristics, so that the model predicts that stimulus dominance will vary if acoustic within-category variations of the stimuli are introduced. This was confirmed by Repp (1976b) within a restricted stimulus set--that of the voiced stop consonants. By varying the initial formant transitions, the dominance relationships between the stimuli from a "place continuum" could be reliably influenced, and the pattern of the data conformed at least qualitatively to the prototype model.

The present experiment investigated the generality of these earlier findings. In order to be useful, the prototype model should explain the response pattern for all dichotic combinations of the six stop consonants, as well as the effects of variations in cues other than the initial formant transitions. Consider first how the model explains blend responses. Two stimuli such as /ba/ and /ta/ will not only activate their correct prototypes (B and T, respectively) but also, to a lesser degree, the blend prototypes, D and P, which are neighbors in perceptual space. Because of the presumed additivity of prototype activation levels, the blend prototypes may reach activation levels comparable to those of the correct prototypes, to which only one of the two stimuli makes a substantial contribution.

In principle, this model allows for variations in the frequencies of blends between individual stimulus pairs, since they depend in a complex way on the arrangement of prototypes and stimuli in the perceptual space. A mathematical formulation of the model should be able to predict their pattern. In the present context, however, we will be content with qualitative predictions concerning changes in the response pattern, leaving quantitative tests to a future study.

Contrary to the feature recombination model, the prototype model predicts variations in the response pattern with changes in the acoustic structure of the stimuli. Consider again the previous example, the stimulus pair /ba/-/ta/. Assume that we delay the voice onset time (VOT, the important acoustic cue for the voicing feature) of /ba/, so that the stimulus is still identified as B, but in the perceptual space it is farther removed from the B prototype and closer to the P prototype. It will now be closer to the boundary between voiced and voiceless sounds, and it will contribute less activation to B and D and more to P and T than the original /ba/. As a result, the frequencies of P and T responses should increase, and that of B and D responses should decrease. Similar predictions may be made for changes of VOT in the other direction or in the other stimulus, or for changes in the formant transitions (the acoustic cue for place of articulation) of either stimulus. A number of other, more detailed, predictions may be derived from the model, some of which will be considered in the Results section of this paper.

The phonetic feature recombination model and the prototype model are not the only possible conceptions of the process of dichotic interaction, but most other plausible models are compromises between these two extremes (see the Discussion

section). The detailed formulation of such models seems less important than the empirical demonstration of within-category effects in dichotic competition; such a demonstration would rule out a whole class of models.

In addition to the primary focus on stimulus dominance in dichotic competition, the present study gave attention to the factor of ear dominance. A new method of calculating ear advantage indices, especially designed for the single-response paradigm (Repp, 1976a, 1976b; Repp and Halwes, in preparation) was applied here for the first time. This experiment constituted part of an ongoing series of studies aimed at developing optimal procedures for assessing lateral asymmetries in dichotic listening.

METHOD

Subjects

The subjects were eight paid volunteers, four women and four men, mostly Yale students. All had normal hearing, except one man who claimed to have a slight (5 dB) hearing loss in the right ear. Two subjects were left-handed, one of them only in writing. All were relatively inexperienced listeners.

Stimuli

The stimulus set comprised 24 syllables which were synthesized on the Haskins Laboratories parallel resonance synthesizer. There were four acoustically different versions of each of the six syllables, /ba/, /da/, /ga/, /pa/, /ta/, and /ka/, resulting from all combinations of four different VOTs with six different (second- and third-) formant transitions, as illustrated in Figure 1. All syllables were 300 msec long, had no initial bursts, the same transition durations (50 msec),[1] and the same constant fundamental frequency (90 Hz).

The experimental tape was recorded using the pulse code modulation (PCM) system at Haskins Laboratories. The tape contained first a list of 120 single syllables consisting of five different random sequences of the 24 stimuli. It was followed by two blocks of dichotic pairs. Each block contained 192 pairs, representing all possible double-feature contrast combinations of the 24 stimuli: six phoneme combinations (/ba/-/ta/, /ba/-/ka/, /da/-/pa/, /da/-/ka/, /ga/-/pa/, /ga/-/ta/) with two channel/ear assignments for each, and sixteen different acoustic combinations within each phonemic contrast. Their sequence was completely random, with interstimulus intervals of 3 seconds. The onsets of the syllables in a dichotic pair were exactly simultaneous (0.125 msec maximal error).

Procedure

The subjects were tested in small groups in a single session lasting about two hours. The single-channel series was presented monaurally for identification, followed by the two dichotic blocks. After a break, the tape recorder channels were reversed electronically and the two dichotic blocks were presented

[1] It was discovered after the experiment that the first-formant transitions of the labial consonants were only 40 msec long. However, this was almost certainly of no consequence.

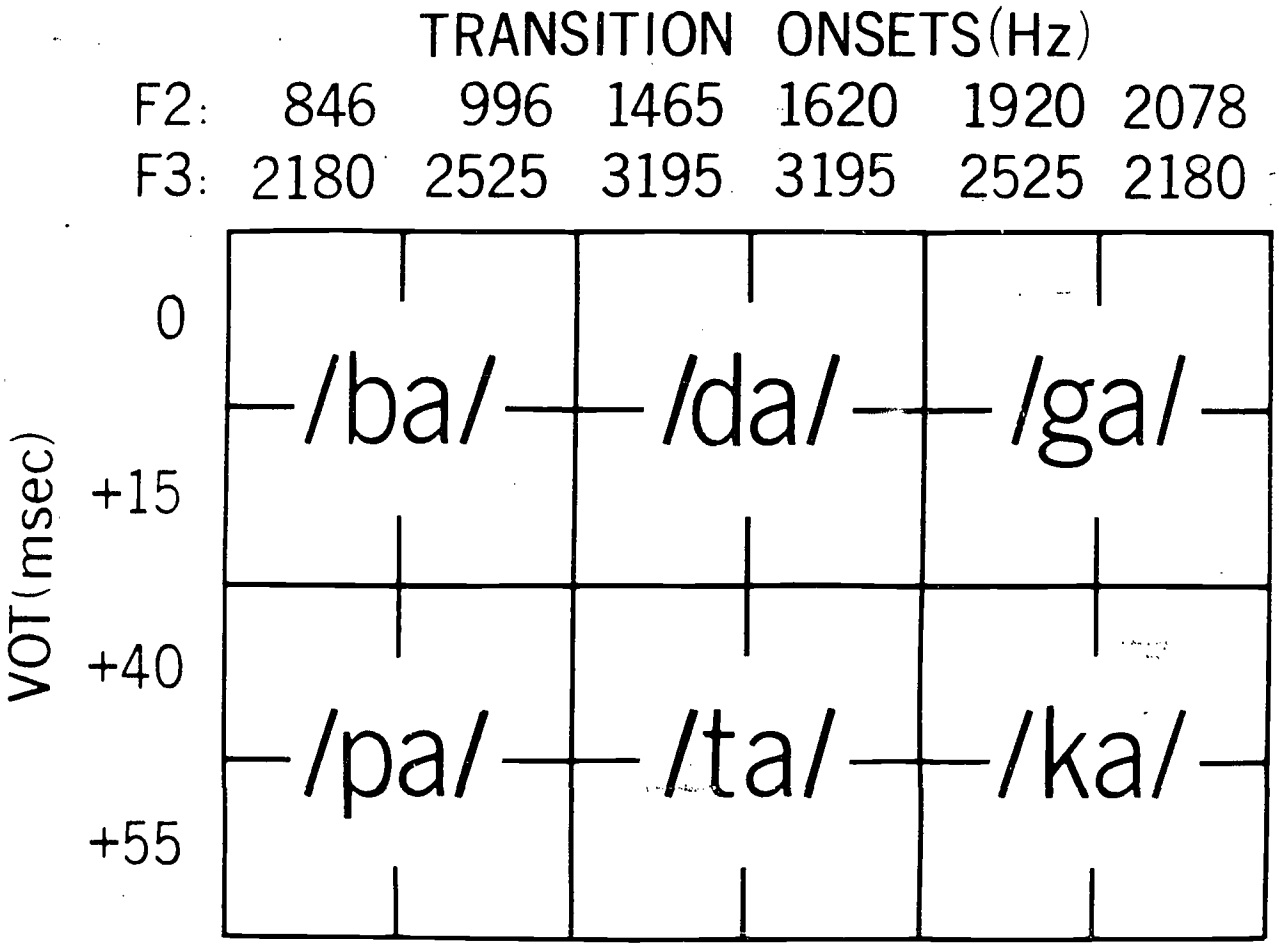


Figure 1: Acoustic stimulus parameters. The steady-state frequencies for /a/ were 1232 Hz (F_2) and 2525 Hz (F_3).

again, followed by the monaural syllables, now in the opposite ear. All in all, each subject listened to ten replications of each single syllable and to four replications of each dichotic pair (eight, if channel/ear assignment is ignored). The tape was played back from an Ampex AG-500 tape recorder through an amplifier/attenuator to Telephonics TDH-39 headphones. The intensities of the two channels were carefully equalized at about 65 dB SPL (peak deflections on a voltmeter).

As part of the instructions, the subjects were first given a talk on the two features--voicing and place--and were told the precise stimulus combinations to expect, with the help of a diagram on the answer sheets. However, they were not informed about the within-category variations until after the experiment. The subjects were asked to write down a single response for each dichotic pair, whatever the fused stimuli sounded most like. Naturally, the responses were restricted to the six stop consonants, with the additional admonition to try to give both voiced and voiceless responses.[2]

RESULTS AND DISCUSSION

Monaural Intelligibility

As is often the case with synthetic syllables, their intelligibility in the experiment turned out to be somewhat poorer than anticipated. The confusion matrix for all eight subjects is shown in Table 1. The problem lay almost exclusively with /da/ and /ta/ which were more often heard as /ga/ and /ka/, respectively. The absence of a burst, which is especially important in alveolar consonants, may have been a factor here. The confusability of these stimuli was not detrimental to the purpose of the experiment, although it had to be dealt with in the analysis of the dichotic data.

Confusions along the voicing dimension were extremely rare and occurred exclusively at the VOTs closer to the boundary. A similar pattern may be seen for /ga/ and /ka/ with respect to place confusions; alveolar responses were more frequent when the velar transitions were closer to the boundary (low). However, for /da/ and /ta/ the opposite was the case; velar responses were more frequent when the transitions were farther away from the alveolar-velar boundary (low). This curious reversal has been confirmed in other studies using the same stimuli (Repp, in preparation); its explanation is far from clear.

---

[2] It was thought that some subjects might give predominantly voiceless responses, which would have reduced the information in the data. This suspicion, derived from pilot observations, was apparently unfounded. For the same reason, four subjects (two old and two new) were (re)tested with the same tape with detection instructions. These instructions restricted the response set to either the voiced consonants (B, D, G) or the voiceless consonants (P, T, K) only, counterbalanced across blocks within subjects. Since the subjects knew that each dichotic pair contained one voiced and one voiceless consonant, this amounted to a detection task. The main purpose of the detection instructions was to force the subjects to give an equal number of voiced and voiceless responses to each pair, and, consequently, only the effects of variations in formant transitions could be assessed. These effects agreed with those under standard instructions, as described in the Results section.

TABLE 1: Confusion matrix of the 24 stimuli (monaural identification).

Stimuli			Responses					
	VOT	F_2	B	D	G	P	T	K
/ba/	0	low	80	-	-	-	-	-
	0	high	80	-	-	-	-	-
	+15	low	76	-	-	4	-	-
	+15	high	79	-	-	1	-	-
/da/	0	low	-	28	52	-	-	-
	0	high	-	34	46	-	-	-
	+15	low	-	27	53	-	-	-
	+15	high	-	39	40	-	-	1
/ga/	0	low	-	5	75	-	-	-
	0	high	-	-	80	-	-	-
	+15	low	-	16	64	-	-	-
	+15	high	-	-	79	-	-	1
/pa/	+40	low	-	-	-	80	-	-
	+40	high	1	-	-	78	-	1
	+55	low	-	-	-	80	-	-
	+55	high	-	-	-	80	-	-
/ta/	+40	low	-	-	-	1	22	57
	+40	high	-	1	-	1	34	44
	+55	low	-	-	-	3	30	47
	+55	high	-	-	-	-	61	19
/ka/	+40	low	-	-	4	-	12	64
	+40	high	-	1	2	-	4	73
	+55	low	-	-	-	2	6	72
	+55	high	-	-	-	2	1	77

The Dichotic Response Pattern

The dichotic response pattern for the six phonemic contrasts, disregarding within-category variations, is shown in Table 2. The underlined percentages represent blends; their total frequencies are given in the last column. It can be seen that blend responses were extremely common but varied in frequency as a function of the stimuli involved: in the two pairs containing /ba/, blend responses comprised almost two-thirds of all responses; in the two pairs containing /pa/, only about one-third; in the remaining two pairs, somewhat less than half. In these two last pairs (alveolar-velar contrasts), the exact proportion of blends was uncertain, as indicated by the parentheses in Table 2. Because of the listeners' uncertainty about the place of articulation of the component stimuli, blend responses could have arisen from either blending or from confusions and, likewise, "correct" responses may have included some true blends.

TABLE 2: Dichotic stimulus-response matrix.

Stimuli	Percentage of responses							
	B	D	G	P	T	K	Correct	Blends
/ba/-/ta/	11.4	3.6 + 3.4		56.7	14.2 + 10.6		36.2	63.8
/ba/-/ka/	13.5	3.5 + 6.8		56.2	3.2 + 16.6		33.5	66.5
/da/-/pa/	4.8	24.6 + 20.0		23.6	12.2 + 14.7		68.2	31.8
/da/-/ka/	1.2 + 16.5		37.9	1.7 + 8.2		34.6	(52.3)	(47.7)
/ga/-/pa/	7.1	8.1 + 38.4		24.6	4.3 + 17.5		71.1	28.9
/ga/-/ta/	0.8 + 10.9		37.5	1.3 + 17.3		32.2	(56.1)	(43.9)

The poor discrimination between alveolar and velar place is also reflected in the responses to the other pairs containing one labial consonant. Since the labials were highly intelligible (Table 1), alveolar and velar responses were therefore simply grouped together in these dichotic pairs, as indicated by the plus signs in Table 2. For example, G responses to /ba/-/ta/ were considered blends, while K responses were considered correct. In alveolar-velar pairs, the few labial responses that occurred (probably random errors) were combined with the alveolar responses. These groupings were maintained in all further data analyses.

Table 2 shows enormous variation in the pattern of blend responses. In the two pairs containing /ba/, P responses predominated and were more than twice as frequent as P responses to pairs actually containing /pa/. In terms of the prototype model, this indicates that /ba/ was far from the B prototype on the voicing dimension but close to it on the place dimension, that is, it was weak on the former but strong on the latter; hence the joint predominance of labial and voiceless responses. This suggests that the response pattern could perhaps be explained in terms of separate and independent competition of the two features--voicing and place--although this would contradict the prototype model. However, in the two pairs containing /pa/, for example, correct responses were much more frequent than predicted by this hypothesis, while in pairs containing /ba/, they were less frequent than predicted. Note that the hypothesis of feature independence predicts that responses in the different place categories should be proportional within voicing categories. However, the stimulus pair /ga/-/pa/, for example, received five times as many G responses as B responses, but actually fewer K than P responses. This result contradicts the hypothesis of feature independence in dichotic competition. In principle, this is compatible with the prototype model, although it is not yet clear whether a more rigorous, quantitative formulation of the model would be able to explain the detailed response pattern. The feature recombination model, on the other hand, cannot explain the variations in the proportions of blend responses for different stimulus pairs or the asymmetries in blend responses to individual pairs, thus confirming Halwes (1969).

Effect of Within-Category Variations in VOT

These results are shown in Table 3. The data are shown as the percentages of voiced and voiceless responses, and of correct and blend responses to the four VOT combinations, averaged over the different phonemic contrasts and the variations in formant transitions.

TABLE 3: Percentages of voiced and voiceless correct responses and blends as a function of VOT combinations.

		Correct		Blends		Total	
	VOT	+40	+55	+40	+55	+40	+55
RESPONSES Voiced	0	42.8	15.1	19.1	5.7	61.9	20.8
	+15	35.1	20.6	18.5	10.4	53.6	31.0
Voiceless	0	16.3	33.1	21.8	46.1	38.1	79.2
	+15	19.5	28.5	26.9	40.5	46.4	69.0
Total	0	59.1	48.2	40.9	51.8		
	+15	54.6	49.1	45.4	50.9		

Obviously, the variations in VOT had a strong effect on the response pattern. The most striking effect was produced by a change in the VOT of the voiceless stimulus. Voiceless stimuli with the shorter VOT (+40) led to a slight predominance of voiced responses, while those with the longer VOT (+55) brought about a predominance of voiceless responses. This is in agreement with the prototype model, since there is good reason to assume that a voiceless stimulus with a VOT of +55 will be closer to its prototype than a stimulus with a VOT of +40. On the other hand, the effect of a change in the VOT of voiced stimuli was less striking and showed an interaction with the VOT of the voiceless competitor. When the VOT of the latter was +40, the effect of a VOT change from 0 to +15 in the voiced stimulus was as predicted, that is, it led to a relative decrease in the percentage of voiced responses. However, when the VOT of the voiceless stimulus was +55, the effect of the same change in the VOT of the voiced stimulus had just the opposite effect. This interaction was unexpected and is difficult to explain.

This pattern of results was highly consistent between individual phoneme combinations and individual subjects. Analysis of variance of the percentages of voiced (voiceless) responses yielded a highly significant effect of the VOT of the voiceless stimulus ($F_{1,7} = 59.14$, $p < .0002$) and a significant interaction between the VOT of the voiced stimulus and the VOT of the voiceless stimulus ($F_{1,7} = 24.63$, $p < .002$). The main effect of the VOT of the voiced stimulus was not significant.

Table 3 also shows that the proportion of correct responses and blends varied as a function of VOT. Correct responses were more frequent where voiced responses were more frequent, while blends tended to accompany voiceless responses. Note that the majority of all voiced responses were correct, while, among the voiceless responses, blends were more frequent than correct responses. This indicates that the place feature of voiceless stimuli was weak in competition with the place feature of voiced stimuli. In terms of the prototype model,

it suggests that noise-excited formant transitions are a less effective cue to place of articulation than voiced transitions. This is plausible since the present stimuli did not contain any bursts--a second important cue to place of articulation that certainly is more important in voiceless plosives.

Effect of Within-Category Variations in Formant Transitions

These results are shown in Table 4 as the percentages of responses with the place of the voiced stimulus and with the place of the voiceless stimulus, and of

TABLE 4: Percentages of correct responses and blends with the place of the voiced (voiceless) stimulus as a function of transition combinations.

		Correct		Blends		Total	
	Voiced	close	far	close	far	close	far
	Voiceless						
Responses with place of voiceless stimulus	close	.9	25.1	29.1	33.0	61.0	58.1
	far	3 0	26.2	36.0	36.8	68.0	63.0
Responses with place of voiced stimulus	close	23.4	28.3	15.6	13.6	39.0	41.9
	far	20.1	25.7	11.9	11.3	32.0	37.0
Tot-	close	54.3	52.6	45.7	47.4		
	far	52.1	51.9	47.9	48.1		

correct responses and blends. The dimensions of each 2 × 2 subtable are the transitions of the voiced stimulus (rows) and of the voiceless stimulus (columns). The transitions were classified according to whether they were close to or far from the category boundary separating the place values of the two competing stimuli. Thus, "close" refers to the higher F_2 transitions for labials and for alveolars paired with velars, but to the lower F_2 transitions for velars and for alveolars paired with labials.

It is evident that the effect of variations in the formant transitions was much smaller than that of VOT, but it was in the direction predicted by the prototype model: responses with the place of the voiced stimulus were most frequent when the transitions of the voiced stimulus were far and those of the voiceless stimulus were close, and they were least frequent when the opposite was the case.[3] This pattern was shown primarily by the correct responses; the blends followed a somewhat different pattern, tending to be least frequent when both stimuli were close and most frequent when both were far.

[3]It may be argued that the within-category effect of the transitions reflected merely changes in the confusion probabilities of alveolar and velar stimuli (cf. Table 1). However, the dichotic effects were only slightly reduced after a correction was applied that took changes in confusion structure into account. Moreover, the transitions of labial consonants (which were rarely confused; see Table 1) had a very pronounced effect.

Analysis of variance of the responses with the place of the voiced (voiceless) stimulus yielded a highly significant effect of the transitions of the voiced stimulus ($F_{1,7} = 27.17$, $p < .002$), but only a marginally significant effect of the transitions of the voiceless stimulus ($F_{1,7} = 4.79$, $p < .07$), with no significant interaction between the two. Thus, the former was more reliable than the latter, which again indicates that the transitions of voiceless stimuli were weak in their perceptual effect.

There were some consistent deviations from the pattern in Table 4, which are in part responsible for the relatively small average effect. Labial-velar pairs, especially /pa/-/ga/, received more labial responses when the velar transitions were far than when they were close. Pairs containing alveolar consonants, on the other hand, conformed to the predictions, despite the inverted pattern of place confusions in monaural presentation (see Table 1).

Within-Category Feature Interactions

It has been pointed out above that the response pattern in Table 2 cannot be explained by independent competition on the two phonetic dimensions (phonetic feature independence). The question of feature independence may also be asked within phonemic combinations (auditory feature independence): Did within-category variations in VOT affect competition on the place dimension, and did within-category variations in the formant transitions influence competition on the voicing dimension?

Responses with the place of the voiced (voiceless) stimulus did not vary significantly as a function of VOT. However, a more detailed analysis showed that the VOT of the voiceless stimulus did have a significant influence in some individual stimulus combinations. The largest of these effects was in /ba/-/ka/ and consisted in a decrease in labial responses and an increase in velar responses as the VOT of /ka/ changed from +40 to +55. This effect is in agreement with the prototype model which predicts a certain amount of positive correlation between features: as a stimulus moves closer to its prototype along one dimension, its overall Euclidean distance from the prototype is reduced, and other dimensions will indirectly benefit from this increase in category goodness.

Voiced (voiceless) responses showed a significant effect of the transitions of the voiceless stimulus ($F_{1,7} = 22.61$, $p < .003$). Voiced responses were more frequent when the voiceless transitions were closer to the boundary, which is again in agreement with the prototype model. The (nonsignificant) effect of the transitions of the voiced stimulus, however, was not in the predicted direction. It was also surprising that the voiceless transitions affected competition on the voicing feature more than competition on the place feature.

The prototype model also predicted variations in the proportion of blend errors (and correct responses) as a function of joint variation in both stimulus dimensions. Correct responses were expected to be most frequent (and blend responses least frequent) when the two competing stimuli were farthest apart in perceptual space--when they were closest to their respective correct prototypes. The opposite result was predicted when the two stimuli were closest in perceptual space, and thus almost as close to the blend prototypes as to the correct prototypes. This hypothesis was most easily tested by considering only the acoustically most similar and the acoustically most dissimilar pair within each phonemic contrast. (For example, in /ba/-/ta/, the most similar pair would be

/ba/ with high F_2 transitions and VOT = +15 paired with /ta/ with low F_2 transitions and VOT = +40, while the most dissimilar pair would be /ba/ with low F_2 transitions and VOT = 0 paired with /ta/ with high F_2 transitions and VOT = +55.) Of the six phonemic contrasts, only one supported the prediction, while four showed differences in the opposite direction. Overall, blends were more frequent when the competing stimuli were acoustically dissimilar. This is in contradiction to the prototype model. However, the result is in agreement, and indeed a consequence of, the earlier observations that variations in the formant transitions had a relatively small effect, and that blends tended to accompany voiceless responses which increased greatly in frequency as VOT changed from +40 to +55.

Ear Dominance[4]

The present experiment offered a first opportunity to apply an improved method for calculating an unbiased index of ear dominance recently proposed by Repp (1976a, 1976b). This new index takes into account the variations in stimulus dominance by applying the methods of signal detection theory and fitting a receiver-operating-characteristic (ROC) curve to the data points for individual stimulus pairs. The index is a linear transformation of the area under the ROC function (cf. Green and Swets, 1966), and it ranges from +1 for a perfect REA to -1 for a perfect left-ear advantage. Its derivation and its advantages over other indices are discussed in a separate paper (Repp and Halwes, in preparation).

The calculation of the unbiased ear advantage index presupposes that the responses can be grouped into two exhaustive categories. Double-feature contrasts present a problem here, because of the large proportion of blend errors which are ambiguous with respect to ear dominance. At present, it is not clear how a valid index could be derived from the responses at the phonemic level. However, the problem can be circumvented by separately considering the two features, voicing and place. Ear dominance indices for voicing only are easily calculated by classifying the responses as voiced and voiceless, ignoring the place feature. These indices (and the corresponding ROC function) were based on 24 data points, representing the four VOT combinations for each of the six phonemic contrasts, ignoring variation in the transitions. The results are shown in the first column of Table 5.

Similar indices were calculated for the place dimension by dichotomizing the responses, using the same grouping of place categories as in the earlier data analysis. Each index was based on 24 data points, representing the four transition combinations for each of the six phonemic contrasts, ignoring variations in VOT. These indices are shown in the second column of Table 5. The third column of Table 5 shows the same indices, but omitting the eight data points for alveolar-velar contrasts.

Table 5 shows that there was a highly significant average REA. Except for one subject on the voicing dimension, all subjects showed REAs. The most striking result is the magnitude of these effects. The average REAs, as well as most of the individual coefficients, are several magnitudes larger than the

[4]The terms, ear dominance and ear advantage, are used interchangeably here, although the former is more appropriate within the single-response paradigm.

TABLE 5: Individual ear advantages [unbiased coefficients based on the method described in Repp (1976a, 1976b) and Repp and Halwes (in preparation)].

Subjects	Voicing	Place	Place[a]
JK[b]	+0.17	+0.09[c]	+0.10[c]
JL	+0.73	+0.52	+0.64
RG	+0.89	+0.57	+0.76
MR	+0.57	+0.82	+0.89
GG	-0.09[c]	+0.35	+0.35
WT[d]	+0.90	+0.76	+0.78
TJ	+0.47	+0.14	+0.26
CW[d]	+0.75	+0.81	+0.98
Average	+0.55	+0.51	+0.60
BHR[e]	+0.96	+0.55	+0.64

[a]Omitting alveolar-velar contrasts.

[b]Claimed a 5-dB hearing loss in the right ear.

[c]Not significant. All other coefficients are significant at $p < .05$ or better [estimated according to the procedure outlined in Repp and Halwes (1976)].

[d]Left-handed (WT for writing only).

[e]Data for the author as a subject; average of three sessions.

advantages reported in earlier studies of normal subjects. (In fact, several subjects show REAs close to the possible maximum.) There are two possible reasons why these indices are so large. One is that some conventional indices, such as the Phi coefficient (Kuhn, 1973; Repp, 1976b), underestimate the "true" size of the ear advantage. For example, the average Phi coefficient on the voicing dimension was +0.30, which is only about half the size of the unbiased index of +0.55. However, this Phi coefficient is still very large compared to those in earlier studies, which required the subjects to give two responses [for example, Shankweiler and Studdert-Kennedy (1975), who reported an average Phi of +0.06]. The reason for this difference may be that the single-response paradigm adopted here eliminates much of the noise that is present in two-response data and therefore reveals the true magnitude of the ear advantage. There is much to be said in favor of this argument (see Repp and Halwes, in preparation). However, Repp (1976b) reported an average Phi coefficient of only +0.06 in a single-response experiment with completely fused syllables that contrasted in place only. Clearly, there must be an additional factor beyond the response requirements and the kind of index used. Although previous studies have not indicated a substantial difference in the REA for completely fused and partially fused syllables, the present results suggest strongly that such a difference exists; it perhaps was obscured by guessing responses in earlier studies requiring two responses.[5]

[5]It may be noted that none of the four subjects (JK, JL, and two new listeners) who received detection instructions (footnote 2) showed a large REA on the place

A comparison between the second and third columns in Table 5 shows that, for all subjects but one, exclusion of alveolar-velar pairs led to an increase in the ear dominance coefficient on the place dimension. This finding illustrates an important methodological point: pairs of stimuli that are highly confusable will tend to show a reduced ear advantage. It follows that high intelligibility of the stimuli in a dichotic test is an important requirement, and that pairs of confusable stimuli should be omitted from consideration when the ear advantage is determined.

Finally, the indices for voicing and place (columns 1 and 3 in Table 5) may be compared. While the average indices are similar, there are substantial individual differences. Some of these may be due to chance, but the larger differences (and especially that for BHR, the author, whose results are based on 2,304 responses) are certainly real. It must be concluded that, for a given individual, the REA on the voicing dimension is not necessarily the same as on the place dimension. Underlying these differences may be individual differences in the perceptual representation of the speech sounds and of their dimensions (for example, in the structure of the subjective perceptual space). This points to a substantial problem in measuring the "true" or "physiological" ear advantage, which we are only now beginning to understand. Future research will have to deal with the possibility of interactions between hemispheric dominance and perceptual organization in individuals.

GENERAL DISCUSSION

The present study demonstrates clear effects of within-category acoustic variations on dichotic stimulus dominance relationships. This finding constitutes conclusive evidence against a simple phonetic feature recombination model, as outlined in the Introduction. It also renders insufficient a more elaborate version of this model incorporating the concept of inherent phonetic feature strength. Rather, the competitive strengths of phonetic feature values are probably a direct function of the acoustic stimulus structure, and changes in the latter lead to changes in the former. Thus, dichotic interaction does not take place at a strictly phonetic level, but at an earlier stage where auditory information is still preserved in some form.

The prototype model provides one possible conception of this auditory representation. According to this model, the dichotic inputs converge in the form of multicategorical vectors, a stage intermediate between continuous auditory and discrete phonetic representation. The multicategorical stage embodies the relationship between the variable auditory input and the more or less fixed phonetic categories. It has proven useful in conceptualizing the process of dichotic interaction and fusion (Repp, 1976b, in press) which so far has been considered only in terms of the auditory-phonetic dichotomy (Studdert-Kennedy, Shankweiler, and Pisoni, 1972; Pisoni, 1975; Cutting, 1976; Studdert-Kennedy, in press). However, the prototype model was only moderately supported by the present data. Below, we will briefly summarize some of its shortcomings, consider some alternative models, and present some theoretical arguments in favor of maintaining the prototype model as a working hypothesis.

---

dimension, and JL showed a marked reduction in her REA. The coefficients for these subjects were +0.05, +0.20, +0.08, and +0.12, respectively (alveolar-velar pairs included).

On the whole, the main prediction of the prototype model was confirmed: a dichotic stimulus tends to gain in competitive strength if its acoustic structure is changed so that it moves closer to its presumed correct prototype and away from category boundaries. However, there were two major exceptions: the inverted effect of a change in VOT from 0 to +15 when the competing stimulus had a VOT of +55 (Table 3), and the inverted effect of a change in the transitions of velars when paired with labials (mentioned in connection with Table 4). Both effects are very difficult to rationalize, but there is no doubt about their reality. A follow-up study of dichotic competition along the VOT dimension has revealed even more bizarre interactions. Note that they cannot be explained by atypical stimulus characteristics (such as synthesis artifacts) or by different assumptions about the location of the prototypes in perceptual space. For example, it has been implicitly assumed that VOT = 0 is closer to the voiced prototype than VOT = +15, and that VOT = +55 is closer to the voiceless prototype than VOT = +40. However, if the obvious hypothesis is introduced that the prototypes represent the modal production values of the corresponding articulatory dimensions, the first part of the assumption is probably false: VOT = +15 is closer to the modal production value than VOT = 0, at least for alveolars and velars (Lisker and Abramson, 1964; Klatt, 1973; Zlatin, 1974). However, even if this were true--and the data permit this interpretation as well as the opposite--it could not explain the interaction obtained; all that would change is the part of the interaction which is considered anomalous. (Note also that the VOT interaction was exhibited by all six phonemic combinations and thus was apparently independent of place of articulation.)

There is little value in discussing the several other respects in which the prototype model has failed. Instead, it seems useful to consider alternative models that perhaps could account for the anomalous findings. Unfortunately, however, the most obvious candidates make rather similar predictions and do not fare better than the prototype model.

It is possible, for example, to consider a pure "auditory averaging model." This model would assume that the dichotic stimuli are integrated at a strictly auditory level of processing, so that a single stimulus, a kind of auditory average of the two components, is phonetically interpreted. In the present context, this model makes predictions that are quite similar to those of the prototype model, but in other contexts differential predictions can be generated and the auditory averaging model has been found insufficient (Cutting, 1976; Repp, 1976b, in press). It is quite possible, however, that some auditory interaction is involved in addition to integration at a higher, multicategorical (and, perhaps, even phonetic) level. Such a multilevel model of dichotic interaction would be of considerable complexity, but it is not clear whether it could explain the anomalies in the present data.

Another alternative model that deserves some discussion is the "feature detector model" which currently enjoys some popularity (Eimas and Corbit, 1973; Cooper, 1974; Cooper and Nager, 1975; Miller, 1975, 1976; Studdert-Kennedy, in press). This model assumes a separate set of detectors for each feature, with one detector corresponding to each value of a feature (Eimas and Corbit, 1973; Cooper, 1974; Miller, 1975). Effectively, this places the prototypes at the level of auditory analysis. Dichotic interaction may be conceptualized as follows: each stimulus passes through separate banks of feature detectors and emerges as an array of multicategorical feature codes (that is, as a multicategorical matrix). These matrices then converge upon a single processor where they

are averaged. Subsequently, separate feature decision mechanisms select the largest detector response for each feature, and finally these categorical feature values are combined into a percept or response. Thus, each feature or dimension has its own little perceptual space and its own set of prototypes.

The predictions of the feature detector model are again rather similar to those of the prototype model, except that, in its simplest form, the former assumes mutual independence of individual features. There are several instances in the present data where this assumption must be rejected, so that rather complex ad hoc assumptions about the interrelations among feature detectors and among feature decisions would have to be introduced. The prototype model, on the other hand, predicts specific interdependencies between different features; some of them were supported by the data but others were not. The data therefore do not permit a choice between these alternative models. However, given that they are equally well (or equally poorly) supported, there are some theoretical reasons why the prototype models might be preferred as a working hypothesis.

The voicing and place features of stop consonants are among the best examples of "integral" dimensions (Lockhead, 1972; Garner, 1974). One cannot exist without the other, and selective attention to one feature is impossible without taking the other feature into account. In fact, there is strong evidence that the whole CV syllable is an integral unit of processing (Pisoni and Tash, 1974; Wood and Day, 1975). Integral units are multidimensional, and their dimensions interact during processing. The feature detector model can deal with such interactions only by some rather strenuous assumptions which, typically, are made post hoc and often are based on assumptions of serial processing, which are inappropriate with integral dimensions (Garner, 1974). The prototype model, by virtue of its multidimensional Euclidean structure, naturally incorporates such interactions, and it makes predictions that can be quantified and falsified. Moreover, it is somewhat counterintuitive and uneconomical to assume a separate categorical decision for each feature, subconscious as these decisions may be. A single phonetic decision is more in line with subjective experience and certainly more parsimonious.

Lockhead (1972) has discussed similar problems with respect to visual stimuli. His views are worth quoting here, since they apply to speech stimuli as well.

> A distinctive feature must be a set of attributes considered in relation to all stimuli; one cannot have distinctive features in a vacuum.... We must determine the space, the set of relations, and not just the features, if we are to understand pattern recognition. The basic hypothesis is that observers first locate an object in some complex psychological space and then analyze that locus according to the needs of the task.... Perhaps a distinctive feature can be defined as an attribute(s), or the value of an attribute(s), of a stimulus which causes that integral object to be distant from other potential stimuli in the psychological space.... [This] directs attention to the possibility that the relations between attributes (which is another way of saying locus in space) may be processed before the values of the attributes themselves are processed. (Lockhead, 1972: 417-418, his emphasis.)

The prototype model is very much in line with Lockhead's views. It adds the assumption of category prototypes, a concept that has been useful in various other areas of perception (e.g., Posner, 1969; Reed, 1972; Rosch, 1973; Smith, Shoben, and Rips, 1974; Hyman and Frost, 1975) but has been neglected in models of speech perception [except perhaps for the work of the Leningrad group; see Galunov and Chistovich (1966) and Galunov (1968)]. Thus, the prototype model has considerable heuristic value, and much more evidence will have to be collected before it can be confidently rejected. The achievement of the present study lies primarily in the rejection of the overly simple phonetic feature recombination model; its contribution to the evaluation of the prototype model remains modest.

The second important result of the present study is the magnitude of the ear advantages obtained. It suggests that the single-response paradigm, together with the unbiased ear dominance index (Repp, 1976a, 1976b; Repp and Halwes, in preparation) is a powerful method for assessing laterality effects, and that it is probably one step closer towards an optimal dichotic test for diagnostic purposes.

REFERENCES

Blumstein, S. E. (1974) The use and theoretical implications of the dichotic technique for investigating distinctive features. Brain Lang. 1, 337-350.

Cooper, W. E. (1974) Adaptation of phonetic feature analyzers for place of articulation. J. Acoust. Soc. Am. 56, 617-627.

Cooper, W. E. and R. N. Nager. (1975) Perceptuo-motor adaptation to speech: An analysis of bisyllabic utterances and a neural model. J. Acoust. Soc. Am. 58, 256-265.

Cutting, J. E. (1974) Two left-hemisphere mechanisms in speech perception. Percept. Psychophys. 16, 601-612.

Cutting, J. E. (1976) Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. Psychol. Rev. 83, 114-140.

Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation of linguistic feature detectors. Cog. Psychol. 4, 99-109.

Galunov, V. I. (1968) Some aspects of speech perception. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 21, 43-47.

Galunov, V. I. and L. A. Chistovich. (1966) Relationship of motor theory to the general problem of speech recognition (review). Soviet Physics-Acoustics 11, 357-365.

Garner, W. R. (1974) The Processing of Information and Structure (Potomac, Md.: Lawrence Erlbaum Assoc.).

Green, D. M. and J. A. Swets. (1966) Signal Detection Theory and Psychophysics (New York: Wiley).

Haggard, M. P. (1971) Encoding and the REA for speech signals. Quart. J. Exp. Psychol. 23, 34-45.

Halwes, T. G. (1969) Effects of dichotic fusion on the perception of speech. Unpublished Ph.D. dissertation, University of Minnesota.

Hyman, R. and N. H. Frost. (1975) Gradients and schema in pattern recognition, In Attention and Performance V, ed. by P. M. A. Rabbitt and S. Dornic (London: Academic Press), pp. 630-654.

Klatt, D. H. (1973) Voice-onset time, frication and aspiration in word-initial consonant clusters. Quarterly Progress Report (Cambridge, Mass.: M.I.T. Quarterly Progress Report), No. 109, 124-136.

Kuhn, G. M. (1973) The Phi coefficient as an index of ear differences in dichotic listening. Cortex 9, 447-457.
Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.
Lockhead, G. R. (1970) Identification and the form of multidimensional discrimination space. J. Exp. Psychol. 85, 1-10.
Lockhead, G. R. (1972) Processing dimensional stimuli: A note. Psychol. Rev. 79, 410-419.
Miller, J. L. (1975) Properties of feature detectors for speech: Evidence from the effects of selective adaptation on dichotic listening. Percept. Psychophys. 18, 389-397.
Miller, J. L. (1976) Properties of feature detectors for VOT. J. Acoust. Soc. Am., Suppl. 59, S41(A).
Pisoni, D. B. (1975) Dichotic listening and processing phonetic features. In Cognitive Theory, vol. 1, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman, and D. B. Pisoni (Hillsdale, N. J.: Lawrence Erlbaum Assoc.), pp. 79-102.
Pisoni, D. B. and J. Tash. (1974) "Same-different" reaction times to consonants, vowels and syllables. In Research on Speech Perception (Department of Psychology, Indiana University), Progress Report No. 1, 129-139.
Posner, M. I. (1969) Abstraction and the process of recognition. In Psychology of Learning and Motivation, vol. 3, ed. by G. H. Bower and J. T. Spence (New York: Academic Press), pp. 93-100.
Reed, S. K. (1972) Pattern recognition and categorization. Cog. Psychol. 3, 382-407.
Repp, B. H. (1976a) Acoustic determinants of "stimulus dominance" in dichotic competition. J. Acoust. Soc. Am., Suppl. 59, S6(A).
Repp, B. H. (1976b) Identification of dichotic fusions. J. Acoust. Soc. Am. 60, 456-469.
Repp, B. H. (in press) Discrimination of dichotic fusions. Percept. Psychophys.
Repp, B. H. and T. G. Hawes. (in preparation) Measuring laterality effects in dichotic listening. (Copies available from Bruno H. Repp.)
Rosch, E. H. (1973) On the internal structure of perceptual and semantic categories. In Cognitive Development and the Acquisition of Language, ed. by T. M. Moore (New York: Academic Press), pp. 111-144.
Shankweiler, D. and M. Studdert-Kennedy. (1975) A continuum of lateralization for speech perception? Brain Lang. 2, 212-225.
Smith, E. E., E. J. Shoben, and L. J. Rips. (1974) Structure and process in semantic memory: A featural model for semantic decisions. Psychol. Rev. 81, 214-241.
Studdert-Kennedy, M. (in press) Speech perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass (New York: Academic Press).
Studdert-Kennedy, M. and D. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Am. 48, 579-594.
Studdert-Kennedy, M., D. Shankweiler, and D. B. Pisoni. (1972) Auditory and phonetic processes in speech perception: Evidence from a dichotic study. Cog. Psychol. 3, 455-466.
Wood, C. C. and R. S. Day. (1975) Failure of selective attention to phonetic segments in consonant-vowel syllables. Percept. Psychophys. 17, 346-350.
Zlatin, M. A. (1974) Voicing contrasts: Perceptual and productive voice onset time characteristics of adults. J. Acoust. Soc. Am. 56, 981-994.

Distance Measures for Speech Recognition--Psychological and Instrumental*

Paul Mermelstein

ABSTRACT

Perceptual confusion among speech sounds can serve as a guide to the selection of appropriate distance metrics for verification of hypotheses in speech-recognition systems. Known results covering psychological representation of speech sounds are first reviewed. Desirable properties for distance measures for verification are stated, and previously proposed distance metrics for word-recognition are evaluated in this light. This paper reports on one experiment that demonstrates the need for assessing the significance of local differences by any distance metric to be used for verification of syllable-sized hypotheses concerning the speech signal.

INTRODUCTION

Analysis of the continuous speech signal to obtain a phonetic transcription is a significant problem for any speech-understanding system. Speech sounds undergo a complex reorganization of their acoustic properties, from their form when uttered in isolation, to their form in a sentence context. This reorganization is generally accompanied by a loss of information; distinctive differences among sounds become reduced and sometimes disappear altogether.

Analytic segmentation and labeling rules may be constructed to extract the segments of speech that are characterized by unchanging features (Mermelstein, 1975). Due to variations in context and speaker, however, these rules are at best probabilistic in nature, as they only select a highly likely hypothesis concerning the underlying segments. The rules are based on acoustic measurements pertaining only to a short-time interval of the signal in and around the hypothesized segment.

To utilize information from a somewhat larger context, one attempts to verify the analysis-derived hypotheses at the syllable or word level. Word boundaries are not readily apparent in fluent speech; therefore one wants to consider the verification of syllable-sized units. By restricting our analysis to admissible syllables of the language, both those found within words and those

---

*This paper was presented at the Joint Workshop on Pattern Recognition and Artificial Intelligence, Hyannis, Mass., 1-3 June 1976.

Acknowledgment: This work was supported in part by the Advanced Research Projects Agency, Department of Defense.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

spanning word boundaries, we can immediately reject a large number of hypotheses. Additionally, knowing the syllable context, we can utilize predictions concerning the effects of neighboring sounds on each other in order to ascertain whether the data in fact support those hypotheses.

We first review some results concerning human perceptual confusions among speech sounds in order to select an appropriate representation on which to compute distance measures. Next, several desirable properties are cited for a distance metric appropriate for the verification of syllable-length hypotheses. Distance measures previously used for limited word-recognition systems possess these properties to a variable extent. Distance-based recognition is generally inappropriate for selecting one of more than a few hundred distant patterns. For a fixed finite probability of error for any individual membership comparison, the recognition probability tends to zero as the number of patterns is increased. Therefore, we suggest that analysis be used to select only a few reasonable hypotheses concerning the phonetic content of a syllable, and conventional word-recognition techniques be limited to verification of such hypotheses. In order that a metric be appropriate for verification as well as recognition, we require not only that the distance to the correct category be a minimum, but also that such minima lie below a fixed threshold, and distances to incorrect categories lie above that threshold. Finally, we cite a simple experiment whose results emphasize the need for weighting the short-time spectral distances according to the significance of the local differences.

Psychological Distance Representation

Experimental data on confusion among speech sounds by human listeners are available from perception and recall experiments. Miller and Nicely (1955) measured perceptual confusions among single initial consonants under various conditions of noise added to the speech signal. Wickelgren (1966) measured confusion among consonants that were perceived correctly in a serial recall experiment. The confusion patterns were generally similar. Essentially the same feature system could explain the confusions in auditory perception as in short-term memory. Where confusion exists, it can be viewed as the result of selective substitution of features such as voicing, nasality, openness, and place. Similarity among consonants was found to be a monotonic function of the number of features they share. Where confusion among consonant-vowel and vowel-consonant sequences was tested, the order was not significant for vowel errors but was a feature of consonant errors.

Shepard (1972) derived a similarity matrix from the Miller-Nicely confusion data and obtained a spatial representation of the speech sounds. He assumed that similarity is an exponentially decreasing function of interclass distance and minimized the error between the similarity and its distance derived representation,

$$\sum_{i>j} \{S_{ij} - (e^{-bD_{ij}} + c)\}^2$$

$S_{ij} = (p_{ij} + p_{ji})/(p_{ii} + p_{jj})$ is a function of the reported confusion matrix.
D_{ij} is the distance between classes i and j in the spatial representation

recovered, given by $\sqrt{\sum_k (X_{ik} - X_{jk})^2}$, where X_{ik} is the projection of the coordinate of the i^{th} class on the k^{th} orthogonal dimension of the underlying perceptual space. Parameters to be determined are b and c. Over 99 percent of the variance for confusion among 16 consonants was accounted for on the basis of two orthogonal dimensions. These dimensions corresponded roughly to the perceptual features of voicing and a combination of nasality and frication.

This spatial representation is shown in Figure 1. A hierarchical clustering procedure which sequentially clusters sound pairs in the order of their

MATERIAL REMOVED DUE TO COPYRIGHT RESTRICTIONS

Figure 1 removed due to copyright restrictions. (Spatial and hierarchical representation of the perceptual similarity between consonants. From Shepard, 1972, McGraw-Hill, Inc.)

similarity yields the clusters indicated. These clusters roughly correspond to those one derives on the basis of confusions at decreasing levels of signal to noise ratio. There appears to be a good correlation between the similarity values under different noise conditions--decreasing signal to noise increases the confusion among similar sounds.

It is significant to note that the sound space is not uniformly populated. A distance sufficiently large to cross the boundary between /p/ and /k/ is probably not significant for variation among different tokens of /s/. The technique relies on confusion data; therefore, the distance between distinct tokens of members of the same phonemic category is assumed to be zero. Since any continuous instrumental measure must be sensitive to both intercategory and intracategory variation, these results can only be used as a guide to the construction of an appropriate distance metric.

A similar spatial distribution can be achieved for vowel sounds and is given in Figure 2. Although the data are shown in three dimensions, which

---

MATERIAL REMOVED DUE TO COPYRIGHT RESTRICTIONS

Figure 2 removed due to copyright restrictions. (Three dimensional spatial representation for 10 vowel phonemes. From Shepard, 1972, McGraw-Hill, Inc.)

account for 99 percent of the variance, the first two dimensions account for 97 percent. While the principal dimensions correspond roughly to the first two formant frequencies of the vowels, the second dimension appears to be compressed roughly logarithmically with frequency. These results correlate well with known data concerning the spacing of critical bands in the human auditory system--the band within which noise effectively masks a signal of fixed frequency. These critical bands are about equally spaced with frequency below 1000 Hz, increasing logarithmically thereafter. The mel-frequency scale reflects that spacing.

Confusion between vowels and consonants seems quite rare, but no data are available. It is unfortunate that the semivowels and glides were not included in the Miller-Nicely confusion experiments since these would have yielded the most interesting consonant-vowel confusion data.

Compound consonants present additional problems. Despite the close fusion in articulation between the component consonants of a compound, the confusions of the compounds can be explained in terms of the confusion of the components (Pickett, 1958). This result may be due to phonological constraints among the compounds. Since stops and fricatives are relatively rarely confused, the classes of compounds in which they participate will also be rarely confused. Confusion predominates among the stop-liquid compounds in initial and the nasal-stop group in final position.

According to Wickelgren (1966) consonant similarity and vowel similarity can be considered as independent dimensions in syllable recall. However, co-articulation effects modify the acoustic cues for consonants, depending on the syllabic vowel. Therefore the possibility of perceptual interactions between consonant and vowel must be recognized.

Desirable Distance Measure Properties

In view of the above results, a distance measure that models human performance should ideally recognize the phonemes, and construct the distance measure from phoneme confusability data. Failing such recognition, we can at best approximate the peripheral, precategorical aspects of human speech perception behavior.

Let us postulate a set of desirable properties for a distance measure for the verification of syllable-sized segments.

1. The measure should operate on time-aligned versions of the tokens to ensure consonant-to-consonant and vowel-to-vowel comparison. Since syllables have but one prominent vowel, the best aligned tokens can be viewed as those that will minimize vowel-vowel differences as well as differences in the prevocalic and postvocalic position.

2. If the final distance measure is a time integral of some distributed distance function, an appropriate weighting function that assesses the significance of the contributions from the individual short-time segments must be used.

3. The distance measure between tokens should be symmetric, $D(X,Y) = D(Y,X)$.

4. It should be possible to utilize the distance measure to determine phonetic equivalence. If X and Y are phonetically equivalent, but X and Z are not, the $D(X,Y) < D(X,Z)$.

5. Let A,B be parametric representations of two tokens, then $M(A,B) = M(B,A) = (A+B)/2$ is a template for the class (A,B) such that $D(A,M) \leq D(A,B)$ and $D(B,M) \leq D(A,B)$.

Templates are used as compact descriptors for equivalence classes. Consider the class of metrics defined as the weighted sum of elemental metric components for short-time segments. Let P be some space of time-warping transformations such as shown in Figure 3:

$$D(X,Y) = \min_{p(\tau)\varepsilon P} \sum_{\tau} w(\tau) \; d[x(\tau),y(\tau)]$$

where $d(\tau) = d[x(\tau), y(\tau)]$ is an elemental metric component over a short-time segment of the path $p(\tau)$ that maps $1 \leq t_x \leq T_x$, and $1 \leq t_y \leq T_y$ onto τ and $w(\tau)$ is some positive semidefinite weighting function that assesses the significance of the contribution from each element of the path.

Among requirements that we may want to impose on the elemental distance metric between any two short-time segments are

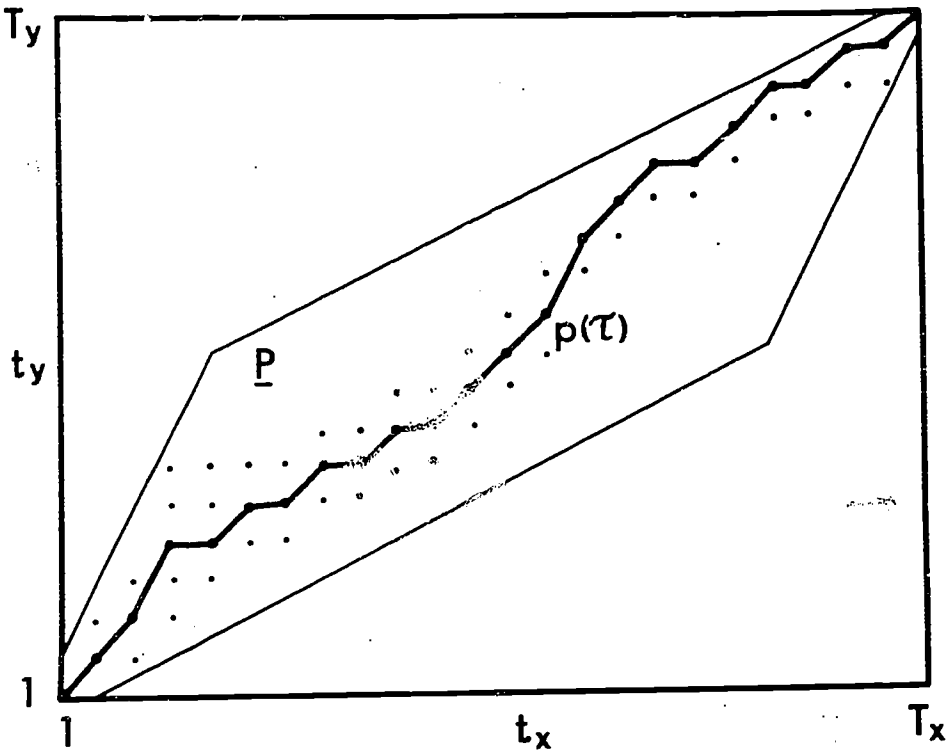
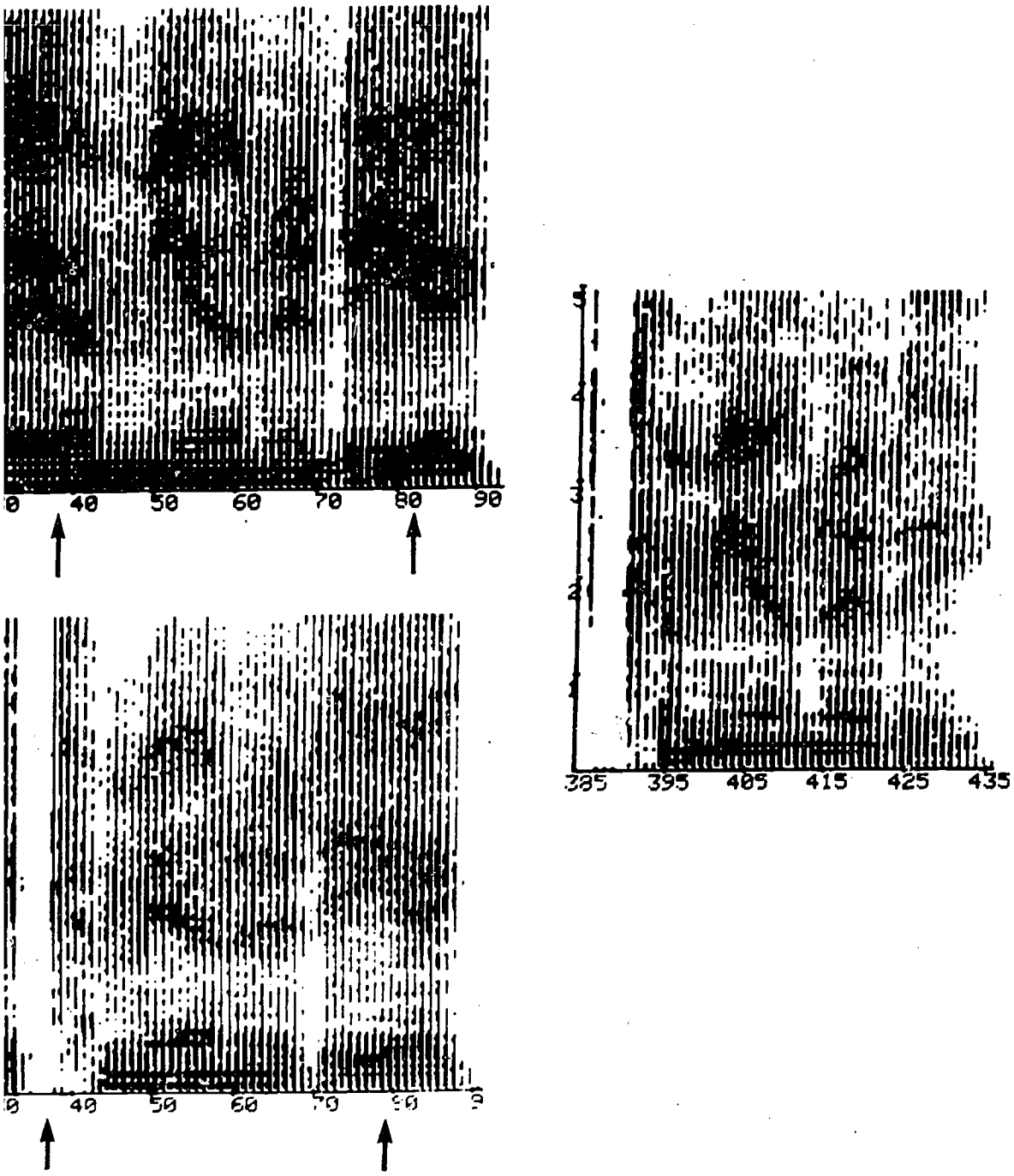


Figure 3: Typical path in the space of time-alignment transformations between two speech segments.



e 4: Spectrograms for the reference words "immunity" (top), "community" (bottom), and unknown (right). Frequency in kHz units, time in increments of 12.8 msec.

97

1. positive semidefinite, $d(x,y) \geq 0$ (this ensures that the global metric is also positive semidefinite),

2. symmetric, $d(x,y) = d(y,x)$,

3. that it satisfies the triangle inequality $d(x,y) + d(y,z) \geq d(x,z)$.

4. that it satisfies a perceptual weighting of the frequency components of the power spectra of the signals. If variation in $s(\omega_1)$, the energy at frequency ω_1, is perceptually more significant than that in $s(\omega_2)$, then $d[x, x + \Delta s\ (\omega_1)] > d[x, x + \Delta s\ (\omega_2)]$.

The need for careful assessment of the significance of spectral variations was realized when we carried out the following experiment (Nye, Copper, and Mermelstein, 1975). Human spectrographic pattern recognizers were asked to match the words of an unknown sentence, presented in spectrographic form, with the same words from a reference library of spectrographic patterns. The reference library was generated by the same speaker, stored in computer retrievable form, and displayed through specification of a list of required features. Since the phonetic transcription of the reference words was not made available, the subjects wide discouraged from using syntax and semantics to assist the pattern matching operation. While the subjects had no problem in rejecting the phonetically dissimilar words, they encountered frequent confusions between similar words. Figure 4 shows the two reference words "community" and "immunity" at left, and the unknown word at the right. In the presence of some uncertainty concerning the word boundary, the disagreement in the unstressed syllable at the top just to the right of the first arrow was accepted by two observers in view of the wide agreement over the rest of the word. The region of significant spectral disagreement between the two extends for no more than 100 msec. Clearly we need a rather sophisticated metric to resolve such distinctions.

Acoustics Based Distance Measures

Let us now examine some distance measures proposed previously in the light of these requirements. Sokoe and Chiba (1971) constructed an Euclidean distance metric on short-time spectral samples obtained from a bank of band-pass filters. When the words were aligned in time through use of a dynamic programming algorithm to minimize the total word-to-word distance, they achieved 99 percent recognition of the 100 two-digit Japanese numbers of five speakers. Klatt (1976) has proposed weighting the spatial distance metric with a function that reflects the increased perceptual importance of differences near the spectral peaks, and reduced perceptual importance of the differences near spectral minima. Itakura (1975) suggested use of the minimum prediction residual as a distance measure for isolated word recognition. This measure computes the ability of the linear predictor that is optimum for the reference-word segment to predict the signal waveform of the target-word segment,

$$d(X/a) = \log (\underline{a}\ \underline{V}\ \underline{a}'/\underline{\hat{a}}\ V\ \underline{a}')$$

That is, the distance between the target segment characterized by process X and the reference segment, having the optimum linear-prediction vector $\underline{a}$, is given by the log-likelihood ratio where $\underline{\hat{a}}$ is the optimum linear predictor of X, and $\underline{V}$ is

the vector of autocorrelation coefficients of X. While this measure can be computed rather quickly from the signal waveform, it is not symmetric between reference and target. To overcome this, Gray and Markel (1975) have suggested a symmetric modification of the linear-prediction residual, namely

$$d_s(X/a) = d(X/a) + d(a/X) .$$

The linear-prediction residual is a measure of the unpredicted signal energy. There is no attempt to assess the significance of the suboptimum prediction of the signal waveform. For some signals even a rough spectrum approximation appears adequate, for others a finer representation is required.

White and Neely (1975) performed a comparative evaluation of the Euclidean spectral distance measure and the one based on the linear-prediction residual. He found them roughly equivalent in terms of performance for recognition of a 36-word and a 91-word vocabulary of one speaker. They concluded that the major improvement over previous results arose from the use of the various dynamic programming algorithms for word alignment. Use of the dynamic programming technique for word recognition was first proposed by Velichko and Zagoruyko (1970).

Atal (1974) has used a non-Euclidean distance measure for speaker recognition, namely

$$d(\underline{\mu}_1, \underline{\mu}_2) = (\underline{\mu}_1 - \underline{\mu}_2)W^{-1} (\underline{\mu}_1 - \underline{\mu}_2)'$$

where the $\underline{\mu}_j$ are parameter vectors to be selected and W is the covariance matrix of $\underline{\mu}$. He explored representations in terms of linear-prediction coefficients, impulse response coefficients, autocorrelation function samples, predictor derived area functions, and cepstral parameters. The cepstral coefficients c_k are related to the linear-prediction parameters by

$$\sum_{k=-n}^{k=n} c_k e^{-jk\theta} = \ln[\sigma/|A(e^{j\theta})|]^2$$

where $1/A(e^{j\theta})$ is the linearly predicted signal spectrum and σ is the rms energy. Among the different parametric representations, the cepstral coefficients gave the highest speaker identification accuracy. Representation in terms of cepstral coefficients has the advantage that a set of coefficients of the same order can be averaged, and the result equals the cepstral representation of the average of the log power spectra (after normalization to unity gain). Use of the covariance matrix normalizes the contributions of the components of the parameter vectors independently of any linear transformations they may undergo.

Bridle and Brown (1974) used a set of 19 weighted spectrum-shape coefficients given by the cosine transform of the outputs of a set of nonuniformly spaced bandpass filters. The filter spacing is chosen to be logarithmic above 1 kHz and the filter bandwidths are increased there as well. We will, therefore, call these the mel-based cepstral parameters. Pols (1971) showed good word recognition results using only the three shape variation components maximally contributing total spectral shape variation. These components resemble the mel-based cepstral parameters rather closely in terms of their frequency variation. The mel-based cepstral parameters have the advantage that generally fewer parameters suffice for an adequate representation of the power spectrum than the

linear-prediction coefficient series. A truncated cepstral representation corresponds to a frequency-smoothed power spectrum, one from which evidence concerning the individual harmonics of the speech signal is missing. To the extent that the spectrum of the excitation signal is invariant between successive voiced segments of the speech signal, the mel-based cepstral measure corresponds to a mel-weighted summation of the difference between the two smoothed vocal tract transfer functions.

Experiments With a Mel-Based Cepstral Distance Measure

I have been concerned with the adequacy of a mel-based cepstral distance measure to discriminate phonetically similar words and syllables. To evaluate the contribution of time-dependent significance functions to an integrated distance measure, I conducted the following experiment: four speakers, two male, two female, recorded one production of each of the twelve phonetically similar words, "stick," "sick," "skit," "spit," "sit," "slit," "strip," "scrip," "skip," "skid," "spick," and "slid" in a reference context "say ___ again." The words were excised from the carrier by listening to a specifiable delimited segment of the signal. Spectra were computed for all the words and reduced to a two-dimensional cepstral respesentation. The respective interword distances were determined for all possible pairs of words by time alignment with Itakura's dynamic algorithm. The unweighted metric used was

$$d(a,b) = \frac{1}{N} \sum_{\tau=1}^{N} \sum_{k=1,2} [C_k^a(\tau) - C_k^b(\tau)]^2$$

$C_k^x(\tau)$, $\tau = 1, \ldots, N$; $x = a,b$; $k = 1,2$ are the time-aligned, two-dimensional, mel-based cepstral coefficient vectors for the two words. Figure 5 shows histograms of the interword distances for the same word spoken by two different speakers, as well as for all other pairs comparing different words spoken by the same or different speakers. The complete overlap between the two comparison categories is surprising. Although the unweighted distance measure is useful to differentiate phonetically distant words, it is clearly not applicable to the discrimination of phonetically similar words.

I next generated templates for each of the words by time warping the words of each speaker onto the one with longest duration using the same dynamic programming algorithm. The mean and variance of the first two cepstral parameters were next computed for the time-aligned versions and used as templates representative of the respective words. Next the weighted distance between each token x and template A was determined using the inverse of the variance for weighting each cepstral coefficient difference, for example,

$$d_w(x,A) = \frac{1}{N_A} \sum_{\tau=1}^{N_A} \sum_{k=1,2} [(C_k^x(\tau) - C_k^A(\tau))/\sigma_k^A(\tau)]^2$$

The time-alignment path, $\tau = 1, \ldots, N_A$ is now a function of the local cepstral variance, $[\sigma_k^A(t)]^2$.

A fixed distance threshold allowed the correct assignment of all but 2 of the 48 tokens to the appropriate word class. The two confusions arose through incorrect assignment of one token of "slit" to "sit" and one token of "spit" to

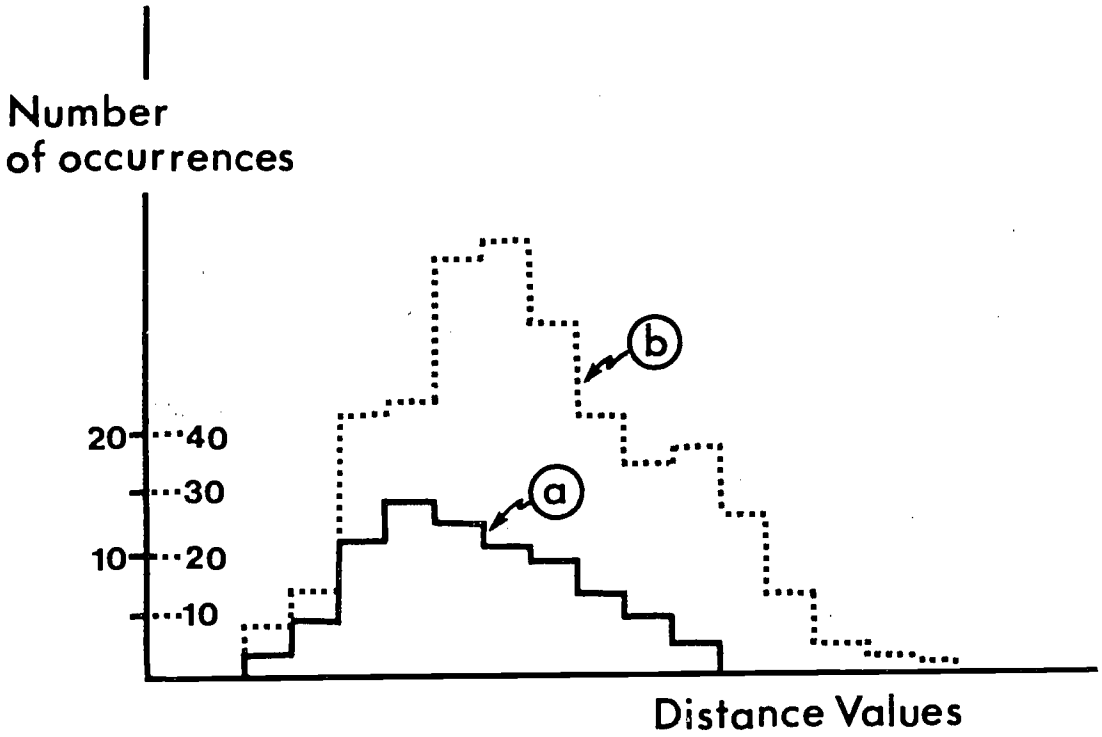


Figure 5: Histograms of computed interword distance values for (a) words from the same category (different speakers), (b) words from different categories (same or different speakers).

"spick." The same tokens were used to generate the template and to test them; therefore, this represents a biased test of discriminability. When I attempted to generate templates from fewer tokens, editing problems near the word edges, such as whether the release of the final stop was properly included, the result was significantly poorer discrimination. Nevertheless, the dramatic difference, as compared to the use of unnormalized distances, underlines the necessity of including appropriate modeling of the significance of the encountered variation of the parameters.

One result of using the inverse of the parameter variance for the weighting function, is to assign more significance to silent segments where the variance was actually zero (assigned a finite nominal value), than to the segments having finite energy. Since all our tokens began with the phoneme /s/, we could not explore the question of the relative weights to be assigned to fricatives and voiced sounds. Presumably, the relative cepstral distance among the class of unvoiced fricatives is larger than that among the vowels. Therefore one would want to tolerate larger differences in fricative regions than in vowellike regions before rejecting a given hypothesis.

A further desirable property of a time dependent weighting function appears to be the assignment of larger weights to regions of high spectral variation than to stationary regions. Otherwise, for steady-state segments the contributions to overall distance are proportional to the durations of the segments. Under those

conditions vowel differences would be overemphasized. No experimental results are as yet available on this point.

Discussion and Conclusions

Synthesis represents an alternative technique for generating the reference templates. Klatt (1975) and Cook (1976) have proposed a word verification procedure based on synthesis of the hypothesized word. Its use offers large potential savings in storage requirements at the costs of a small increment in processing requirements.

The prime motivation of using templates derived from actual productions at this point is the need to establish quantitatively the amount of speaker and context dependent variation for which verification techniques must provide. While synthesis procedures generally give us a perceptually acceptable representative of the class to which the token may be assigned, they provide no information concerning the admissible variation in the individual parameters. As we gain more insight into the relative significance of short-time variations in speech spectra and achieve an ability to model the process adequately, synthesis will undoubtedly become a more cost-effective procedure for the generation of templates. Until that time, however, one must resort to the generation of templates from actual productions in the exploration of hypothesis verification techniques.

Our attempt to utilize insights from speech perception processes as an aid to improved speech verification techniques suffers from an inability to separate the peripheral and central processes in human speech perception. There remains a large gap in our knowledge concerning the transformations that the signal undergoes before the segmental information is extracted. We do not yet have an adequate model of the extent of acceptable variation among tokens that belong to a segmental equivalence class. Nevertheless, known properties of perception may be used to guide us toward perceptually relevant representations of the speech signal. We have some evidence that improved verification results are obtainable by focusing on those representations of the speech signal which have proven to be of interest for human speech perception.

REFERENCES

Atal, B. S. (1974) Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am. 55, 1304-1312.

Bridle, J. S. and M. D. Brown. (1974) An experimental automatic word recognition system. JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England.

Cook, C. (1976) Word verification in a speech understanding system. Conference Record, IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, 553-556. (Available from IEEE, 345 East 47th Street, New York, N. Y. 10017.)

Gray, A. H., Jr. and J. D. Markel. (1975) COSH measure for speech processing. J. Acoust. Soc. Am., Suppl. 58, S97(A).

Itakura, F. (1975) Minimum prediction residual principle applied to speech recognition. IEEE Trans. Acoust. Speech Sig. Proc. ASSP-23, 67-72.

Klatt, D. (1975) Word verification in a speech understanding system. In Speech Recognition, ed. by D. R. Reddy (New York: Academic Press), pp. 321-341.

Klatt, D. (1976) A digital filter bank for spectral matching. Conference Record, IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, 573-575. (Available from IEEE, 345 East 47th Street, New York, N. Y. 10017.)

Mermelstein, P. (1975) A phonetic-context controlled strategy for segmentation and phonetic labeling of speech. IEEE Trans. Acoust. Speech Sig. Proc. ASSP-23, 79-82.

Miller, G. A. and P. T. Nicely. (1955) An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. 27, 338-352.

Nye, P. W., F. S. Cooper, and P. Mermelstein. (1975) Interactive experiments with a Digital Pattern Playback. J. Acoust. Soc. Am., Suppl. 58, S105(A).

Pickett, J. M. (1958) Perception of compound consonants. Lang. Speech 1, 288-304.

Pols, L. C. W. (1971) Real-time recognition of spoken words. IEEE Trans. Computers 20, 972-978.

Sakoe, H. and S. Chiba. (1971) A dynamic-programming approach to continuous speech recognition. Reports of the 7th International Congress on Acoustics, Budapest, 20-C-13, 65-68.

Shepard, R. N. (1972) Psychological representation of speech sounds. In Human Communication, a Unified View, ed. by E. E. David and P. B. Denes (New York: McGraw-Hill), pp. 67-113.

Velichko, V. M. and N. G. Zagaruyko. (1970) Automatic recognition of 200 words. Intl. J. Man-Machine Studies 2, 223-234.

White, G. M. and R. B. Neely. (1975) Speech recognition experiments with linear prediction, bandpass filtering and dynamic programming. IEEE Trans. Acoust. Speech Sig. Proc. ASSP-24, 173-188.

Wickelgren, W. A. (1966) Distinctive features and errors in short-term memory for English consonants. J. Acoust. Soc. Am. 39, 388-398.

Laryngeal Timing in Consonant Distinctions*

Arthur S. Abramson+

ABSTRACT

The concept of voice onset time (VOT) is reviewed with attention to recent misunderstandings. Although it was procedurally convenient and linguistically interesting to focus for some time on word-initial stop consonants, VOT is properly viewed as a particular manifestation of a more general phenomenon, laryngeal timing.

The timing of the valvular action of the larynx may be said to be a physiological mechanism that underlies such acoustic phonetic features as the onset and offset of voice pulsing, intensity of plosive release, amount of aspiration noise, attenuation of the first formant, onset of voice-excited formant transitions, and perturbations of fundamental frequency. These features intersect in various combinations to furnish the phonetic basis of phonologically relevant voicing and aspiration.[1] These features also seem to cover most instances of the vaguely defined term "tense" or "fortis," as applied to consonants.[2]

In our early approach to these matters (Lisker and Abramson, 1964, 1965; Abramson and Lisker, 1965),[3] Leigh Lisker and I focused our attention on stop-consonant distinctions in word-initial position. For our cross-language investigations, this choice made sense, because the richest sets of contrasts are most often found in initial stops. We hypothesized that temporal variations in

---

*Under the editorship of Celia Scully and Gunnar Fant, this is to be published as one of a group of papers based on the seminar on "The Larynx and Language" held at the Eighth International Congress of Phonetic Sciences, Leeds, England, 17-23 August 1975.

+Also of University of Connecticut, Storrs.

---

[1]The phonemic use of voiced aspiration is not fully handled by laryngeal timing alone; it also requires a dimension of glottal aperture.

[2]For example, in English and Spanish.

[3]In this short review, I shall cite mainly work done in collaboration with a few of my colleagues. Certain references needed to document controversial matters and theoretical points will also be given.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

glottal settings for phonation would differentiate most homorganic consonants said to be distinguished phonologically by such features as voicing, aspiration, and tensity. Since, in those days—and to a great extent to this day--it was difficult to make extensive physiological observations of the action of the larynx, we used instrumental displays of the acoustic signal for analysis. The most convenient acoustic index to the closing of the glottis for phonation in initial position was the beginning of regular vertical striations corresponding in a wide-band spectrogram to the quasi-periodic voice pulses of speech. We proposed the term voice onset time (VOT) which we defined as the temporal relation between the onset of glottal pulsing and the release of the initial stop consonant. Specifically, voicing detected before the release, that is, during the stop occlusion, was called voicing lead, while voicing starting after the release was called voicing lag.

By and large, we found that VOT is indeed a very good index to laryngeal timing for the types of homorganic stop consonants in question. The measure provided rather good separation for labial, dental, alveolar, retroflex, and velar stops across a variety of languages that have two or three distinct classes at each place of articulation (Lisker and Abramson, 1964, 1967).[4] Adopting the convention of assigning a timing value of zero to the moment of stop release, negative values to voicing lead, and positive values to voicing lag, we found an essentially trimodal distribution of VOT values for eleven languages that were examined. The first mode centers at -100 msec for a range of values representing voiced unaspirated stops. The second mode centers at +10 msec and corresponds most generally to voiceless unaspirated stops. The third mode centers at +75 msec and corresponds to voiceless aspirated stops. Voicing lag, seemingly occurring for the most part with an open glottis, was regularly accompanied by turbulent excitation of the upper vocal tract (aspiration); in addition, attenuation of the first formant was often visible in the spectrogram [for example; F_1 cutback (Liberman, Delattre, and Cooper, 1958)].

It is clear then that VOT is not defined as an acoustic continuum, although it may be viewed as an articulatory or physiological continuum. In using techniques of speech synthesis to validate our findings, we varied values of voicing lead and voicing lag, with the latter including increments of cutback of the first formant and noise excitation of the upper formants. With stimuli simulating labial, apical, and dorsal CV syllables, we demonstrated the perceptual efficacy of the VOT dimension across a few languages (Abramson and Lisker, 1965, 1970a, 1973; Lisker and Abramson, 1970).

Since this work has stimulated many studies on the part of others, gratifyingly too numerous to list here, it is important to stress that psychological and linguistic discussions of VOT should not give the impression that it is an acoustically simple dimension. It is radically different from many other continua in the literature in that there is an abrupt qualitative discontinuity at the point of stop release. Discussions of special mechanisms for the processing of speech, feature detectors, and other related matters must make it

---

[4]For those with a fourth laryngeal class, see footnote 1. Ejectives, not considered here, may also be said to involve laryngeal timing; however, it is the timing of the tight closing of the vocal folds relative to oral closure that is relevant.

clear that we are dealing with profound psychoacoustic shifts.[5] Voicing lead presents the ear with a low-amplitude, low-frequency spectrum during the initial part of the stimulus. In the absence of lead, we have the sudden full unfolding of the formant pattern for the syllable. For appreciable values of voicing lag, the noise excitation of the formant pattern with its sudden shift to a train of voicing pulses has been shown by our data to be psychoacoustically easier to process.

I fear that our coining of the term voice onset time with its popular acronym VOT, handy as it was for much of our research, has led some colleagues astray. A more appropriate concept is simply that of voice timing--that is, laryngeal timing--which subsumes VOT as a special case. Some scholars, finding VOT very useful for their purely perceptual speculations, have perhaps found little interest in our more psysiological endeavors, which, I think, put our acoustic and perceptual data into proper perspective. Transillumination of the larynx (Lisker, Abramson, Cooper, and Schvey, 1969), fiberoptic observations (Lisker, Sawashima, Abramson, and Cooper, 1970; Sawashima, Abramson, Cooper, and Lisker, 1970; Cooper, Sawashima, Abramson, and Lisker, 1971), and electromyographic recordings combined with fiberoptic observations (Hirose, Lisker, and Abramson, 1972) all show that in running speech the dimension of laryngeal timing is a powerful differentiator of homorganic consonants.

I cannot refrain from alluding to two serious misunderstandings of our concept of VOT. In a purported demonstration of the unimportance of VOT for English initial stops, Winitz, LaRiviere and Herriman (1975) manipulate the onset of voice timing, that is, the beginning of simulated glottal pulsing, as a completely independent variable. Thus, VOT values were altered in real speech recordings in such a way as to yield improbable and even impossible temporal combinations and sequences of voice pulsing and aspiration. Using the resulting "syllables" as stimuli in perception tests, they claimed to show that aspiration is the major cue to voicing distinctions, while VOT is a secondary cue. Clearly these investigators have not grasped the central point that VOT is a physiological dimension which generates a complex set of intersecting, overlapping or even discrete acoustic cues. To take, for example, an original English /du/ and move the consonant burst back so that there is a silent gap of 35 msec between it and the onset of voicing (Winitz, LaRiviere, and Herriman, 1975:Figure 1) and say that this is the equivalent of a VOT value of plus 35 msec in conformity with the conventional model (Lisker and Abramson, 1964; 1971), is simply untenable. An honest use of our concept and test thereof would reveal that such a value of VOT would include turbulent excitation of the upper formants and attenuation of the first formant. These authors (Winitz et al, 1975) have the perfect right to tease out any of the acoustic cues associated here with laryngeal timing, and perhaps others not yet mentioned, and to test the perceptual efficacy of any one of them, as has been done, for example, for the completion of

---

[5] After all, even chinchillas have been trained to perceive VOT differences (Kuhl and Miller, 1975).

formant transitions before or after the onset of voicing by Stevens and Klatt (1974)[6] and the role of fundamental frequency by Haggard, Ambler, and Callow (1970) and Fujimura (1971). Although I readily concede that our terminology needs elaboration to cover the separate acoustic aspects of laryngeal timing,[7] it hardly behooves other investigators to cite us in denigrating VOT without reading closely to see that we mean much more than the mere timing of voice pulsing as a feature orthogonal to other consequences of laryngeal timing.

The other recent instance of misunderstanding I have in mind is a study of voicing and aspiration in Hindi final stop consonants by Bhatia (1976). The author somehow interprets the work on VOT by Lisker and me (1964) and on the related matter of the size of glottal opening by Kim (1970), to predict the neutralization of aspirated and unaspirated stops in final position. To the extent that certain statements by Kim may be vulnerable to Bhatia's criticism, I have no wish to enter into the argument; nevertheless, degrees of glottal opening seem clearly relevant to the final distinctions in Hindi. Except for the special states of the glottis required for such features as murmur and creak, we would argue that the degrees of glottal opening needed for voicing distinctions including voiceless aspiration go with laryngeal timing. I must protest that here too an investigator (Bhatia, 1976) has failed to grasp the point that VOT is an utterance-initial manifestation of the more general phenomenon of laryngeal timing. Indeed, one could go further and argue reasonably that word-final aspiration is an instance of voice onset time. Consider that in an English word like potato the unstressed first syllable is likely to have no voicing at all; that is to say, it is completely aspirated so that VOT proper does not take place until well after the beginning of the second syllable. The result is a voiceless vowel in the first syllable. This is a case, if you will, of a voicing lag so extreme as to deprive a whole syllable of voiced excitation. To produce aspiration in final position, it is necessary to release the stop, thus articulating an unstressed additional syllable (or perhaps "pseudosyllable"). This additional unstressed "syllable" includes a noise-excited vowel appropriate to the vocal tract configuration of the moment. Bhatia's remarks on the predictive powers of phonetic theories (1976:73) are quite gratuitous!

It must not be supposed, one early critic notwithstanding (Kim, 1965), that we have ever claimed that even in utterance-initial position the dimension of laryngeal timing will explain every distinction of homorganic consonants that apparently involves laryngeal features of one sort or another (Lisker and Abramson, 1964, 1971, 1972; Abramson and Lisker, 1970b). VOT may be said to distinguish the voiced aspirated (murmured) stops of such languages as Hindi and Marathi from voiceless stops but certainly not from the voiced unaspirated stops. Here VOT intersects with the kind of glottal opening that permits weak but audible phonation to occur with simultaneous turbulence (Hirose et al., 1972). For the three stop categories of Korean, VOT gives mixed results (Lisker and Abramson, 1964). In word-initial position, two of the categories show a fair amount of overlap although the two of them are well separated from the

---

[6] Lisker (1975) clarifies the matter in experiments in which he pits a literal interpretation of VOT against "voiced transition duration."

[7] Celia Scully: personal communication.

third. These data taken with the rather complicated response patterns of perceptual experiments with VOT (Abramson and Lisker, 1972) led us to conclude that the timing of glottal adjustments relative to supraglottal articulation does contribute to the Korean distinctions, but that there must be another dimension that works with VOT in distinguishing the stop categories. The latter conclusion has been borne out by fiberoptic and electromyographic studies (Kagaya, 1974; Hirose, Lee, and Ushijima, 1974).

Shifts in three extralaryngeal features are commonly adduced in descriptions of the voicing distinction: the volume of the supraglottal tract, stop closure duration in medial position, and vowel duration before a final stop. For phonation to be sustained during an occlusion of the supraglottal vocal tract, it is necessary to prevent equalization of transglottal air pressure. Rothenberg (1968:91) calculates that without any special adjustment this equalization would occur in four msec, which would allow only one or two glottal oscillations. With passive expansion of the pharyngeal walls, voiced closures could be accommodated up to 20-30 msec (pp. 93-94). Active expansion of the pharynx, according to Rothenberg's calculations (pp. 94-99) might give voiced closure durations of 80-90 msec. The even longer voiced closure durations often observed (Lisker and Abramson, 1964) might be explained by incomplete velopharyngeal closure (Rothenberg, 1968:99-106).[8]

Expansion of the pharynx during voiced occlusions has been observed by a number of investigators, at least for citation forms. Apparently because of a conviction that English voiced stops are "lax" and voiceless stops, "tense," some of them, for example Perkell (1969), assumed that the pharyngeal walls expanded passively to help maintain the transglottal air flow for voicing, while the walls were tensed to prevent voicing for the voiceless stops. Electromyographic examination of the relevant musculature (Bell-Berti, 1975; Bell-Berti and Hirose, 1975) reveals that one cannot predict for a given subject whether active or passive control, or some combination of the two, will be exercised for variations in the volume of the supraglottal cavity for voicing distinctions in English. The feature of pharyngeal expansion is linked with laryngeal timing, yet it may be independent. This is not known. For that matter, we do not know how reliable the feature of pharyngeal expansion itself is in running speech.

For some time (Lisker, 1957), it has been known that spectrograms of English medial voiceless stops before unstressed syllables show longer closure durations than do voiced stops, and that manipulation of this feature, providing that no voiced pulsing is present during the closure, furnishes a sufficient cue for the perception of the voicing distinction. Whether this feature is independent or somehow has a dependency relationship with laryngeal timing is not known at this time. Comparison of closure durations across all principal environments, using oral air pressure traces (Lisker, 1972), shows that this feature is likely to be present only in medial poststressed positic . and thus much less useful as an index to the voicing distinction than is laryngeal timing.

The final nonlaryngeal feature to be considered here is the well documented observation that in English and some other languages, vowels preceding final

---

[8] See, for example, data for Sindhi (Nihalani, 1975).

voiced consonants are longer than those preceding final voiceless consonants. This durational difference is perceptually relevant (Denes, 1955; Raphael, 1972). One attempt (Halle and Stevens, 1967) has been made to tie this feature directly to the laryngeal control needed to maintain voicing during consonant closure. Since, however, voicing distinctions in final position are likely to be characterized by differences in laryngeal timing, namely voice offset time, the question arises as to whether the concomitant difference in vowel duration is completely independent of laryngeal timing.

In order to distinguish classes of consonants, many languages make extensive use of the timing of the valvular action of the larynx relative to supraglottal articulation. Certain nonlaryngeal features accompany laryngeal timing, but it remains to be determined whether any of them are controlled by the same mechanism. Laryngeal timing underlies a complex set of interrelated acoustic features any one of which may have perceptual efficacy. The total set varying rather predictably with changes in laryngeal timing has differentiating power in speech perception. The focus of attention for many years on utterance-initial position, reflected in the widely used term Voice Onset Time (VOT), seems to have led some investigators to fail to understand VOT and its acoustic complexity as a positional manifestation of the more general phenomenon of laryngeal timing.

REFERENCES

Abramson, A. S. and L. Lisker. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. Proceedings of the 5th International Congress on Acoustics, A51. (Liege: G. Thone).

Abramson, A. S. and L. Lisker. (1970a) Discriminability along the voicing continuum: Cross-language tests. Proceedings of the 6th International Congress of Phonetic Sciences, Prague, 1967. (Prague: Academia), pp. 569-573.

Abramson, A. S. and L. Lisker. (1970b) Laryngeal behavior, the speech signal and phonological simplicity. Actes du X^{e} Congress Internationale des Linguistes, IV, Bucarest, 1967. (Bucarest: L'Academie), pp. 123-129.

Abramson. A. S. and L. Lisker. (1972) Voice timing in Korean stops. Proceedings of the 7th International Congress of Phonetic Sciences, Montreal 1971. (The Hague: Mouton), pp. 439-446.

Abramson, A. S. and L. Lisker. (1973) Voice-timing perception in Spanish word-initial stops. J. Phonetics 1, 1-8.

Bell-Berti, F. (1975) Control of pharyngeal cavity size for English voiced and voiceless stops. J. Acoust. Soc. Amer. 57, 456-461.

Bell-Berti, F. and H. Hirose. (1975) Palatal activity in voicing distinctions; A simultaneous fiberoptic and electromyographic study. J. Phonetics 3, 69-74.

Bhatia, T. K. (1976) On the predictive role of the recent theories of aspiration. Phonetica 33, 62-74.

Cooper, F. S., M. Sawashima, A. S. Abramson, and L. Lisker. (1971) Looking at the larynx during running speech. Ann. Otol. Rhinol. Laryngol. 80, 678-682.

Denes, P. (1955) Effect of duration on the perception of voicing. J. Acoust. Soc. Amer. 27, 761-764.

Fujimura, O. (1971) Remarks on stop consonants--synthesis experiments and acoustic cues. Form and substance: Phonetic and Linguistic Papers Presented to Eli Fischer-Jørgensen, ed. by L. L. Hammerich, R. Jakobson, and E. Zwirner. (Copenhagen: Akademisk), pp. 221-232.

Haggard, M., S. Ambler, and M. Callow. (1970) Pitch as a voicing cue. J. Acoust. Soc. Amer. 47, 613-617.
Halle, M. and K. N. Stevens. (1967) On the mechanism of glottal vibration for vowels and consonants. Quarterly Progress Report 85. (Research Laboratory of Electronics, MIT), pp. 267-271.
Hirose, H., L. Lisker, and A. S. Abramson. (1972) Physiological aspects of certain laryngeal features in stop production. Haskins Laboratories Status Report on Speech Research SR-31/32, 183-191.
Hirose, H., C. Y. Lee, and Ushijima, T. (1974) Laryngeal control in Korean stop production. J. Phonetics 2, 145-152.
Kagaya, R. (1974) A fiberscopic and acoustic study of the Korean stops, affricatives and fricatives. J. Phonetics 2, 161-180.
Kim, C-W. (1965) On the autonomy of the tensity feature in stop classification (with special reference to Korean stops). Word 21, 339-359.
Kim, C-W. (1970) A theory of aspiration. Phonetica 21, 107-116.
Kuhl, P. K. and J. D. Miller. (1975) Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. Science 190, 69-72.
Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1958) Some cues for the distinction between voiced and voiceless stops in initial position. Lang. Speech 1, 153-167.
Lisker, L. (1957) Closure duration and the intervocalic voiced-voiceless distinction in English. Language 33, 42-49.
Lisker, L. (1972) Stop duration and voicing in English. Papers in Linguistics and Phonetics to the Memory of Pierre Delattre, ed. by A. Valdman. (The Hague: Mouton), pp. 339-343.
Lisker, L. (1975) Is it VOT or a first-formant detector? J. Acoust. Soc. Amer. 57, 1547-1551.
Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.
Lisker, L. and A. S. Abramson. (1965) Stop categorization and voice onset time. Proceedings of the 5th International Congress of Phonetic Sciences, Münster, 1964. (Basel: Karger), pp. 389-391.
Lisker, L. and A. S. Abramson. (1967) Some effects of context on voice onset time in English stops. Lang. Speech 10, 1-28.
Lisker, L. and A. S. Abramson. (1970) The voicing dimension: Some experiments in comparative phonetics. Proceedings of the 6th International Congress of Phonetic Sciences, Prague, 1967. (Prague: Academia), pp. 563-567.
Lisker, L. and A. S. Abramson. (1971) Distinctive features and laryngeal control. Language 47, 767-785.
Lisker, L. and A. S. Abramson. (1972) Glottal modes in consonant distinctions. Proceedings of the 7th International Congress of Phonetic Sciences, Montreal, 1971. (The Hague: Mouton), pp. 366-370.
Lisker, L., A. S. Abramson, F. S.Cooper, and M. H. Schvey. (1969) Transillumination of the larynx in running speech. J. Acoust. Soc. Amer. 45, 1544-1546.
Lisker, L., M. Sawashima, A. S. Abramson, and F. S. Cooper. (1970) Cinegraphic observations of the larynx during voiced and voiceless stops. Haskins Laboratories Status Report on Speech Research SR-21/22, 201-210.
Nihalani, P. (1975) Velopharyngeal opening in the formation of voiced stops in Sindhi. Phonetica 32, 89-102.
Perkell, J. S. (1969) Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study. (Cambridge: MIT Press).

Raphael, L. J. (1972) Preceding vowel duration as a cue to the perception of voicing of word-final consonants in American English. J. Acoust. Soc. Amer. 51, 1296-1303.

Rothenberg, M. (1968) The breath-stream dynamics of simple-released-plosive production. Bibliotheca Phonet. No. 6. (Basel: Karger).

Sawashima, M., A. S. Abramson, F. S. Cooper, and L. Lisker. (1970) Observing laryngeal adjustments during running speech by use of a fiberoptics system. Phonetica 22, 193-201.

Stevens, K. N. and D. H. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Amer. 55, 653-659.

Winitz, H., C. LaRiviere, and E. Herriman. (1975) Variations in VOT for English initial stops. J. Phonetics 3, 41-52.

Phonetic Aspects of Time and Timing*

Leigh Lisker+

ABSTRACT

By a definition narrow enough to exclude acoustic and physiological aspects of speech behavior, phonetics is reduced to the descriptive practice of linguists, whose judgments on the physical nature of a speech signal are primarily auditory and sympathetic proprioceptive. These judgments are for the most part embodied in a special alphabet of indeterminate size, each element of which is defined with reference to some particular state of the vocal apparatus. In general, a dimension of time is not included in the set of auditory and articulatory properties by which the different states are specified. Since, in all but a negligible number of cases, speech signals are said to not involve a single state of the vocal apparatus, but rather a sequence of such states, this sequential ordering is explicit recognition of a temporal dimension. But the time-ordered elements are themselves "timeless" unless the linguist determines that varying the duration of one or more of them serves to signal a semantic--that is, a linguistic--distinction. At this point, one of the two segments said to differ significantly in duration will often be judged to have a duration "inherently" determined by its other properties, while the other will be characterized as "long" or even "overlong." Aside from duration as a property ascribed to the segments constituting a speech signal, there are temporal aspects of speech that are less often given an explicit representation in the linguist's transcription; these are at best indirectly indicated by the so-called "junctural" marks and stress markers. One temporal aspect of speech that is regularly ignored is the feature of rate of articulation, for within certain ill-defined limits speech tempo is ad libitem.

Let me begin with a preamble to explain my understanding of "phonetic aspects of time and timing," in the present context. That understanding is to a considerable degree determined by a factor that is itself temporal, or at least temporal at one remove. I have in mind the spacial arrangement of the discussion titles on the program I was given, and my belief that it follows the conventions of written English and thus signals our chairman's wish that I speak

---

*This paper was presented by invitation of the 100th meeting of the American Speech and Hearing Association, Washington, D.C., 21-24 November 1975.

+Also University of Pennsylvania, Philadelphia.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

first, with Katherine Harris, Dennis Klatt, and Peter MacNeilage to follow in that order. Therefore, I supposed, in preparing for this morning's business, that our discussion of the temporal organization of speech activity would begin with a consideration of certain aspects to be called "phonetic" and then go on to physiological and acoustical data and theories bearing on our topic. This order seems to imply that physiological and acoustical data comprise what are in some sense nonphonetic aspects of the speech process, and while I prefer to think of phonetics as deserving a much broader definition, it is both convenient here, and regrettably close to the general practice of some language scholars in discussing language behavior, to restrict the scope of my contribution so as to exclude in particular the subjects which Katherine Harris and Dennis Klatt will be addressing. That leaves the perceptual aspect for me to talk about--itself a broad enough subject to include a good deal that many of us might want to exclude from phonetics.

Since the scholarly caste that has for longest concerned itself with speech is the one of linguists, even though some who are called linguists would deny the study of speech activity a place in their discipline, I will for now take as reports on the phonetic aspects of speech timing those observations on the temporal properties of speech that linguists include, more or less systematically, in their language descriptions. Such observations, like most others referring to physical properties of a language in its speech guise, reflect judgments by the observing linguist as to the physical attributes characterizing speech events as readings of particular strings of linguistic items. Physical properties are most often defined in articulatory terms, sometimes in acoustic, but the judgments are almost entirely based on auditory input without overt reliance on any observational data obtained under laboratory conditions. That linguists' phonetic judgments are to an extent based on such data seems undeniable, but it is not usual to find them informed by a knowledge of the latest laboratory findings. This is understandable when we remember that many linguists are not primarily interested in precise physical descriptions, but rather in devising spelling systems that meet certain criteria, only one of which is that its letters bear a statable relation to physically describable aspects of the classes of speech signals they are designed to represent.

However, this does not make the linguist's phonetic transcription a fully explicit physical description. First of all, it represents speech by a linear array of discrete letters, so that as description it misrepresents speech in a serious way. Second, the physical properties represented by the transcription are primarily those to which distinctive function is attributed by the linguist; if some others are represented as well, this is, as Bloomfield (1933) put it, "due merely to chance observations...by an observer with a good ear," exercising a skill of "little scientific value." The linguist's representation embodies a partial physical description, but despite a possible implication of Bloomfield's comment, the disclaimer of completeness is no gesture of modesty. The linguist claims to know, from observing speech behavior in the interview situation, just what it is in the physical signal that the native speaker-hearer must produce and attend to in order that the signal be correctly interpreted. The incompleteness of the physical specification is dictated by the linguist's assertion that not all features of the signal and the signal-generating activity are linguistically significant, and that the linguist's technique of observation and analysis suffices to identify those features that are. In one undoubtedly influential view, that of Chomsky and Halle (1968), it is asserted that it is indeed linguistically irrelevant whether the linguist's phonetic statements correspond to physically

attested fact; in their Sound Pattern of English a hypothetical speaker-hearer is invoked whose beliefs concerning the phonetic properties of his language are what a phonetic transcription should represent. We may be permitted to note, though only in passing, that the ideal speaker-hearer whose phonetic intuitions are to be represented seems to be very well aware of the acoustic and physiological studies to be found in MIT's Quarterly Progress Reports, and may even bear a suspicious resemblance to one of the authors of The Sound Pattern of English. It might well be the case, therefore, that the phonetic notions of this speaker-hearer are not immutable.

At the heart of the linguist's practice of phonetic transcription, and serving as the principal bearer of his phonetic judgments or assessments of the ideal speaker-hearer's intuitions, is an alphabet of unknown but possibly finite size, to each letter of which is assigned a function as the referential of a physically defined set of vocal-tract configurations or its acoustic consequences. In defining the value of each letter of this alphabet, reference is made to a smaller set of parameters by which the state of the vocal tract is to some degree specified. In using this alphabet to spell a speech signal, successive vocal-tract configurations are identified and appropriate letters are arranged in a linear left-to-right order, which corresponds to the temporal order of the observed vocal-tract states. Each state represented has a temporal-order relation to every other state represented by the letter sequence, and the expression of this temporal relation is obligatory. This is trivially so because the only allowed spacial relation between letters is either left or right placement. Despite the fact that the letters stand for incomplete specifications of vocal-tract shape, no two of them may be simultaneously applicable, each being appropriate for a unique and unspecified time interval. Or, if you prefer, the duration is specified as being equal to that of one "segment," the duration of which is not further specified. Presumably each vocal-tract state represented is maintained over the duration of the segment, though it is not clear that this is necessarily the claim in all cases. Only the order of succession of the different states is represented--of necessity--with one segment succeeding another without overlap and without the intervention between any two immediate neighbors of a third requiring representation on linguistic grounds.

In addition to the letters that represent temporal segments, there are others that have, along with grammatical and intonational meanings, some significance as temporal markers. These are the several so-called juncture signs, as well as those indicating levels of stress. The juncture marks, which correspond very roughly to word-space and the punctuation marks of standard orthography, indicate places in the temporal sequence where, together with other phenomena, there may also occur ritardandos and even brief pauses, especially if they coincide with certain grammatical boundaries. But none of these so-called suprasegmental indicators is exclusively or even primarily temporal in reference, and demonstrations of the need to employ them in phonetic transcription generally focus on variations in pitch and loudness. Marks for stress, which for many linguists mean relative loudness, also have secondary temporal meaning; the presence of a mark of high stress usually can be taken to imply a local increase in segment durations, and, at least for English, the intervals between successive high stresses in a speech stretch are said to be of roughly constant duration. Thus, the placement of high stress marks may be said to govern the relative tempo with which the segments are produced within the utterance, in the same way that the vertical lines marking off the measures of musical notation

tell us that all the notes of one measure are to be performed within a time span equal to that occupied by all the notes within any other measure in the same text. Modern musical notation is more explicit on the matter of timing, of course, and stress placement in musical performance is not rigidly tied to the measure, but some phoneticians occasionally make use of the musical measure to represent temporal regularities observed in speech. In the view of many linguists, however, such regularities are not distinctive in language, and hence have no place in a phonetic transcription, however important they might be for the global characterization of the phonetic properties of speech generally, or of one language as against others. Except insofar as juncture and stress markers provide some guidance to tempo, the task of performing a piece of phonetic transcription is very like that of the musician asked to sight-read an unfamiliar piece from a medieval neumatic score, which indicates nothing of the individual notes but their relative pitch and sequencing. The lack of explicit timing information or instructions has its advantages for both kinds of performance, allowing scope for individual variety of expression. For the performers of speech and music the freedom of choice implied by the notations is probably wide enough to permit readings of the same score that are different enough to convey different messages to a listener. For the musician a notation that fails to specify segment duration allows one kind of temporal latitude if the musician is a flutist--the segments can be given durations at will. For the slide-trombonist segment durations are also ad libitum, and there is the additional freedom to determine how rapidly to shift from one pitch to the next in glissando playing. Producing speech is more like playing the trombone than the flute, and phonetic transcription does not prescribe how rapidly the shift from one vocal-tract state to the next is to be accomplished. I have probably pushed the analogy much too far, for it is fair to object that musical notation is a set of instructions for performance more than a description, while phonetic transcription is more a description than a performance. As a set of instructions, the phonetic transcription will have an adequacy that depends, I suspect, less on its degree of specificity than on whether or not the "score" it presents is familiar to the reader. Even if the score as a whole is novel, it must be made up of parts that <u>are</u> familiar if it is to be performed correctly. At the very least, the reader must be a practiced producer of fluent speech in order to implement the score as intended by the transcriber.

As a model of speech, the linguist's graphical representation suffers from inadequacies that are well-known: a speech signal does <u>not</u> consist of a sequence of sounds, each fixed for some unspecified duration and separated from its closest neighbors by intervals of near-zero duration; but it is perhaps unfair to charge the linguist with responsibility for such a model merely because his transcription practice seems to presuppose it. In fact, likely enough, the linguist is well aware that it is wide of the mark, and is only too ready to accept the contrary view of speech as a process, everywhere continuous, which possesses no properties that provide a physical basis for segmentation. The static definitions of vocal-tract states that he provides as interpretations of the transcription represent, then, outputs of a particular kind of sampling of this continuous signal, where the number of sampling points is determined by the number of perceived "change points" in the signal, but is pretty much independent of duration. In short, the transcription is the output of a special kind of "A to D" converter whose sampling rate is not temporally specified, the interval from one sampling point to the next depending roughly on when a perceived change in signal quality comes along. Instead of supposing the speech signal to consist of a succession of states, each maintained for some finite

time interval corresponding to the segment, one can instead say that each segment or letter of the transcription represents a state of the vocal tract which must be achieved or approximated within some time interval, and that this interval, though not necessarily the state which characterizes it, has both a finite duration and a specified place in the temporal sequence of states. The duration over which the state characterizing the segment is maintained may or may not be as great as the total duration of the segment, whatever that might be defined to be; the linguist as auditor will order segments with respect to duration, independently of what the laboratory phonetician may say about the duration over which the specified vocal-tract state is maintained. Because in fluent speech it is not unusual for that duration to be close to zero, it seems clear that we cannot hope to account for the linguist's judgments (and those of the rest of us as well) of segment duration simply by measuring the durations of steady-state intervals that might be discovered here and there in the speech signal. Since a good deal of the literature on speech timing is devoted to reporting durations of phonetic segments--when, in fact, what is being talked about are durations measured between acoustically specified change-points in the speech signals--a close relation between these measurements and the listener's judgments of segment duration must be established before those measurements can be claimed to reflect phonetic aspects of speech activity, at least in the narrow definition of phonetics I am assuming at the moment. In other words, before we can justify referring to durations between physically specified events as equivalent to vowel durations, for example, we must do what the psychophysicists did to establish the nature of the relation between pitch and fundamental frequency or to connect loudness with sound pressure level and frequency. In short, we must confront our old friend the segmentation problem. As we know, this is not so much a question of how to segment a signal, which is everywhere continuous, but rather where to cut the signal, amply possessed of discontinuities, so that the pieces derived can be claimed to correspond reasonably to the listener's segments.

Let us look now at the linguist's representation of speech as a sequence of segments, defined by reference to states either aimed at or manifested by the vocal tract or the homunculus that runs it, with segment durations not specified. This reticence as to the temporal dimension of the segment is tacit admission of the freedom to perform what is linguistically the same speech piece with tempos varying over a considerable range; moreover, no claim is made that the relative durations of segments are constant with changes of tempo (Gaitenby, 1965). But is it in fact true that relative duration is never specified by the linguist's description? Of course not. In the description of some languages, the linguist finds it useful to distinguish members of a particular phonetic class with respect to a temporal dimension; for example, Thai is said to distinguish between short and long vowels (Abramson, 1962); for Estonian both vowels and stop consonants come in three grades of duration (Lehiste, 1970); English vowels are either short and lax or long and tense. Where a difference in length is considered to be distinctive, the linguist may elect to represent the longer of a pair as a sequence of two like segments, thus by implication recognizing that a single segment possesses one unit of duration. Sometimes, however, a special kind of segment is devised, whose only characteristic is that it has the length of one segment, all other properties being given by the specification of an immediate neighbor, usually the one directly preceding it. The long vowel or consonant involves, then, a particular kind of reduplication. Whether the observation that a particular vocal-tract state is maintained sometimes for a longer interval and sometimes for a shorter one to be represented by one spelling device or another, has most often been decided by criteria not primarily phonetic in nature.

(Sometimes there may be some phonetic basis for asserting that the extra-long duration of an articulatory position must be analyzed as a sequence of repeated gestures.)

Uncertainty as to the number of segments over which a single position is maintained is not the only problem encountered in dealing with a temporal dimension at the segmental level; the same uncertainty may arise in connection with the evaluation of what are clearly recognizable sequences of--at least--one level of phonetic description. The notorious example of this is the case of stop-fricative sequences that may be accorded the status of one-segment-long affricates if there seem to be strong phonotactic (that is, distributional) reasons to do so, in which case an especially close temporal relation between its sequential components may also be discovered. There are other such examples: Are the so-called "prenasalized" stops of certain west African languages "really" one or two segments? Are the Russian palatalized consonants sequences of consonant and /y/? Sometimes, but apparently not very often, there is a genuine convergence of phonetic and phonotactic considerations; in Polish two linguistically distinct stop-fricative sequences differ phonetically in ways that allow some justification for calling one of them a single segment and the other a sequence. Thus it would appear that recognizably different vocal-tract states in immediate succession are not invariably allotted to two segments with only one possible temporal relation; that relation may be characterized as one of "close" or "open" transition, or, in the case of vocal-consonant sequences, "close" versus "loose nexus." Now perhaps we should be inclined to look for and find differences in degree of coarticulation to support a particular answer to the question of "one segment or two?"

Apart from cases where the linguist is forced to recognize a temporal feature because it appears to play a linguistically distinctive role quite like the features by which vocal-tract shape is specified, there are occasions where contextually conditioned variations in segment duration are recognized. The greater duration of vowels preceding voiced stops is marked in phonetic transcription, but that added duration (as compared with the durations of the same vowels before voiceless consonants) is not said to constitute another segment, and both the linguist and the phonetician are motivated to discover some basis, phonetic in the broad sense, for considering it to be a consequence of coarticulation. Similarly, the durational difference between the English vowels /ɪ,ʊ/ and /i,u/ is ascribed to the laxness of the first pair as contrasted with the tenseness of the second. Similarly, the brevity of the apical flap of American English is a consequence of the small force of articulation exerted in its production. Some kinds of temporal variation at the level of the segment that have been reported appear to have escaped attention; for example, the greater durations of initial fricatives as compared to final, or the greater durations of final nasals as compared to initial.

Observations of this last kind, which relate relative duration to position within the segment sequence, are in effect, assertions that there must be postulated units larger than the individual segment for which temporal regularities may be stated. The smallest of these is the syllable (only phoneticians who look at physiological and acoustic data worry about the organization of consonant-vowel and vowel-consonant sequences), a unit whose usefulness in phonetic description is acknowledged in the same measure as its resistance to definition is deplored. Linguists tend to solve this problem by believing in the syllable as a phonotactic unit with no phonetic standing, while phoneticians incline to

describe it as the basic element of speech organization. In this view the positional variation to which a phonetic unit conforms is, first of all, that of position within the syllable. In fact, it would seem that the durations of the segments composing a single syllable are changed sufficiently from their hypothesized "inherent" durations for the syllable to be the elementary temporal unit, with "inherent" or baseline durations assigned to segments defined purely with respect to their status within this unit. This, for the linguist, is so far from being a controversial statement that I think I might justly be charged with beating a horse that was stillborn; no linguist's discussion of speech timing has ever proposed a direct relation between utterance duration and the number of segments composing it. But a fairly direct relation between duration and syllable number appears to be intuitively acceptable, whether or not linguists make an explicit statement on the matter. The acceptance of this relation underlies the practice of defining the somewhat elusive quality of speech that we presume is chiefly temporal in nature, namely speech tempo, as corresponding more to a measure of syllables than to segments per unit time.

The same belief--I would suppose--underlies the distinction made between languages like Spanish, which exhibit this feature of "syllable timing," and languages such as English, whose contrary tendency to "stress timing" seems to require explanation as a departure from the expected. In English, we are told, the constant duration intervals into which an utterance can be analyzed are marked by stress (as has already been mentioned). In effect, this says that the durations of utterances are determined by syllable count, but not all syllables count. So far as I know, however, no one has proposed that speech tempo for English be equated with a measure of the duration separating adjacent stressed syllables. What has sometimes been reported (and this makes such a measure less appealing) is that syllables that are stressed at one tempo may be produced with noticeably reduced stress when tempo is increased, suggesting a tendency to keep interstress durations constant over a range of tempos. Of course, with all the importance that has been ascribed to the syllable as a unit of speech organization, both in production and in perception, it is remarkable that the linguist's writing system fails to represent this unit any more directly than do the more widely known alphabetic orthographies (and I suspect it would create at least as large a class of problem readers if put to more general use). Perhaps the fact that linguists have followed the alphabetic rather than the syllabic model in their writing practice comes from the general exclusion of temporal aspects in specifying speech, but it does seem odd, nevertheless, that a fundamental unit is not explicitly represented.

We come, finally, to the aspect of speech referred to as its "rhythmic" quality, which everyone seems to agree is an all-pervasive feature. From time to time linguists appeal to rhythm as a factor that determines stress placement in the case of, for example, lexical items like fourteenth, whose stress contour is variable with context; and linguists have sometimes, for example, characterized languages as "machine-gun-like" in effect. The basis for this conviction that we all share--namely, that speech can be described as rhythmic and that it is profitable to discuss the temporal organization of the process without first deciding whether any exists--doesn't seem so obvious as to be undeserving of a final remark. It is this; if all speech is rhythmic, it is certainly true that some speech is more obviously rhythmic than other, the most well-regulated speech being the perfectly metrical performance of a child's chant or poetry reading. If prose speech differs from these in temporal organization, is the difference one of kind or degree of regularity in timing? The poet's art bends speech to

aesthetic effect; some of the stuff it fashions probably is unchanged from nature (that is, the phoneme stock), but some is creative transformation or even possibly additive. How much of our conviction that rhythm is a characteristic of natural speech represents a metricization, and how much a metrification, of the object of all our attention? I wish I might end on the note of that ringing question, but more soberly suppose that we shall learn, from studies that examine data and not just the sometimes stray observations of the linguist--that speech activity may be described as at least, or at best, "quasi-rhythmic" in nature.

REFERENCES

Abramson, A. S. (1962) The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments (Bloomington: Indiana University Research Center in Anthropology, Folklore, and Linguistics), Publication 20.

Bloomfield, L. (1933) Language (New York: Henry Holt).

Chomsky, N. and M. Halle. (1968) The Sound Pattern of English (New York: Harper & Row).

Gaitenby, J. H. (1965) The elastic word. Haskins Laboratories Status Report on Speech Research SR-2, 3.1.

Lehiste, I. (1970) Suprasegmentals (Cambridge, Mass.: MIT Press), pp. 45-48.

Static and Dynamic Acoustic Cues in Distinctive Tones*

Arthur S. Abramson+

ABSTRACT

It is conventional to classify phonemic tones into dynamic or contour tones and static or level tones. The perceptual relevance of this impressionistic dichotomy is considered here for Central Thai, which has two dynamic tones (falling and rising pitches) and three static tones (high, mid, and low). A fundamental-frequency range appropriate to an adult male voice was used to synthesize three series of tonal variants on a syllable type available for five tonally differentiated words: (1) sixteen F_0 levels at intervals of 4 Hz, (2) sixteen F_0 movements from a mid origin to end points ranging from top to bottom of the range in steps of 4 Hz, and (3) seventeen variants rising from the bottom to end points from top to bottom in steps of 4 Hz. The stimuli were played to native speakers for identification. The results indicate that level variants contain sufficient cues for identification as static tones but with considerable overlap. Identification, however, is enhanced by slow F_0 movement. Rapid F_0 movement is required for dynamic tones. Although imprecise, the typological dichotomy is useful.

In a tone language, part of the specification of each morpheme or word is a distinctive pitch pattern. Although some tones may have additional phonetic features,[1] the major characteristics of a tone system are fundamental-frequency states and movements.

Some linguists refer to level tones, which are heard as having no pitch movement, and gliding tones which audibly rise or fall (Pike, 1948). In

---

*This is a slightly revised version of a paper presented at the 91st meeting of the Acoustical Society of America, Washington, D.C., 4-9 April 1976.

+Also University of Connecticut, Storrs.

Acknowledgment: While most of the analysis of data was performed at Haskins Laboratories, the data themselves were collected while the author was on sabbatical leave in Thailand on research fellowships from the American Council of Learned Societies and the Ford Foundation Southeast Asia Fellowship Program. I gratefully acknowledge the hospitality of Dr. Udom Warotamasikkhadit, Dean, the Faculty of Humanities, Ramkhamhaeng University, and Mrs. Mayuri Sukwiwat, Director, the Central Institute of English Language, both in Bangkok.

[1]Example: creaky voice.

phonological analysis, the question may arise as to whether glides should be treated simply as whole pitch movements or as movements between level tones that are otherwise present in the system (Gandour, 1975). Here I am more interested in the validity or usefulness of the distinction between gliding or dynamic tones and level or static tones. The question is examined in Thai, the official language of Thailand.

Some years ago I published typical fundamental-frequency contours of the five tones of Thai, as shown in Figure 1 (Abramson, 1962). The tones are con-

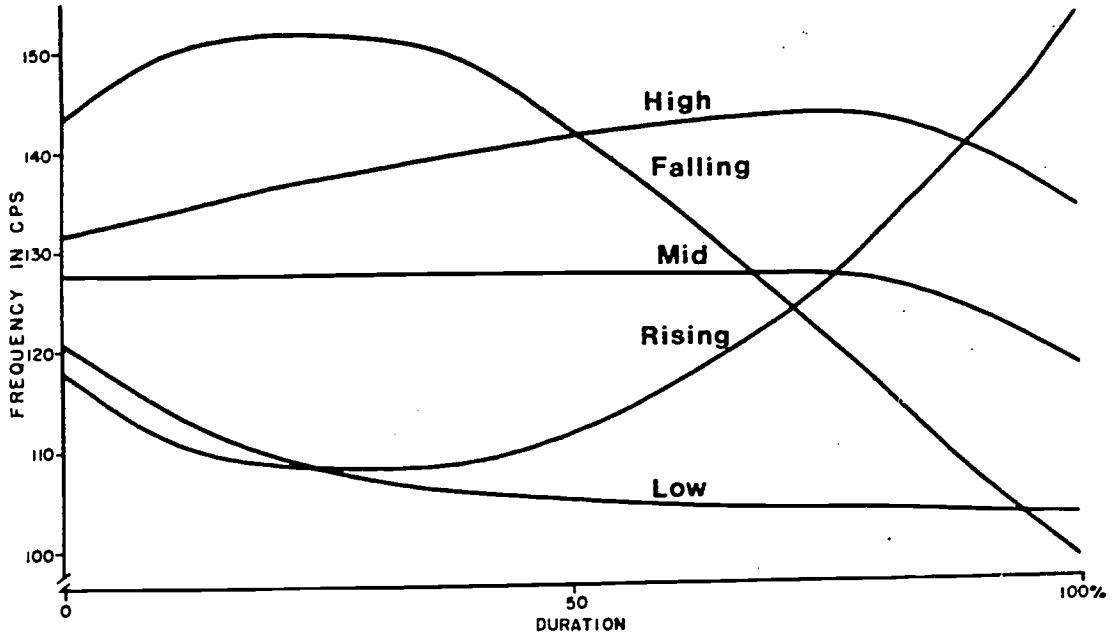


Figure 1: Average F_0 contours of the tones of Thai on long vowels (from Abramson, 1962:Figure 3.6).

ventionally labeled from top to bottom: high, falling, mid, rising, and low. Perceptual experiments with synthetic speech showed that these contours carried sufficient information for high intelligibility in the labeling of monosyllabic words. More recent experiments with the present formant synthesizer of Haskins Laboratories have again demonstrated the sufficiency of these contours (Abramson, 1975a). Moreover, these findings provide a baseline for the experiments to be discussed here. Note that all the tones show at least some movement. The only one that may really be level is the mid tone because its final drop appears to be an intonational phenomenon before a pause.[2] In the experiments, I sought a basis for a division between dynamic and static tones in these curves. The falling and rising tones with their abrupt changes in frequency showed considerably

[2]Speakers of Thai may find a prepausal mid tone without a final drop abnormal, but they identify such a contour nearly as well as the normal one (Abramson, 1975a).

more movement than the others. I labeled the falling and rising tones dynamic and the high, mid, and low tones, static.

In the past few years, further acoustic analysis of the Thai tones (Erickson, 1974, 1976; Abramson, 1975b) has suggested that, especially in running speech, the static tones are not very different from the dynamic tones. The high tone can be described as a high rising tone, while the rising tone can be described as a low rising tone. The low tone tends to fall to the bottom of the speaker's voice range and stay there, although this fall starts at a somewhat lower point than of the falling tone. It is only the mid tone that does not make extreme excursions into the high and low regions of the voice range, although it seldom has the ideal level shape of Figure 1. The following three experiments are intended to shed light on the perceptual validity of the distinction between static and dynamic tones.

A syllable of the type [kha:] was prepared on the Haskins Laboratories formant synthesizer. Sixteen variants were made by superimposing sixteen level fundamental-frequency trajectories ranging from 152 Hz down to 92 Hz in steps of 4 Hz. Each stimulus had a flat amplitude except for a slight rise at the beginning and a slight fall at the end. In Test 1, these were played in several randomizations to 37 native speakers of Thai for identification as one of five possible words.[3] The question considered was the following: Do fundamental-frequency levels carry enough information for identification of the static tones, or must there be some movement for acceptability? The results in Figure 2 show that only the three static tones are used as response categories. Note that nowhere is 100 percent identification reached. A peak of 90 percent for the low tone is about the same as the peak shown in the baseline test (Abramson, 1975a) by the same subjects for the typical low tone displayed in Figure 1. The high tone at the left reaches a peak of only 88 percent compared with 98 percent for the typical high tone in the baseline test. The mid tone in the middle reaches 73 percent as compared with 82 percent in the baseline test.

It is also true that all three tones elicit responses throughout the range. Most of the latter effect was caused by three subjects who used only two labeling categories, high and low or mid and low. Even in isolated monosyllables, then, flat fundamental-frequency trajectories can elicit static-tone responses. For this to happen in natural speech, there must be some auditory accommodation to the speaker's pitch range as well as to the immediate tonal context. At the time of this test, the subjects had become used to the voice and frequency range of the synthesizer. Lack of F_0 movement did cause some confusion for the subjects, and for three of them it was rather disrupting. It is not surprising that the dynamic tones were not used as response categories.

In Figure 3 we see the tonal variants used in Test 2. They all start from a common mid origin and end at the same points as in Test 1. I wondered whether the static-tone responses would be increased by the moderate amount of movement in most of these variants and at the same time, whether at least the extreme values in the continuum would yield mainly dynamic responses. These stimuli

---

[3]The capable and efficient selection and supervision of the test subjects by Miss Panit Chotibut of the Faculty of Humanities, Ramkhamhaeng University, is much appreciated. The subjects were college students who were native speakers of the Central Thai dialect of Bangkok and its environs.

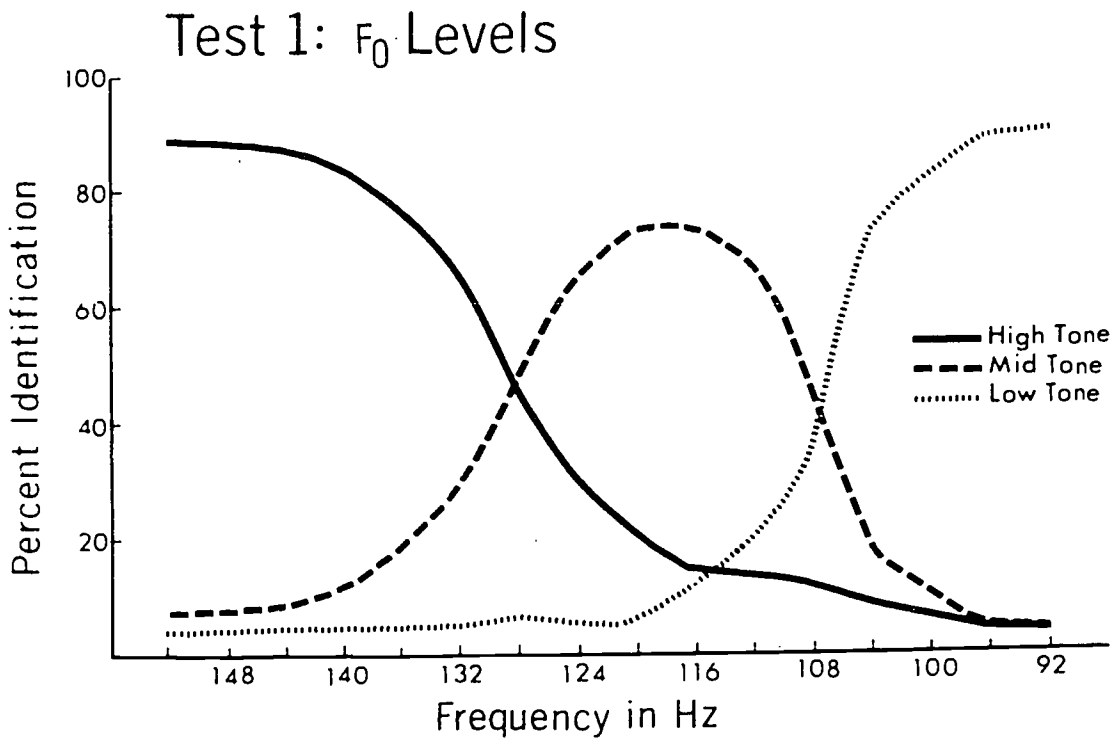


Figure 2: Identification functions for fundamental-frequency levels as static tones.

were played to 31 of the original subjects, and the results are shown in Figure 4. A few stimuli at either end do indeed yield dynamic responses, but no greater than a peak of almost 14 percent for the rising tone at the high end, and almost 5 percent for the falling tone at the low end. Otherwise, the static tones are again the predominant responses. Except for the low tone, there is somewhat better labeling here. The high tone goes from 88 percent in Test 1 to 94 percent in Test 2, and the mid tone improves from 73 percent to 84 percent. In fact, it is a slightly downward movement from 120 to 116 Hz that yields 84 percent,[4] while the flat variant at 120 Hz yields only 72 percent![5] It seems safe to say that fundamental-frequency movements increase the acceptability of synthesized syllables as static tones. For the low tone, a more appropriate movement would start somewhat lower in the voice range.

In Figure 5 we see the variants for Test 3. All the variants start from a low origin at 90 Hz and reach the same end points as before except for a flat

[4]This should be compared with the 82 percent for the mid tone of the baseline test (Abramson, 1975a). That stimulus did not slope downward from its onset as does the one described for Test 2 here, but it did have a final drop.

[5]Compare it with the flat variant at 120 Hz in Test 1 which yielded 73 percent.

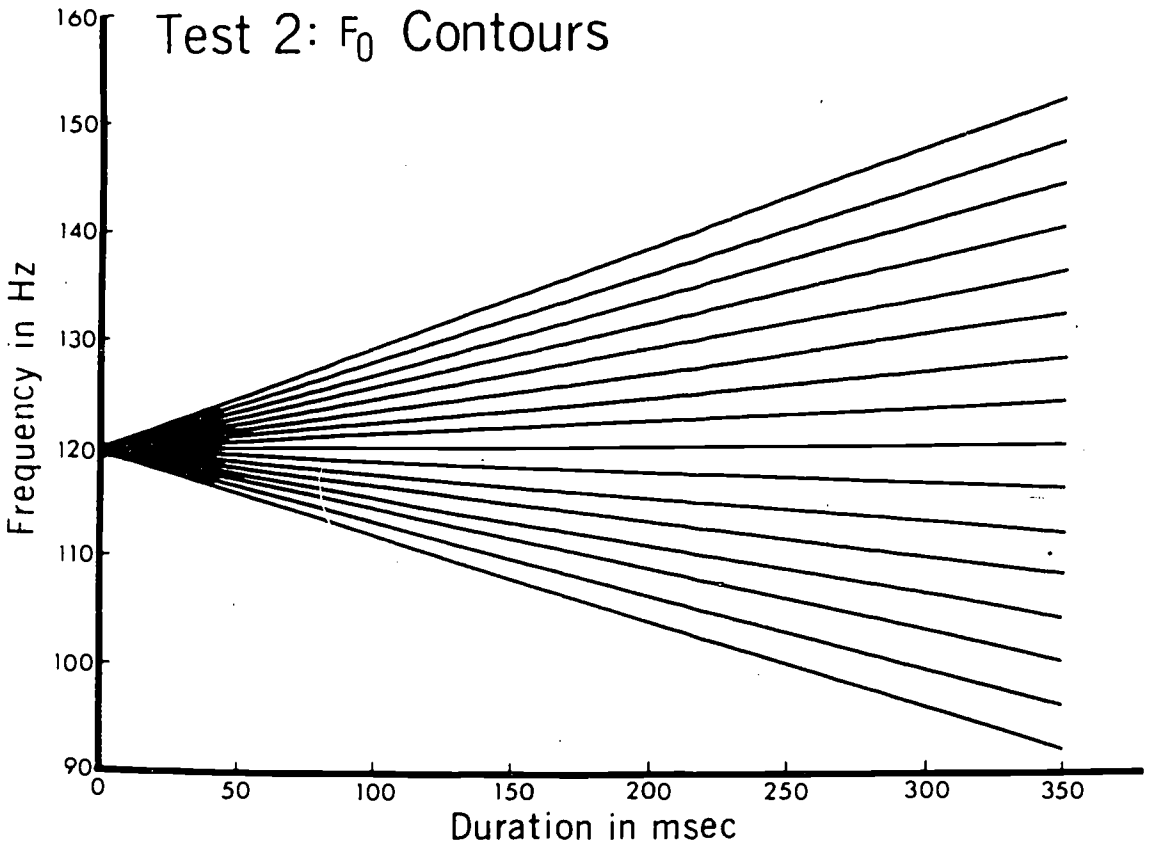


Figure 3: Fundamental-frequency contours from a mid origin.

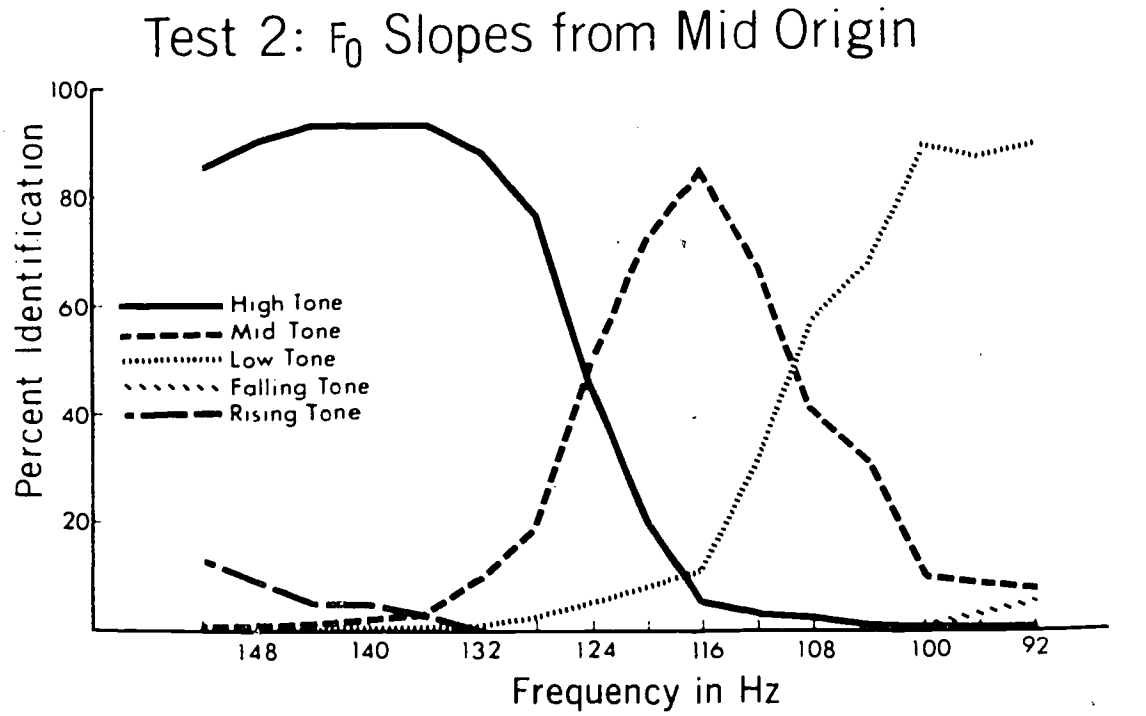


Figure 4: Identification functions for the contours of Figure 3.

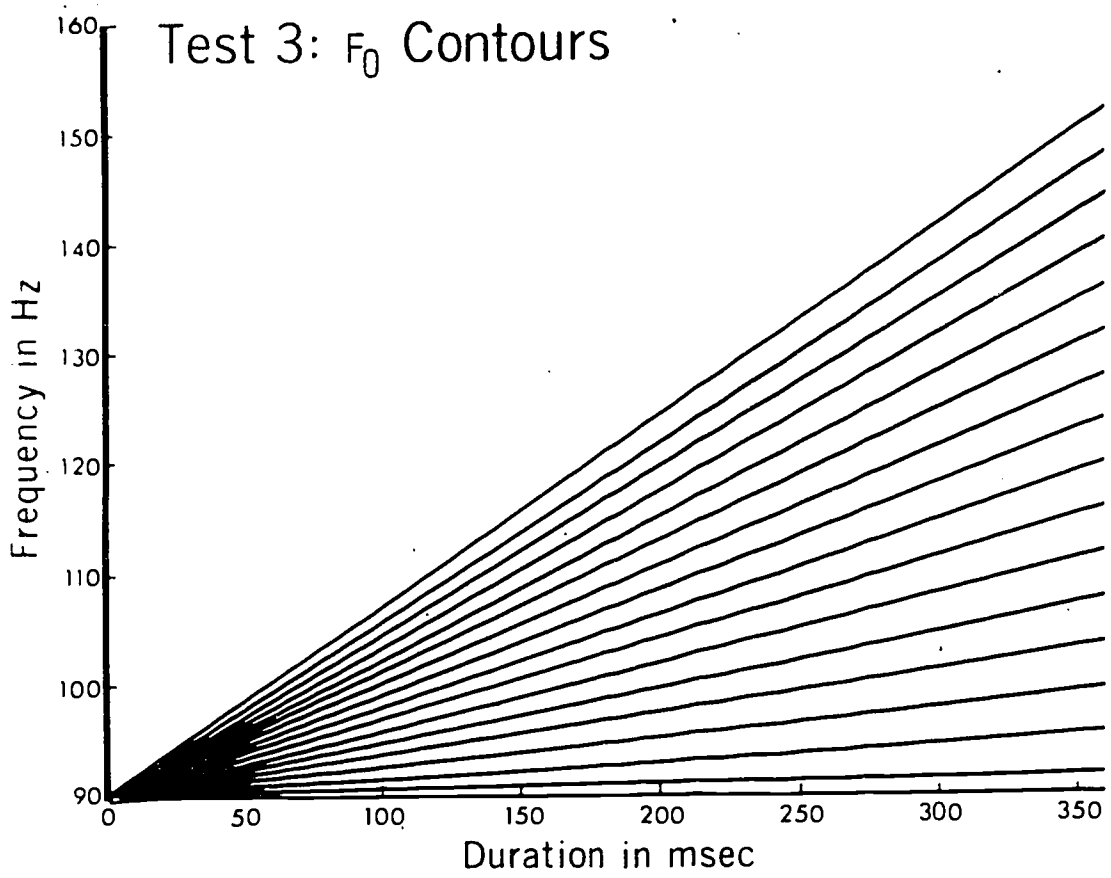


Figure 5: Fundamental-frequency contours from a low origin.

variant ending at 90 Hz.[6] In the test these 17 stimuli were played to the 31 subjects for identification. It was expected that the sharply rising variants would be heard as a dynamic tone, namely the rising tone, with the others divided among the static tones with some preference for the low tone. The results are shown in Figure 6. With a peak at 91 percent, the rising tone is clearly favored. The low tone reaches a peak of 88 percent only at the very bottom of the range. It would be more convincing if it started higher and drifted downward. The third response category is the high tone which peaks at 38 percent. For this tone, a more appropriate movement would start higher. The mid tone which peaks at just under 12 percent, is negligible.

We may conclude that fundamental-frequency levels do carry much information on the static tones, although they improve with movement. For the dynamic tones, as exemplified here by the rising tone, a rather abrupt movement is required. Other continua that bear on this question have been tested but are not yet ready

[6]For a reason that is hard to reconstruct, possibly no more than an oversight, the low point was set at 90 Hz instead of 92 Hz as in Test 2. It is not likely that the downward shift of 2 Hz has any bearing on the outcome.

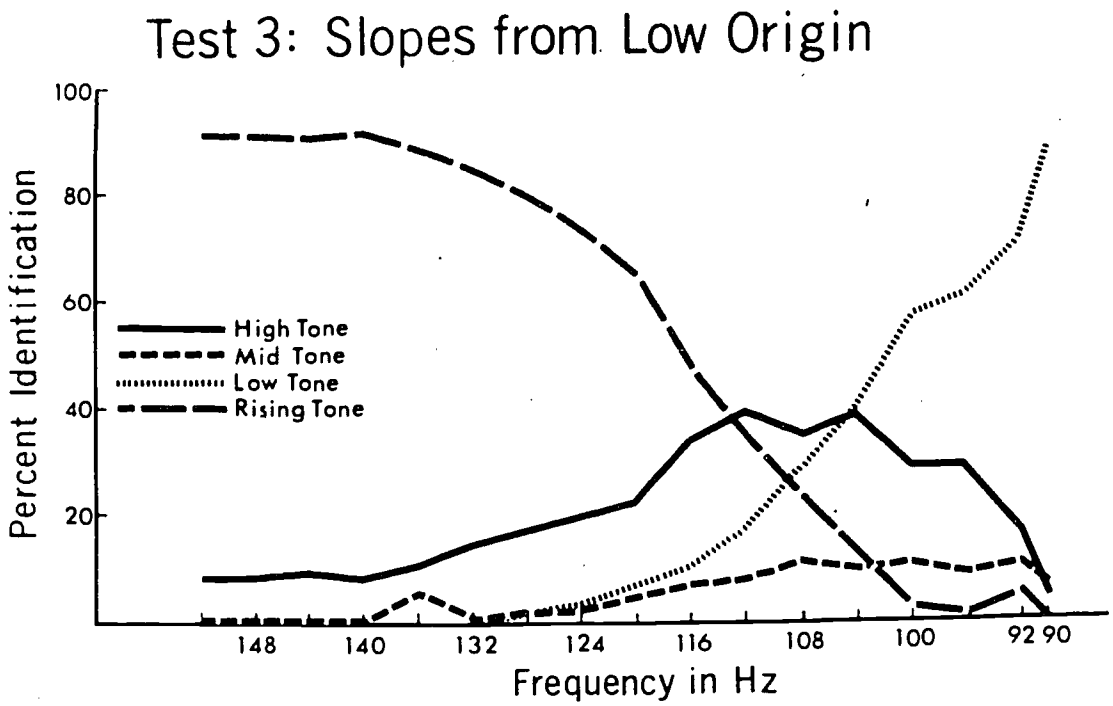


Figure 6: Identification functions for the contours of Figure 5.

for presentation. Although the dichotomy between static and dynamic tones is imprecise and unstable, more so in production (Abramson, 1975b) than perception, it is still useful as a rough classification of tone production and as an index to the types of acoustic cues used in recognition of tones.

REFERENCES

Abramson, A. S. (1962) The vowels and tones of Standard Thai: Acoustical measurements and experiments. Intl. J. Am. Ling. 28(2, pt. 3) (Indiana University Research Center in Anthropology, Folklore, and Linguistics, Publication 20).

Abramson, A. S. (1975a) The tones of Central Thai: Some perceptual experiments. In Studies in Tai Linguistics in Honor of William J. Gedney, ed. by. J. G. Harris and J. R. Chamberlain (Bangkok: Central Institute of English Language), pp. 1-16.

Abramson, A. S. (1975b) The coarticulation of tones: An acoustic study of Thai. Haskins Laboratories Status Report on Speech Research SR-44, 119-125. (To appear in Proceedings of the 8th International Congress of Phonetic Sciences, Leeds.)

Erickson, D. (1974) Fundamental frequency contours of the tones of Standard Thai. Pasaa: Notes and News about Language Teaching and Linguistics in Thailand 4, 1-25.

Erickson, D. (1976) A physiological analysis of the tones of Thai. Unpublished Ph.D. dissertation, University of Connecticut.

Gandour, J. (1975) On the representation of tone in Siamese. In Studies in Tai Linguistics in Honor of William J. Gedney, ed. by J. G. Harris and J. R. Chamberlain (Bangkok: Central Institute of English Language), pp. 170-195.

Pike, K. (1948) Tone Languages (Ann Arbor: University of Michigan).

The Effects of Selective Adaptation on Voicing in Thai and English

S. L. Donald[+]

ABSTRACT

Native Thai speakers and native English speakers took part in a selective adaptation experiment. The stimuli were a labial series of 25 stimuli from a voice-onset-time continuum. This series spanned three phonological categories for the Thai-speaking subjects but only two categories for the English-speaking subjects. The data suggest that three feature detectors mediate the perception of voicing contrasts for the Thai-speaking subjects, whereas only two feature detectors appear to be active in the English-speaking subjects' perception of voicing contrasts. Implications of this difference are considered.

Several selective adaptation experiments have examined the perception of the voicing distinction (for example, Eimas and Corbit, 1973; Cooper, 1974). The variable in the experiments was voice onset time (VOT), or the interval between stop release and the onset of phonation (Lisker and Abramson, 1964). Eimas and Corbit (1973), as well as later investigators, have suggested that two feature detectors mediate the perception of voicing contrasts. Since native English speakers were used, these adaptation experiments were limited to the distinction between voiced stops and voiceless aspirated stops, with a continuum being tested ranging from 0- to 80-msec VOT. These detectors are hypothesized to respond to a slightly overlapping range of VOT values. The category boundary lies at the point at which both detectors respond with equal strength. Repetitive stimulation of either detector is said to fatigue that detector, resulting in weakened output. The unadapted detector will thus respond to boundary stimuli with relatively greater strength than the adapted detector, resulting in a shift in the phonetic boundary.

Thai, in contrast to English, has three voicing categories: prevoiced, voiceless unaspirated, and voiceless aspirated stops. (This is true for the labial and alveolar places of articulation. Thai lacks a prevoiced velar stop.) Abramson and Lisker (1965) found the category boundaries here to occur at -20 msec and at +40 msec in comparison to the single English boundary at +25 msec.

[+]Also University of Connecticut, Storrs.

Acknowledgement: I would like to thank Arthur Abramson and Ignatius Mattingly for their comments and criticisms of earlier versions of this paper.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

Following Eimas and Corbit's reasoning, three feature detectors might operate in the perception of three voicing categories, as found in languages like Thai. The purpose of the first experiment reported here is to investigate this issue. Native speakers of Thai were used as subjects. Three adaptation conditions were presented: adaptation to a prevoiced stimulus, adaptation to a voiceless unaspirated stimulus, and adaptation to a voiceless aspirated stimulus. The VOT continuum examined here ranged from -80 to +70 msec.

In a second experiment, native English speakers responded to the same stimuli with only two, rather than three voicing labels. The purpose of this experiment was to examine the effect of linguistic experience by comparing the effect of adaptation with the two stimuli the English speakers labeled as voiced, with the effects these stimuli produced on Thai speakers for whom they were phonologically distinct.

EXPERIMENT I

Subjects

Five native Thai speakers of Central Thai (Siamese) served as subjects.

Stimuli

The stimuli were a labial series of 25 stimuli from a VOT continuum prepared by Lisker and Abramson (1970). The variations in VOT were produced by varying the onset of the first formant relative to the onset of the second and third formants. During the absence of the first formant, the upper formants are excited by a noise source rather than by a periodic source. The VOT values ranged from -80 to +70 msec in 5 and 10 msec steps. Table 1 contains a list of the values of these stimuli. The adapting stimuli were -80 msec for the prevoiced adaptation condition, +5 msec for the voiceless unaspirated condition and +70 msec for the voiceless aspirate condition.

TABLE 1: Stimulus values.

Stimulus	VOT value	Stimulus	VOT value
0*	-80 msec	13*	5 msec
1	-70 msec	14	10 msec
2	-60 msec	15	15 msec
3	-50 msec	16	20 msec
4	-45 msec	17	25 msec
5	-40 msec	18	30 msec
6	-35 msec	19	35 msec
7	-30 msec	20	40 msec
8	-25 msec	21	45 msec
9	-20 msec	22	50 msec
10	-10 msec	23	60 msec
11	-5 msec	24*	70 msec
12	0 msec		

* denotes stimulus used as adaptor

Procedure

Each subject participated in four experimental sessions. The first session consisted solely of an initial identification tape, in which each stimulus was presented 16 times in random order. The stimuli were presented in blocks of 15, with three seconds separating the presentation of each stimulus. Ten seconds separated the presentation of each block. Each of the adaptation sessions started with a short identification tape which presented each stimulus in the continuum eight times in random order. Thus, a total of 40 responses for each stimulus was obtained in an unadapted condition. Under each adaptive condition the listeners were exposed to 60 presentations of the adapting stimulus (with an interstimulus interval of 30 msec). After the period of adaptation the subjects were asked to identify five stimuli. Each stimulus in the continuum was presented eight times, in random order, for such postadaptation identification. Subjects responded with the three labial stops written in Thai orthography.

Results

The unadapted boundaries for both the boundary between the prevoiced and voiceless unaspirated stops, and the boundary between the voiceless unaspirated and voiceless aspirate stops were extrapolated for each subject from the pooled identification responses from all sessions. The boundary was defined as that point on the stimulus scale which would, by extrapolation, receive 50 percent responses from either category involved. The boundaries for the adapted responses were estimated in the same manner. The boundary shifts are the differences between the unadapted boundary and the adapted boundary. The results are displayed in Tables 2 and 3. The shifts predicted by the hypothesis of three

TABLE 2: Thai subjects [b]-[p] boundary.

Subject	Original boundary	Shift after [b] adaptation*	Shift after [p] adaptation	Shift after [ph] adaptation
1	-25 msec	-8	+1	+2
2	-19 msec	-2	+2	+1
3	-18 msec	-14	+10	-7
4	-23 msec	-11	+1	--
5	-24 msec	-9	+3	-4

*significant boundary shifts

TABLE 3: Thai subjects [p]-[ph] boundary.

Subject	Original boundary	Shift after [b] adaptation	Shift after [p] adaptation*	Shift after [ph] adaptation*
1	+27 msec	-2	-3	+6
2	+22 msec	-3	--	+1
3	+20 msec	-4	-7	+11
4	+28 msec	-1	-2	--
5	+30 msec	-4	-6	+16

*significant boundary shifts

feature detectors mediating the perception of the voicing contrasts occurred. With the [b] adapting stimulus, fewer [b] responses were obtained. With the [ph] adapting stimulus, fewer [ph] responses were obtained. The results for the [p] adaptation condition are somewhat less definitive. The [p]-[ph] boundary shift is fairly robust, and in the predicted direction: fewer [p] responses were obtained. The [b]-[p] boundary, on the other hand, is small. However, one subject who generally produced relatively large boundary shifts, did respond with a large boundary shift. Except for the [b]-[p] boundary shift after [p] adaptation, all these boundary shifts are significant (by a t test for two related groups, $p < .05$). Neither the [b]-[p] boundary after [p] adaptation nor the [p]-[ph] boundary after [b] adaptation are significant.

EXPERIMENT II

Subjects

Four native American English speakers served as subjects.

Stimuli

The same stimuli were used in this experiment as were used in Experiment I.

Procedure

The same procedure was followed as was followed in Experiment I, except that these subjects responded with only two answers--voiced stops or voiceless aspirated stops.

Results

The data obtained in Experiment II were analyzed in the same manner as were the data from Experiment I. When subjects were adapted to the prevoiced stimulus, fewer voiced responses were given. When subjects were adapted to the voiceless unaspirated stimulus, which they categorized as voiced, again fewer voiced responses were given. When subjects were adapted to the voiceless aspirated stimulus, fewer voiceless responses were obtained. All conditions produced significant boundary shifts (by a t test for two related groups, $p < .05$). These results are displayed in Table 4.

TABLE 4: English subjects [b]-[p] boundary.

Subject	Original boundary	Shift after [b] adaptation*	Shift after [p] adaptation*	Shift after [ph] adaptation*
1	+15 msec	-4	-5	+8
2	+15 msec	-8	-2	+10
3	+18 msec	-8	-5	+9
4	+10 msec	-12	-7	+14

*significant boundary shifts

Discussion

The results of Experiment I suggest that Eimas and Corbit's (1973) original assertion--that two phonetic feature detectors mediate the perception of voicing distinctions--ought to be amplified by the addition of a third detector sensitive to negative VOT values. By this hypothesis one detector would be sensitive primarily to voiceless aspirated cues, a second would be sensitive to voiceless inaspirate cues, and the third to cues of prevoicing. The two boundaries between these three voicing distinctions would occur at those VOT values to which two of the feature detectors were equally responsive.

A somewhat surprising result of Experiment I was that the voiceless aspirate detector was more resistant to adaptation than the other two detectors. Recall that in Eimas and Corbit's (1973) experiment, and also in Eimas, Cooper, and Corbit's (1973) findings, the voiceless detector was more susceptible to adaptation than the voiced detector. Similarly, for the English speakers of Experiment II, adaptation of the voiceless detector produced a more robust boundary shift than adaptation of the voiced detector. A possible explanation for this discrepancy is that adaptation takes place at both the auditory and the phonetic level. In a set of two experiments, the first involving place of articulation and the second voicing, Tartter and Eimas (1975) found that the greater the acoustic overlap between the adapting stimulus and the test continuum, the greater the adaptation effect. For example, the addition of first-formant, or steady-state information to adapting stimuli that contained all the relevant place of articulation information, produced a substantially larger boundary shift than was obtained by an adapting stimulus lacking this first formant. This fact defies the acoustic theorist. If the existence of some higher level feature is accepted, however, a clear explanation is possible: after adapting to a complete stimulus, with all three formants present, all three of the auditory feature detectors will have been fatigued. In contrast, after adapting to a stimulus lacking the first formant, only the F_2 and F_3 detectors will have been adapted. Ades (1976), however, objects to this proliferation of levels of detectors, saying that "In general, two strengths of adaptation do not necessarily indicate two levels of adaptation: it could be that there is just one level, more engaged by the full syllable than by parts of it."[1] Obviously this is not a resolved issue, and the present experimentation does not help in its resolution. In light of these claims, consider again the present discrepancy.

Although adequate information is present in the stimuli used to allow the voicing distinctions to be perceived--according to Tartter and Eimas's explanation--some information present in natural speech is not present. The VOT information present in the stimuli is picked up by low-level detectors sensitive to certain aspects of voicing distinctions, which yield their output to higher level voicing detectors. These higher level detectors fail to receive input from low-level detectors sensitive to other acoustic features cueing voicing distinction in natural speech. The lack of adaptation of these low-level detectors accounts for the so-called "resistance" to adaptation in Thai and English subjects. That English speakers and Thai speakers are resistant to adaptation by different adaptors is due to differences in the production of voicing distinctions in the two languages.

[1] Ades, A. E. (1976) Adapting the property detectors for speech perception; preprint sent to author, p. 30.

Ades's viewpoint allows an almost identical explanation, differing primarily in terminology. Here again the synthetic stimuli used in these experiments lack some acoustic information present in natural speech, for example, variations in release-burst intensity. The lack of relevant information in certain stimuli would decrease the strength of adaptation. Furthermore, the contrast in strengths of adaptation in Thai and English subjects is due to differences in the production of voicing distinctions in the two languages.

At any rate, it is apparent that linguistic environment has substantial effect on the development of feature detectors. First, as discussed above, the discrepancy of degree of adaptation in the Thai and English subjects indicate differences in the perception of the same acoustic stimuli by subjects from differing linguistic backgrounds. Second, the present two experiments demonstrate that the perceptual boundaries between voicing categories are different in Thai and in English, confirming earlier work (Abramson and Lisker, 1965; Lisker and Abramson, 1970). Not only does Thai have one more category of voicing than does English, but the exact location of the boundary common to the two languages is different; in English at approximately +15 msec, as opposed to approximately +25 msec in Thai.

This point is also supported by the fact that adaptation with a prevoiced stimulus produced a boundary shift between the voiced-voiceless categories of English-speaking subjects, but no boundary shift between the voiceless categories of Thai subjects. This discrepancy is striking evidence that phonetic feature detectors are subject to effects of language detectors.

The end result is clear: for English-speaking subjects, stimuli--that through different linguistic experience would be perceived as belonging to separate categories--are perceived as belonging to the same category, and when serving as adaptive stimuli, produce the same effects. What has happened to the hypothesized third detector? Perhaps through lack of stimulation the detector has atrophied. The evidence from studies of infants' perception of voicing contrasts supports this view.

Streeter (1976) investigated the discrimination of VOT by infants from a linguistic environment that distinguishes between prevoiced and voiceless unaspirate, but not voiceless aspirated stops. She found that these infants discriminated both the prevocied/voiceless unaspirated distinction and the voiceless unaspirate/voiceless aspirated distinction. Lasky, Syrdal-Lasky, and Klein (1975), studying infants born to Spanish-speaking parents, also found three categories of discrimination comparable to those Streeter found. And yet the single-phoneme boundary separating Spanish stops corresponds to neither of the boundaries which Lasky et al. found. Apparently infants, like chinchillas (Kuhl and Miller, 1975), and adults discriminating among nonspeech stimuli differing in relative onset time (Miller, Pastore, Wier, Kelly, and Dooling, 1974; Pisoni, 1976[2]) distinguish between three categories characterized by leading, simultaneous, and lagging temporal events.

A second alternative explanation for the disappearance of the third detectors, not incompatible with the first, is that the detectors sensitive to

[2]Pisoni, D. B. (1976) Identification and discrimination of the relative onset time of two-component tones: Implications for voicing perception in stops; preprint sent to author.

prevoicing are present but that linguistic experience affects the labeling of the output of the detectors. This alternative would also account for the varying locations of the exact voicing boundaries in different languages.

In summary, the present experiments suggest the existence of a phonetic feature detector sensitive to cues of prevoicing. They also demonstrate that language learning has a substantial effect on the detectors mediating the perception of voicing contrasts.

REFERENCES

Abramson, A. S. and L. Lisker. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. In Proceedings of the 5th International Congress on Acoustics, Liege, Belgium, 7-14 September, ed. by D. E. Commins, abstract A51.

Cooper, W. E. (1974) Selective adaptation for acoustic cues of voicing in initial stops. J. Phonetics 2, 303-313.

Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation of linguistic feature detectors. Cog. Psychol. 4, 247-252.

Kuhl, P. K. and J. D. Miller. (1975) Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar consonants. Science 190, 69-72.

Lasky, R. E., A. Syrdal-Lasky, and R. E. Klein. (1975) VOT discrimination by four to six and a half month old infants from Spanish environments. J. Exp. Child Psychol. 20, 213-225.

Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.

Lisker, L. and A. S. Abramson. (1970) The voicing dimension: Some experiments in comparative phonetics. In Proceedings of the 6th International Congress of Phonetic Sciences, Prague, 1967 (Prague: Academia), pp. 563-567.

Miller, J. D., R. E. Pastore, C. C. Wier, W. J. Kelly, and R. J. Dooling. (1974) Discrimination and labeling of noise-buzz sequences with varying noise-lead times. J. Acoust. Soc. Am. 55, 390(A).

Streeter, L. A. (1976) Language perception of two-month-old infants shows effects of both innate mechanisms and experience. Nature 259, 38-41.

Tartter, V. C. and P. D. Eimas. (1975) The role of auditory feature detectors in the perception of speech. Percept. Psychophys. 18, 293-298.

Perception of Nonspeech by Infants*

Peter W. Jusczyk,[+] Burton S. Rosner,[+] James E. Cutting,[++] Christopher F. Foard,[+] and Linda B. Smith[+]

ABSTRACT

According to recent investigations, adult listeners perceive rise-time differences in both speech and nonspeech stimuli in a categorical manner (Cutting and Rosner, 1974). Adults labeled sawtooth-wave stimuli as either plucked or bowed. The present study used the high amplitude sucking technique to explore the two-month-old infant's perception of rise-time differences for sawtooth stimuli. Infants discriminated rise-time differences that marked off the different nonspeech categories but did not discriminate equal differences within either category. Thus, the present study shows that infants, like adults, can perceive nonspeech stimuli in a categorical manner.

INTRODUCTION

Considerable evidence indicates that many speech sounds are perceived categorically. With these stimuli, subjects are no better at discriminating sounds than they are at differentially labeling them. This claim is supported by experimental findings from a number of different paradigms including: (a) accuracy (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Mattingly, Liberman, Syrdal, and Halwes, 1971; Pisoni, 1971, 1973); (b) reaction time (Pisoni and Tash, 1974); and (c) average evoked potentials (Dorman, 1974). These results contrast with those observed for a wide variety of nonspeech sounds, varying along such physical continua as frequency, amplitude, and duration, for which the subject's ability to discriminate between stimuli far outstrips his ability to label them differentially (Miller, 1956).

---

*To be published in Perception and Psychophysics.

[+]University of Pennsylvania.

[++]Also Wesleyan University.

Acknowledgment: This research was supported by NICHD grant 5T01 HD0037 to the Department of Psychology, University of Pennsylvania, under which the first and fifth authors served as trainees, and NICHD grants HD-01994 and RR-05596 to the Haskins Laboratories.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

There is also a growing body of evidence that shows that human infants are capable of discriminating speech segments on the basis of minimal phonetic cues. To date, infants have displayed an ability to perceive subtle differences in voicing (Eimas, Siqueland, Jusczyk, and Vigorito, 1971; Streeter, 1974; Eimas, 1975b; Lasky, Syrdal-Lasky, and Klein, 1975), place of articulation (Morse, 1972; Eimas, 1974), initial burst cues (Miller, Morse, and Dorman, 1975), and third format cues for the /ra/-/la/ distinction (Eimas, 1975a). Not only do infants make fine distinctions between speech sounds, but they do so in a categorical manner (for example, they make interphonemic distinctions but not intraphonemic ones). Further, Eimas (1974, 1975b) has shown that infants, like adults (Mattingly et al., 1971), perceive certain acoustic cues categorically in speech contexts but not in nonspeech contexts. On the basis of these findings, Eimas (1975b and elsewhere) has suggested that the actual mechanisms which underlie the categorical perception of speech may be part of the biological makeup of the human infant.

Thus, speech appears to be perceived in a quite different fashion from nonlinguistic auditory stimuli. However, several recent developments may require us to reexamine the claim that categorical perception is evidence for the distinctive nature of speech perception. Categorical perception has now been observed in a number of instances of nonspeech sounds (Locke and Kellar, 1973; Cutting and Rosner, 1974; Cutting, Rosner, and Foard, in press; Miller, Wier, Pastore, Kelly, and Dooling, in press). In particular, Cutting and Rosner (1974) have reported categorical perception for nonspeech sounds varying in rise-time. They have explored the perception of rise-times in both sawtooth-wave and sine-wave stimuli (as well as for affricate-fricatives in speech). Adult listeners usually reported that these nonspeech stimuli sound as though they were produced by a musical stringed instrument. Sounds with rapid rise-times (less than 40 msec) were perceived as coming from a plucked string, whereas sounds with more gradual rise-times (greater than 40 msec) were perceived as being produced by a bowed string. The listeners easily identified the stimuli as either "pluck" or "bow." Moreover, the perception of these stimuli was categorical.

In a related study, Cutting, Rosner, and Foard (in press) extended the findings for the sawtooth-wave stimuli by demonstrating selective adaptation effects with them. These effects were similar to those observed with speech stimuli (Eimas and Corbit, 1973) both in direction and degree of shift. Moreover, as in the case of speech stimuli, adaptation shifts for the sawtooth stimuli were greatest when the adapting stimulus shared all dimensions with the test continuum.

Although the claim for the distinctive nature of categorical perception in speech has been weakened by these lines of research, there has been no indication that infants might exhibit categorical perception for nonspeech sounds. In fact, Eimas (1974, 1975b) has reported that two- and three-month-old infants tend not to perceive nonspeech cues categorically. However, the cues which Eimas studied were acoustic features that adults do not perceive categorically (Mattingly et al., 1971; Miyawaki, Strange, Verbrugge, Liberman, Jenkins, and Fujimura, 1975). The sawtooth-wave stimuli employed by Cutting and Rosner (1974) would seem to be a better choice for such a test. Not only do adults perceive these sounds categorically, but rise-time is also an important acoustic cue in various contexts. Accordingly, the present study explored the perception of rise-time differences in sawtooth-wave stimuli by two-month-old infants.

METHOD

Procedure

Each infant was tested in a mobile laboratory. The infants were placed in a reclining seat which faced a loudspeaker approximately two feet away. Each subject sucked on a blind nipple which one of the experimenters held in place.

The experimental procedure was a modification of the high amplitude sucking technique developed by Siqueland and DeLucia (1969). For each infant, the high amplitude sucking criterion and the baseline rate of high amplitude nonnutritive sucking were established before presentation of any stimuli. The criterion for high amplitude sucking was adjusted to produce sucking rates of 10 to 20 sucks per minute. After a baseline rate was established, the presentation of stimuli was made continuent upon the rate of sucking. If the time between criterion responses was two seconds or more, then each response produced one presentation of the stimulus, which had an average duration of 1050 msec, followed by 950 msec of silence. If the infant produced a burst of high amplitude responses within this two-second interval, the timing apparatus was automatically reset and the two-second interval began again.

The criterion for habituation to the first stimulus was a decrement in sucking rate of 25 percent or more over two consecutive minutes, compared to the rate in the immediately preceding minute. At this point, the auditory stimulation was changed without interruption by switching channels on the tape recorder. For infants in the experimental conditions, the change resulted in the presentation of a second acoustically distinct stimulus. For the infants in the control condition, the channels on the tape recorder were switched but no acoustic change was made. The postshift period lasted for four minutes. The infants' sensitivity to the change in the auditory stimulation was inferred from comparisons of the response rates of subjects in the experimental and control conditions during the postshift period.

Stimuli

The stimuli were sawtooth waves generated on the Moog synthesizer at the Presser Electronic Studio of the University of Pennsylvania. The four stimuli were synthesized at 440 Hz and differed solely in their onset characteristics. Amplitude envelopes reached maximum in 0, 30, 60, and 90 msec after onset. By 0 msec rise-time, we mean that a stimulus reached maximum amplitude in one-fourth of a period. Previous research by Cutting and Rosner (1974) indicated that adults easily label the rapid onset (0 and 30 msec) sounds as "plucks." The more gradual onset stimuli (60 and 90 msec) were easily labeled as "bows." The durations of the four nonspeech stimuli were 1020, 1050, 1080, 1110 msec, varying according to rise-time. The decay period of each stimulus was 1020 msec.

All the stimuli were prerecorded on three 30-minute audio tapes for presentation to the subjects. Tape #1 (pluck-pluck) was composed of 0 msec rise-time stimuli on channel A and of 30 msec rise-time stimuli on channel B. Tape #2 (bow-bow) was composed of 60 msec rise-time stimuli on channel A and of 90 msec rise-time stimuli on channel B. Tape #3 (pluck-bow) was composed of 30 msec stimuli on channel A and of 60 msec stimuli on channel B.

Design

Table 1 shows the within-subjects design for the present experiment. All subjects were seen for two experimental sessions. (Mean interval between ses-

TABLE 1: Design.

	Session A	Session B
Group 1 (n=6)	Pluck-bow	Pluck-pluck
Group 2 (N=6)	Pluck-bow	Bow-bow
Group 3 (n=6)	Pluck-bow	NO CHANGE

sions was 8 days; range was 5 to 14 days.) In one session, all subjects heard the pluck-bow tape. The other session differentiated the three groups of subjects. Subjects in Group 1 heard the pluck-pluck tape. Subjects in Group 2 heard the bow-bow tape. Subjects in Group 3 were randomly assigned one of the four rise-time stimuli for the entire session (the NO CHANGE condition). The order of sessions and the order of stimuli within a session were each counterbalanced.

Apparatus

A blind nipple was connected to a Grass PT5 volumetric pressure transducer which was coupled in turn to a Beckman Type RS Dynograph. An integrator-coupler provided a digital output of criterial high amplitude sucking responses. Additional equipment included a 4-track Hitachi tape recorder with speakers, a Hunter digital timer, two relays, and a counter. Each criterion response activated the digital timer for a two second period or restarted the period. Auditory stimulation at a level of 72 ± 2 dB SPL was available to the infant whenever the timer was in an active state.

Subjects

The subjects were 18 infants, nine males and nine females. Mean age was eight weeks (range: five to ten weeks). In order to obtain complete data on 18 infants, it was necessary to test 25. Seven infants were dropped from the study for the following reasons: two infants fell alseep prior to shift, three cried excessively prior to shift, and the mothers of two infants were unable to keep the second appointment.

RESULTS

Figure 1 displays the mean number of high amplitude sucking responses as a function of minutes and experimental groups. For purposes of statistical comparisons, we examined each subject's rate of high-amplitude sucking during five intervals: baseline minute, third minute before shift, average of minutes one and two before shift, average of minutes one and two after shift, and average of

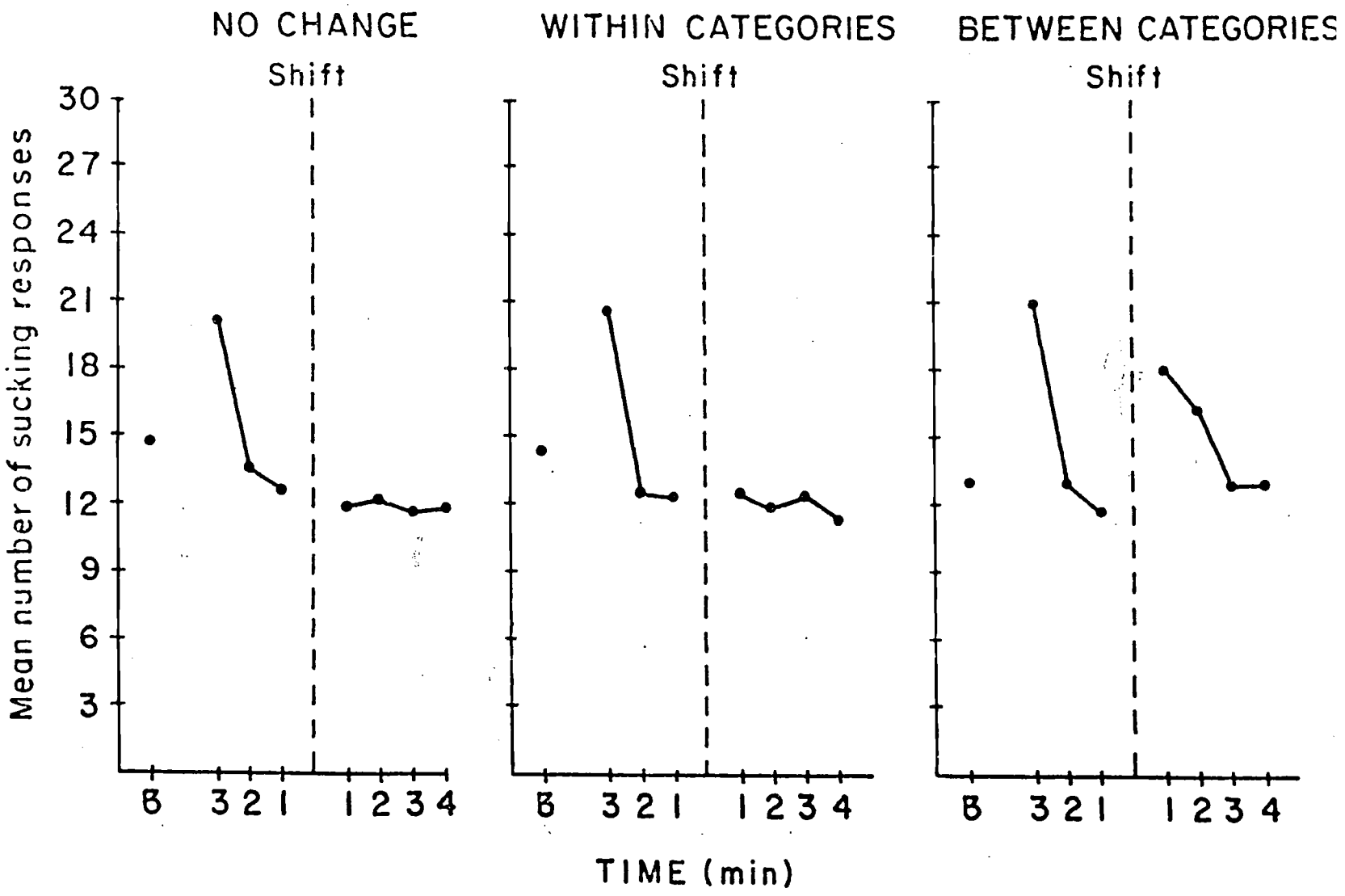


FIGURE 1

Figure 1: Mean sucking rates as a function of time and experimental session. Time is measured with reference to the moment of the stimulus shift, marked by the vertical dashed line. The baseline rate of sucking is indicated by the letter "B".

minutes three and four after shift. Difference scores were calculated for each subject for the following rate comparisons: (1) acquisition of the sucking response: third minute before shift less baseline; (2) habituation: third minute before shift less average of last two minutes before shift; (3) dishabituation: average of first two minutes after shift less average of first two minutes before shift; (4) dishabituation during third and fourth minutes: average of third and fourth minutes after shift less average of last two minutes before shift; and (5) rehabituation: average of first two minutes after shift less average of third and fourth minutes after shift.

Kruskal-Wallis one-way analyses of variance (Seigel, 1956) were employed to determine if the data for the pluck-bow sessions could be collapsed across the three experimental groups. No significant differences were observed between groups for any of the five comparisons [$\chi^2(2)$ ranged from 0.37 to 4.10]; accordingly the data for the pluck-bow sessions were collapsed across groups in further analyses. Additionally, Kruskal-Wallis tests indicated no differences [$\chi^2(1)$ ranged from 0.03 to 0.92] for the bow-bow and pluck-pluck subgroups, whose data were similarly combined for further treatment.

Wilcoxon matched-pairs signed-ranks tests (Seigel, 1956) were used to analyze performance within each type of session. The results of these analyses, presented in Table 2, indicated that in all sessions subjects acquired the

TABLE 2: T-values for Wilcoxon matched-pairs signed-ranks test.

Comparison	Experimental session: Pluck-bow (n=18)	Pluck-pluck or bow-bow (n=12)	NO CHANGE (n=6)
Acquisition: third minute before shift versus baseline.	0**	0**	0*
Habituation: third minute before shift versus average of last two minutes before shift.	0**	0**	0*
Dishabituation: first two minutes after shift versus last two minutes before shift.	-1**	-19	0*[a]
Late dishabituation: third and fourth minutes after shift versus last two minutes before shift.	-52.5	23	4
Rehabituation: first two minutes after shift versus third and fourth minutes after shift.	-12**	-26	8

** p < .01
* P < .04
[a] indicates reliable decrease in sucking

conditioned high-amplitude sucking response and habituated the response prior to shift. However, only in the pluck-bow condition did subjects display a reliable increase in sucking after the shift. Moreover, these subjects showed a reliable increase in sucking during the first two minutes after shift followed by a reliable decrease in rate between that period and the next two minutes, thus indicating rehabituation. By contrast, subjects in the other three conditions showed no evidence of any increase in sucking after shift. Subsequent analysis of the data for the pluck-pluck, bow-bow, and NO CHANGE sessions by Kruskal-Wallis tests indicated no reliable differences in the pattern of responding by subjects in these sessions. Randomization tests on within-subjects data across conditions confirmed these findings.

Butterfield and Cairns (1974) have reported that asymmetrical order effects are sometimes observed for speech stimuli which cross phonetic boundaries (a shift from a voiced to a voiceless stop producing greater dishabituation than from voiceless to voiced). We tested for such asymmetries with the present stimuli. None were discovered, as Kruskal-Wallis tests for the pluck-bow sessions yielded no reliable differences [$\chi^2(1)$ ranged from 0.02 to 1.73] between the two presentation orders.

DISCUSSION

The present data incidate that infants as young as two months of age perceive rise-time cues in sawtooth-wave stimuli in a categorical manner, as do adults (Cutting and Rosner, 1974). This constitutes the first demonstration that infants perceive acoustic stimuli other than speech in a categorical fashion. Our results are consistent with those observed for speech stimuli (for example, Eimas et al., 1971; Eimas, 1974, 1975a, b), since infants displayed a reliable increase in sucking only for stimuli chosen from opposite sides of the adult categorical boundary.

How can we explain the two-month old's propensity to categorize "plucks" and "bows"? One relevant result (Cutting and Rosner, 1974) is that rise-time is a sufficient cue for the categorical perception of [ʃa] and [tʃa] as in "shop" and "chop." One possible explanation for the present results, then, is that the sawtooth-wave stimuli are perceived categorically just because rise-time is a salient dimension in speech perception. By one interpretation of this linguistic hypothesis, however, every acoustic dimension which is perceived categorically in speech should also be perceived categorically in nonspeech sounds. Yet, Mattingly et al. (1971) reported that second formant transitions which are perceived categorically in speech are not perceived categorically when heard in isolation. These results undercut the strong version of the linguistic hypothesis. An alternative formulation would hold that all dimensions perceived categorically in nonspeech sounds also are perceived categorically in speech sounds. Locke and Kellar's (1973) report of categorical perception of triadic chords seems to contradict this view. Acceptance of this weak version of the hypothesis also leaves open the question of why some dimensions and not others are perceived categorically outside of speech.

A second hypothesis can account for our results. This acoustic hypothesis argues that the categorical perception of [ʃa] and [tʃa] is merely a special case of the categorical perception of the acoustic dimension of rise-time. Indeed, many other nonspeech stimuli may also be perceived categorically. According to this view, the categorical perception of speech sounds is a consequence

of general properties of the auditory system rather than of a special system devoted entirely to the perception of speech. This is supported by a number of results. For example, Lisker and Abramson (1964), Cooper (1974), and Stevens and Klatt (1974) have demonstrated that voice-onset-time (VOT) is actually composed of several acoustic cues. Selective adaptation with the individual acoustic cues from these dimensions produced boundary shifts along the VOT continuum (Lisker, 1975). Similarly, Tartter and Eimas (1975) demonstrated that a number of acoustic cues produced selective adaptation effects for the place-of-articulation continuum as well as for the VOT continuum. Their investigations led Tartter and Eimas to conclude that some selective adaptation effects previously thought explicable only by a phonetic model (for example, Eimas, Cooper, and Corbit, 1973) can be more simply handled by reference to acoustic features. Thus, these recent studies tend to show that more and more of the presumably unique features of human speech can be explained in terms of the acoustic properties of the sounds. Perhaps the particular combination of information available for auditory analysis determines the activation of higher level analyzers which possibly deal only with phonetic information. Thus, the human's tendency to perceive categorically is not limited to speech sounds. The number and variety of nonspeech sounds which are preceived categorically remains to be determined.

REFERENCES

Butterfield, E. C. and G. F. Cairns. (1974) Whether infants perceive linguistically is uncertain, and if they did, its practical importance would be equivocal. In Language Perspectives: Acquisition, Retardation and Intervention, ed. by R. L. Schiefelbush and L. L. Lloyd (Baltimore: University Park Press).

Cooper, W. E. (1974) Selective adaptation for acoustic cues of voicing in initial stops. J. Phonetics 2, 303-313.

Cutting, J. E. and B. S. Rosner. (1974) Categories and boundaries in speech and music. Percept. Psychophys. 16, 564-571.

Cutting, J. E., B. S. Rosner, and C. Foard. (in press) Perceptual categories for musiclike sounds: Implications for theories of speech perception. Quart. J. Exp. Psychol.

Dorman, M. F. (1974) Auditory evoked potential correlates of speech sound discrimination. Percept. Psychophys. 15, 215-220.

Eimas, P. D. (1974) Auditory and linguistic processing of cues for place of articulation by infants. Percept. Psychophys. 16, 513-521.

Eimas, P. D. (1975a) Auditory and phonetic coding of the cues for speech: Discrimination of the [r-l] distinction by young infants. Percept. Psychophys. 18, 341-347.

Eimas, P. D. (1975b) Speech perception in early infancy. In Infant Perception, ed. by L. B. Cohen and P. Salapatek (New York: Academic Press).

Eimas, P. D., W. E. Cooper, and J. D. Corbit. (1973) Some properties of linguistic feature detectors. Percept. Psychophys. 13, 247-252.

Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation of linguistic feature detectors. Cog. Psychol. 4, 99-109.

Eimas, P. D., E. R. Siqueland, P. Jusczyk, and J. Vigorito. (1971) Speech perception in infants. Science 171, 303-306.

Lasky, R. E., A. Syrdal-Lasky, and R. E. Klein. (1975) VOT discrimination by four and six and a half month old infants from Spanish environments. J. Exp. Child Psychol. 20, 215-225.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
Lisker, L. (1975) Is it VOT or a first-formant transition detector. J. Acoust. Soc. Am. 57, 1547-1551.
Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.
Locke, S. and L. Kellar. (1973) Categorical perception in a nonlinguistic mode. Cortex 9, 355-369.
Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.
Miller, C. L., P. A. Morse, and M. F. Dorman. (1975) Infant speech perception, memory, and the cardiac orienting response. Paper presented at the meeting of the Society for Research in Child Development, Denver, Col., April.
Miller, G. A. (1956) The magical number seven, plus or minus two, or some limits on our capacity for processing information. Psychol. Rev. 63, 81-96.
Miller, J. D., C. C. Wier, R. E. Pastore, W. M. Kelly, and R. M. Dooling. (in press) Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. J. Acoust. Soc. Am.
Miyawaki, K., W. Strange, R. Verbrugge, A. M. Liberman, J. J. Jenkins, and O. Fujimura. (1975) An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. Percept. Psychophys. 18, 331-340.
Morse, P. A. (1972) The discrimination of speech and nonspeech stimuli in early infancy. J. Exp. Child Psychol. 14, 477-492.
Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Ph.D. thesis, University of Michigan. [Published in Dissertation Abstracts International, 1972, 32, 6693B (University Microfilms no. 72-14, 964).]
Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253-260.
Pisoni, D. B. and J. Tash. (1974) Reaction times to comparisons within and across phonetic categories. Percept. Psychophys. 15, 285-290.
Seigel, S. (1956) Nonparametric Statistics for the Behavioral Sciences (New York: McGraw-Hill).
Siqueland, E. R. and C. A. DeLucia. (1969) Visual reinforcement of non-nutritive sucking in human infants. Science 165, 1144-1146.
Stevens, K. N. and D. H. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Am. 55, 653-659.
Streeter, L. A. (1974) The effects of linguistic experience on phonetic perception. Ph.D. dissertation, Columbia University. [Published in Dissertation Abstracts International, 1973, 35, 4696B (University Microfilms no. 75-5257).]
Tartter, V. C. and P. D. Eimas. (1975) The role of auditory feature detectors in the perception of speech. Percept. Psychophys. 18, 293-298.

Categorical Perception Along an Oral-Nasal Continuum*

Roland Mandler+

ABSTRACT

Dental and labial nasal consonants were constructed using two methods of synthesis, one employing the nasal branch resonances, and one the oral branch resonances of the OVE III in simulation of period of closure. Oral-nasal continua were generated for both places (/da/ to /na/ and /ba/ to /ma/) for both methods of synthesis. Identification and same-different discrimination tests from all four resulting sets were administered to thirteen subjects. Their responses yielded strong evidence for categorical perception along the oral-nasal dimension.

Extensive analysis of the structure of nasal consonants by Fujimura (1962) and other researchers has revealed the predominant cue value of two basic components: one, the presence of a low amplitude noise through the period of closure, and two, a stoplike transition following the closure. Using a tape-splicing technique, Malécot (1956) confirmed that place cues were carried in the transition and nasality was cued by the low amplitude noise through the period of closure. A study by Liberman, Delattre, Cooper, and Gerstman (1954) using synthetic speech employed identical transitions for nasals and oral stops to get labeling judgments for continua across place of articulation. Both sets yielded good identification. These and other perception experiments suggested that listeners perceived such stimuli categorically, that is, distinguished stimuli across phonetic categories but not within categories, despite identical acoustic variations. The present experiment was undertaken to determine whether categorical perception could be evidenced along an oral-nasal continuum.

METHOD

Stimulus Specifications

The specifications of the OVE III serial synthesizer at the Haskins Laboratories allowed for two methods of construction of nasals from stops:

---

*Presented at the 91st meeting of the Acoustical Society of America, 4 April 1976, Washington, D.C.

+Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

the first method, henceforth referred to as the oral branch method, simulates nasals on the oral branch of the synthesizer by making use of wider bandwidth settings for the first and second formants through the period of closure; the second method, henceforth referred to as the nasal branch method, preceded and overlayed oral branch stop transitions with the output of the nasal branch of the synthesizer, with one variable formant and a number of higher fixed formants.

Consonant-vowel nonsense syllables in configuration were chosen as the basic stimuli of the experiment. The neutral vowel /a/ was employed for both methods, with continua constructed for bilabial (that is, /ma/ to /ba/) and alveolar (/na/ to /da/) places of articulation. Figure 1 illustrates the extreme ends of the bilabial continua for both methods, with shading through the portions varied.

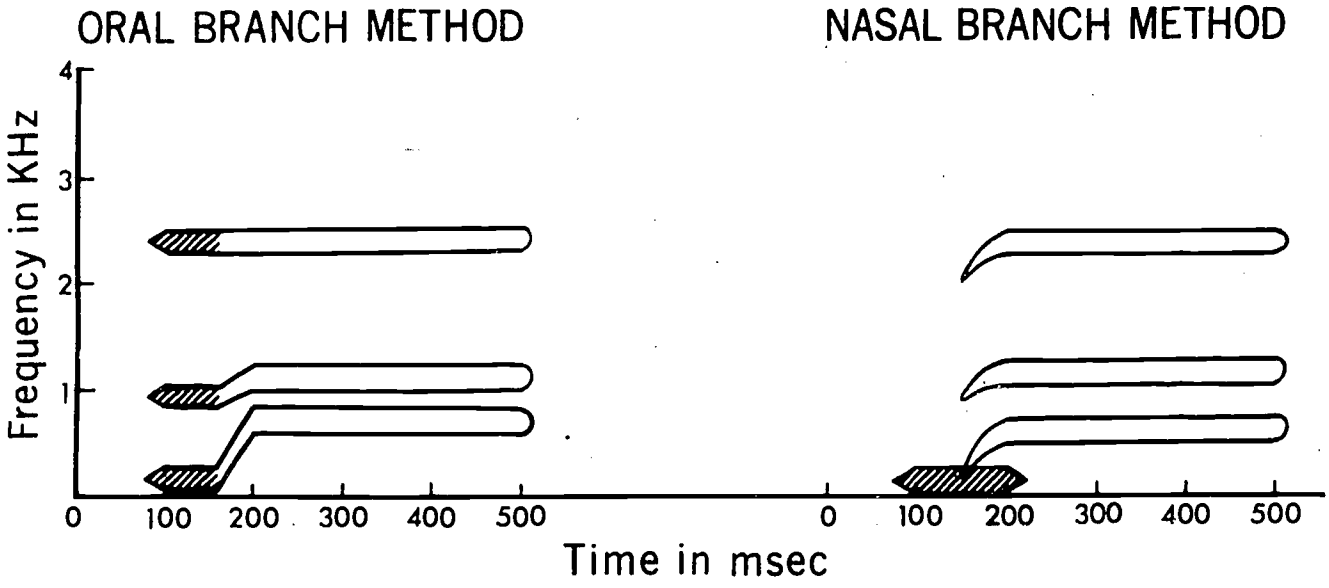


Figure 1: Schematic diagram of extreme nasal bilabial stimuli using both methods. Shading indicates portions varied through continua.

The extreme nasal stimulus of the oral branch method bilabials had the following structure: an 80-msec period of closure with F_1 at 240 Hz, F_2 at 1000 Hz, and F_3 at 2600 Hz, followed by a 40-msec transition to the vowel steady-state values of 820 Hz for F_1, 1180 Hz for F_2, and 2630 Hz for F_3. The vowel duration was 280 msec, resulting in a total stimulus length of 400 msec. Over the nine stimuli of the continuum, three parameters were varied in equal

steps during the period of closure: oral branch amplitude was varied from 0 through 24 dB relative to the amplitude of the vowel steady-state; F_1 bandwidth was varied through the range 205 to 70 Hz; F_2 bandwidth was varied through the range 350 to 80 Hz. The alveolar set consisted of the same acoustic variations with differences only in the initial formant frequencies, namely F_2 at 2000 Hz and F_3 at 2800 Hz.

The extreme nasal stimulus of the nasal branch method bilabials had the following structure: nasal branch excitation with its lowest resonance at 240-Hz through the 80-msec period of closure and the 40-msec stop transition; oral branch resonances initiated the transition from values of 240 Hz for F_1, 1000 Hz for F_2, and 2200 Hz for F_3. The schematic does not illustrate the fixed upper formants of the nasal branch to prevent confusion with the upper formants of the oral branch. Only nasal branch amplitude was varied in this set, through the range 0 dB to -14 dB relative to the amplitude of the vowel steady-state. This variation resulted in an eight-member continuum. The alveolar set contained the same nasal formants and variations, with initial oral branch formant values identical to those of the oral branch method alveolar set.

Procedure

Pilot free-choice labeling tests were given to 20 subjects to determine that end points represented the nasals and stops intended, and to assure that only those two categories were perceived through all continua.

Final forced-choice identification tests, one randomly arranged test for each continuum, consisted of four presentations of each stimulus. Each presentation contained two samples of the stimulus with a one-second interstimulus interval. The interval between presentations was four seconds. Thus, each test set for the oral branch method contained 36 presentations, while each test set for the nasal branch method contained 32 presentations.

Same-different discrimination tests, one for each of the four continua, had the following form: as sames, four randomly arranged pairs of each stimulus were included; as differents, four randomly arranged pairings of adjacent stimuli, two in the order AB and two in the order BA, appeared. The interstimulus interval was 500 msec, and the interval between pairs was four seconds. The resulting oral branch method test sets consisted of 64-pair presentations for each place, while the nasal branch method sets consisted of 56 pairs each.

Subjects

Fifteen paid subjects, all University of Connecticut students, took all eight tests. All were native speakers of American English who claimed normal hearing ability and phonetic naiveté. Eight were right-handed males and seven were right-handed females. The test sets were presented binaurally through headphones in a soundproofed room. Up to five subjects took the tests at one time. All four identification tests were presented first; the four discrimination tests followed.

RESULTS

Figure 2 shows bilabial identification and discrimination results on the oral branch method for subject DC. This subject was given the tests again

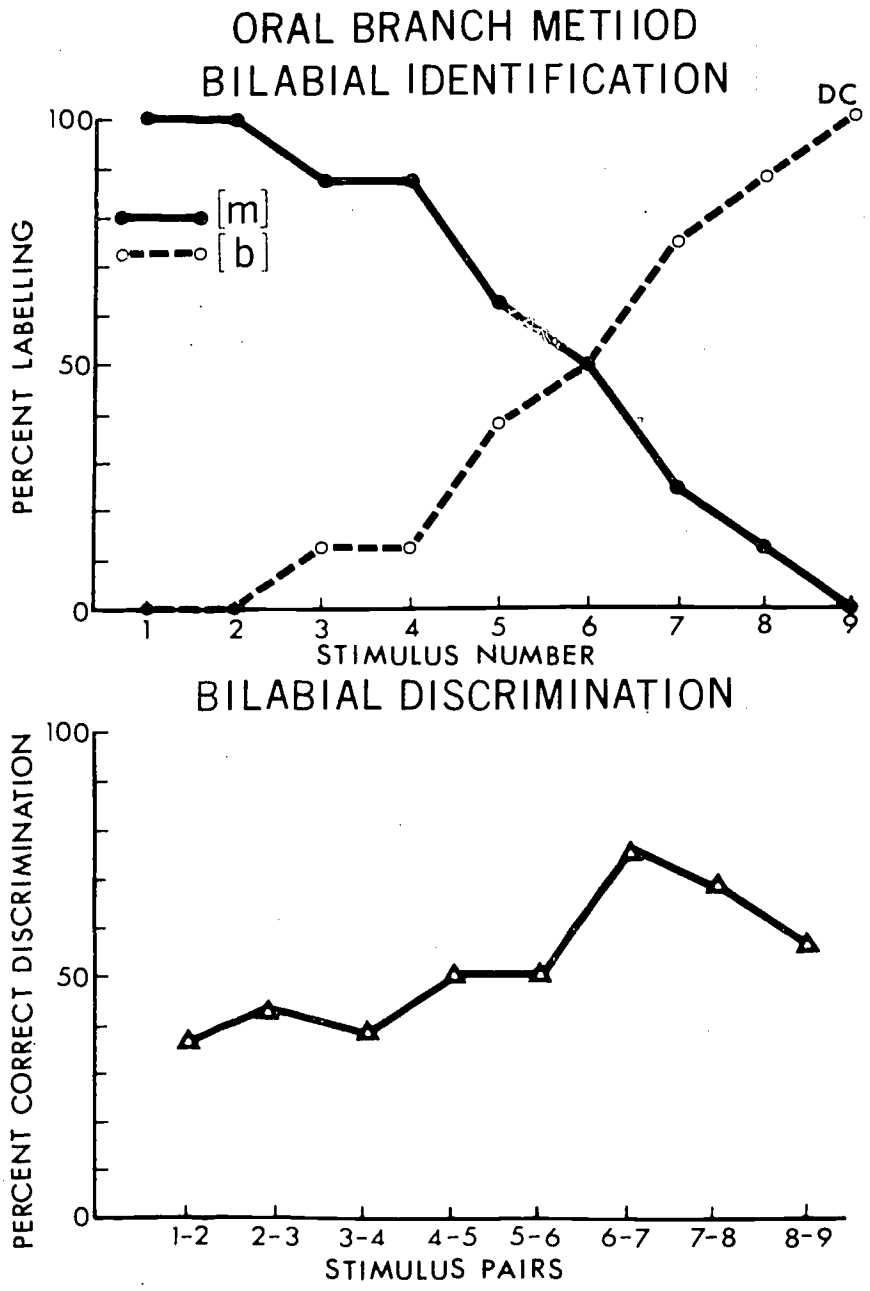


Figure 2: Oral branch method bilabial results for subject DC.

later in order to get more reliable identification and discrimination curves. Each point on the labeling graphs thus represents eight judgments, while each discrimination point represents sixteen responses. Standard methods of data analysis were used.

The labeling crossover occurs in the region of stimulus 6. The discrimination curve rises from a baseline value of approximately 50 percent to a peak of 75 percent in the region between stimuli 6 and 7. Thus, stimuli 1 through 5 formed the nasal category, and stimuli 7 through 9 made up the stop category for this subject. The alveolar data given by DC showed a similar pattern of agreement between labeling crossover and discrimination peak. Crossover was at stimulus 6 and the discrimination peak of 82 percent occurred for the pairing of stimuli 6 and 7.

Figure 3 shows the nasal branch method bilabial results for the same subject. The labeling crossover corresponds to stimulus 5 (which had nasal amplitude of -8 dB relative to the vowel steady-state amplitude). The discrimination peak of 89 percent corresponded to the pairing of stimuli 5 and 6. Again, good agreement is evident between crossover and discrimination peak. The alveolar data for this stimulus type showed a labeling crossover between stimuli 4 and 5, with a discrimination peak of 89 percent in that same region.

Figure 4 depicts discrimination responses for all four test sets given by DC. Correspondences across place for peak and baseline data show good consistency. Notable in the oral branch results is the fact that the slope toward the peak from the nasal category is greater than that away from the peak toward the oral-stop category. This suggests that a peculiarity of the method introduced greater cue value for within-category discrimination of oral stops, than for nasals.

The nasal branch method did not yield consistent correspondences between peak values across place. The peak for the bilabial set peak occurred in the region of stimuli 5 and 6, while the alveolar peak was in the region of stimuli 4 and 5, despite the fact that the acoustic variations were identical. Both sets, nonetheless, yielded well established nasal and stop categories at either end.

Group data for both stimulus types shows essentially the same patterns of distribution as were evident in this individual's responses. Perceptual variation across subjects causes a wider region of indecision around crossovers and wider and somewhat depressed discrimination peaks. What is apparent in even the group data, however, is the consistent correspondence between crossovers and discrimination peaks. These correspondences offer substantial evidence for categorical perception along an oral-nasal continuum.

A relevant problem brought out by the results of this experiment concerns a hypothesis of Fujisaki and Kawashima (1968, 1969, 1970). They proposed that consonants are perceived categorically, and vowels less so, due to the acoustically transient character of the consonantal cues. The continua of this study, however, relied on an 80-msec steady-state noise with varying amplitude for cue value. The categorical perception observed here therefore cannot be attributed to transience of the distinctive oral-nasal cue.

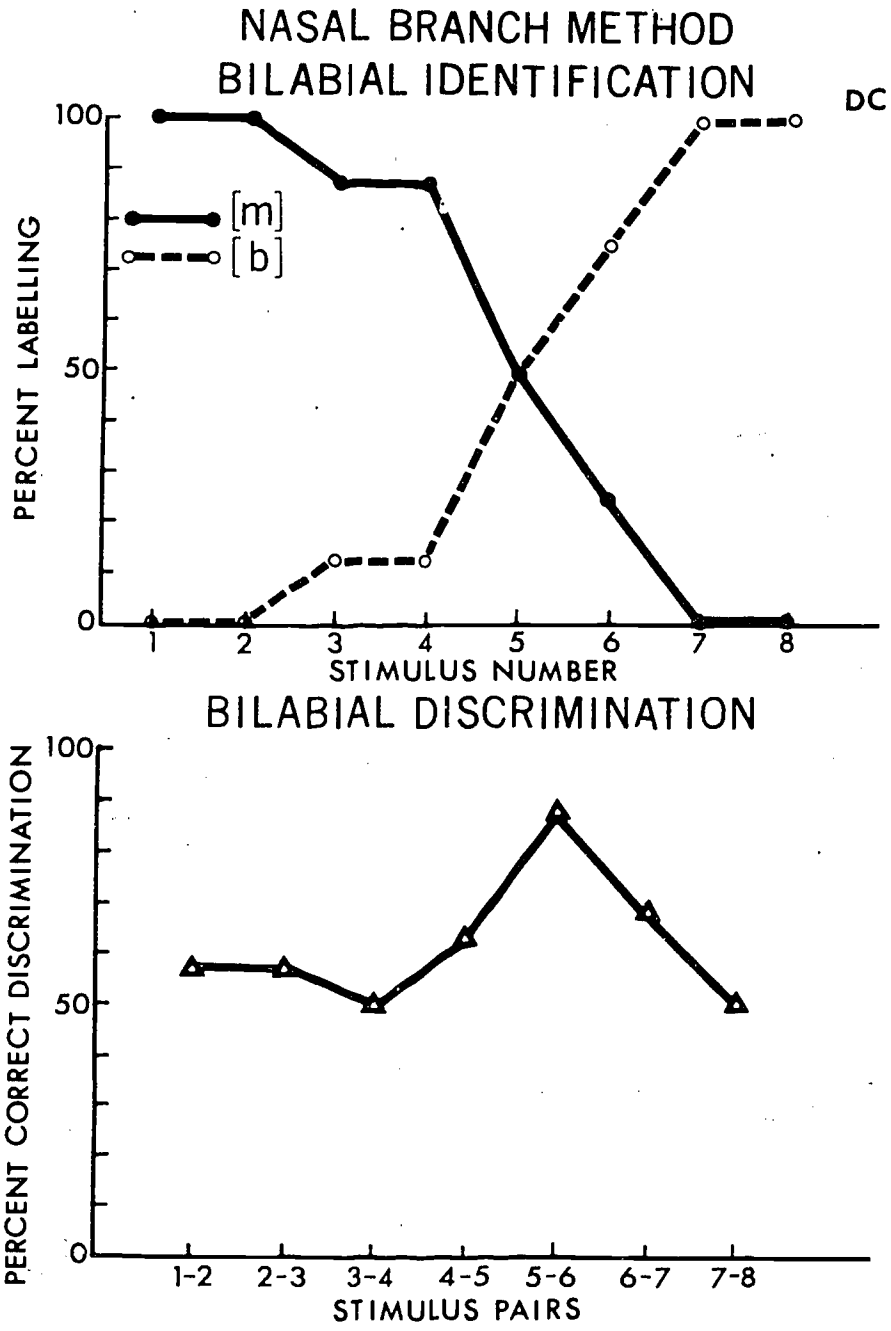


Figure 3: Nasal branch method bilabial results for subject DC.

152

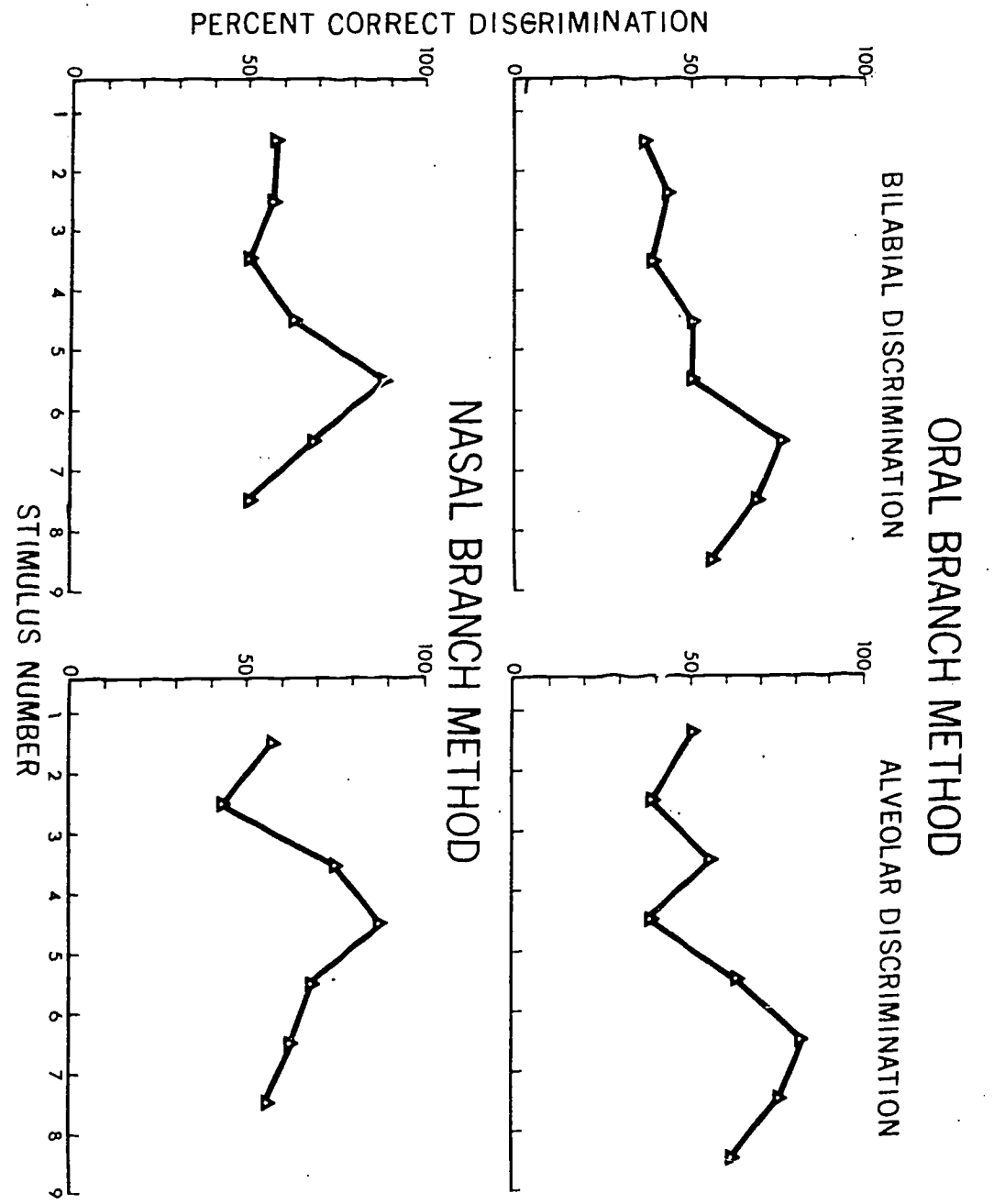


Figure 4: Discrimination results for all four test sets given by DC.

REFERENCES

Fujimura, O. (1962) Analysis of nasal consonants. J. Acoust. Soc. Am. 34, 1865-1875.

Fujisaki, H. and T. Kawashima. (1968) The influence of various factors on the identification and discrimination of synthetic speech sounds. Paper given at the 6th International Congress on Acoustics, Tokyo, Japan, August.

Fujisaki, H. and T. Kawashima. (1969) On the modes and mechanisms of speech perception. Annual Report of the Engineering Research Institute 28, 67-73.

Fujisaki, H. and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute 29, 207-214.

Liberman, A., P. Delattre, F. Cooper, and L. Gerstman. (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monogr. 379, 1-14.

Liberman, A., K. Harris, H. Hoffman, and B. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.

Malécot, A. (1956) Acoustic cues for nasal consonants; an experimental study involving a tape-splicing technique. Language 32, 274-284.

Stop Voicing Production: Natural Outputs and Synthesized Inputs*

Leigh Lisker+

ABSTRACT

In recent years the initial stop consonants of English have been subjected to the relentless attention of speech researchers concerned with the basis for their division into the two category-sets /p,t,k/ and /b,d,g/. The data which suggest the several hypotheses currently entertained have two main sources: natural production of "normal" speakers of the language operating in "normal" fashion, and the responses of persons of like description to synthesized speech stimuli designed to measure the effect of systematic variation of selected acoustic features. The responses required of subjects in tests of synthetic speech can hardly be considered representative of their behavior in responding to natural speech; what the testing of synthetic speech demonstrates is the capability of the perceptual system to deal with the features selected for study, not that this capability is necessarily exploited in the perception of speech. Two kinds of information of relevance to the question of speech cues have not been collected: (1) the extent to which features having potential cue value show variations in natural speech that match the magnitudes tested in synthesis, and (2) the extent to which features for which distinctive function is claimed may be subjected to experimental manipulation by skilled speakers without significantly reducing intelligibility. Experimental data are presented to indicate that at least one acoustic feature that affects stop-voicing perception in synthetic speech is of marginal or less importance in the perception of natural speech.

Let us consider the hypothesis that listeners, in making a stop voicing decision, attend primarily to that part of the signal produced by a stop-vowel articulation which comes immediately after onset of voicing, and that they report /b,d,g/ if they detect a first-formant frequency shift and otherwise report /p,t,k/.

The experiments from which this hypothesis derives were reported by Stevens and Klatt (Stevens and Klatt, 1974), who established that the (Voice Onset Time)

---

*This paper was presented orally at the 90th meeting of the Acoustical Society of America, San Francisco, 3-7 November 1975.

+Also University of Pennsylvania.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

VOT boundary (Lisker and Abramson, 1964) between synthetic /da/ and /ta/ syllables shifts with increase in the duration of the formant transitions, and that when the interval of voiced F_1 transition is decreased to a point where it may no longer be detected by the listener, the reported stop is /t/. A replication and extension of their experiment showing the shift in VOT boundary with transition duration was reported to the Acoustical Society of America by Lisker, Liberman, Erickson, and Dechovitz (1975), and those data, shown in Figure 1, are in close agreement with the finding of Stevens and Klatt, so far as demonstrating that the VOT boundary is not fixed.

However, as emerges more clearly in the lower display of Figure 2, it appears that increasing the transition duration effects an even more drastic shift in the boundary duration of the voiced F_1 transition (VTD) than in VOT. Moreover, the patterns in both the M.I.T. and the Haskins experiments just referred to might be as well described by reference to at least two other measures, namely the F_1 onset frequency and the frequency range of transition. To be sure, of the 20 subjects who provided these data, there was one whose judgments make better sense if described as responses to voiced F_1 transition duration, but for the subjects as a group, VOT seems to have been a more compelling cue.

The data of Figure 1 are replotted in another way in Figure 3 to answer the following question: How effective is varying overall transition duration (or slope), and thereby altering VTD for fixed VOT values, as a factor affecting stop labeling behavior? From this display we see that judgments shift category with increasing transition duration for only three values of VOT, that is, +25, +35, and +45 msec. No transition durations yield more than a negligible number of /ta/ judgments for VOT less than +25 msec, or /da/ judgments for VOT greater than +45 msec. For the three curves of Figure 3 which cross the 50 percent line, the VTD values at the crossover are respectively about 10, 30, and 50 msec, and this shift in VTD boundary value is just double the amount of shift in VOT boundary placement.

It should be remembered in connection with this comparison of VOT and VTD measures, that they are not independent for any particular stimulus, since their sum is, of course, simply the combined durations of burst and transition. VTD is just another measure of voice onset timing, differing from VOT only in that it takes as the temporal reference point the onset of the steadystate vowel instead of the burst. The fact that this point is much less reliably determined in spectrograms of natural speech than in synthetic speech patterns designed with this measure in mind, does not make implausible the hypothesis that a detector which registers the presence of voiced F_1 transition provides the basis for the stop-voicing decision; it does make VTD a rather less convenient measure to apply in the acoustic analysis of stop-vowel sequences.

However, there are other questions with respect to this hypothesis when we consider some other experimental data. If the transition detector fails to sense a voiced F_1 transition, we should regularly obtain a /p,t,k/ judgment; when a stimulus has a transition which under some circumstances provokes /b,d,g/ responses, we should expect it regularly to trigger the detector, barring possibly the special circumstance of fatiguing that is alleged to explain the adaptation effect (Eimas and Corbit, 1973). In Figure 4 we have labeling responses to stimuli derived from naturally produced syllables cut and recombined by an electronic segmentation procedure (Cooper and Mattingly, 1969). The upper

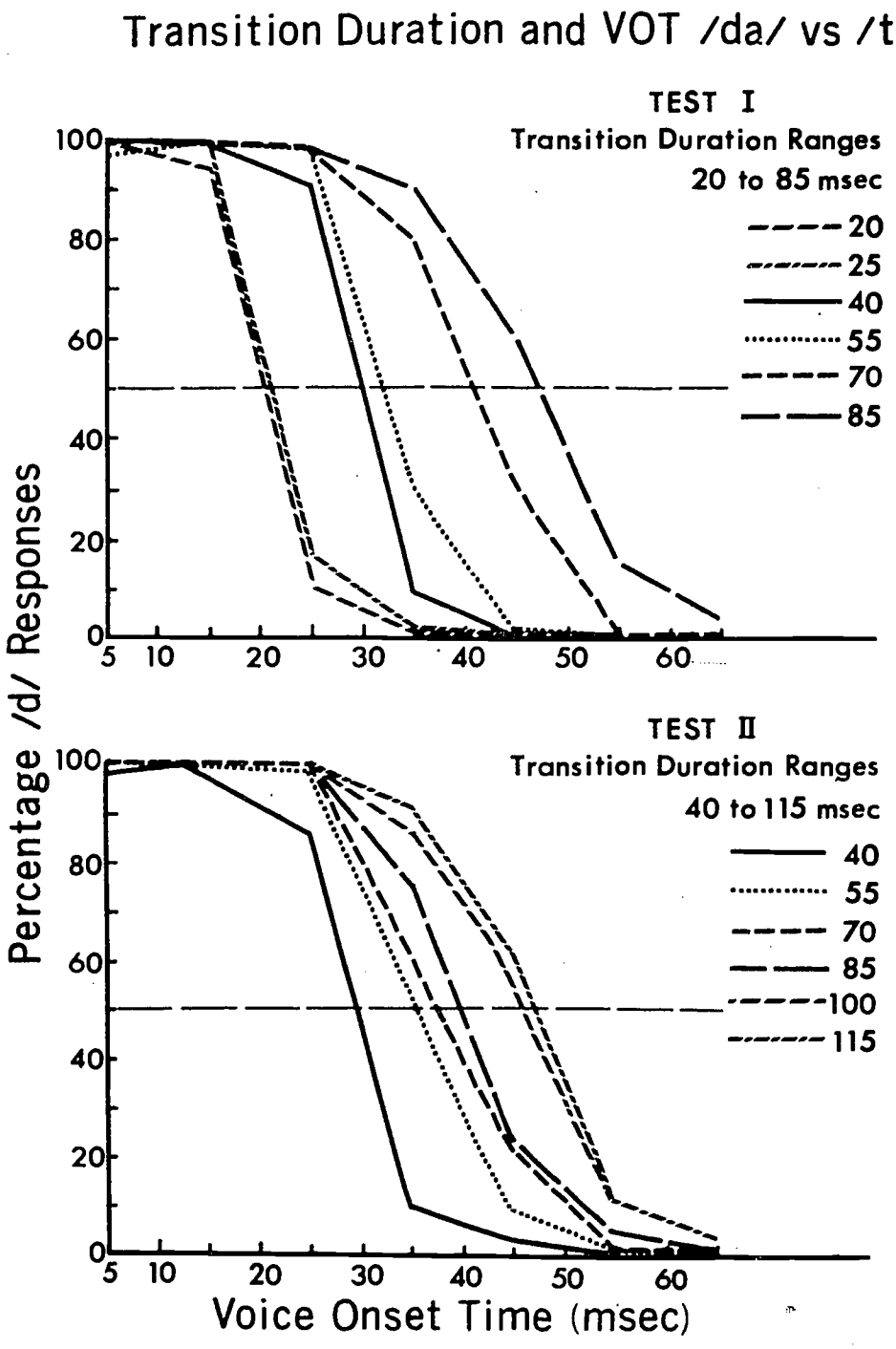


Figure 1: Labeling responses of 20 Ss (8 trials) in a forced-choice task. VOT values were varied in 10 msec steps over a 5-65 msec range (voice onset lagging release-burst onset); transition durations ranged from 20 to 115 msec. To mitigate subject fatigue, the 56 different stimuli were presented in two tests of 42 stimuli each, with the four midrange transition durations (40-85 msec) presented in both tests.

FIGURE 1

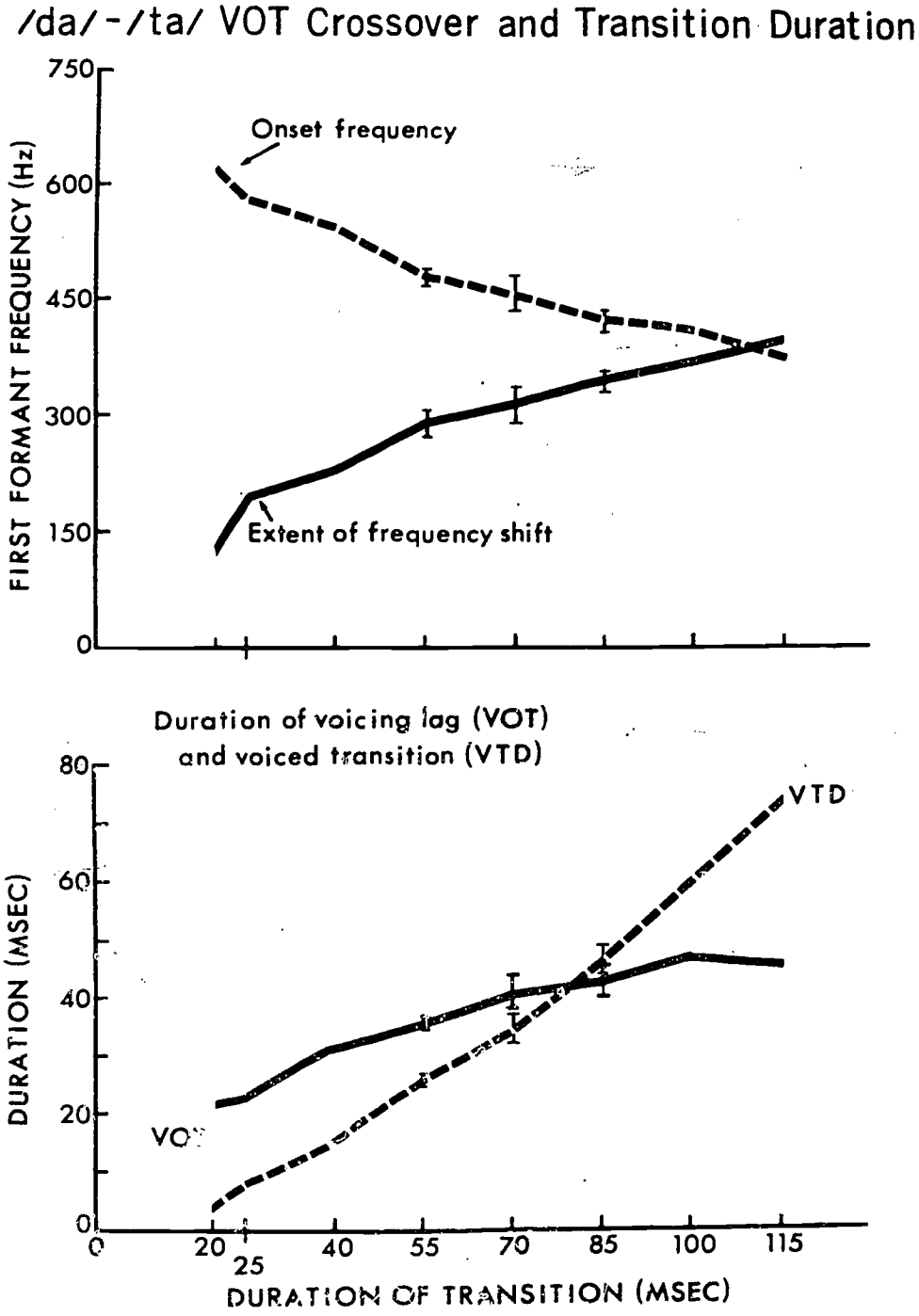


Figure 2: The data of Figure 1 are represented in the four curves shown. For the transition durations tested twice, the curves show overall mean values; the short vertical lines indicate the magnitude of the differences in the means of Test 1 and Test 2.

158

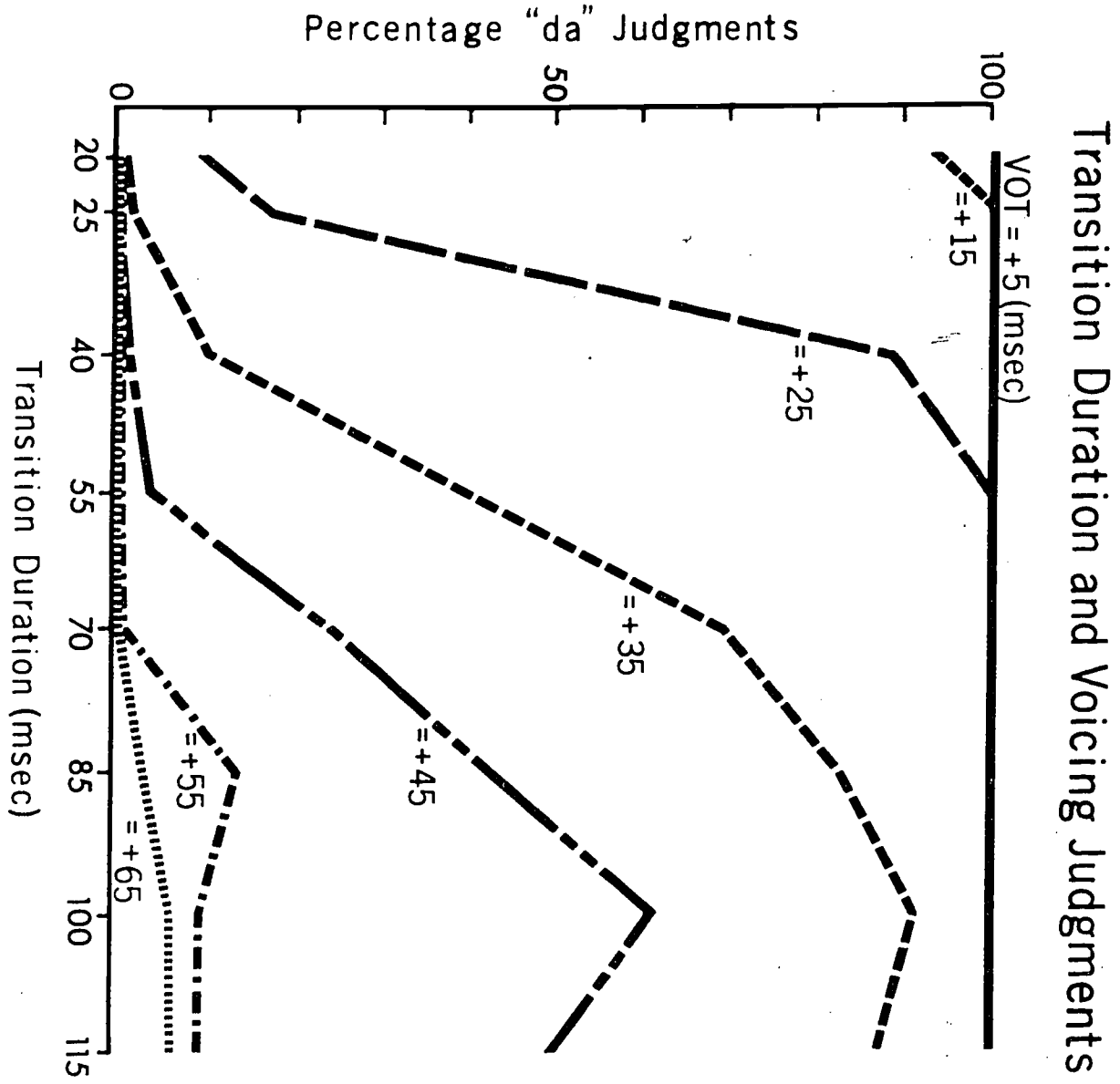


Figure 3: The data of Figure 1 are represented here to show the effect of varying transition duration on the /da/-/ta/ labelings, with VOT as the parameter.

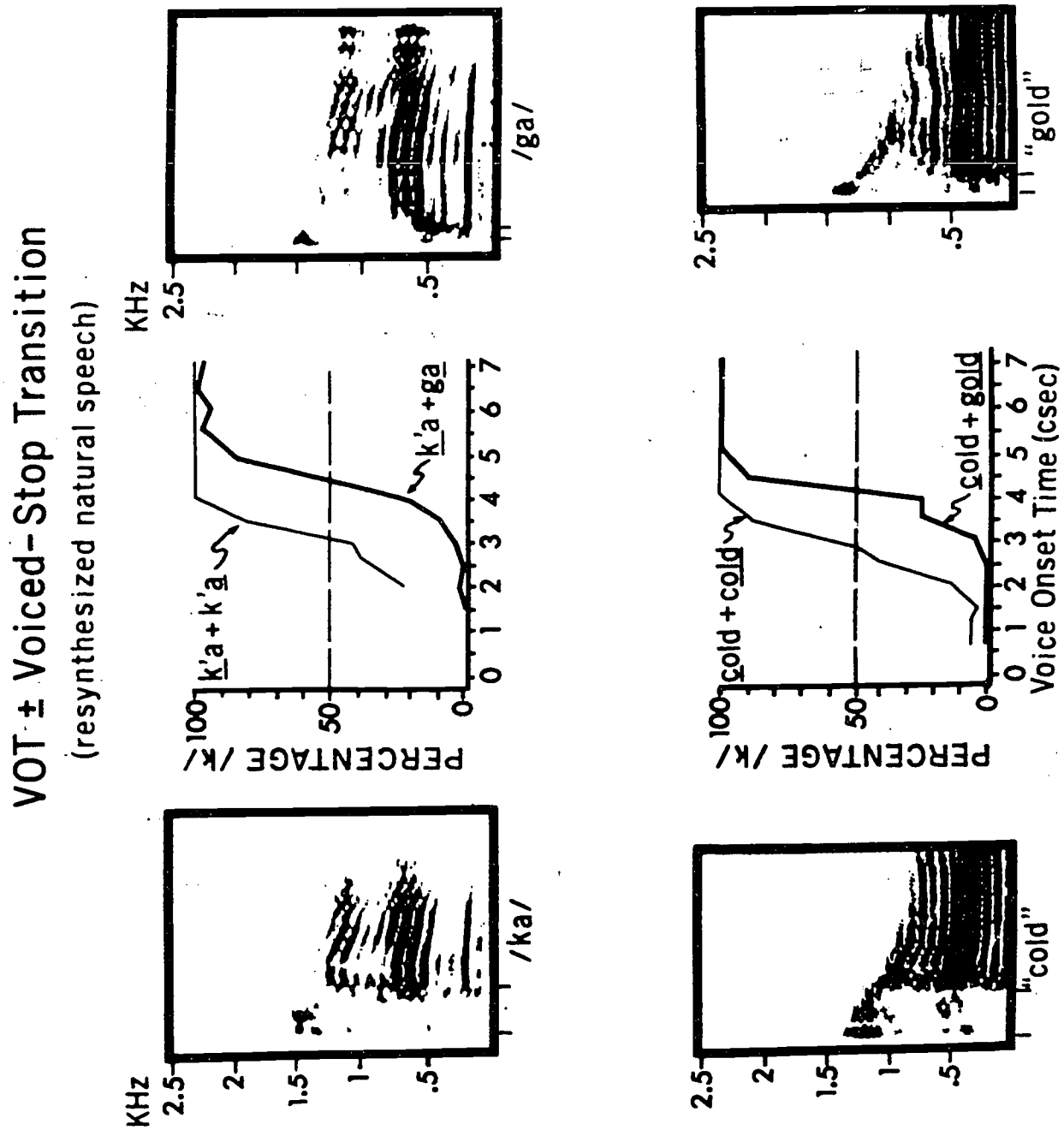


Figure 4: Labeling data obtained with stimuli derived from natural productions of the nonsense syllables /ka/ and /ga/ and the isolated words cold and gold. These syllables were digitalized for computer storage and editing to produce various combinations of initial and final segments of the original signals. The curves in the upper graph show labelings of stimuli composed (1) by varying amounts of the voiceless onset of /ka/ combined with the voiced residue of the same syllable, and (2) by combining these same voiceless /ka/ onsets with the voiced residue of /ga/ obtained by deleting the /g/-burst. The same operations on the monosyllabic words yielded the labelings given in the lower graph. Twelve Ss gave ten responses each to each of 23 stimuli derived from /ka,ga/ and 28 stimuli from cold-gold.

curves give percent /ka/ judgments to two stimulus sets. In one, the VOT of a syllable /ka/ was varied by reducing to varying degrees the duration of the voiceless aspiration, preserving both the original burst and the voiced portion of the syllable; in the other, the same operation was performed on the same /k/ aspiration, but the voiced portion of the stimuli was derived from a spoken /ga/ by deleting the /g/ release transient. Both stimuli yielded, for appropriate VOT values, both /ka/ and /ga/ judgments, with a difference in crossover values of about 15 msec. It is true that the first set produced, to judge from the curves, rather less convincing /ga/ syllables than the second did /ka/ syllables, but despite the absence of a /g/ voiced transition in the first stimulus set, more than half the responses reflected the presence of the short VOT rather than the absence of /g/ transition. Similar operations were used to obtain the stimuli for a second experiment in which subjects were asked to make word rather than phoneme identifications. Here too the absence of /g/ transition did not block "gold" responses to stimuli with VOT values of less that +30 msec, nor did the presence of /g/ transition prevent "cold" judgments for VOT greater than about +42 msec.

For the last experiments to be reported we return to pure synthesis. In the first of these experiments, a stimulus set was generated whose end points are illustrated by the schematic spectrograms on the left and upper right in Figure 5. The first formant, transition and all, was retracted by varying amounts with a maximum delay of 50 msec after onset of the upper formants. The F_1 voiced transition detector should fire equally for any one of the set to produce a /b/ response. The labeling curves of the upper data display show that judgments shifted from /ba/ to /pla/, with an intermediate zone in which both /pla/ and /bla/ were reported. When VOT exceeds +35 msec, it appears that the presence of the buzz-excited F_1 transition is interpreted, not entirely as a /b/ cue, but as a cue also to the presence of an additional phonetic segment preceding the vowel. If /ba/ and /bla/ responses are summed, a /b/-/p/ boundary can be located at about VOT = +40 msec. Recalling that for patterns incorporating F_1 cutback of more orthodox type (Liberman, Delattre, and Cooper, 1958), the boundary value is generally placed near VOT = +25 msec, we find it interesting that the effect of preserving the F_1 transition intact is equal to the VOT boundary difference attributable to presence vs. absence of the /g/ voiced transition in the experiments involving manipulation of naturally produced syllables.

In the two remaining experiments represented in Figure 5, the stimulus sets also contained the left pattern at one extreme, with one of the two lower patterns on the right at the other. In neither of these sets is there any hiss excitation, and no /pa/ or /pla/ responses were elicited. For the set with F_1 retraction, /b/ responses were registered for amounts of retraction up to a magnitude of 50 msec lag behind the voiced upper formants, at which point responses shifted to /bla/. In the final stimulus set tested there was no F_1 cutback, but only a variable delay in shifting F_1 frequency from its low onset value to the steadystate value for the vowel; in this case a shift from /ba/ to /bla/ occurred when the transition was delayed about 25 msec relative to the higher formants. If the same feature detector said to operate in the /b,d,g/ vs. /p,t,k/ decision is also at work here, it seems that the phonetic interpretation of its output is <u>not</u> independent of the temporal relation between the activating feature and the other acoustic properties which signal stop articulation.

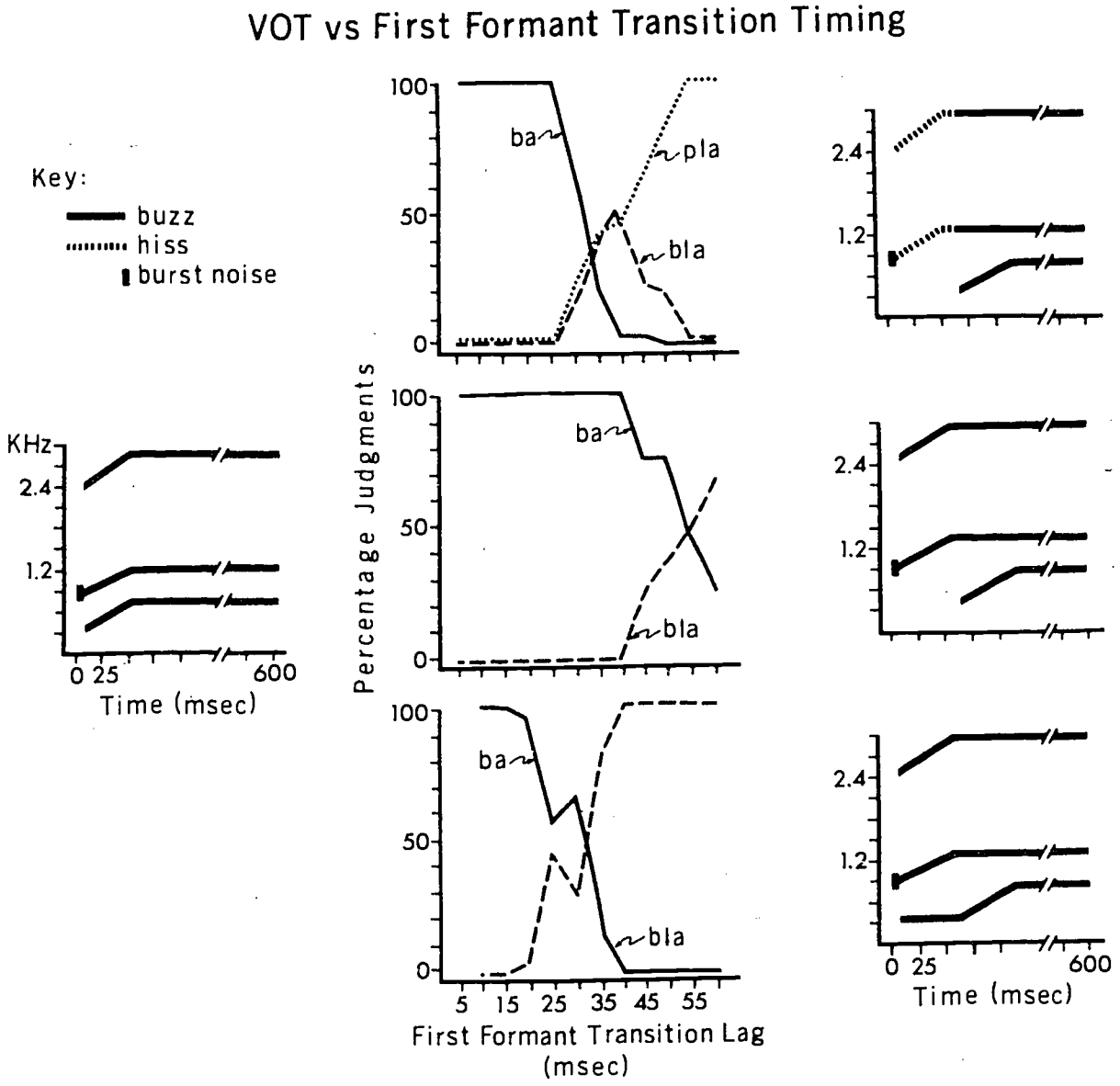


Figure 5: Labeling responses of 10 Ss (8 trials) to three sets of test patterns, all having as variable the timing of first-formant transition. The upper graph shows responses to a stimulus set in which the signal preceding F_1-transition onset was hiss-excited; the mid graph gives data for a set in which only buzz excitation was present; the lower display is of data derived from a set in which the first-formant onset was simultaneous with that of the upper formants, and the first formant was maintained at the onset frequency until onset of the transition.

If variability in VOT boundary location observed in the data from experiments in synthesis means that the voicing decision depends on features in addition to VOT, this by no means implies that some other more stable feature must turn up. Speech being what it is in the temporal dimension generally, it is not totally unexpected that VOT resists any very simple description in its perceptual aspects. It is, I think, also worth mentioning that whereas experimentally determined cue value of VOT and the boundary values of that feature are both consistent with measurements on natural speech, the same cannot be said for transition duration as a significant variable. There is no evidence so far that natural speech exhibits a matching variation correlated with the linguistic difference. In fact, somewhat oddly, one well-known study reporting an extensive set of transition duration data (Lehiste and Peterson, 1961) found consistently shorter durations for /b,d,g/ than for /p,t,k/. In a well-regulated world the reverse relation would allow an occasional imprecision in voice onset timing to be compensated for by a longer duration of voiced transition in /b,d,g/ production, or its shorter duration in /p,t,k/. Demonstrations that features such as fundamental frequency and transition duration are available as stop voicing cues are not invalidated by any evidence that they are not provided by natural speech signals. However, we should be wary of a too ready acceptance of the Panglossian view that speech productive behavior matches perfectly the properties of the auditory-phonetic perceptual mechanism. A good enough match, by definition, "yes." A perfect one? Perhaps yes, but only "perhaps."

REFERENCES

Cooper, F. S. and I. G. Mattingly. (1969) Computer-controlled PCM system for investigation of dichotic speech perception. Haskins Laboratories Status Report on Speech Research SR-17/18, 17-21.

Eimas, P. D. and J. E. Corbit. (1973) Selective adaptation of linguistic feature detectors. Cog. Psychol. 4, 99-109.

Lehiste, I. and G. E. Peterson. (1961) Transitions, glides, and diphthongs. J. Acoust. Soc. Am. 33, 268-277.

Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1958) Some cues for the distinction between voiced and voiceless stops. Lang. Speech 1, 153-167.

Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: acoustical measurements. Word 20, 384-422.

Lisker, L., A. M. Liberman, D. M. Erickson, and D. Dechovitz. (1975) On pushing the voice-onset-time (VOT) boundary about. J. Acoust. Soc. Am. 57 (Suppl. 1), S50 (abstract).

Stevens, K. N. and D. H. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Am. 55, 653-659.

Shifts in Vowel Perception as a Function of Speaking Rate*

Robert R. Verbrugge[+], Donald Shankweiler[++], Winifred Strange[+++], and Thomas R. Edman[+++]

ABSTRACT

In rapid speech, acoustic analysis reveals that steady-state vowel targets characteristically are not reached. Lindblom and Studdert-Kennedy (1967) found in an experiment with synthetic speech that listeners showed a shift in the boundary between medial vowels /ɪ/ and /ʊ/ with variations in the rate and direction of formant transitions. Apparently, perceivers compensate for simulated articulatory undershoot by perceptual overshoot. An experiment with natural speech demonstrated shifts in the acoustic criteria listeners employed in vowel recognition as a function of perceived rate of utterance. Nine American English vowels in /p-p/ environment were produced by a panel of 15 talkers in a fixed sentence frame. The destressed, rapidly-articulated /p-p/ syllables were excised from the tape recording and assembled into listening tests. Errors on vowels in the excised syllables averaged 23.8 percent. Errors jumped to 28.6 percent when point-vowel precursors were introduced, while presentation of the syllables in their original sentence context reduced errors to 17.3 percent. The results suggest that sentence

---

* This paper was presented at the 91st meeting of the Acoustical Society of America, Washington, D.C., 4-9 April 1976. A more complete description of this research may be found in Verbrugge, Strange, Shankweiler, and Edman (in press). An extended discussion of the problem of perceptual constancy in speech perception may be found in Shankweiler, Strange, and Verbrugge (in press).

[+] Also Department of Psychology, University of Michigan, Ann Arbor.

[++] Also University of Connecticut, Storrs.

[+++] Center for Research in Human Learning, University of Minnesota, Minneapolis.

Acknowledgement: The work was supported by grants to the Center for Research in Human Learning, University of Minnesota, Minneapolis, and to Haskins Laboratories, New Haven, Connecticut, from the National Institute of Child Health and Human Development, by grants awarded to D. Shankweiler and J. J. Jenkins by the National Institute of Mental Health, and by a fellowship to R. R. Verbrugge from the University of Michigan Society of Fellows.

[HASKINS LABORATORIES: Status Report on Speech Research SR-47 (1976)]

context aids vowel identification by allowing adjustment primarily to a talker's tempo, rather than to the talker's vocal tract characteristics.

Acoustic measurements of vowels in continuous speech often show a deviation of formant frequencies from the steady-state values typical of slow citation-form speech. Lindblom (1963) characterized this effect as an "undershoot" of "target" frequencies in rapid speech. He argued that the degree of undershoot is a systematic function of the talker's tempo; thus, the underlying target may be fully specified by the formant contours even though the target value is never reached. Lindblom went on to suggest that listeners could compensate for the undershoot and infer the underlying target if they had information about the tempo of articulation. This information would presumably be carried by the formant contours and by syllable duration. Lindblom and Studdert-Kennedy (1967) found some support for this idea in a study with synthetic speech.

In this study, we used natural speech to determine the extent of the perceptual problem posed by rapid articulation. We were interested in what information allows listeners to achieve constancy of vowel perception across different speaking rates. In particular, we wondered whether the formant contours of a single syllable are sufficient to specify a talker's tempo, or whether longer stretches of speech are necessary.

Imagine snatching a syllable from running speech and presenting it to a listener for identification. It seems reasonable to suppose that the vowel in such a syllable would be less identifiable than the same vowel in a syllable spoken in citation form; the syllable will be shorter and there may be no region approximating a steady state.

To test this supposition, we asked a panel of fifteen talkers to produce vowels at two different tempos: (1) in /p-p/ syllables spoken in citation-form, and (2) in /p-p/ syllables spoken in destressed position in the context of a full sentence. In the citation-form syllable test, each of the nine English monophthongs was represented five times, spoken by different talkers. Thus, listeners heard a total of 45 /p-vowel-p/ syllables. For the destressed syllable test, corresponding /p-p/ syllables for each talker were excised from the carrier sentences and assembled into a comparable test series. Separate groups of listeners heard the two tests.

The results are shown in Figure 1. On the average, listeners misidentified 17 percent of the vowels in citation-form syllables and 24 percent of the vowels in destressed syllables. In a sense, the 24 percent error rate for the excised syllables is surprisingly low since the talkers varied from trial to trial, the syllables contained little or no steady-state energy, the syllable centers deviated from target values, and the syllables were very short in duration. Even so, the error rate was significantly greater than that for citation-form syllables.

There are two possible reasons for the increase in errors on destressed syllables. The increase may reflect a greater overlap of cross-sectional formant frequency values for the destressed vowels or it may be a result of misperceiving the talkers' tempo.

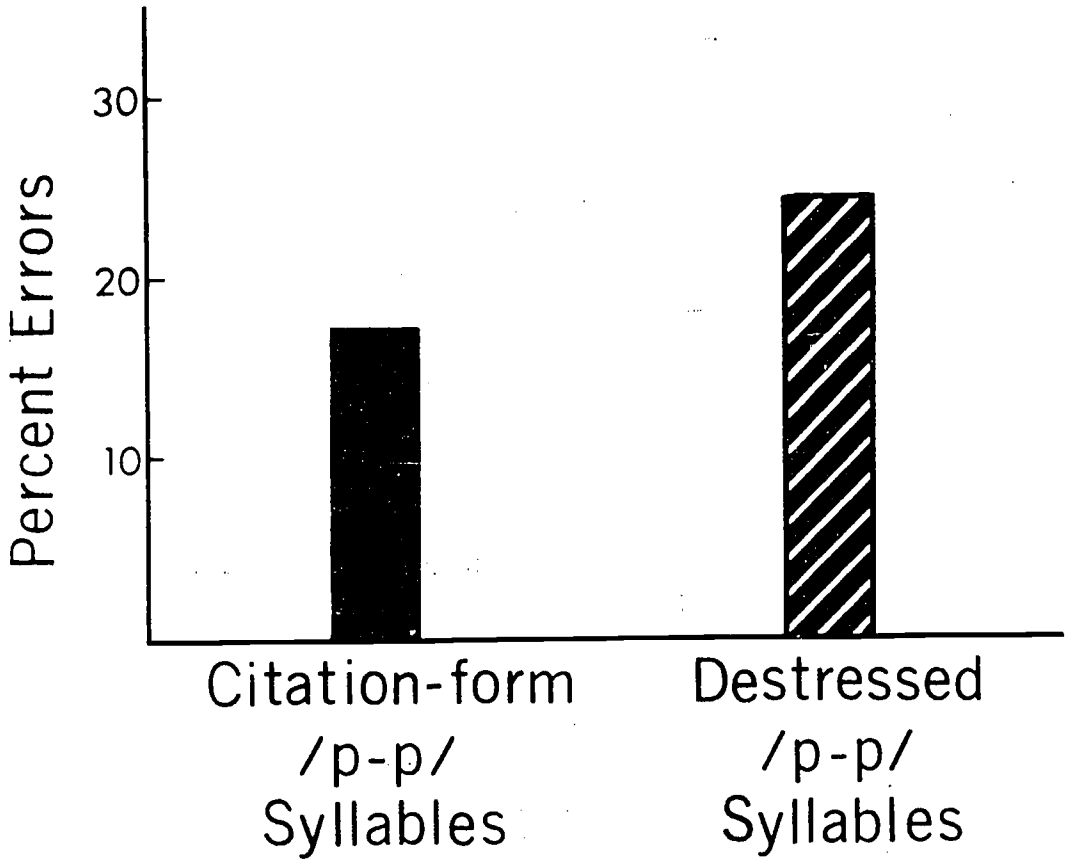


Figure 1: Mean percent errors in identifying vowels in citation-form /p-p/ syllables and in destressed /p-p/ syllables excised from sentence context.

An analysis of listeners' errors provides one means of answering this question. We applied an extension of Luce's Choice Axiom to the confusion matrices for each condition. The Luce model assumes that response probabilities are a function of two types of parameters: (1) similarities between each pair of stimulus categories, and (2) response biases for the various categories. If the increased errors on destressed syllables were due primarily to increased spectral overlap, we would expect a widespread increase in pairwise similarity values. No such increase was found. The major difference between citation-form and destressed syllables was found in response biases. Listeners were biased toward hearing the shorter vowel alternatives: for example, hearing /pæp/ as /pɛp/, /pɑp/ as /pʌp/, and /pup/ as /pʊp/. These bias shifts suggest that listeners treated the excised syllables as if they had originally been spoken in isolation, that is, as if they had been spoken more slowly in citation form.

Thus, the error pattern suggests that information about a talker's tempo is critical in achieving constancy and that the information is not completely specified at the single syllable level. To make a more direct test of this, we prepared two additional listening conditions in which the same destressed syllables were embedded in longer stretches of speech. These contexts were intended to establish two different rates of articulation. In one condition, we preceded each test syllable with the precursors /hi/ha/hu/ spoken at a slow rate by the same talker. In the second condition, we presented the syllables in their original sentence context: "The little /p-p/'s chair is red."

Results for the two context conditions are presented in Figure 2. The three bars on the right depict average error rates for vowels in the destressed syllables. Following the point-vowel precursors, errors rose significantly to 29 percent, compared to 24 percent errors for the syllables heard in isolation. In contrast, errors dropped significantly to 17 percent when listeners heard the syllables embedded in their original sentence context.

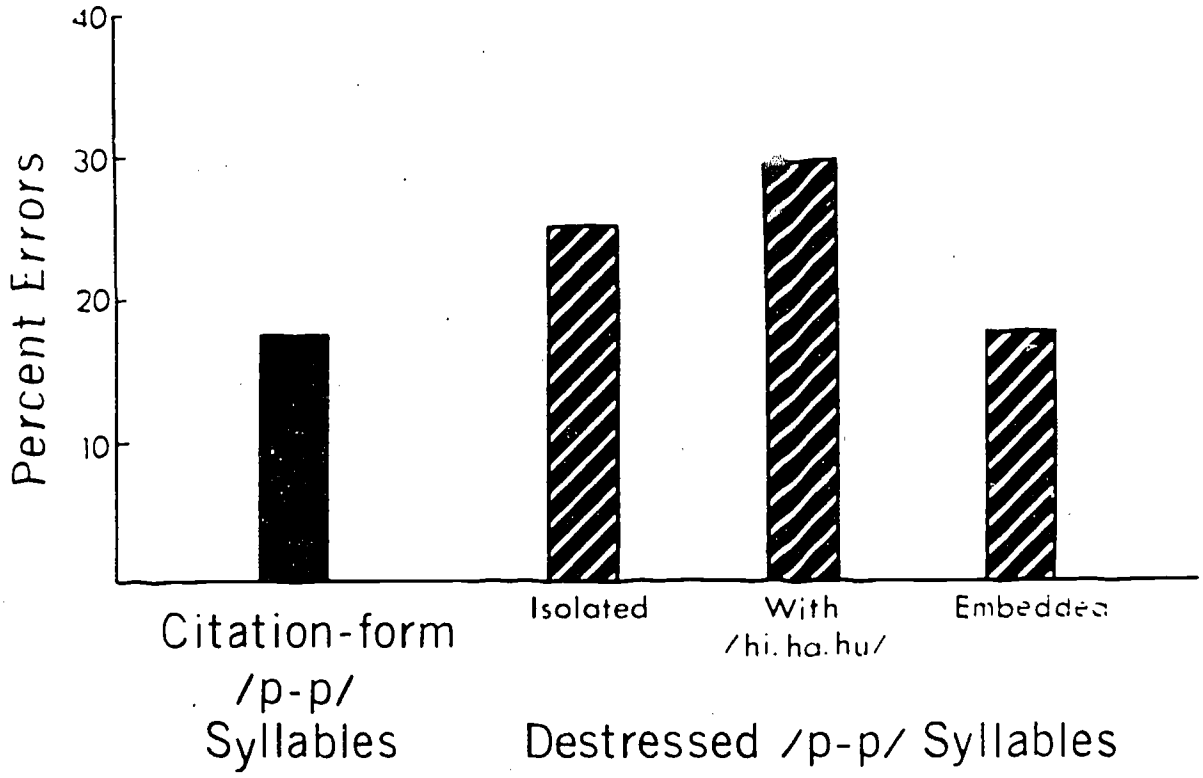


Figure 2: Mean percent errors in identifying vowels under four conditions: citation-form /p-p/ syllables, destressed /p-p/ syllables excised from sentence context, the excised syllables when preceded by point-vowel precursors, and the destressed syllables embedded in their original sentence context.

An analysis of listeners' errors again proves very instructive. The dominant effect of the point-vowel precursors was to enhance the pattern of errors found when destressed syllables were heard in isolation. Response biases toward /pɛp/ and /pʌp/ were even larger than before. Apparently listeners treated the test syllables as if they had been spoken slowly, like the precursors, in citation form. Thus, the mismatch between perceived and actual tempo was even greater than it had been for the isolated syllables.

When listeners heard the test syllables embedded in sentence context, there were no major response biases of the kind found for isolated syllables. The biases toward /pɛp/, /pʌp/, and /pʊp/ were substantially smaller in sentence context, and as a consequence, there were fewer errors for /pæp/, /pap/, /pɔp/, and /pup/. The original sentence context apparently contained sufficient information to specify tempo accurately and to preclude the kinds of errors we found in the other two conditions.

It is interesting to note that the error rate for destressed vowels in sentence context was very close to the 17 percent rate for vowels in citation-form syllables; the difference between the two conditions was not significant. This suggests that a vowel will be identifiable to the same degree whenever the full natural utterance is available to define the tempo. In the case of short sentences, the whole sentence is probably the natural unit of articulation. In the case of citation-form syllables, the syllable is a self-contained unit of articulation. There seems to be a stable level of identifiability when the full natural unit is available to the listener. Failure to reach steady-state target frequencies does not necessarily make a syllable more ambiguous. If a listener is tuned to the ongoing tempo, a short destressed syllable is as fully determinate as a citation-form syllable.

The results for the two context conditions raise a further possibility: a carrier sentence may aid identification more by defining tempo and stress than by defining the spectral range for a given talker. In the point-vowel precursor test, the information about rate of utterance was of greater significance for perception than the range of spectral values provided by the precursor string. Some researchers have proposed that experience with a talker's point vowels should reduce the ambiguity of subsequent utterances (cf. Lieberman, 1973). In the present study, at least, misinformation about tempo clearly outweighed any helpful information to be gained from exposure to the point vowels.

A similar conclusion seems appropriate for the sentence context condition: prosodic information was of greater perceptual significance than any available information about the talker's spectral range. As before, listeners' errors provide a useful means for distinguishing these alternatives. If a carrier sentence mainly adjusts listeners to a talker's spectral range, we would expect extensive reductions in vowel similarities (specifically, reductions in the ambiguities due to talker differences). On the other hand, if the sentence mainly adjusts listeners to the talker's rate of speech, we would expect changes in the response biases for short and long vowel alternatives--and this is what we observed. Thus, the identification of vowels in sentence context was more sensitive to the transformation produced by tempo and stress than to the transformation produced by varying talkers.

In general, these results point to the importance of dynamic properties of speech in the perception of vowels. The effects of prosody on the perception of phonemic segments deserves fuller exploration.

REFERENCES

Lieberman, P. (1973) On the evolution of language: A unified view. Cognition 2, 59-94.

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773-1781.

Lindblom, B. E. F. and M. Studdert-Kennedy. (1967) On the role of formant transitions in vowel recognition. J. Acoust. Soc. Am. 42, 830-843.

Shankweiler, D., W. Strange, and R. R. Verbrugge. (in press) Speech and the problem of perceptual constancy. In Perceiving, Acting, and Knowing: Toward an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Hillsdale, N.J.: Lawrence Erlbaum Associates).

Verbrugge, R. R., W. Strange, D. P. Shankweiler, and T. R. Edman. (1976) What information enables a listener to map a talker's vowel space? J. Acoust. Soc. Am. 60, 198-212.

II. PUBLICATIONS AND REPORTS

III. APPENDIX

PUBLICATIONS AND REPORTS

Bailey, Peter J. (1976) Some properties of the auditory component of selective adaptation. Journal of the Acoustical Society of America, 59, Suppl. 1, 526(A).

Geffner, D. S. and M. F. Dorman. (1976) Hemispheric specialization for speech perception in four-year-old children from low and middle socioeconomic classes. Cortex, 12, 1.

Healy, A. F. and J. E. Cutting. (1976) Units of speech perception: Phoneme and syllable. J. Verbal Learn. Verbal Behav., 15, pp. 73-83.

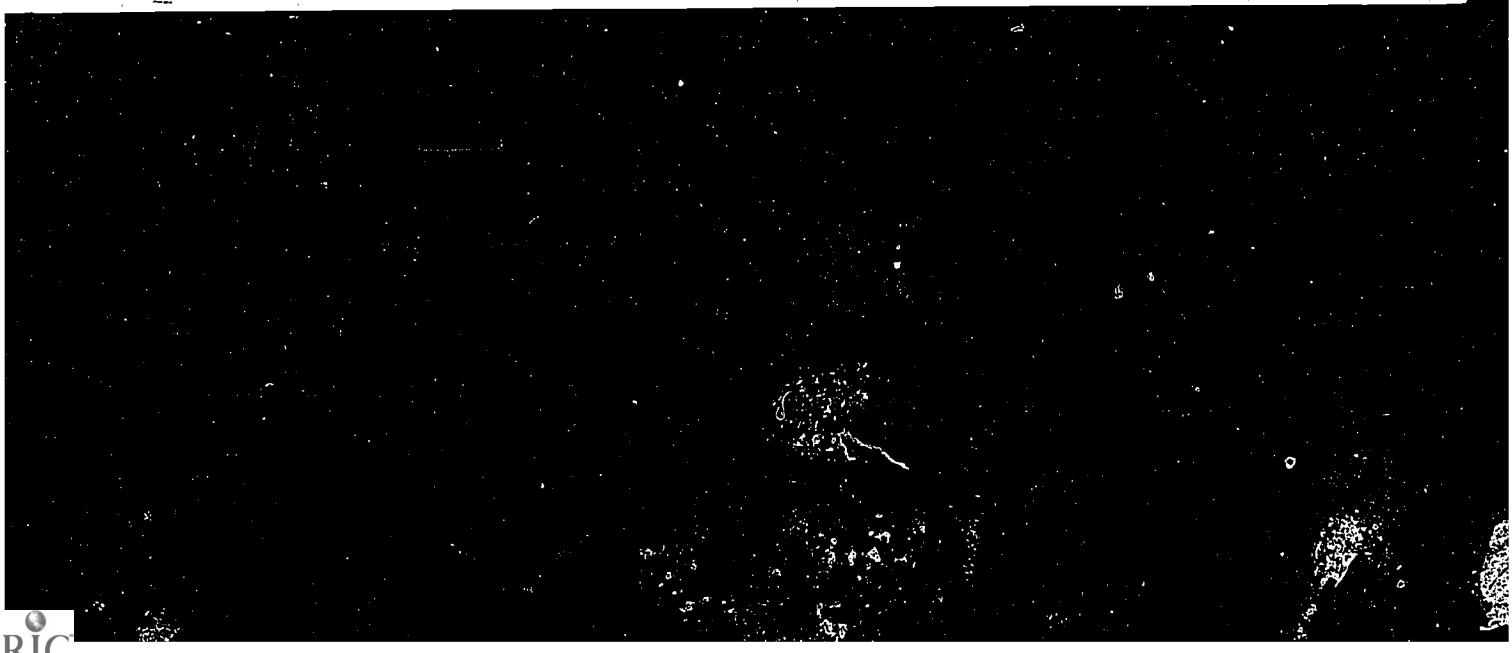
Raphael, Lawrence J., Michael F. Dorman, and Alvin M. Liberman. (1976) Some ecological constraints on the perception of stops and affricates. Journal of the Acoustical Society of America, 59, Suppl. 1, 525(A).

Repp, Bruno H. (1976) Effects of fundamental frequency contrast on identification and discrimination of dichotic CV syllables at various temporal delays. Memory and Cognition, 4, 75-90.

Repp, Bruno H. (1976) Identification of dichotic fusions. Journal of the Acoustical Society of America, 60, 456-469.

Rubin, P., M. T. Turvey, and P. van Gelder. (1976) Initial phonemes are detected faster in spoken words than in spoken nonwords. Perception and Psychophysics, 19, no. 5, 394-398.

Studdert-Kennedy, M. (1976) Speech perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Luss. (New York: Academic Press)



UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Haskins Laboratories, Inc. 270 Crown Street New Haven, Connecticut 06510	Unclassified
	2b. GROUP N/A

3. REPORT TITLE

Haskins Laboratories Status Report on Speech Research, No. 47, July - September 1976

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Interim Scientific Report

5. AUTHOR(S) (First name, middle initial, last name)

Staff of Haskins Laboratories; Alvin M. Liberman, P.I.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
September 1976	175	294

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
DE-01774 RR-5596 HD-01994 V101(134)P-342 N00014-76-C-0591 DAAB03-75-C-0419(L433) N01-HD-1-2420	SR-47 (1976)
	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.*

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
N/A	See No. 8

13. ABSTRACT

This report (1 July - 30 September 1976) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation of its investigation, and practical applications. Manuscripts cover the following topics:

-Stop Consonant Recognition: Release Bursts and Formant Transitions
-Modes of Perceiving: Abstracts, Comments and Notes
-Discrimination Intensity Differences Carried on Formant Transitions Varying in Extent and Duration
-Discrimination Functions Predicted from Categories in Speech and Music
-Right-ear Advantage for Musical Stimuli Differing in Rise Time
-Dichotic Competition of Speech Sounds: Role of Acoustic Stimulus Structure
-Distance Measures for Speech Recognition - Psychological and Instrumental
-Laryngeal Timing in Consonant Distinctions
-Phonetic Aspects of Time and Timing
-Static and Dynamic Acoustic Cues in Distinctive Tones
-Effects of Selective Adaptation on Voicing in Thai and English
-Perception of Nonspeech by Infants
-Categorical Perception Along Oral-Nasal Continuum
-Stop Voicing Production: Natural Outputs and Synthesized Inputs
-Shifts in Vowel Perception as Function of Speaking Role

DD FORM 1 NOV 65 1473 (PAGE 1)
S/N 0101-807-6811

UNCLASSIFIED
Security Classification

A-31408

*This document contains no information not freely available to the general public. It is distributed primarily for library use.

APPENDIX

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers SR-21/22 to SR-45/46:

Status Report		DDC	ERIC
SR-21/22	January - June 1970	AD 719382	[illegible]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 A103325	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

AD numbers may be ordered from: U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22151

ED numbers may be ordered from: ERIC Document Reproduction Service
Computer Microfilm International Corp. (CMIC)
P.O. Box 190
Arlington, Virginia 22210

Haskins Laboratories Status Report on Speech Research is abstracted in Language and Behavior Abstracts, P.O. Box 22206, San Diego, California 92122.