DOCUMENT RESUME

ED 102 633 CS 501 001

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-December 31, 1974. Report No. SR-39/40 (1974).

INSTITUTION Haskins Labs., New Haven, Conn.

REPORT NO SR-39/40(1974)

PUB DATE Dec 74

NOTE 265p.; See related document ED 094 445

EDRS PRICE MF-$0.76 HC-$13.32 PLUS POSTAGE

DESCRIPTORS *Educational Research; Higher Education; *Language Skills; Language Usage; Linguistic Patterns; *Research Design; *Speech; Speech Skills; *Word Recognition

ABSTRACT

This report, covering the period from July 1 to December 31, 1974, is one of a regular series on the status and progress of studies concerning the nature of speech, instrumentation for its investigation, and practical applications. The manuscripts in this report include: speech perception, speech recognition through spectrogram matching, phonetic segmentation and recording in the beginning reader, word recall in aphasia, pitch control in speech, physiological control of durational differences between vowels preceding voiced and voiceless consonants in English, jaw movement during speech, and laryngeal activity in Danish consonant production. (RB)

SR-39/40 (1974)

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 - 31 December 1974

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.)

ACKNOWLEDGMENTS

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

Research and Development Division of the Prosthetic and Sensory Aids Service, Veterans Administration
Contract V101(134)P-71

Office of Naval Research, Information Systems Branch
Contract N00014-67-A-0129-0001

Advanced Research Projects Agency, Information Processing Technology Office, under contract with the Office of Naval Research, Information Systems Branch
Contract N00014-67-A-0129-0002

National Institute of Child Health and Human Development
Contract NIH-71-2420

National Institutes of Health
General Research Support Grant RR-5596

HASKINS LABORATORIES

Personnel in Speech Research

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

Investigators

Arthur S. Abramson*
Fredericka Bell-Berti*
Gloria J. Borden*
Earl Butterfield[1]
James E. Cutting*
Christopher J. Darwin[2]
Ruth S. Day*
Michael F. Dorman*
Peter Eimas[3]
Jane H. Gaitenby
Thomas J. Gay*
Katherine S. Harris*
Philip Lieberman*
Leigh Lisker*
Ignatius G. Mattingly*
Paul Mermelstein
Seiji Niimi[4]
Patrick W. Nye
Lawrence J. Raphael*
Donald P. Shankweiler*
George N. Sholes
Michael Studdert-Kennedy*
Michael T. Turvey*
Tatsujiro Ushijima[4]

Technical and Support Staff

Eric L. Andreasson
Elizabeth P. Clark
Cecilia C. Dewey
Janneane F. Gent
Donald S. Hailey
Harriet G. Kass*
Diane Kewley-Port*
Sabina D. Koroluk
Christina R. LaColla
Roderick M. McGuire
Agnes McKeon
Terry F. Montlick
Susan C. Polgar*
Loretta J. Reiss
William P. Scully
Richard S. Sharkany
Edward R. Wiley
David Zeichner

Students*

David Agard
Mark J. Blechner
Susan Brady
David Dechovitz
Susan Lea Donald
G. Campbell Ellison
Donna Erickson
F. William Fischer
Carol A. Fowler
Frances J. Freeman
Gary M. Kuhn
Andrea G. Levitt
Roland Mandler
Terrance M. Nearey
Barbara R. Pick
Barbara Pober
Robert F. Port
Robert Remez
Philip E. Rubin
Helen Simon
Elaine E. Thompson
James M. Vigorito

*Part-time

[1] Visiting from the University of Kansas, Lawrence.

[2] Visiting from the University of Sussex, Brighton, England.

[3] Visiting from Brown University, Providence, R. I.

[4] Visiting from University of Tokyo, Japan.

CONTENTS

I. Manuscripts and Extended Reports

I. MANUSCRIPTS AND EXTENDED REPORTS

Speech Perception*

Michael Studdert-Kennedy+
Haskins Laboratories, New Haven, Conn.

"The understanding of speech involves essentially the same problems as the production of speech.... The processes...have too much in common to depend on wholly different mechanisms" (Lashley, 1951:120).

INTRODUCTION

We can listen to speech at many levels. We can listen selectively for meaning, sentence structure, words, phones, intonation, chatter, or even, at a distance, Auden's "high, thin, rare, continuous hum of the self-absorbed." This paper is concerned solely with phonetic perception, the transformation of a more-or-less continuous acoustic signal into what may be transcribed as a sequence of discrete, phonetic symbols. The study of speech perception, in this sense, has in recent years begun to adopt the aims, and often the methods, of the information-processing models of cognitive psychology which have proved fruitful in the study of vision (Neisser, 1967; Haber, 1969; Reed, 1973). The underlying assumption is that perception has a time-course, during which information in the sensory array is "transformed, reduced, elaborated" (Neisser, 1967:4) and brought into contact with long-term memory (recognized). The experimental aim is to intervene in this process (either directly or by inference) at various points between sensory input and final percept, in order to discover what transformations the original information has undergone. The ultimate objective is to describe the process in terms specific enough for neurophysiologists to search for neural correlates.

Let us begin by considering how speech perception differs from general auditory perception. It does so in both stimulus and percept. First, the sounds of speech constitute a distinctive class, drawn from the set of sounds that can be produced by the human vocal mechanism. They can be described, to an approximation, as the output of a filter excited by an independent source. The source is the flow of air from the lungs, modulated at the glottis to produce a quasi-periodic sound, or above the glottis to produce a noisy turbulence. The filter

*Chapter prepared for Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass (Springfield, Ill.: C. C Thomas, in press).

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

Acknowledgment: I thank Alvin Liberman, Ignatius Mattingly, and Donald Shankweiler for their valuable comments and criticism, David Pisoni for much fruitful conversation and for drawing my attention to the work of Eleanor Rosch.

is the supralaryngeal vocal tract, whose varying configurations give rise to varying resonances (formants). The resulting sound wave may be displayed as an oscillogram or, after spectral analysis, as a spectrogram. It is important to bear in mind that the spectrogram does not display the sensory input, but a transformation of that input, often presumed to represent the output at an early stage of auditory analysis. [For accounts of the speech signal and its mechanisms of production, see Fant, 1960; Stevens and House, 1972; Kent (in Lass, in press); Babcock (in Lass, in press).]

Here our main concern is to stress functional differences between speech and nonspeech acoustic structure in perception. Speech does not lie at one end of an auditory (psychological) continuum which we can approach by closer and closer acoustic (physical) approximation. The sounds of speech are distinctive. They form a set of "natural categories" similar to those described by Rosch (1973). She studied form and color perception among the Dani, a Stone-Age people of New Guinea, whose language contains "only two color terms which divide the color space on the basis of brightness rather than hue" (p. 331), and no words for the Gestalt "good forms" of square, circle, and equilateral triangle. She found that her subjects were significantly faster in learning arbitrary names for the four primary hue points than for other hues, and for the three "good forms" of Gestalt psychology than for others. She points to the possible physiological underpinnings of these "natural prototypes." Her work is reminiscent of a study by House, Stevens, Sandel, and Arnold (1962). They constructed several ensembles of sounds along an acoustic continuum from clearly nonspeech to speech. The time taken by subjects to learn associations between sounds and buttons on a box was least for the speech ensemble, and did not decrease with the acoustic approximation of the ensembles to speech. In short, a signal is heard as either speech or nonspeech, and once heard as speech, elicits characteristic perceptual functions that we shall discuss below.

The second peculiarity of speech perception, as we are viewing it, is in perceptual response. The final percept is a phonetic name, and the name (unlike those for "natural categories" of form and color) bears a necessary, rather than an arbitrary, relation to the signal. In other words, speech sounds "name themselves." Notice that this is not true of the visual counterparts of phonetic entities: the forms of the alphabet are arbitrary, and we are not concerned that, for example, the same visual symbol, P, stands for /p/ in the Roman alphabet, for /r/ in the Cyrillic. Nothing comparable occurs in the speech system: the acoustic correlates of [p] or [r] can be perceived as nothing other than [p] or [r]. A central problem for the student of speech perception is to define the nature of this inevitable percept.

LEVELS OF PROCESSING

Implicit in the foregoing is a distinction between auditory and phonetic perception. As a basis for future discussion, we will lay out a rough conceptual model of the perceptual process (cf. Studdert-Kennedy, 1974; also Day, 1968, 1970). We can conceive the signals of running speech as climbing a hierarchy through at least these successive transformations: (1) auditory, (2) phonetic, (3) phonological, (4) lexical, syntactic, and semantic. The levels must be at least partially successive, to preserve aspects of temporal order in the signal. They must also be at least partially parallel, to permit higher decisions to guide and correct lower decisions [cf. Turvey's (1973) discussion of peripheral and central processes in vision].

The auditory level is itself a series of processes (Fourcin, 1972). Early work (Licklider and Miller, 1951) showed that the speech waveform could be vastly distorted without serious loss of intelligibility. Spectrographic analysis (Potter, Kopp, and Green, 1947; Joos, 1948) and speech synthesis (Liberman, 1957) showed that patterns of speech important to its perception lay not in its wave-form, but in its time-varying spectrum as revealed by the spectrogram. We may imagine, therefore, an early stage of the auditory display, soon after cochlear analysis, as the neural correlate of a spectrogram. Notice in Figure 1: regions of high energy concentration (formants, usually labeled from the bottom up as F1, F2, F3); different formant patterns associated with the vowels of *read* and *book*, for example; intervals of silence during stop consonant closure; a sharp scatter of energy (noise burst) upon release of the voiceless stop in *to*, and fainter bursts following release of the voiced stops in *began*; rapid formant movements (transitions) as articulators move into and out of vowels; a nasal formant (between F1 and F2) at the end of *began*; a broad band of noise associated with the fricative of *she*; and finally, regular vertical striations, reflecting a series of glottal pulses, from which fundamental frequency can be derived. A later, perhaps cortical, stage of auditory analysis may entail detection of just such features in the spectrographic display. Whether there are acoustic feature analyzers specially tuned to speech is an open question that we consider below. In any event, the signal has not yet been transformed into the message, and may indeed have passed through the same processes as any other auditory input.

The phonetic level is abstract in the sense that its output is a set of properties not inherent in the signal. They derive from the auditory display by processes that must be peculiar to humans, since they can only be defined by reference to the human vocal mechanism. These properties correspond to the linguistic entities of distinctive feature (Jakobson, Fant, and Halle, 1963) and phoneme. For the psychological reality of these units, there is ample evidence, discussed below. There is also evidence that extraction of these units from the auditory display calls upon specialized decoding mechanisms (Studdert-Kennedy and Shankweiler, 1970). In any event, the output from this level is now speech, although much variability remains to be resolved.

Resolution is accomplished at the phonological level, where processes peculiar to the listener's language are engaged. Here, the listener merges phonetic variations that have no function in his language, treating, for example, both the initial segment of [p^hIt] and the second segment of [spIt] as instances of /p/. Here, too, the listener may shift distinctions across segments, interpreting English vowel length before a final stop, for example, as a phonetic cue to the voicing value of the stop. In short, this is the level at which phonetic variability is transformed into phonological system. Of course, for untrained listeners all of the time, and for phoneticians most of the time, the distinction between phonetic and phonological levels has little import. Listeners usually hear speech in terms of the categories of their native language (e.g., Lotz, Abramson, Gerstman, Ingemann, and Nemser, 1960; Scholes, 1968; Day, 1968, 1969, 1970a, 1970b). However, since they may learn (at some pain) to make phonetic distinctions, we must assume that phonetic information is available in the system, though unattended in normal listening. Most of the research to be discussed has concerned itself with a single language and has not distinguished between phonetic and phonological levels. (For extended discussion of experimental paradigms that serve to reflect several levels of processing from auditory to phonological, see Cutting, 1973, in press-a.)

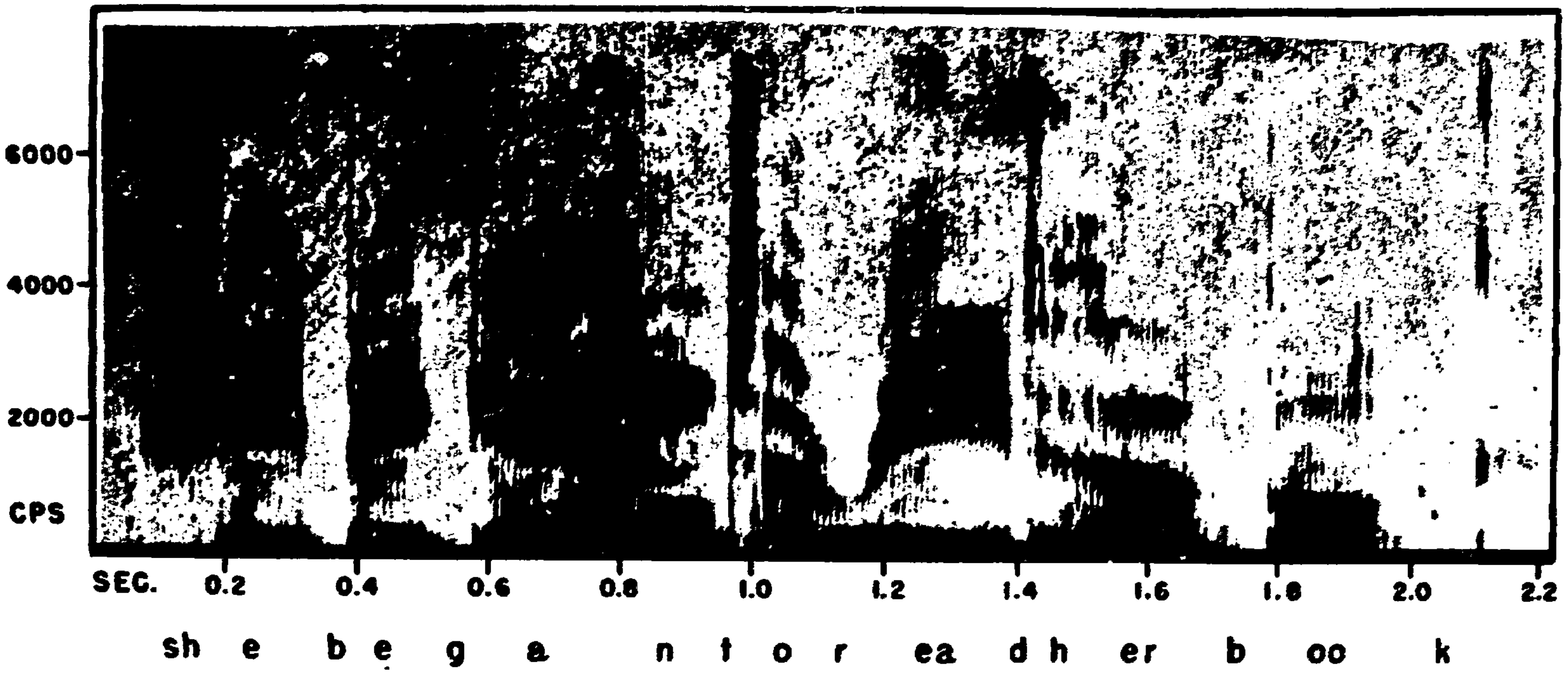


Figure 1: Spectrogram of a natural utterance: She began to read her book. Frequency is plotted against time, with relative intensity represented by degree of blackness. For fuller description, see text.

The upper levels of lexical, syntactic, and semantic processing complete the normal process of speech perception. There is good evidence that outputs from these levels can affect phonological and phonetic perception. Miller, Heise, and Lichten (1951), for example, showed that words were more intelligible in a sentence than in a list. Pollack and Pickett (1963) and Lieberman (1963) found that words excised from sentences and presented to listeners without syntactic and semantic context were often not recognized. Several writers (e.g., Jones, 1948; Chomsky and Miller, 1963; Chomsky and Halle, 1968) have placed a heavy load on the syntactic structure and semantic content of an utterance in their accounts of speech perception. However, while these higher levels may serve to "clean" the message when phonetic lapse is slight (cf. Warren, 1970; Warren and Obusek, 1971, Cole, 1973a), and may even be deliberately brought to bear while conversing with a foreigner in a railway tunnel, their control is not sufficient to disguise all slips of the tongue (cf. Fromkin, 1971). Unambiguous perception is possible in spite of context, and, as will be seen, presents sufficient theoretical problems. Bearing in mind our primary distinction between auditory and phonetic levels, we turn now to a brief review of acoustic cues and of the problems that emerge for perceptual theory.

THE ACOUSTIC CUES

Many of the acoustic cues to the phonetic message have been uncovered over the past twenty years by the complementary processes of analysis and synthesis. Spectrographic analysis of natural speech suggests likely candidates, such as formant frequency, formant movement, silent interval, or burst of noise. Synthesis then permits these "minimal cues" (Liberman, 1957) to be checked for perceptual validity. Results of this work are described elsewhere (Liberman, 1957; Fant, 1960, 1968; Mattingly, 1968, 1974; Flanagan, 1972; Stevens and House, 1972). Here, we do no more than summarize its outcome and frame the problems it raises for speech perception.

The problems are those of invariance and segmentation. The speech signal carries neither invariant acoustic cues nor isolable segments that reliably correspond to the invariant segments of linguistic analysis and perception. The speech signal can certainly be segmented. Fant (1968) and his colleagues have outlined a procedure for dividing the signal in both frequency and time, and have developed a terminology to describe its segments. But these do not correspond to the phonetic segments of distinctive feature or phoneme. There are exceptions: fricatives and stressed vowels, for example, may present stable and more-or-less isolable patterns. But, in general, as Fant (1962) has remarked, a single segment of sound contains information concerning several neighboring segments of the message, and a single segment of the message may draw upon several neighboring segments of sound. In short, the sounds of speech are not physically discrete, like letters of the alphabet, but rather are shingled into an intricate, continuously changing pattern (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967).

Whether the source of this shingled pattern is to be found in mechanical constraints, neuromuscular inertia, and temporal overlap of successive commands to the articulators (Öhman, 1967), or in elegantly controlled, yet variable responses to fixed articulatory instructions (MacNeilage, 1970), the result is not only a loss of segmentation, but also a loss of acoustic invariance. The cues to a given phonetic segment display enormous variability as a function of phonetic context, stress, and speaking rate (e.g., Kozhevnikov and Chistovich, 1965; Stevens, House, and Paul, 1966).

As a simple instance, consider the acoustic structure of a mirror image consonant-vowel-consonant (CVC) syllable such as [bæb]. Experiments with synthetic speech have demonstrated the importance of second and third formant transitions as cues for distinguishing among labial, alveolar, and velar stops (Liberman, Delattre, Cooper, and Gerstman, 1954). Here, the two formants rise rapidly over the first 40 msec or so into the vowel, and then, after a relatively sustained formant pattern for, say, 200 msec, drop rapidly back to their starting points. The acoustic cues to initial and final allophones of [b] are mirror images, and, separated from the syllable, are heard as distinct nonspeech sounds. Experiments with tone glissandos matching such patterns in duration and frequency range reveal no psychoacoustic basis for the perceived phonetic identity (Klatt and Shattuck, 1973).

Similar discrepancies occur as a function of vowel context. Initial formant transitions in a CV syllable reflect the changing resonances of the vocal tract as the articulators move from consonant closure, or constriction, into a more open position for the following vowel. Since vowels are distinguished by the positions of their first two or three formant centers on the frequency scale (Delattre, Liberman, Cooper, and Gerstman, 1952; Peterson and Barney, 1952), consonantal approach varies with vowel: for example, both second and third formants fall in the syllable [dæ]; the second rises and the third falls in the syllable [de]. Yet listeners fail to detect these acoustic differences, and phonetic identity of the initial segments is preserved.

As a final example, consider vowels. Each stressed vowel, spoken in isolation, has its characteristic set of formant frequencies. However, in running speech, these values are seldom reached, particularly if speech is rapid and vowels unstressed (Lindblom, 1963). If vowel portions are excised from running speech and presented without their surrounding formant transitions, identifications shift (Fujimura and Ochiai, 1963). This suggests (as do the consonantal examples given above) that listeners track formants over at least a syllable in order to make their phonetic decisions. (For other examples of phonetic identity in face of acoustic variance, see Shearme and Holmes (1962), Lindblom (1963), Öhman (1966), Liberman et al. (1967), and Stevens and House (1972).)

A different class of acoustic variability is instanced by interspeaker variations. Here differences in acoustic quality can be clearly heard, but are disregarded in phonetic perception. Center frequencies of vowel formants vary widely among men, women, and children (Peterson and Barney, 1952), with the result that acoustically identical patterns may be judged phonetically distinct, while acoustically distinct patterns may be judged phonetically identical. "Normalization" probably cannot be accomplished by application of a simple scale factor (Peterson, 1961) because male-female formant ratios are not constant across the vowel quadrilateral (Fant, 1966).

A favored belief is that listeners judge vowels by reference to other vowels uttered by the same speaker. This notion originated with Joos (1948) and was tested by Ladefoged and Broadbent (1957). They demonstrated that the same synthetic vowel pattern could be judged differently, depending on the formant pattern of a precursor phrase. Gerstman (1968) developed an algorithm, derived from the formant frequencies of [i,a,u] for each speaker, that correctly identifies 97.5 percent of the Peterson and Barney (1952) vowels. And Lieberman (1973) claims that unless a listener has heard "calibrating signals," such as the vowels [i,a,u] or the glides [y] and [w], from which to assess the size of a particular speaker's vocal tract, "it is impossible to assign a particular acoustic signal into the correct class" (p. 91).

However, an algorithm is not a perceptual model, and remarkably little is actually known in this area: there is a dearth of data on how listeners judge the varied vowel patterns of different speakers. Furthermore, the phenomenon of normalization is not confined to vowels. Fourcin (1968) demonstrated that a synthetic "whispered" syllable with a constant formant pattern could be heard as a token of [d] if preceded by a man's hallo, of [b] if preceded by a child's. Rand (1971) showed a similar systematic shift, without benefit of precursor, when formant frequencies of synthetic CV syllables were increased by 20 percent above the "male" base. Evidently, normalization can be accomplished within a syllable, presumably from information provided by formant structure and fundamental frequency (cf. Fujisaki and Nakamura, 1969). This is precisely what is suggested by recent work of Strange, Verbrugge, and Shankweiler (1974) and Verbrugge, Strange, and Shankweiler (1974). They find that a speaker's precursor vowels, whether [i,a,u] or [I,æ,ʌ], do little to reduce listener error in judging following vowels spoken by a panel of men, women, and children. Far more effective in reducing error is presentation of the vowel within a consonantal frame. Of course, formant reference is clearly involved in studies where consonantal context is held constant (Summerfield and Haggard, 1973). However, the results again suggest perceptual tracking of an entire syllable, and emphasize that invariant acoustic segments matching the invariants of perception are not readily found. [For a recent review of the normalization problem, see Shankweiler, Strange, and Verbrugge (in press).]

Nonetheless, the search for acoustic invariance has not been abandoned. A main reason for this is the obvious worth of some form of feature theory in linguistic description and, incidentally, in the description of listener behavior (see next section). Distinctive-feature theorists have always maintained that correlates of the features are to be found at every level of the speech process--articulatory, acoustic, auditory--(Jakobson and Halle, 1956; Jakobson, Fant, and Halle, 1963; Chomsky and Halle, 1968), and a good deal of current research is directed toward grounding features in acoustics and physiology (cf. Ladefoged, 1971a, 1971b; Lindblom, 1972).

Before giving examples, we should emphasize the redundancy of the speech signal. A given feature may be signaled by several different cues. Studies of synthetic speech have tended to emphasize "sufficient" cues and to disregard their interaction. Harris (1958) provides an exception, in her study of noise bands and formant transitions as cues to English fricatives. So, too, do Harris, Hoffman, Liberman, Delattre, and Cooper (1958) and Hoffman (1958), who examined the relative weights of second and third formant transitions in the perception of English voiced stops.

Finally, exceptions are also provided by Lisker and Abramson (1964, 1967, 1970, 1971) and by Abramson and Lisker (1965, 1970; see also Zlatin, 1974) in an extensive series of studies of voicing in many languages. Noting that voicing in initial stops may be cued by explosion energy, degree of aspiration, and first formant intensity, they sought a cover variable that would encompass all these cues. They found it in voice onset time (VOT), the interval between release of stop closure and the onset of laryngeal vibration. Figure 2 displays spectrograms of synthetic stops in which VOT is a sufficient cue for the distinction between [ba] and [pa]. Notice that VOT is not a simple variable, either articulatorily or acoustically: it refers to a temporal relation between

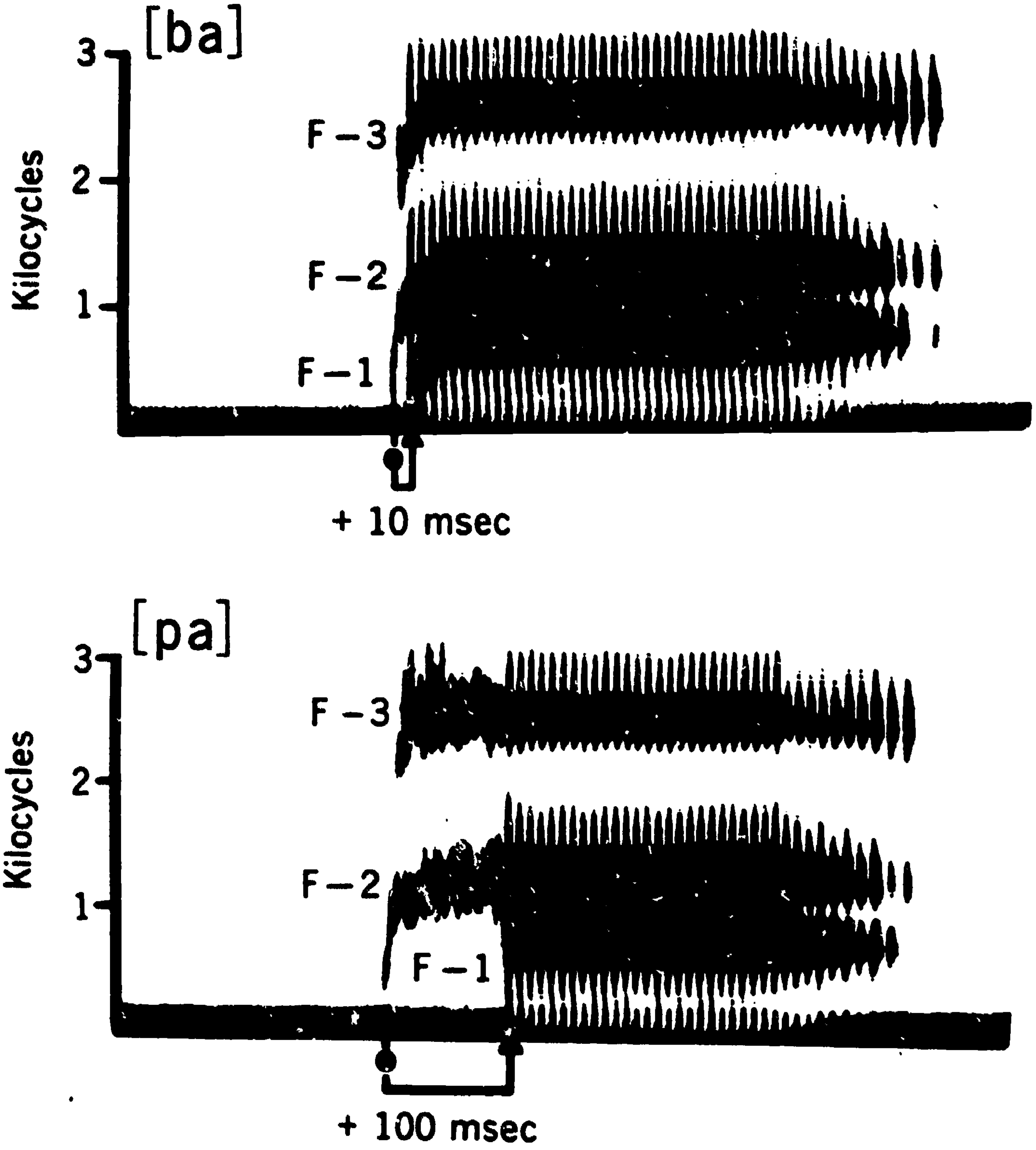


Figure 2: Spectrograms of synthetic syllables, [ba] and [pa]. The interval between release and voicing (vertical striations) (VOT) is 10 msec for [ba], 100 msec for [pa]. During this interval, F1 is absent and the regions of F2 and F3 are occupied by "aspirated" noise. [After Lisker and Abramson, with permission.]

two distinct events. In production, it calls for precise timing of a laryngeal gesture (approximation of the vocal cords) in relation to supralaryngeal release; in perception, it calls for judgment of a complex acoustic pattern arrayed in time. Nonetheless, within these limits, VOT offers a relatively invariant physical display and a relatively invariant sequence of coordinated articulatory gestures that might serve to define a feature, albeit not a feature within the generally accepted system (Chomsky and Halle, 1968). [For a full account of the underlying rationale, the reader is referred to the publications cited above and, for discussions of the approach, to Stevens and Klatt (1974) and Summerfield and Haggard (1972); see also Haggard, Ambler, and Callow (1970).]

A second example of the search for feature invariants is provided by the work of Stevens (1967, 1968a, 1968b, 1972a, 1972b, 1973). In a recent paper, for example (Stevens, 1973), he approaches an acoustic definition of [+ Consonantal], describing consonants as displaying "a rapid change in the acoustic spectrum" (p. 157) in the region of F2, following release (cf. Fant, 1962). He develops this description, emphasizing the entire spectrum rather than individual formants, into an acoustic account of place features [+ Coronal], [+ Labial], and [+ Velar], for which he posits "property detectors." The acoustic description is based on spectrographic analysis and computations from an idealized vocal tract model. The model reveals certain "quantal places of articulation which are optimal from the point of view of sound generation" (Stevens, 1968a: 200) since they permit relatively imprecise articulation without serious perturbation of the signal. Obviously, these tract shapes can be correlated with articulatory gestures to provide the needed feature correlates.

Finally, less ambitious attempts to discover feature invariants are instanced by tape-cutting experiments with natural speech, in which consonantal portions of a syllable are removed and presented for identification alone or with vowels other than the original (Fischer-Jørgensen, 1972; Cole and Scott, 1974). If this approach leads to precise definition of acoustic invariants, it will have proved valuable. However, if experiments merely demonstrate that transposing initial portions of two CV syllables, for example, yields no change in perception of initial consonant, we have not advanced. The transposed patterns remain different both acoustically and, if removed from the speech stream, psychoacoustically, and the demonstrated source of invariance is still the listener. The ultimate test of all these attempts will be in control of a speech synthesizer from a set of invariant articulatory or acoustic feature specifications (Mattingly, 1971).

THE PHONETIC PERCEPT

Up to this point we have simply assumed the units of speech perception. However, research has sporadically puzzled over their definition for the past 25 years. The puzzle arises, as we have seen, from the mismatch between the acoustic signal and the abstract entities of linguistic analysis, distinctive features, and phonemes. Nonetheless, each of these units has been shown to have psychological reality. Perhaps the most direct evidence comes from studies of speaking errors. Fromkin (1971) has analyzed many utterances for errors of metathesis (spoonerism). She finds that speakers may metathesize not only words and phrases, but syllables (clarinet and viola → clarinola), phonemes (far more → mar fore), and features (clear blue → glear plue) (cf. Boomer and Laver, 1968; MacKay, 1970; Cairns, Cairns, and Williams, 1974). Of particular interest is her observation that speakers may exchange consonant for consonant and vowel for vowel, but never consonant for vowel. This reflects a distinction in production between phonetic elements of the syllable that are, as we shall see, repeatedly distinguished in perception. In any event, errors of metathesis logically require that the speaker have independent control over the unit of error. And if these units are independently produced, it is reasonable to believe that they are independently perceived.

Evidence from perceptual studies is not lacking. Errors of subjects listening to speech through noise (Miller and Nicely, 1955; Mitchell, 1973) or dichotically (Studdert-Kennedy and Shankweiler, 1970; Studdert-Kennedy, Shankweiler, and Pisoni, 1972; Blumstein, 1974), are patterned according to some form of feature system. Scaling studies, in which the experimenter attempts to determine the psychological space occupied by a set of consonants or vowels, repeatedly reveal a structure parsimoniously described by feature theory (Greenberg and Jenkins, 1964; Singh, 1966; Hanson, 1967; Singh and Woods, 1970; Shepard, 1972). A new paradigm has recently provided further evidence. Goldstein and Lackner (in press), adapting a technique devised by Warren and Gregory (1958; also Warren, 1968, and in Lass, in press), played a 200 msec nonsense syllable over and over (200 times per minute), asking listeners to report what they heard. After a few repetitions, listeners began to hear different words (verbal transformation). The new words were systematically related to the originals: they entailed changes in value of only one or two distinctive features, and reflected phonological constraints of English as described by distinctive feature theory. Finally, errors in short-term memory studies also follow a feature pattern (Sales, Cole, and Haber, 1969; Wickelgren, 1965, 1966). Several of these studies have used their perceptual data to compare the predictive power (and so the validity) of different feature systems. Such work is particularly important if linguistics is to be regarded as a branch of human psychology (Chomsky, 1972), and if the abstract units of phonology are to be grounded in human articulatory and perceptual capacities (Ladefoged, 1971a, 1971b; Liljencrants and Lindblom, 1972; Lindblom, 1972).

The perceptual status of the columns in a feature matrix has proved more controversial. Functionally, the column (phone) represents the grouping of distinctive features within a syllable, specifying the domain within which a particular feature is to apply. We recognize this perceptually in alliteration (big boy) and in rhyme (bee and see), where two syllables are perceived as identical at their beginning, but not at their end, or vice versa. Listeners reveal this function when asked to judge similarities among words. Vitz and Winkler (1973) found, in fact, that the number of phones shared by a pair of words was a more satisfactory predictor of their judged similarity than the number of shared features. In the verbal transformation study described above (Goldstein and Lackner, in press), transformations were best described in terms of phones and features rather than syllables and features: consonant transforms and vowel transforms, for example, were independent, reflecting feature shifts within, but not across, phones. Finally, several studies (Kozhevnikov and Chistovich, 1965; Savin and Bever, 1970; Day and Wood, 1972) have shown reaction time differences in identification of consonants and vowels within the same syllables. These differences would not occur if the syllable were an unanalyzed perceptual entity.

Despite such evidence and despite the clear role of phoneme-size phonetic segments in speaking and in writing systems, students have been tempted to regard these segments as "nonperceptual" (Savin and Bever, 1970) or as "fictitious units" based on the historical accident of alphabet invention [Warren (in Lass, in press)]. Among the arguments for this conclusion seem to be three solid facts, two (or more) pieces of ambiguous evidence, and one false belief. The facts are: first, that no phoneme-size segment can be isolated in the acoustic signal; second, that some phonemes (stop consonants) cannot be spoken in isolation; third, that we do speak in syllables and that syllables are the carriers of stress and speech rhythm. The ambiguous evidence comes from reaction time studies suggesting that syllables, and even higher order units, may be identified before the elements of which they are composed. Savin and Bever (1970) and Warren (1971) showed that the reaction time of listeners monitor'ng a monosyllabic list for syllables is faster than their reaction time when monitoring the same list for the initial phoneme of the syllable. Subsequently, Foss and Swinney (1973) showed that, under similar conditions, listeners responded more rapidly to words than to their component syllables, while Bever (1970) revealed that listeners responded more rapidly to three-word sentences than to their component words. It was left to McNeill and Lindig (1973) to release us from this "Looking Glass" world, in which the trial precedes the crime, by demonstrating that reaction time was always fastest to the largest elements of which a list was composed. In other words, listeners' response is most rapid at the level of linguistic analysis to which context has directed their attention.

Finally, the false belief is that invariance and segmentation problems would disappear if the syllable were an unanalyzed unit of perception. This belief is no better founded than Wickelgren's (1969) attempt to solve the invariance problem by positing context-sensitive allophones. and is open to many of the same objections. These objections have been well summarized by Halwes and Jenkins (1971), and we will not review them here. However, it is worth adding that the syllable has resisted acoustic definition only somewhat less than the phoneme-size phonetic segment. Its nucleus may be detected by amplitude and fundamental frequency peak picking (Lea, 1974), and Malmberg (1955) drew attention to the possible role of formant transitions in defining syllable boundaries, but no fully satisfactory definition has yet emerged. Furthermore, coarticulation and perceptual context effects across syllables, though less marked than across phones, still occur. Öhman (1966), for example, found drastic variations in vowel formant transitions on either side of stop closure, as a function of the vowel on the opposite side of the closure. And Treon (1970) has demonstrated contextual effects in perception extending across two to three syllables. In fact, as Fodor, Bever, and Garrett (1974) hint, an account of syllable perception may well require the same theoretical apparatus as an account of phone perception.

Much of the confusion over units of speech perception might be resolved if the distinctions between signal and message, and among acoustic, phonetic, and higher levels were strictly maintained. There is wide agreement among writers, whose views may otherwise diverge, that the basic acoustic unit of speech perception (and production) is of roughly syllabic length [e.g., Liberman, Delattre, and Cooper, 1952; Liberman, 1957; Kozhevnikov and Chistovich, 1965; Öhman, 1966; Ladefoged, 1967; Liberman et al., 1967; Savin and Bever, 1970; Massaro, 1972; Stevens and House, 1972; Cole and Scott, 1973; Kirman, 1973; McNeill and Repp, 1973; Warren (in Lass, in press); Studdert-Kennedy, in press]. This is not to deny that there are longer stretches of the signal over which the perceptual apparatus must compute relations, but simply to say that the smallest stretch of signal on which it goes to work is produced by the articulatory syllabic gesture (Stetson, 1952). This does not mean [as Massaro (1972), for example, seems to suppose] that the syllable is the basic linguistic and perceptual unit.

We may clarify by conceptualizing the process of constructing an utterance from a lexicon of morphemes. The abstract entity of the morpheme is the fundamental unit in which semantics, syntax, and phonology converge. Each morpheme is constructed from phonemes and distinctive features. At this level, the syllable does not exist. But morphemic structure is matched to (and must ultimately derive from) the articulatory capacities of the speaker. Both universal and language-specific phonotactic constraints ensure that a morpheme will eventuate in pronounceable sequences of consonants and vowels. Under the control of a syntactic system governing their order and prosody, the morphemes pass through the phonetic transform into a sequence of coarticulated gestures. These gestures give rise to a sequence of acoustic syllables, into which the acoustic correlates of phoneme and distinctive feature are woven. The listener's task is to recover the features and their phonemic alignment, and so the morpheme and meaning. In short, perception entails the analysis of the acoustic syllable, by means of its acoustic features, into the abstract perceptual structure of features and phonemes that characterize the morpheme. We now turn to some theoretical accounts of how this might proceed.

MODELS OF PHONETIC PERCEPTION

We have no models specified in enough detail for serious test. But a brief account of two approaches that have influenced recent research may serve to summarize the discussion up to this point. The two approaches are those of the Haskins Laboratories investigators and of Stevens and his colleagues at the Massachusetts Institute of Technology. Both groups are impressed, in varying degrees, by the invariance and segmentation problem. Both have therefore rejected a passive template- or pattern-matching model in favor of an active or generative model. (For a review, see Cooper, 1972.)

Liberman et al. (1967), reformulating a theme that had appeared in many earlier papers from the Haskins group, proposed a "motor theory of speech perception." The crux of their argument was that an articulatory description of speech is not merely simpler, but is the only description that can rationalize the temporally scattered and contextually variable patterns of speech. They argue that phonetic segments undergo, in their passage through the articulatory system, a process of "encoding." They are restructured acoustically in the syllabic merger, so that cues to phonetic identity lose their alignment and are distributed over the entire syllable (Liberman, 1970). Not all phonetic segments undergo the same degree of restructuring: there is a hierarchy of encodedness, from the highly encoded stop consonants, through nasals, fricatives, glides, and semivowels, to the relatively unencoded vowels. Nonetheless, recovery of phonetic segments from the syllable calls for parallel processing of both consonant and vowel; neither can be decoded without the other. And this demands a specialized decoding mechanism, in which reference is somehow made to the articulatory gestures that gave rise to the encoded syllables.

Liberman et al. (1967) assume, reasonably enough, that "at some level...of the production system there exist neural signals standing in one-to-one correspondence with the various segments of the language," and th . for the phoneme "the invariant is found far down in the neuromotor system, at the level of the commands to the muscles" (p. 454). It is important to note that actual motor engagement is not envisaged. Liberman (1957) has written: "We must assume that the process is somehow short-circuited--that is, that the reference to articulatory movements and their sensory consequences must somehow occur in the brain without getting out into the periphery" (p. 122).

A virtue of the model is that it accounts for a fair amount of data and has generated a steady stream of research. Also, the concept of encoding, though descriptive rather than explanatory, draws attention to a process at the base of language analogous to syntactic processes suggested by generative grammar, and hints at formal similarities in the physiological processes underlying phonetic and syntactic performance (Mattingly and Liberman 1969; Liberman, 1970; Mattingly, 1973, 1974). Conspicuously absent is any account of first-language acquisition. The child may be presumed to be born with some "knowledge" of vocal tract physiology and an incipient capacity to interpret the output of an adult tract in relation to that of its own (Mattingly, 1973), but a detailed account of the process is lacking.

Stevens (1973) has concerned himself with this problem, and addresses it in the most recent version of his analysis-by-synthesis model (Stevens, 1972a; cf. Stevens, 1960; Stevens and Halle, 1967). The model is far more explicit than that of the Haskins group. The perceptual process is conceived as beginning with some form of peripheral spectral analysis, acoustic feature and pitch extraction. Pitch and spectral information, over a stretch of several syllables, is placed in auditory store. Acoustic feature information undergoes preliminary analysis by which a rough matrix of phonetic segments and features is extracted and passed to a control system. On occasion, this matrix may provide sufficient information for the control (which knows the possible sequences of phonetic segments and has access to the phonetic structure of earlier sections of the utterance) simply to pass the description on to higher levels. If this is not possible, the control guesses at a phonetic description on the basis of its inadequate information and sends the description to a generative rule system, the same that in speaking directs the articulatory mechanism. The rule system generates a version of the utterance and passes it to a comparator for comparison with the spectral description in temporary auditory store. The comparator computes a difference measure and feeds it back to the control. If the "error" is small enough, the control system accepts its original phonetic description as correct. If not, it makes a second guess and the cycle repeats until an adequate match is reached.

This rough account does no justice to the model's elegance and subtlety, but it may serve to focus attention on several points. First, the solution to the invariance problem is a more abstract and more carefully specified version of a motor theory. Second, the model emphasizes the necessity of at least a preliminary feature analysis, to ensure that the system is not doomed to an infinity of bad guesses, and that the child, given a set of innate "property detectors," can latch onto the utterance. At the same time, no account is offered of how the invariant acoustic properties are transformed into phonetic segments and features (the process is simply consigned to "a preliminary analysis"), nor of the precise form that the phonetic description takes. Finally, the model emphasizes the need for a short-term auditory store. As we shall see, the form and duration of such a store is currently the focus of a great deal of research.

THE PROCESSING OF CONSONANTS AND VOWELS

Preliminary

To brace ourselves for a fairly prolonged discussion of consonants and vowels, let us consider why they are interesting. For theory, the answer is that they lie at the base of all phonological systems. All languages are syllabic, and all languages constrain syllabic structure in terms of consonants and vowels. If we are to ground phonological theory in human physiology, we must understand why this path was taken. Lieberman (1970) has argued that phonological features may have been selected through a combination of articulatory constraints and "best matches" to perceptual capacity. One purpose of current research is to understand the nature and basis of the best match between syllables, constructed from consonants and vowels, and perceptual capacity.

For experiment, the interest of consonants and vowels is that they are different. If all speech sounds were perceived in the same way, we would have no means of studying their underlying relations. Just as the biologist could not study the genetics of eye-color in Drosophila melanogaster until he had found two flies with different eyes, so the student of speech had no means of analyzing syllable perception until he had found portions of the syllable that reflected different perceptual processes (cf. Stetson, 1952). Fortunately, the interests of theory and research converge.

Categorical Perception

Study of sound spectrograms reveals that portions of the acoustic patterns for related phonetic segments (segments distinguished from one another by a single feature) often lie along an apparent acoustic continuum. For example, center frequencies of the first two or three formants of the front vowels /i,I, ε,æ/ form a monotonic series; syllable-initial voice-voiceless pairs /b,p/, /d,t/, /k,g/ differ systematically in voice onset time; voiced stops /b,d,g/ before a particular vowel, differ primarily in the extent and direction of their formant transitions.

To establish the perceptual function of such variations speech synthesis is used. Figure 3 sketches a schematic spectrogram of a synthetic series in which changes of slope in F2 transition effect perceptual changes from /b/ through /d/ to /g/. Asked to identify the dozen or so sounds along such a continuum, listeners divide it into distinct categories. For example, a listener might consistently identify stimuli -6 through -3 of Figure 3 as /b/, stimuli -1 through +3 as /d/, and stimuli +5 through +9 as /g/. In other words, he does not, as might be expected on psychophysical grounds, hear a series of stimuli gradually changing from one phonetic class to another, but rather a series of stimuli, each of which (with the exception of one or two boundary stimuli) belongs unambiguously in a single class. The important point to note is that, although steps along the continuum are well above nonspeech auditory discrimination threshold, listeners disregard acoustic differences within a phonetic category, but clearly hear equal acoustic differences between categories.

To determine whether listeners can, in fact, hear the acoustic differences belied by their identifications, discrimination tests are carried out, usually in ABX format. Here, on a given trial, the listener hears three stimuli, separated by a second or so of silence: the first (A) is drawn from a point on the continuum two or three steps removed from the second (B), and the third (X) is a repetition of either A or B. The listener's task is to say whether the third stimulus is the same as the first or the second. The typical outcome for a stop consonant continuum, is that listeners hear few more auditory differences than phonetic categories: they discriminate very well between stimuli drawn from different phonetic categories, and very poorly (somewhat better than chance) between stimuli drawn from the same category. The resulting function displays peaks at

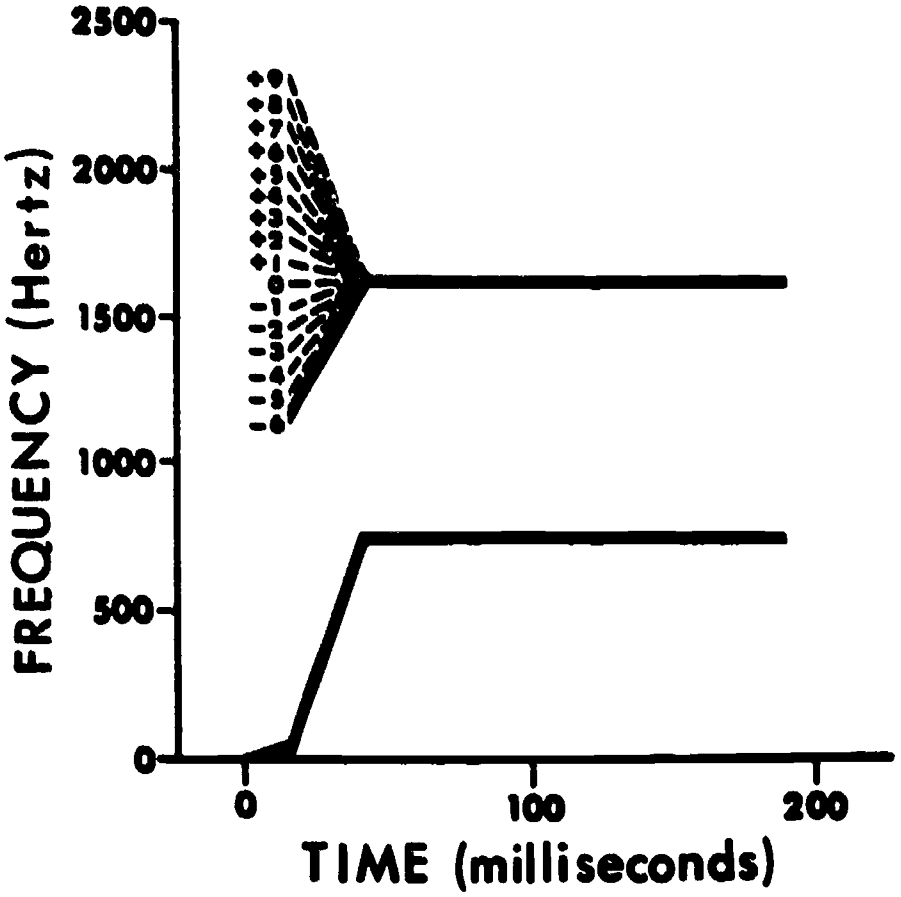


Figure 3: Schematic spectrogram for a series of synthetic stop-vowel syllables varying only in F2 transition. F2 steady-state, F1 transition, and steady-state remain constant. As F2 transition changes from -6 to +9, perception of initial consonant shifts from [b] through [d] to [g].

phonetic boundaries, troughs within phonetic categories. In fact, discriminative performance can be predicted with fair accuracy from identifications: the probability that acoustically different syllables are correctly discriminated is a positive function of the probability that they are differently identified (Liberman, Harris, Kinney, and Lane, 1961). This close relation between identification and discrimination has been termed "categorical perception": that is to say, perception by assignment to category. Figure 4 (left side) illustrates the phenomenon. Note that, although prediction from identification to discrimination is good, it is not perfect: listeners can sometimes discriminate between different acoustic tokens of the same phonetic type. Note, further, that neither identification nor discrimination functions display quantal leaps across category boundaries. This is not a result of data averaging, since the effect is given by individual subjects. Evidently auditory information about consonants is slight, but not entirely lacking.

We may now contrast categorical perception of stop consonants with "continuous perception" of vowels. Figure 4 (right side) illustrates the effect. There are two points to note. First, the vowel identification function is not as clear-cut as the consonant. Vowels, particularly those close to a phonetic boundary, are subject to context effects: for example, a token close to the /i-I/ boundary will tend to be heard, by contrast, as /i/, if preceded by a clear /I/, as /I/, if preceded by a clear /i/. The second point to note is that vowel discrimination is high across the entire continuum. Phonetic class is not

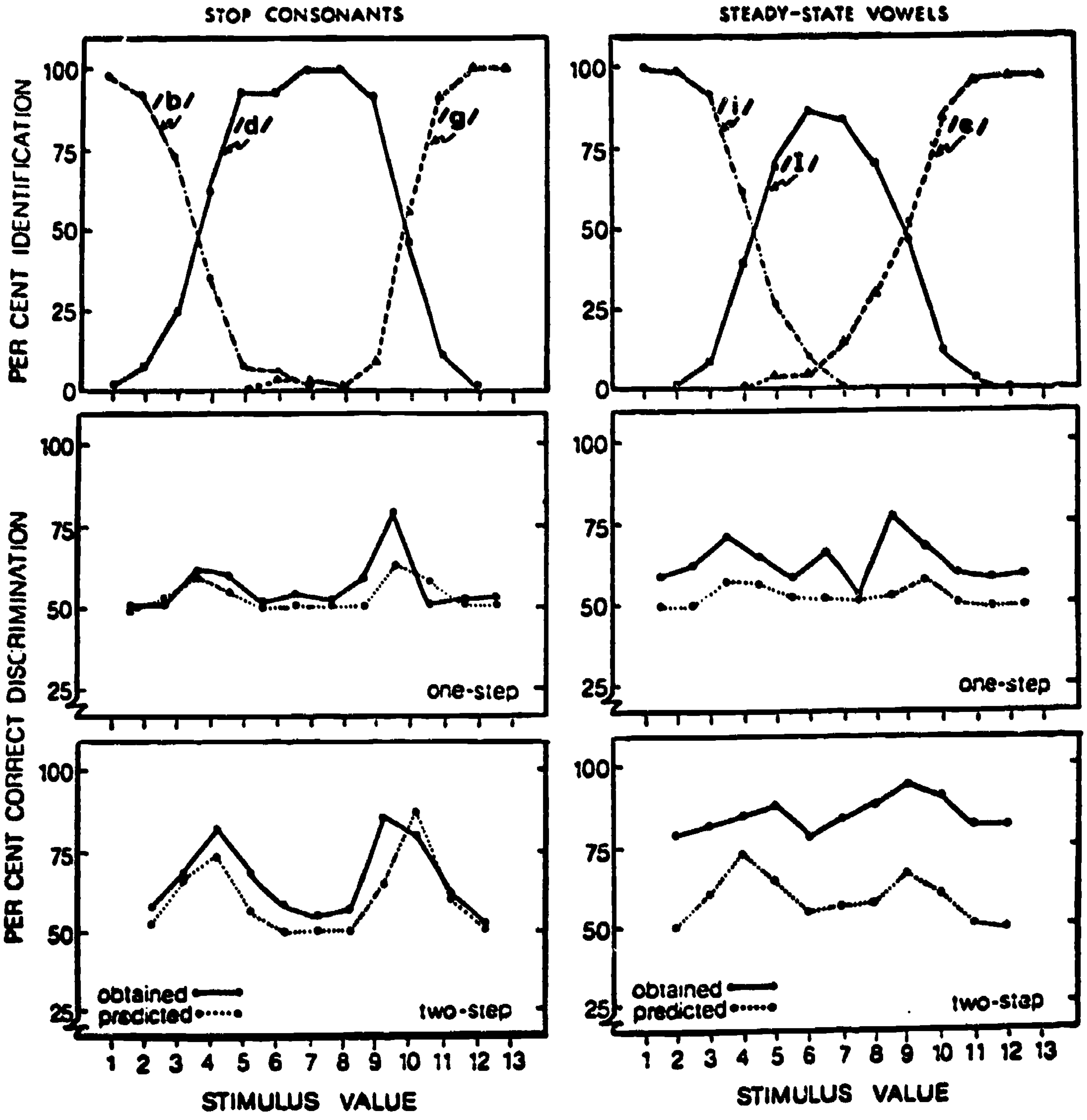


Figure 4: Average identification functions for synthetic series of stop consonants and vowels (top). Average one-step (middle) and two-step (bottom) predicted and obtained ABX discrimination functions for the same series. [After Pisoni (1971), with permission of the author.]

totally irrelevant (there is a peak in the discrimination function at the category boundary), but both within and between categories listeners discriminate many more differences than they identify. Their perception is said to be "continuous." [For fuller discussion, see Studdert-Kennedy, Liberman, Harris, and Cooper (1970a) and Pisoni (1971).]

Continuous perception is typical not only of vowels, but also of many nonspeech psychophysical continua along which we can discriminate more steps than we can identify (Miller, 1956). This fact has been taken as evidence both that categorical perception is peculiar to speech, and that stop consonants and vowels engage fundamentally different perceptual processes (Liberman et al., 1967; Studdert-Kennedy et al., 1970a). In fact, an early account of the phenomenon invoked a motor theory of speech perception (Liberman et al., 1967). As we have seen, there are independent grounds for hypothesizing that speech is perceived by reference to its articulatory origin. Here seemed to be additional evidence: the discrete articulatory gestures of stop consonants yielded discrete perceptual categories; the more variable gestures of vowels, more variable categories. But this account has several weaknesses, and recent work has largely eroded it. For one thing, we now know that categorical perception is not confined to speech (Locke and Kellar, 1973; Miller, Pastore, Wier, Kelly, and Dooling, 1974; Cutting and Rosner, in press).

However, this discovery in no way diminishes the importance of the phenomenon, as will become clear in the following sections. Here, we merely note two facts. First, the acoustic patterns distributed along a speech continuum are not arbitrary. They have been selected from the range of patterns that the articulatory apparatus can produce and that the auditory system can analyze. The categories are therefore natural, in the sense that they reflect physiological constraints on both production and perception. As Stevens (1972b) has pointed out, our task is to define the joint auditory and articulatory origin of phonetic categories.

Second, categorical perception reflects a functionally important property of certain speech sounds. The initial sound of /da/, for example, is difficult, if not impossible, to hear: the sound escapes us and we perceive the event, almost instantly, as phonetic. Rapid sensory decay and transfer into a nonsensory code is probably crucial to an efficient linguistic signaling system. Study of categorical perception has, in fact, revealed functional differences between stop consonants and vowels that are central to the syllabic structure of speech. At the same time it has provided basic evidence for the distinction between auditory and phonetic levels of processing.

In the following sections we consider two main aspects of categorical perception: first, the division of a physical continuum into sharply defined categories, and the assignment of names to the categories; second, listeners' apparent inability to discriminate among members of a category.

The Bases of Phonetic Categories

Phonetic categories do not arise from simple discriminative training, as proposed by Lane (1965). Subjects may certainly learn to divide a sensory continuum into clear-cut categories, with a resultant small peak in the discrimination function at the category boundary. But discrimination within categories remains high (Parks, Wall, and Bastian, 1969; Studdert-Kennedy et al., 1970a;

Pisoni, 1971): training may increase, but not obliterate discriminative capacity. Furthermore, the learned boundary is likely to be unstable. The process is familiar to tha psychophysicist. For example, if we present a subject with a series of weights and ask him to judge each weight as either heavy or light, he will, with a *minimum* of practice, divide the range cleanly around its balance point (see Woodworth and Schlosberg, 1954:Ch. 8). However, the boundary between *heavy* and *light* can be readily shifted by a change in experimental procedure. If an extreme token is presented for judgment with a probability several times that of other stimuli along the continuum, it comes to serve as an *anchor* with which other stimuli contrast: the result is a shift in category boundary toward the anchoring stimulus. Pisoni and Sawusch (in press) have shown that such shifts occur for a series of tones, differing in intensity, and for vowels, but not for stop consonants distributed along a voice-onset time continuum. They suggest that response criteria for voicing categories are mediated by internal rather than external references. By thus reframing the observation that stop consonant categories are not subject to context effects, they invite us to consider the nature of the internal reference.

Such a reference must be some distinctive perceptual quality shared by all members, and by no nonmembers, of a category. There is, of course, no reason to suppose that distinctive perceptual qualities are confined to speech continua. They will emerge from any physical continuum for which sensitivity is low within restricted regions and, by corollary, high between these regions. However, while the distinctive perceptual quality of a nonspeech event (such as a click, a musical note, or a flash of light) has the character of its sensory mode, the distinctive perceptual quality of a speech sound is phonetic. It is into a phonetic code that speech sounds are rapidly and automatically transferred for storage and recall.

With this in mind, we turn to several studies of nonspeech continua. We begin with Cutting and Rosner (in press), who determined an auditory boundary between rapid and slow stimulus onsets. Variations in stimulus onset, or rise time, are known to contribute to the affricate/fricative distinction, /tʃa/ versus /ʃa/ (Gerstman, 1957). The authors varied rise time from 0 to 80 msec for sawtooth wave trains, generated by a Mocg synthesizer, and for synthetic affricate/fricatives. The rapid-onset sawtooth waves sounded like a plucked guitar string, the slow-onset waves like a bowed string. Cutting and Rosner presented their two classes of stimuli for identification (*pluck* - *bow*, /tʃa/ - /ʃa/) and for ABX discrimination. Both speech and nonspeech yielded category boundaries at a 40-50 msec rise time, with appropriate peaks and troughs in the discrimination functions.

A second instance of nonspeech categorical perception is provided by Miller et al. (1974). These investigators constructed a rough nonspeech analog of the voice-onset time continuum. They varied the relative onset times of bursts of noise and periodic buzz, over a range of noise-leads from -10 to +80 msec, and presented them to subjects for labeling and discrimiuation. Listeners divided the continuum around an average noise-lead of approximately 16 msec, displaying clear discrimination troughs within *no noise-lead* and *noise-lead* categories, and a discrimination peak at the category boundary. The boundary value agrees remarkably well with that reported by Abramson and Lisker (1970) for the English labial VOT continuum, though not with the systematically longer perceptual boundaries associated with English apical and velar VOT continua (Lisker and Abramson, 1970). The authors conclude that the categories of their experiment (and, presumably, of at least the English labial VOT continuum) lie on either side of a "difference limen" for duration of the leading noise. While possibly correct, their conclusion places a misleading emphasis on the boundary between categories rather than on the categories themselves.

The emphasis is reversed in a recent study of Stevens and Klatt (1974). Following Liberman, Delattre, and Cooper (1958), they examined auditory discrimination of two acoustic variables along the stop consonant voice-voiceless continuum: delay in formant onset and presence/absence of F1 transition. For their first experiment they constructed a nonspeech analog of plosive release and following vowel: a 5 msec burst of noise separated from a vowel-like buzz by between 0 and 40 msec of silence. Listeners' "threshold" for detection of silence between noise and buzz was approximately 20 msec, a close match with the value for detection of noise lead found by Miller et al. (1974). Stevens and Klatt (1974) imply that the unaspirated/aspirated stop consonant perceptual boundary in the 20-40 msec VOT range may represent "a characteristic of the auditory processing of acoustic stimuli independent of whether the stimuli are speech or nonspeech" (p. 654).

We will not pursue the details of their second experiment. However, they were able to confirm the contribution of a detectable F1 transition to the voice-voiceless distinction. Furthermore, by hewing to the articulated speech signal and by focusing on acoustic properties within categories rather than on acoustic differences between them, Stevens and Klatt were able to offer a fully plausible account of systematic increases in the voice-voiceless perceptual boundary that are associated with shifts from labial to apical to velar stop consonants (Lisker and Abramson, 1970; Abramson and Lisker, 1973).

If the argument of the last few pages has given the impression that auditory boundaries between phonetic categories are readily determined, the impression must be dispelled. The criterion for such boundaries is that they be demonstrated in a nonspeech analog, a feat that has proved peculiarly difficult for the voiced stop consonants. The typical outcome of studies in which formant patterns controlling consonant assignments are removed from context and presented for discrimination is that they are perceived continuously (e.g., Mattingly, Liberman, Syrdal, and Halwes, 1971). A striking instance is provided by the work of Popper (1972). He manipulated F2 transitions within a three-formant pattern (cf. Figure 3) to yield a synthetic series from /ab/ to /ad/. He then measured energy passed by a 300 Hz band-width filter, centered around the F2 steady-state frequency, and noted a sharp drop at the /b-d/ boundary both for isolated F2 and for the full formant pattern. However, subjects evinced the expected discrimination peak only for the full pattern: the isolated F2, despite its acoustic discontinuity, was continuously perceived.

In short, no simple notion of fixed regions of auditory sensitivity serves to account for categorical division even of the /ba,da,ga/ continuum, let alone for perceptual invariance across phonetic contexts, for the normalizing shifts in category boundary associated with speaker variation (cf. Fourcin, 1968; Rand, 1971), or for cross-language differences in boundary placement. The problem is not confined to articulatory place distinctions. Consider, for example, the fact that Spanish speakers typically yield a somewhat shorter labial VOT boundary than do English (Lisker and Abramson, 1964) and that their perceptual boundary shows a corresponding reduction (Lisker and Abramson, 1970). We can hardly account for the perceptual shift by appeal to an inherently sharp threshold. Precise category position along a continuum is clearly a function of linguistic experience (see also Stevens, Liberman, Studdert-Kennedy, and Öhman, 1969). Popper (1972) proposes, in fact, that "people who speak different languages may tune their auditory systems differently" (p. 218). Differential "tuning" could result from cross-language differences in selective attention to aspects of the signal, and in criterion levels for particular phonetic decisions. Given the close match between perception and production (Stevens et al., 1969; Abramson and Lisker, 1970; Lisker and Abramson, 1970), it seems plausible that such differences should arise from complex interplay between speaking and listening during language acquisition (see below, From Acoustic Feature to Phonetic Percept).

The notion of "tuning" presupposes the existence of acoustic properties to which the auditory system may be attuned. The first steps toward definition of these properties have been taken by Stevens (see especially 1972b, 1973). As earlier remarked, Stevens has used spectrographic analysis and computations from an idealized vocal tract model to describe possible acoustic correlates of certain phonetic features. He finds, for example, that the spectral patterns associated with continuous changes in place of articulatory constriction along the vocal tract do not themselves change continuously. Rather, there are broad plateaux, within which changes in point of constriction have little acoustic effect, bounded by abrupt acoustic discontinuities. These acoustic plateaux tend to correlate with places of articulation in many languages. In short, Stevens is developing the preliminaries to a systematic acoustic account of phonetic categories and their boundaries. His work is important for its emphasis on the origin of phonetic categories in the peculiar properties of the human vocal tract. Furthermore, as will be seen below, his approach meshes neatly with recent work on auditory feature analyzing systems as the bases of phonetic categories.

Auditory and Phonetic Processes in Categorical Perception

We turn now to the second main aspect of categorical perception--listeners' failure to discriminate among members of a category--and to the contrast between continuously perceived vowels and categorically perceived stop consonants. A long series of experiments over the past few years has shown that listeners' difficulty in discriminating among members of a category is largely due to the low energy transience of the acoustic signal on the basis of which phonetic categories are assigned. Lane (1965) pointed to the greater duration and intensity of the vowels and showed that they were more categorically perceived if they were degraded by being presented in noise. Stevens (1968b) remarked the brief, transient nature of stop consonant acoustic cues, and showed, as did Sachs (1969) (1969), that vowels were more categorically perceived if their duration and acoustic stability were reduced by placing them in CVC syllables.

The role of auditory memory, implicit in the work just cited, was made explicit by Fujisaki and Kawashima (1969, 1970) in a model of the decision process during the ABX trial. If a listener assigns A and B to different phonetic categories (i.e., if A and B lie on opposite sides of a phonetic boundary), his only task is to determine whether X belongs to the same category as A or as B: his performance is then good and a discrimination peak appears in the function for both consonants and vowels. However, if a listener assigns A or B to the same phonetic category, he is forced to compare X with his auditory memory of A and B: his performance is then slightly reduced for vowels, for which auditory memory is presumed to be relatively strong, but sharply reduced for consonants, for which auditory memory is presumed to be weak. Evidence for the operation of such a two-step process within phonetic categories in man, but not in monkey, has recently been reported by Sinnott (1974).

Before we proceed, let us spell out some distinctions between auditory and phonetic memory stores. The auditory store, or trace, is usually assumed to be rather like an echo: a faint simulacrum, if not of the waveform, at least of its neural correlates at an early stage of processing. Like an echo, the trace is an analog of its original, decays rapidly, and may be displaced if another sound arrives to interfere before decay is complete. The phonetic store, on the other hand, is a set of discrete features, its decay is a good deal slower, and interference can only be accomplished by another phonetic entity with similar phonetic features.

With this in mind, we turn to several experiments by Pisoni (1971, 1973a, 1973b) in which he tested and supported Fujisaki and Kawashima's hypothesis concerning auditory memory for consonants and vowels. In the first (Pisoni, 1973a) he varied the A-to-X delay interval from zero to two seconds in an AX <u>same</u> - <u>different</u> task for vowel and stop consonant continua. Between-category performance (presumably based on phonetic store) was high and independent of delay interval for both consonants and vowels; within-category performance (presumably based on auditory store) was low and independent of delay interval for consonants, but for vowels was high and declined systematically as delay interval increased. In subsequent experiments, Pisoni (1973b) demonstrated that the degree of categorical (or continuous) perception of vowels can be manipulated by the memory demands of the discrimination paradigm and by the amount of interference from neighboring stimuli (Glanzman and Pisoni, 1973).

Changing tack, Pisoni and Lazarus (1974) sought methods of increasing apparent auditory memory for stop consonants. This is more difficult, but by a particular combination and sequence of experimental conditions, they were able to demonstrate improved within-category discrimination on a voice-voiceless continuum. The same continuum (/ba-pa/) also elicited reaction time differences in a pair-matching task (Pisoni and Tash, 1974; cf. Posner, Boies, Eichelman, and Taylor, 1969). Here, listeners were asked to respond <u>same</u> or <u>different</u> to pairs of stimuli drawn from the continuum. <u>Same</u> reaction times were faster for identical pairs than for acoustically distinct pairs, drawn from the same phonetic category; <u>different</u> reaction times decreased as acoustic differences between pairs from different categories increased. This last result recalls Barclay's (1972) finding that listeners can correctly and reliably judge acoustic variants of /d/, drawn from a synthetic continuum, as more similar to /b/ or /g/. If we add these studies to our earlier observation that listeners always display a margin of within-category discrimination for consonants, and that discrimination functions do not display a quantal leap between categories, we must conclude that the auditory system does retain at least some trace of consonantal passage. At the same time, there is little question that this trace is fainter than that for vowels.

The conclusion of all these studies is pointed up by the work of Raphael (1972). He studied voice-voiceless VC continua, manipulating initial vowel duration as the acoustic cue to voicing of the final stop. Here, where the perceptual object was consonantal, but the acoustic cue vocalic, perception was continuous. In short, consonants and vowels are distinguished in the experiments we have been considering, not by their phonetic class or the processes of assignment to that class, but by their acoustic characteristics and by the duration of their auditory stores. If the longer store of the vowels is experimentally reduced, their membership in the natural class of segmental phonetic entities is revealed by their categorical perception.

Stages of Auditory Memory

Several independent lines of research, drawing on different experimental paradigms, have recently begun to converge on perceptual and memorial processes below the level of phonetic classification. Experimenters often share neither terminology nor theoretical framework, but we can discern two, not entirely overlapping, lines of division in the perceptual process. The first divides short-term memory into a brief store lasting some hundreds of milliseconds, and a longer store lasting several seconds. The second divides peripheral from central processes; this is important, but we will not consider it in detail here, since the cut cannot be as surely made in audition as in vision (due to incomplete decussation of auditory pathways), and most of the processes to be discussed are certainly central.

Short-Term Auditory Stores

Store I. As a step toward further analysis of auditory memory for speech, consider the concept of parallel processing. Liberman et al. (1967) used this term to describe the decoding of a CV syllable, in which acoustic correlates of consonant and vowel are distributed over an entire syllable (Liberman, 1970). Obviously, the process requires a store at least as long as the syllable to register auditory information, and presumably somewhat longer to permit transfer into phonetic code.

Direct evidence of this type of parallel processing comes from several sources. Liberman et al. (1952) showed that the phonetic interpretation of a stop release burst varied with its following vowel, and concluded that we perceive speech over stretches of roughly syllabic length (cf. Schatz, 1954). Lindblom and Studdert-Kennedy (1967) demonstrated that the phonetic boundary for a series of synthetic vowels shifted as a function of the slope and direction of initial and final formant transitions: listeners judged vowels in relation to their surrounding consonantal frames (cf. Fujimura and Ochiai, 1963; Strange et al., 1974). More recently, Pisoni and Tash (1974) have studied reaction time to CV syllables: they called for same - different judgments on vowels or consonants of syllable pairs in which nontarget portions of the syllables were also either the same or different. Whether comparing consonants or vowels, listeners were consistently faster when target and nontarget portions of the syllable were redundant (i.e., both same, or both different). In other words, information from an entire syllable contributed to listeners' decisions concerning "segments" of the syllable. In a related study by Wood and Day (in press), listeners identified either the vowel or the consonant of synthetic CV syllables, /ba,da,bæ,dæ/. If all test items were identical on the nontarget dimension (i.e., if all had the same vowel on a consonant test, or all the same consonant on a vowel test), subjects' reaction times were significantly faster than if both target and nontarget dimensions varied. In the latter case, the unattended vowel (consonant) retarded listeners' decisions on the attended consonant (vowel). In short, we have a variety of evidence that, for at least some syllables, consonant and vowel recognition are interdependent, parallel processes, requiring a short-term auditory store of at least syllabic duration.

Massaro (1972) made the functional distinction between such a "perceptual auditory image" and a longer "synthesized" auditory store; he initiated attempts to estimate duration of the "image" by backward masking studies. First discovered in visual experiments (Werner, 1935), the paradigm takes advantage of the fact that perception of a stimulus may be blocked if a second stimulus is presented some hundreds of milliseconds later; it has been used to good effect in vision to separate and describe peripheral and central processes (Turvey, 1973). However, the belief that the critical interstimulus interval (ISI), at which the first stimulus is freed from interference by the second, may be taken as an estimate of the duration of primary auditory display (Massaro, 1972) is difficult to sustain, and application of the technique to the study of speech perception has proved problematic for several reasons.

To begin with, auditory information is displayed over time, so that perception of a target CV syllable of natural duration (say, 200-300 msec) can be interrupted only by a masking syllable that begins before the first syllable is complete. Temporal relations between syllables must then be expressed in terms of stimulus onset asynchrony (SOA) rather than in terms of ISI, and the effectiveness of the mask is reduced because it is itself masked by the first syllable (forward masking). For example, Studdert-Kennedy, Shankweiler, and Schulman (1970b) found that the first syllable was completely freed from masking by the second at a SOA of 50 msec, certainly an underestimate of display time, since it is no more than the duration of the critical consonant information in the formant transitions of the target CV syllable.

There are two solutions to this impasse: make the syllables unnaturally short, or present target and mask to opposite ears (dichotically), thus evading peripheral masking of the second syllable. Several investigators (Massaro, 1972; Pisoni, 1972; Dorman, Kewley-Port, Brady-Wood, and Turvey, 1973) have attempted the first solution. Results are difficult to interpret because both the degree of masking and the critical ISI for release from masking vary with target (consonant or vowel), size and range (acoustic or phonetic) of target set, target and mask energy, relations between target and mask structure (acoustic or phonetic), and individual listeners, many of whom show no masking whatever even for brief (15.5 msec) vowels (Dorman et al., 1973). Where masking could be obtained, the shortest critical ISI observed in these studies (80 msec) was for 40 msec steady-state vowels, and the longest (250 msec) for 40 msec CV syllables (Pisoni, 1972). Note, incidentally, that complete absence of masking has been observed only with vowels, and just as categorical perception of vowels can be induced by degrading them with noise, so too can their masking (Dorman, Kewley-Port, Brady, and Turvey, 1974).

In any event, these variable results do not encourage one to believe that critical ISI is measuring the fixed duration of auditory display. And the case is no better when we turn to dichotic masking paradigms. Pisoni and McNabb (1974), for example, observed a critical SOA for release from dichotic backward masking of between 20 and 150 msec, depending upon target and mask vowel relations. A somewhat longer estimate of 200-250 msec can be extrapolated from the data of Studdert-Kennedy et al. (1970b). A narrower estimate comes from McNeill and Repp (1973b). They studied forward masking of dichotically presented CV syllables, determining the SOA necessary for features of the leading syllable to have no further effect on errors in the lagging syllable, and so presumably to have passed out of the phonetic processor. Their estimate of 80-120 msec may be more realistic for running speech than others, since their procedure eliminated a component present in all previous studies, namely, time taken to prepare a response, a period during which effective interruption may still occur (Repp, 1973).

However, it is more likely that the entire endeavor is misguided. It seems intuitively plausible that syllable processing time is not constant, but varies, under automatic attentional control, with speaking rate and other factors. The studies reviewed are simply measuring time required for release from masking under a variety of more or less adverse conditions. This is certainly not without interest, particularly if we can show it to be a function of well-specified target-mask relations. But we shall then be turning attention away from the notion of a primary auditory store, and toward the more important question of what acoustic dimensions are extracted in the very earliest stage of processing, and how they interact to determine the phonetic percept.

Store II. Nonetheless, some form of auditory store is clearly necessary. We would otherwise be unable to interpret the prosody of running speech, and there is ample experimental evidence of cross-syllabic auditory interaction (Hadding-Koch and Studdert-Kennedy, 1964; Studdert-Kennedy and Hadding, 1973; Atkinson, 1973). Detailed analysis of this longer store, perhaps lasting several seconds, was made possible by the work of Crowder and Morton (1969; see also Crowder, 1971a, 1971b, 1972, 1973). They were the first experimenters to undertake a systematic account of what they termed "precategorical acoustic storage" (PAS).

Evidence for the store comes from studies of immediate, ordered recall of span-length digit lists. Typically, error probability increases from beginning to end of list, with some slight drop on terminal items (recency effect). The terminal drop is significantly increased, if the list is presented by ear rather than by eye (modality effect). Crowder and Morton (1969) argue that these two effects reflect the operation of distinct visual and auditory stores for precategorical (prelinguistic) information, and of an auditory store that persists longer than the visual. Support comes from demonstrations that the recency effect is significantly reduced, or abolished, if subjects are required to recall the list by speaking rather than by writing (Crowder, 1971a), or if an auditory list is followed by a redundant, spoken suffix (such as the word zero), as a signal for the subject to begin recall (suffix effect). That the suffix interferes with auditory, rather than linguistic, store is argued by the facts that the effect (1) does not occur if the suffix is a tone or burst of noise; (2) is unaffected if the spoken suffix is played backward; (3) is unaffected by degree of semantic similarity between suffix and list; (4) is reduced if suffix and list are spoken in different voices; and (5) is reduced if suffix and list are presented to opposite ears.

Of particular interest in the present context is that all three effects (modality, recency, suffix) are observed for CV lists, of which members differ in vowel alone, or in both vowel and consonant (spoken letter names), but not for voiced stop consonant CV or VC lists, of which members differ only in the consonant (cf. Cole, 1973b). Crowder (1971a:595) concludes that "vowels receive some form of representation in PAS while voiced stop consonants receive none." Liberman, Mattingly, and Turvey (1972:329) argue further that phonetic classification "strips away all auditory information" from stop consonants.

However, this last claim is unlikely to be true. First, there is no good reason why the process of categorization should affect vowels and consonants differently. Second, we have a variety of evidence that listeners retain at least some auditory trace of stop consonants (see previous section). Third, consonant and vowel differences in PAS can be reduced by appropriate manipulation of the signal array (Darwin and Baddeley, 1974). These investigators demonstrated a recency effect for tokens of a stop CV, /ga/, and two highly discriminable CV syllables in which the consonantal portion is of longer duration, /ʃa/, /ma/. They also demonstrated that the recency effect for vowels can be eliminated if the vowels are both very short (30 msec of a 60 msec CV syllable) and close neighbors on an F1-F2 plot. They conclude that "the consonant-vowel distinction is largely irrelevant" (p. 48) and that items in PAS cannot be reliably accessed if, like /ba,da,ga/ or /I,ɛ,æ/, they are acoustically similar. The effect of acoustic similarity is, of course, to confound auditory memory. As we shall see shortly (The Acoustic Syllable, below) and, as Darwin and Baddeley (1974) themselves argue, it is to the more general concept of auditory memory that we must have recourse, if we are to understand the full range of experiments in which consonant-vowel differences have been demonstrated.

We turn now to the duration of PAS and the mechanisms underlying its reflection in behavior. Notice, first, that if an eight-item list is presented at a rate of two per sec and is recalled at roughly the same rate, time between presentation and recall will be roughly equal for all items. Therefore, the recency effect cannot be attributed to differential decay across the list, but is due rather to the absence of "overwriting" or interference from succeeding items. Second, since the degree of interference (i.e., probability of recall error) decreases as the time between items increases, and since the suffix effect virtually disappears if the interval between the last item and suffix is increased to 2 sec, we are faced with the paradox that performance improves as time allowed for PAS decay increases. Crowder's (1971b) solution is to posit an active "read-out" or rehearsal process at the articulatory level. Time for a covert run through the list is "...a second or two" (p. 339). If a suffix occurs during this period, PAS for the last couple of items is spoiled before they are reached; if no suffix occurs, the subject has time to check his rehearsal of later items against his auditory store, and so to confirm or correct his preliminary decision. Crowder (1971b) goes on to show that there is, in fact, no evidence for any decay in PAS: in the absence of further input, PAS has an infinite duration. This is intuitively implausible, but we will not pursue the matter here.

Notice, however, that the term precategorical refers to the nature of the information stored, not to the period of time during which it is stored. A preliminary (or even final) articulatory, if not phonetic, decision must have been made before PAS is lost, if rehearsal is to permit cross-check with the store. We are thus reminded of the temporary auditory store hypothesized in the analysis-by-synthesis model of Stevens (1960, 1972a). Crowder's account, with its preliminary analysis and generative rehearsal loop, is so similar to Stevens' model that we may be tempted to identify the two, and to see evidence for PAS function as support for Stevens' hypothesis.

We may remark, however, one important difference. Stevens introduced a synthesis loop to handle the invariance problem, a problem at its most acute for stop consonants. But these are precisely the items excluded from PAS, and all our evidence for consonantal auditory memory suggests a store considerably less than infinite, probably less than a second. We may, of course, assume that a synthesis loop goes into operation very early in the process, while consonant auditory information is still available, and that the PAS rehearsal loop is simply a sustention beyond the point at which stop consonantal auditory information can be accessed. We would then be forced to posit the decay of consonantal information from auditory store. Continuation of the loop might be automatic during running speech, enabling prosodic pattern to emerge, but under attentional control for special purposes, such as listening to poetry and remembering telephone numbers. But we have, at present, no direct evidence for the earlier stage of the loop.

Stages of Processing

Nor, as we have seen, do we have direct evidence for the primary auditory store inferred from parallel processing. We may, in fact, do well to dismiss division of the process into hypothetical stores, and concentrate attention on the types of information extracted during early processing, and their interactions. Several experimental paradigms have already been applied.

Day and Wood (1972) and Wood (1974) have reported evidence for parallel extraction of pitch (fundamental frequency) and spectral information bearing on segmental classification. For the first experiment they synthesized two CV syllables, /ba,da/, each at two pitches, and prepared two types of random test order. In one, they varied a single dimension, either fundamental frequency or phonetic class; in the other, they varied both dimensions independently. They then called on subjects to identify, with a reaction-time button, either pitch or phonetic class, each in its appropriate one-dimensional test and also in the two-dimensional test. Reaction times were longer for both tasks on the two-dimensional test than on the one-dimensional test, but the increase was significantly greater on the phonetic test than on the pitch task: unpredictable pitch differences interfered with phonetic decision more than the reverse. The authors took this finding as evidence for separate nonlinguistic and linguistic processes, the first mandatory, the second optional. In a follow-up experiment, Wood (1974) substituted a two-dimensional test in which fundamental frequency and phonetic class variations were correlated rather than independent. Reaction times were now significantly shorter for both tasks on the two-dimensional test: subjects drew on both pitch and phonetic information for either pitch or phonetic classification. Wood (1974) concludes that the two types of information are separately and simultaneously extracted [as required, incidentally, by Stevens' (1960) model].

There is more to these experiments. The phonetic task called for a decision on the consonant (/ba/ vs /da/), but pitch information was primarily carried by the vowel. In fact, had fundamental frequency differences been carried solely by initial formant transitions, it is doubtful whether they would have interacted with phonetic decision. Dorman (1974) has shown that listeners are unable to discriminate intensity differences carried by the 50 msec initial transitions of a voiced stop CV syllable, but are well able to discriminate identical differences carried by isolated transitions, or by the first 50 msec of a steady-state vowel. While the experiment has not been done, it seems likely that Dorman's results would have held had he used fundamental frequency instead of intensity. We would then be forced to conclude that, in Wood's (1973b) experiment, subjects were using adventitious pitch information carried by the vowel to facilitate judgment of the consonant, and vice versa. The experiments thus reflect parallel processing, both of linguistic and nonlinguistic information and of consonant and vowel.

Experimental separation of auditory and phonetic processes has also been attempted in dichotic studies. Consider, for example, the following series. Shankweiler and Studdert-Kennedy (1967; also Studdert-Kennedy and Shankweiler, 1970) found that listeners were significantly better at identifying the consonants of dichotically competing CV or CVC syllables if the consonants shared a phonetic feature than if they did not. Since the effect was present both for pairs sharing vowel (e.g., /bi,di/, /du,tu/, etc.) and for pairs not sharing vowel (e.g., /bi,du/, /di,tu/, etc.), and since the latter pairs differ markedly in the auditory patterns by which the shared features are conveyed, Studdert-Kennedy, Shankweiler, and Pisoni (1972) concluded that the effect had a phonetic rather than an auditory basis. In another experimental paradigm, Studdert-Kennedy et al. (1970b) presented CV syllables at various values of SOA and demonstrated dichotic backward masking. They attributed the masking to interruption of central processes of speech perception, but left the level at which the interruption occurred uncertain (cf. Kirstein, 1971, 1973; Porter, 1971; Berlin, Lowe-Bell, Cullen, Thompson, and Loovis, 1973; Darwin, 1971a).

Recently, Pisoni and McNabb (1974) have combined and elaborated the two paradigms in a dichotic feature-sharing study, varying both masks and SOA. Their targets were /ba,pa,da,ta/; their masks were /ga,ka,gæ,kæ,gɛ,kɛ/. If target and mask consonants shared voicing, little or no masking was observed. If they did not share voicing, masking of the target consonant increased both as the masking-syllable vowel approached target-syllable vowel from /ɛ/ through /æ/ to /ɑ/, and as the mask intensity increased. In other words, identification of the target consonant was facilitated by similarity of the masking consonant, but, in the absence of facilitation, was impeded by similarity of the masking vowel, particularly if the vowel was of relatively high intensity. In a theoretical discussion of these results, Pisoni (in press) concludes that masking and facilitation occur at different stages of the perceptual process: masking reflects integration (rather than interruption) at the auditory level, while facilitation reflects integration at the phonetic level.

However, these results are also open to a purely auditory interpretation. They seem, in fact, to be consistent with a system that extracts the acoustic correlates of voice onset time separately for each vowel context (cf. Cooper, 1974b). We are thus led to consider the possible role of discrete acoustic feature analyzing systems, tuned to speech. This has proved among the most fruitful approaches to analysis of early processing, but we defer discussion to a later section (see below, Feature-Analyzing Systems).

The Acoustic Syllable

We have now touched on some half dozen paradigms--categorical perception, backward masking, short-term memory, reaction time studies, and others--in which consonant and vowel perception differ. As a final example, we may mention dichotic experiments [Berlin (in Lass, in press)]. Shankweiler and Studdert-Kennedy (1967; also Studdert-Kennedy and Shankweiler, 1970) showed a significant right-ear advantage for dichotically presented CV or CVC syllables differing in their initial or final consonants, but little for steady-state vowels or CVC syllables differing in their vowels. Day and Vigorito (1973) and Cutting (in press-b) reported a hierarchy of ear advantages in dichotic listening from a right-ear advantage for stop consonants through liquids to a null or small left-ear advantage for vowels. Recently, Weiss and House (1973) have demonstrated that a right-ear advantage emerges for vowels, if they are presented at suitably unfavorable signal-to-noise ratios, while Godfrey (1974) has shown that the right-ear advantage for vowels may be increased by adding noise, reducing duration,or using a more confusable set of vowels (cf. Darwin and Baddeley, 1974).

The pattern is familiar. In virtually every instance, a consonant-vowel difference can be reduced or eliminated by taxing the listener's auditory access to the vowel, or by sensitizing his auditory access to the consonant. These qualifications only serve to emphasize the contrast between them, and to pinpoint its source in their acoustic structure. The consonant is transient, low in energy, and spectrally diffuse; the vowel is relatively stable, high in energy, and spectrally compact.

Together they form the syllable, each fulfilling within it some necessary function. Consider, first, vowel duration. Long duration is not necessary for recognition. We can identify a vowel quite accurately and very rapidly from little more than one or two glottal pulses, lasting 10 to 20 msec. Yet in running speech, vowels last ten to twenty times as long. The increased length may be segmentally redundant, but it permits the speaker to display other useful information: variations in fundamental frequency, duration, and intensity within and across vowels offer possible contrasts in stress and intonation, and increase the potential phonetic range (as in tone languages). Of course, these gains also reduce the rate at which segmental information can be transferred, increase the duration of auditory store, and open the vowel to contextual effects--the more so, the larger the phonetic repertoire. A language built on vowels, like a language of cries, would be limited and cumbersome.

Adding consonantal "attack" to the vowel inserts a segment of acoustic contrast between the vowels, reduces vowel context effects, and increases phonetic range. The attack, itself part of the vowel [the two produced by "...a single ballistic movement" (Stetson, 1952:4)], is brief, and so increases the rate of information transfer. Despite its brevity, the attack has a pattern arrayed in time, and the full duration of its trajectory into the vowel is required to display the pattern. To compute its phonetic identity, time is needed, and this is provided by the segmentally redundant vowel. Vowels are the rests between consonants.

Finally, rapid consonantal gestures cannot carry the melody and dynamics of the voice. The segmental and suprasegmental loads are therefore divided over consonant and vowel: the first, with its poor auditory store, taking the bulk of the segmental load; the second taking the suprasegmental load. There emerges the syllable, a symbiosis of consonant and vowel, a structure shaped by the articulatory and auditory capacities of its user, fitted to, defining, and making possible linguistic and paralinguistic communication.

SPECIALIZED NEURAL PROCESSES

Cerebral Lateralization

That the left cerebral hemisphere is, in most persons, specialized for language functions is among the most firmly established findings of modern neurology. That one of those functions may be to decode the peculiar acoustic structure of the syllable into its phonetic components was first suggested by the results of dichotic studies. Kimura (1961a, 1961b, 1967) discovered that if different digit triads were presented simultaneously to opposite ears, those presented to the right ear were more accurately recalled than those presented to the left. She attributed the effect to functional prepotency of contralateral pathways under dichotic competition, and to left-hemisphere specialization for language functions. Later experiments have amply supported her interpretation.

Shankweiler and Studdert-Kennedy (1967) applied the technique to analysis of speech perception. They demonstrated a significant right-ear advantage for single pairs of nonsense syllables differing only in initial or final stop consonant, and separable advantages for place of articulation and voicing (Studdert-Kennedy and Shankweiler, 1970; cf. Halwes, 1969; Darwin, 1969; Haggard, 1971). Among the questions raised by these studies was whether the left hemisphere was specialized only for phonetic analysis, or also for extraction of speech-related acoustic properties, such as voice onset, formant structure, temporal relations among portions of the signal, and so on. We will not rehearse the argument here, but simply state the conclusion that "while the auditory system common to both hemispheres is equipped to extract the auditory parameters of a speech signal, the dominant hemisphere may be specialized for the extraction of linguistic features from those parameters" (Studdert-Kennedy and Shankweiler, 1970:594).

Striking evidence in support of this conclusion has recently been gathered by Wood (1975) and Wood, Goff, and Day (1971). This work deserves careful study, as an exemplary instance of the use of electroencephalography (EEG) in the study of language-related neurophysiological processes. Wood synthesized two CV syllables, /ba/ and /ga/, each at two fundamental frequencies, 104 Hz (low) and 140 Hz (high). From these syllables he constructed two types of random test order: in one, items differed only in pitch [e.g., /ba/ (low) vs /ga/ (high)]; in the other, they differed only in phonetic class [e.g., /ba/ (low) vs /ga/ (low)]. Subjects were asked to identify either the pitch or the phonetic class of the test items with reaction-time buttons. While they did so, evoked potentials were recorded from a temporal and a central location over each hemisphere. Records from each location were averaged and compared for the two types of test. Notice that both tests contained an identical item [e.g., /ba/ (low)], identified on the same button by the same finger. Since cross-test comparisons were made only between EEG records for identical items, the only possible source of differences in the records was in the task being performed, auditory (pitch) or phonetic. Results showed highly significant differences between records for the two tasks at both left-hemisphere locations, but at neither of the right-hemisphere locations. A control experiment, in which the "phonetic" task was to identify isolated initial formant transitions (50 msec), revealed no significant differences at either location over either hemisphere. Since these transitions carry all acoustic information by which the full syllables are phonetically distinguished, and yet are not recognizable as speech, we may conclude that the original left-hemisphere differences arose during phonetic, rather than auditory, analysis. We will discuss the adequacy of isolated formant transitions as control patterns in the next section. However, the entire set of experiments strongly suggests that different neural processes go on during phonetic, as opposed to auditory, perception in the left hemisphere, but not in the right hemisphere (cf. Molfese, 1972).

The distinctive processes of speech perception would seem then to lie in linguistic rather than acoustic analysis. Two other types of evidence suggest the same conclusion. First, visual studies have repeatedly shown a right-field (left-hemisphere) advantage for tachistoscopically presented letters and, by contrast, a left-field (right-hemisphere) advantage for nonlinguistic geometric forms (for a review, see Kimura and Durnford, 1974). Second, Papçun, Krashen, Terbeek, Remington, and Harshman (1974) and Krashen (1972) have shown a right-ear advantage in experienced Morse code operators for dichotically presented Morse code words and letters. If the arbitrary patterns of both a visual and an auditory alphabet can engage left-hemisphere mechanisms, there might seem tu be little ground for claiming special status for the speech signal.

However, alphabets are secondary, and while their interpretation may well engage specialized linguistic mechanisms, analysis of their arbitrary signal patterns clearly should not. The speech signal, on the other hand, is primary, its acoustic pattern at once the natural realization of phonological system and the necessary source of phonetic percept. Given its special status and peculiar structure, we should perhaps be surprised less if there were, than if there were not, specialized mechanisms adapted to its auditory analysis.

Hints of such processes have begun to appear. Halperin, Nachshon, and Carmon (1973), for example, showed a shift from left-ear advantage to right-ear advantage for dichotically presented tone sequences as a function of the number of alternations in the sequence. Their stimuli were patterned permutations of brief (200 msec) tone bursts, presumably not unlike those of Papçun et al. (1974), who showed a right-ear advantage in naive subjects for Morse code patterns up to seven units in length. Both studies suggest left-hemisphere specialization for assessing the sort of temporal relations important in speech. Both studies suffer from having called upon subjects to label the patterns, a process that might well invoke left-hemisphere mechanisms.

This weakness is avoided in recent work by Cutting (in press-b). He synthesized two normal CV syllables, /ba/ and /da/, and two phonetically impossible "syllables" identical with the former except that their first formant transitions fell rather than rose along the frequency scale, so that they were not recognized as speech. In a nonlabeling dichotic task, subjects gave equal right-ear advantages for both types of stimulus. The outcome suggests a left-hemisphere mechanism for extraction of formant transitions and is reminiscent of a study by Darwin (1971b), who found a right-ear advantage for synthetic fricatives when formant transitions from fricative noise into vowel were included, but no ear advantage when transitions were excluded.

There are, then, grounds for believing that the left hemisphere is specialized not only for phonetic interpretation of an auditory input, but also for extraction of auditory information from the acoustic signal. The evidence is tenuous, but systematic study of feature-analyzing systems--whether lateralized or not remains to be seen (cf, Ades, 1974a)--has opened up a new range of possibilities.

Feature-Analyzing Systems

Neurophysiological systems of feature detectors, selectively responsive to light patterns, were first reported by Lettvin, Maturana, McCulloch, and Pitts (1959). They found receptive fields in the visual ganglion cells of frog that responded, under specific conditions, to movement. The biological utility of the system to an animal that preys on flies is obvious. Moving up the nervous system, and the evolutionary scale, Hubel and Wiesel (1962) reported yet more complex detectors: single cells in the visual cortex of cat that responded selectively to the orientation of lines, to edges, and to movement in a certain direction. Since then, work on visual feature-detecting systems has proliferated (see Julesz, 1971:58-68, for a review).

Complex auditory feature detectors in the cortex of cat were reported by Evans and Whitfield (1964): single cells responsive to specific gradients of intensity change, and others ("miaow" cells) to the rate and direction of frequency change (Whitfield and Evans, 1965). Similar cells were reported by Nelson, Erulkar, and Bryan (1966) in the inferior colliculus of cat. Other research has borne directly on acoustic signaling systems. Frishkopf and Goldstein (1963) and Capranica (1965) reported single units in the auditory nerve of bullfrog responsive only to the male bullfrog's mating call. Recently, Wollberg and Newman (1972) have described single cells in the auditory cortex of squirrel monkey which answer to that species' "isolation peep." Stimulus and response were isomorphic: presentation of the "peep" with portions gated out yielded a response in which corresponding portions were absent. Furthermore, the remaining portions were no longer normal: if a central portion of the signal was missing, the response pattern to the final portion changed. Interaction of this kind is particularly interesting in light of the contextually variant cues of speech, for which interpretation may demand details of a complete pattern, such as the syllable.

The relevance of all this to speech has not gone unnoticed. The possible role of feature-detecting systems in speech perception was scouted briefly by Liberman et al. (1967), by Studdert-Kennedy (1974), and, at considerable length, by Abbs and Sussman (1971). However, advance awaited a telling experimental procedure. This was found in "adaptation" studies, a method with a long history in visual research (Woodworth and Schlosberg, 1954). The paradigm is simple enough. For example, after prolonged fixation of a line curved from the median plane, a vertical line, presented as a test stimulus, appears curved in the opposite direction: there is a "figural after-effect" in which portions of the image are displaced (Köhler and Wallach, 1944). Related effects in color and tilt also occur. While none of these effects is understood in any detail, they are frequently interpreted in terms of specific receptors or of feature-analyzing systems. Prolonged stimulation "fatigues" or "adapts" one system, and relatively "sensitizes" a physically adjacent or related (perhaps opponent) system. On this interpretation, to demonstrate perceptual shifts upon prolonged exposure to a particular physical (or psychological) "feature" is to demonstrate the presence of analyzing systems for that feature, and its relative.

The method was first used by Warren and Gregory [1958; see also Warren, 1968 (in Lass, in press); Perl, 1970; Clegg, 1971; Lass and Golden, 1971; Lass and Gasperini, 1973; Lass, West, and Taft, 1973; Obusek and Warren 1973], yielding an effect that they termed "verbal transformation." Subjects listen to a meaningful word played repeatedly once or twice per second for several minutes, and are asked to report any changes in the word that they hear. They report a large number of transformations, usually meaningful words and not always closely related phonetically to the original. However, Goldstein and Lackner (in press) refined the method by using nonsense syllables to reduce semantic influence (CV, V, VC) and by presenting them monaurally. They analyzed transforms phonetically, and showed that each was confined to a single phone, usually on one or two distinctive features (as defined by Chomsky and Halle, 1968), and were independent of their syllabic context. Furthermore, the right ear gave significantly more transforms than the left ear on consonants, but not on vowels, and the transforms followed the phonological constraints of English. These last two points are among the arguments that the authors present for suspecting that the effects result from adaptation of phonetic, rather than auditory, analyzing systems.

In a further refinement, Lackner and Goldstein (in press) used a natural CV syllable, repeated monaurally 36 times in 30 sec, and a final test item presented to either the same or the opposite ear. Both adapting and test items were drawn from the set of six English stop consonants, followed by the same vowel (either /i/ or /e/). Subjects reported the last adapting item and the test item. Transforms in the test item occurred on both cross-ear (30 percent) and same-ear (40 percent) trials. They were significantly more likely to occur if the final adapting item was also transformed, and to be on the same feature(s) (place and/or voice) as the adapting item transform, a result that again hints at phonetic feature-detecting systems. The authors conclude from the cross-ear trials that adaptation is central, rather than peripheral, but, unable in this study to distinguish phonetic effects from the acoustic effects that underlie them, they withhold judgment on whether the transforms are auditory or phonetic.

This last is, of course, the crucial question. It can be approached only by use of synthetic speech in which acoustic features can be specified precisely and, within limits, manipulated independently of phonetic category. Eimas, working independently of the previous authors, took this step in a series of experiments growing out of his work on infants (discussed below), and has concluded that the effect is phonetic. We will consider his work in some detail because it introduced a fruitful paradigm that has already been put to good use by others.

In the first experiment (Eimas and Corbit, 1973), the authors used two voice-voiceless series synthesized along the VOT continuum, one from /ba/ to /pa/, the other from /da/ to /ta/ (Lisker and Abramson, 1964). On the assumption of two voicing detectors, each differentially sensitive to VOT values that lie clearly within its phonetic category, and both equally sensitive to a VOT value at the phonetic boundary, the authors reasoned that adaptation with an acoustically extreme token of one phonetic type should desensitize its detector, and relatively sensitize (a metaphor, not an hypothesis) its opponent detector, to boundary values of VOT, with a resulting displacement of the identification function toward the adapting stimulus. They, therefore, collected unadapted and adapted functions for both labial and alveolar series. The adapting stimuli were drawn from the extremes of both series, and their effects were tested within and across series. Figure 5 shows the results for one of their three subjects (the experiment is taxing and prohibits large samples, but the other two subjects gave similar functions). The predicted results obtain. Furthermore, the effect is only slightly weaker across series than within. (This result was replicated in an experiment, briefly reported in their next paper, for which they used eight subjects to demonstrate boundary shifts on alveolar and velar stop consonant VOT continua after adaptation with labial stops.) In a supporting experiment the authors showed that, following adaptation, the peak in an ABX discrimination function is neatly shifted to coincide with the adapted phonetic boundary (cf. Cooper, 1974a).

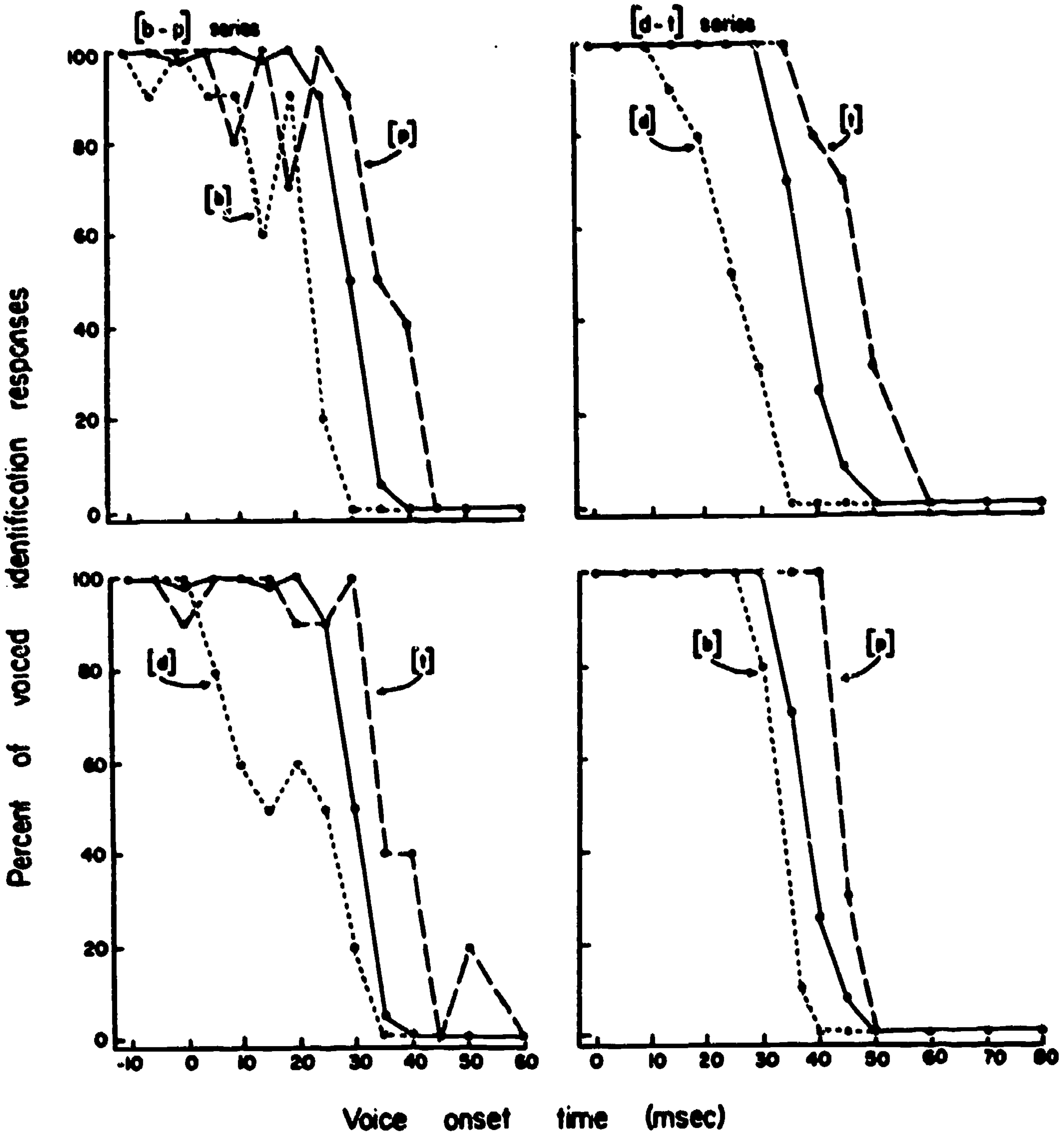


Figure 5: Percentages of voiced identification responses ([b or d]) obtained with and without adaptation, for a single subject. The functions for the [b,p] series are on the left and those for the [d,t] series are on the right. The solid lines indicate the unadapted identification functions; the dotted and dashed lines indicate the identification functions after adaptation. The phonetic symbols indicate the adapting stimulus. [From Eimas and Corbit (1973) with permission of authors and publishers.]

In a second study (Eimas, Cooper, and Corbit, 1973), the authors report three experiments. The first demonstrates that the site of the adaptation effect is probably central rather than peripheral: it obtains as strongly when the adapting stimulus is presented to one ear and the test stimulus to the other, as when both are presented binaurally (cf. Ades, 1974a). The second demonstrates that the effect is not obtained if the adapting stimulus is simply the first 50 msec of the syllable /da/, an acoustic pattern that contains all the voicing information, but is not heard as speech. The third experiment assesses the relative strengths of the two hypothesized detectors, finding that, as in the first study (see Figure 5), voiced stops tend to be more resistant to adaptation (yield smaller boundary shifts) than voiceless. The result encourages the hypothesis of separate detectors for each phonetic value along an acoustic continuum, a notion with obvious relevance to categorical perception. Additional support comes from the work of Cooper (1974a), who found evidence of three distinct detectors along a /b-d-g/ continuum: adaptation with /b/ shifted only the /b-d/ boundary, adaptation with /g/ shifted only the /d-g/ boundary, adaptation with /d/ shifted both neighboring boundaries.

Let us remark first the striking achievement of these studies. Whatever the underlying mechanism, Eimas and his colleagues have demonstrated in a novel, direct, and peculiarly convincing manner the operation of some form of feature-analyzing system in speech perception. The outcome was not foregone. There might, after all, have been no adaptation effect at all. Alternatively, the effect might have been on the whole syllable or on the unanalyzed phonemic segment. But these possibilities were ruled out by the cross-series results. The effect proved to be on a feature within the phonemic segment, and so has provided the strongest evidence to date of a physiologically grounded feature system (cf. Cooper and Blumstein, 1974).

What now is the evidence for phonetic rather than auditory adaptation? First, the cross-series effect: phonetic tokens drawn from labial, alveolar, or velar VOT continua differ acoustically in the extent and direction of their second and third formant transitions, yet they are mutually effective adaptors. If the effect were acoustic, the argument runs, the acoustic differences should eliminate the effect. Note, however, that the differences were in acoustic cues to place of articulation, while the feature being tested was voice onset time. The cues to this feature are complex and, as we have seen, relational. Furthermore, Cooper (1974b) has recently shown that VOT adaptation may be vowel-specific: simultaneous adaptation with [da] and [t^{h}i] produced opposite shifts on [ba-p^{h}a] and [bi-p^{h}i] series. Nonetheless, if outputs from such detectors funneled into acoustic analyzers, tuned to presence or absence of energy in the region of the first formant at syllable onset, we would expect precisely the results that were obtained (cf. Stevens and Klatt, 1974).

The second piece of evidence is the failure of the truncated /da/, not heard as speech, to "sensitize" the supposed /ta/ detector. Here the main problem is the status of the truncated /da/ as a control (cf. Wood, 1975). There are two possible types of design that may throw light on the auditory-phonetic issue. In one, control and test items are acoustically identical (on dimensions relevant to the phonetic dimension under test), but phonetically distinct; in the other, they are acoustically distinct, but phonetically identical. The first design, chosen by Eimas and his colleagues, may yield ambiguous results. If adaptation with the control item shifts the phonetic boundary, we have evidence for the existence of auditory detectors tuned to acoustic features of speech.

Precisely this outcome has, in fact, been reported by Ades (1973), using the first 38 msec of the extreme test stimuli to shift the /bæ/-/dæ/ boundary. If, on the other hand, the control item does not shift the boundary, the outcome is ambiguous. It may mean, as Eimas and his colleagues concluded, that the hypothetical detector is phonetic. But it may also mean that an acoustic detector tuned to features of speech is only adapted if stimulated by a complete (i.e., phonetically identifiable) signal (cf. Wollberg and Newman, 1972). It is not, after all, implausible to suppose that the human cortex contains sets of acoustic detectors tuned to speech and capable of mutual inhibition. Each detector may respond to a particular acoustic property, but may be inhibited from output to the phonetic system in the absence of a collateral response in other detectors. The auditory system would then be immune to adaptation by an incomplete signal.

The second type of design calls for control and test items that are acoustically distinct (on dimensions relevant to the phonetic dimension under study), but phonetically identical. This design rests, of course, on the fact that the speech signal may carry several acoustic cues, each a more or less effective determinant of a particular phonetic percept. The procedure is then to synthesize two acoustic continua, manipulating in each a different acoustic cue to the same phonetic distinction. If now the two series are mutually effective in shifting the phonetic boundaries of the other, we have some preliminary support for the hypothetical phonetic detector. This was the outcome of studies by Ades (1974b), Bailey (1973), and Cooper (1974a), all of whom demonstrated cross-series adaptations for /b-d/ continua with different vowels. The use of different vowels meant that formant transitions cueing a given phonetic type could be falling in one token (e.g., /dæ/), rising in another (e.g., /de/). Thus, adaptation of simple acoustic detectors responsive only to rising or only to falling formants (cf. Whitfield and Evans, 1965) was ruled out. Of course, a more complex "acoustic invariance," derived from some weighted ratio of F2 and F3 transitions, might be posited (Cooper, 1974a). But the conclusion that the detectors are phonetic was tempting enough for both Ades and Cooper to draw. Ades qualified his conclusion because, in a previous experiment (Ades, 1974b), he had found no cross-series adaptation of CV and VC continua (/bæ-dæ/, /æb-æd/): the phonetic detector, unlike phonetic listeners and phonological theory, evidently distinguishes between initial and final allophones. A funnel into a second level of phonetic analysis, possibly the point of contact with an abstract generative system, would be needed to account for the listener's inability to make this distinction.

For Bailey (1973), the phonetic conclusion was less compelling. He pointed to spectral overlap in the transitions of his two series, and suggested an acoustic system involving "...some generalizing balanced detectors of positive and negative transitions" (p. 31) (cf. Cooper, 1974a). To test for the effect of spectral overlap, he constructed two /ba-da/ series, one with a fixed F2 and all place cues in F3, the other with no F3 and all place cues in F2. This, by far the most stringent version of the phonetically identical-acoustically distinct design, yielded cross-adaptation from the F2 cues series to the fixed F2, but none from the F3 cues series to no F3. This argues strongly for auditory adaptation, and Bailey concluded that the system contains "...central feature extractors which process the phonetically relevant descriptors of spectral patterns" (p. 34).

Clearly, the issue of auditory versus phonetic detectors is not resolved. But let us consider implications of each possible resolution for speech perception theory and research. First, if discrete auditory detectors are being isolated by the adaptation technique, we may be in a position to begin more precise definition of the acoustic correlates of distinctive feature systems, ultimately essential if phonological theory is to be given a physical and physiological base. To the extent that this proved possible, we could be isolating invariants in the speech signal, thus aligning speech perception with that of other "natural categories," such as those of color and form (Rosch, 1973). But it is not inevitable that acoustic features be invariant correlates of phonetic features: both the work of Ades (1974b) on initial and final stop consonants and the work of Cooper (1974b) on vowel-specific VOT analyzers suggest that invariance may lie at some remove from the signal. And, in either event, to isolate acoustic features is not to define them phonetically, nor to explain how they are gathered from syllables of the signal into phonemes, each with its peculiar, nonarbitrary name: the auditory to phonetic transformation would remain obscure.

If, on the other hand, the adaptation technique isolates discrete phonetic detectors, its unequivocal achievement will have been to undergird the psychological and physiological reality of features in speech perception. Salutary though this may be for those of little faith, the outcome would be disappointing for research. For again, the process by which these features are drawn from the acoustic display and granted phonetic dimension will be hidden. To analysis of the analyzer a new technique must then be brought.

Finally, we should not discount the possibility that the auditory-phonetic distinction is misleading in this context, and that the adapted systems are both auditory and phonetic. Indeed, recent work (Cooper, in press-a, in press-b) suggests that each system can be adapted selectively, yet is intimately related to the other. The closeness of the relation is revealed by Cooper's (1974c) extension of the adaptation technique to the study of relations between perceptual and motor aspects of speech. He has shown that adaptation on a [bi-pi] continuum yields not only shifts in the perceptual boundary, but correlated shifts in subjects' characteristic VOT values in production. If his findings are replicable, we have here clear evidence for the frequently hypothesized link between perception and production, and one that may supersede the auditory-phonetic distinctions we have been attempting to establish for these adaptation studies. To the origin of this link in the processes of language acquisition we turn in the final section.

From Acoustic Feature to Phonetic Percept

As we have seen, template-matching models of speech perception are not in good standing. Faced with gross acoustic variations as a function of phonetic context, rate, stress, and individual speaker, theorists have had recourse to motor, or analysis-by-synthesis, accounts of speech perception: they have sought invariance in the articulatory control system. Nonetheless, there are grounds for believing that some form of template-matching may operate in both speaking and listening, and there are more fundamental grounds than lack of acoustic invariance for positing a link between production and perception.

Consider the infant learning to speak. Several writers (e.g., Stevens, 1973; Mattingly, 1973) have pointed out that the infant must be equipped with some mechanism by which it plucks from the stream of speech just those acoustic cues that convey the phonetic distinctions it will eventually learn to perceive and articulate. This fact motivates, in part, Stevens' (1973) pursuit of acoustic invariants and his hypothesized property detectors. Evidence for the existence of such detectors comes from the work of Eimas and his colleagues (Eimas, Siqueland, Jusczyk, and Vigorito, 1971; Eimas, 1974; for a review, see Cutting and Eimas, in press). They have investigated the capacity of infants as young as one month to discriminate synthetic speech sounds. We will not describe their method in detail, but broadly, it employs operant conditioning, a synthetic speech continuum, an adapting stimulus, and a test item. The results are reliable and striking: infants discriminate between pairs of stimuli drawn from different adult phonetic categories, but not between pairs drawn from the same phonetic category. The effect has been repeatedly demonstrated on both voicing and place of articulation continua (cf. Moffitt, 1971; Morse, 1972). Furthermore, the effect is absent for truncated control syllables, not heard by adults as speech, exactly as in the adult adaptation studies. Eimas and his colleagues interpret the effect as evidence for the operation of phonetic feature detectors, presumably innate. Unfortunately, the outcome is ambiguous for the same reasons as is the adult outcome: there is no way of assuring that the adapted detectors are phonetic rather than auditory (see Cutting and Eimas, 1974, for further discussion of this point). The more cautious, and perhaps more plausible, view is that they are auditory (cf. Stevens and Klatt, 1974:657-658).

We are then faced with two questions. First, do the acoustic features extracted by such detector systems bear an invariant relation to phonetic features? This is an empirical question and we will say no more here than that given the inconstancy of the speech signal, it is unlikely that they do. Second, and more importantly, how does the infant "know" that the extracted properties are speech? This, of course, is simply another version of the question: how are we to define the phonetic percept? But, asked in this form, an answer immediately suggests itself: the infant learns that sounds are speech by discovering that it can make them with its own vocal apparatus.

Before elaborating this point, let us consider the work of Marler (1970, in press). He has proposed a general model of the evolution of vocal learning, based on studies of the ontogenesis of male "song" in certain sparrows (see also Marler and Mundinger, 1971). Briefly, the hypothesis is that development of motor song-pattern is guided by sensory feedback matched to modifiable, innate auditory templates (cf. Mattingly, 1972). Marler describes three classes of birds. The first (for example, the dove or the chicken) needs to hear neither an external model nor its own voice for song to emerge: crowing and cooing develop normally, if the birds are reared in isolation and even if they are deafened shortly after birth. The second (for example, the song sparrow) needs no external model, but does need to hear its own voice: if reared in isolation, song develops normally, unless the bird is deafened in early life, in which case song is highly abnormal and insect-like.

An example of the third class of bird is the white-crowned sparrow, which needs both an external model and the sound of its own voice. Reared in isolation, the white-crown develops an abnormal song with "...certain natural characteristics, particularly the sustained pure tones which are one basic element in the natural song" (Marler and Mundinger, 1971:429). If the bird is deafened in early life, even this rudimentary song does not develop. There emerges instead a highly abnormal song "...rather like that of a deafened song sparrow...perhaps the basic output of the syringeal apparatus with a passive flow of air through it" (Marler, in press). However, reared in isolation, but exposed to recordings of normal male song during a critical period (10-50 days after birth), the male (and the female, if injected with male hormone) develops normal song some 50 or more days after exposure. Exposure to the songs of other species will not serve, and deafening either before or after exposure to conspecific song prevents normal development [Konishi (1965), cited by Marler, in press].

Marler (in press) proposes that the rudimentary song of the undeafened, isolated white-crown reflects the existence of an auditory template, "...lying in the auditory pathway, embodying information about the structure of vocal sounds." The template matches certain features of normal song, and serves to guide development of the rudimentary song, as well as to "...focus...attention on an appropriate class of external models." Exposure to these models modifies and enriches the template, which then serves to guide normal development, through subsong and plastic song, as the bird gradually discovers the motor controls needed to match its output with the modified template. [Several studies have reported evidence for the "tuning" by experience of visual detecting systems in cat (Hirsch and Spinelli, 1970; Blakemore and Cooper, 1970; Pettigrew and Freeman, 1973) and man (Annis and Frost, in press), and of auditory detecting systems in rhesus monkey (Miller, Sutton, Pfingst, Ryan, and Beaton, 1972).]

Marler (in press) draws the analogy with language learning. He suggests that sensory control of ontogenetic motor development may have been the evolutionary change that made possible an elaborate communicative system as pivot of avian and human social organization. He argues that "new sensory mechanisms for processing speech sounds, applied first, in infancy, to analyzing sounds of others, and somewhat later in life to analysis of the child's own sounds, was a significant step toward achieving the strategy of speech development of *Homo sapiens*." On the motor side, he points out, vocal development must have become dependent on auditory feedback, and there must have developed "neural circuitry necessary to modify patterns of motor outflow so that sounds generated can be matched to preestablished auditory templates."

Certainly, human and avian parallels are striking. Deafened at birth, the human infant does not learn to speak: babbling begins normally, but dies away around the sixth month (Marvilya, 1972). Whether this is because the infant has been deprived of the sound of its own voice, of an external model,or of both, we do not know. But there does seem to be an (ill-defined) critical period during which exposure to speech is a necessary condition of normal development (Lenneberg, 1967; but see Fromkin, Krashen, Curtiss, Rigler, and Rigler, 1974). And the work of Eimas and his colleague has demonstrated the sensitivity of the infant to functionally important acoustic features of the speech signal. At least one of these features, the short VOT lag associated with stops in many languages (Lisker and Abramson, 1964), is known to be among the first to appear in infant babble (Kewley-Port and Preston, 1974). Finally, Sussman (1971) and his colleagues (Sussman, MacNeilage, and Lumbley, 1974; Sussman and MacNeilage, in press) have reported evidence for a speech-related auditory sensorimotor mechanism that may serve to modify patterns of motor outflow, so as to match sounds generated by the vocal mechanism against some standard. In short, Marler's account is consistent with a good deal of our limited knowledge of speech development. Its virtue is to emphasize sensorimotor interaction and to accord the infant a mechanism for discovering auditory-articulatory correspondences.

Paradoxically, if we are to draw on this account of motor development for insight into perceptual development, we must place more emphasis on the relatively rich articulatory patterns revealed in early infant babble. The infant is not born without articulatory potential. In fact, the work of Lieberman and his colleagues would suggest quite specific capacities (Lieberman, 1968, 1972, 1973; Lieberman and Crelin, 1971; Liebe*r*man, Harris, Wolff, and Russell, 1971; Lieberman, Crelin, and Klatt, 1972). They have developed systematic evidence for evolution of the human vocal tract from a form with a relatively high larynx, opening almost directly into the oral cavity, capable of producing a limited set of schwa-like vowel sounds, to a form with a lowered larynx, a large pharyngeal cavity, and a right-angle bend in the supralaryngeal vocal tract, capable of producing the full array of human vowels. Lieberman (1973) argues that this development, taken with many other factors, including the capacity to encode and decode syllables, paved the way for development of language. Associated with changes in morphology must have come neurological changes to permit increasingly fine motor control of breathing and articulation, including in all likelihood, cerebral lateralization (cf. Lenneberg, 1967; Geschwind and Levitsky, 1968; Nottebohm, 1971, 1972). The outcome of these developments would have been a range of articulatory possibilities as determinate in their form as the patterns of manual praxis that gave rise to toolmaking. The inchoate forms of these patterns might then emerge in infant babble under the control of rudimentary articulatory templates.

In short, we hypothesize that the infant is born with both auditory and articulatory templates. Each embodies capacities that may be modified by, and deployed in, the particular language to which the infant is exposed. Presumably, these templates evolved more or less _pari passu_ and are matched, in some sense, as key to lock. But they differ in their degree of specificity. For effective function in language acquisition the auditory template must be tuned to specific acoustic properties of speech. The articulatory template, on the other hand, is more abstract, a range of gestural control, potentially isomorphic with the segmented feature matrix of the language by which it is modified (cf. Chomsky and Halle, 1968:294).

Among the grounds for this statement are the results of several studies of adult speech production. Lindblom and Sundberg (1971), for example, found that, if subjects were thwarted in their habitual articulatory gestures by the presence of a bite block between their front teeth, they were nonetheless able to approximate normal vowel quality, even within the first pitch period of the utterance. Bell-Berti (1975) has shown that the pattern of electromyographic potentials associated with pharyngeal enlargement during medial voiced stop consonant closure varies from individual to individual and from time to time within an individual. Finally, Ladefoged, DeClerk, Lindau, and Papçun (1972) have demonstrated that different speakers of the same dialect may use different patterns of tongue height and tongue root advancement to achieve phonetically identical vowels. They do not report formant frequencies for their six speakers, so that the degree of acoustic variability associated with the varied vocal-tract shapes is not known. But since individuals obviously differ in the precise dimensions of their vocal tracts, it would be surprising if they accomplished a particular gesture and a particular acoustic pattern by precisely the same pattern of muscular action. In short, it seems likely that both infant and adult articulatory templates are control systems for a range of functionally equivalent vocal tract shapes rather than for specific patterns of muscular action. In fact, it is precisely to exploration of its own vocal tract and to discovery of its own patterns of muscular action that the infant's motor learning must be directed.

We should emphasize that neither template can fulfull its communicative function in the absence of the other. Modified and enriched by experience, the auditory template may provide a "description" of the acoustic properties of the signal, but the description can be no different in principle than that provided by any other form of spectral analysis: alone, the output of auditory analysis is void. Similarly, babble without auditory feedback has no meaning. The infant discovers phonetic "meaning" (and linguistic function) by discovering auditory-articulatory correspondences, that is, by discovering the commands required by its own vocal tract to match the output of its auditory template. Since the articulatory template is relatively abstract, the infant will begin to discover these correspondences before it has acquired the detailed motor skills of articulation: perceptual skill will precede motor skill. In rare instances of peripheral articulatory pathology the infant (like the female white-crowned sparrow who learns the song without singing) may even discover language without speaking (cf. MacNeilage, Rootes, and Chase, 1967).

We hypothesize then that the infant is born with two distinct capacities, and that its task is to establish their links. Auditory feedback from its own vocalizations serves to modify the articulatory template, to guide motor development, and to establish the links. The process endows the communicatively empty outputs of auditory analysis and articulatory gesture with communicative significance. In due course the system serves to segment the acoustic signal and perhaps, as analysis-by-synthesis models propose, to resolve acoustic variability. But its prior and more fundamental function is to establish the "natural categories" of speech. To perceive these categories is to trace the sound patterns of speech to their articulatory source and recover the commands from which they arose. The phonetic percept is then the correlate of these commands.

REFERENCES

Abbs, J. H. and H. M. Sussman. (1971) Neurophysiological feature detectors and speech perception: A discussion of theoretical implications. J. Speech Hearing Res. 14, 23-36.

Abramson, A. S. and L. Lisker. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. In Proceedings of the 5th International Congress of Acoustics, ed. by D. E. Commins. (Liege: Imp. G. Thone) A-51.

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

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

Ades, A. E. (1973) Some effects of adaptation on speech perception. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 111, 121-129.

Ades, A. E. (1974a) Bilateral component in speech perception? J. Acoust. Soc. Amer. 56, 610-616.

Ades, A. E. (1974b) How phonetic is selective adaptation? Experiments on syllable position and vowel environment. Percept. Psychophys. 16, 61-66.

Annis, R. C. and B. Frost. (in press) Human visual ecology and orientation anisotropies in acuity. Science.

Atkinson, J. E. (1973) Aspects of intonation in speech: Implications from an experimental study of fundamental frequency. Unpublished Doctoral dissertation, University of Connecticut.

Bailey, P. (1973) Perceptual adaptation for acoustical features in speech. Speech Perception (Department of Psychology, The Queen's University of Belfast) Series 2, 2, 29-34.

Barclay, R. (1972) Noncategorical perception of a voiced stop. Percept. Psychophys. 11, 269-274.

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

Berlin, C. I., S. S. Lowe-Bell, J. K. Cullen, C. L. Thompson, and C. F. Loovis. (1973) Dichotic speech perception: An interpretation of right-ear advantage and temporal offset effects. J. Acoust. Soc. Amer. 53, 699-709.

Bever, T. G. (1970) The influence of speech performance on linguistic structure. In Advances in Psycholinguistics, ed. by G. B. Flores D'Arcais and W. J. M. Levelt. (Amsterdam: North-Holland) 4-30.

Blakemore, C. and G. F. Cooper. (1970) Development of the brain depends on visual environment. Science 168, 477-478.

Blumstein, S. (1974) The use and theoretical implications of the dichotic technique for investigating distinctive features. Brain Lang. 4, 337-350.

Boomer, D. S. and J. D. M. Laver. (1968) Slips of the tongue. Brit. J. Dis. Communic. 3, 1-12.

Cairns, H. S., C. E. Cairns, and F. Williams. (1974) Some theoretical considerations of articulation substitution phenomena. Lang. Speech 17, 160-173.

Capranica, R. R. (1965) The Evoked Vocal Response of the Bullfrog. (Cambridge, Mass.: MIT Press).

Chomsky, N. (1972) Language and Mind, enlarged ed. (New York: Harcourt Brace Jovanovich).

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

Chomsky, N. and G. A. Miller. (1963) Introduction to the formal analysis of natural languages. In Handbook of Mathematical Psychology, ed. by R. D. Luce, R. R. Bush, and E. Galanter. (New York: Wiley) 269-321.

Clegg, J. M. (1971) Verbal transformations on repeated listening to some English consonants. Brit. J. Psychol. 62, 303-309.

Cole, R. A. (1973a) Listening for mispronunciations: A measure of what we hear during speech. Percept. Psychophys. 13, 153-156.

Cole, R. A. (1973b) Different memory functions for consonants and vowels. Cog. Psychol. 4, 39-54.

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

Cooper, F. S. (1972) How is language conveyed by speech? In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

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

Cooper, W. E. (1974b) Contingent feature analysis in speech perception. Percept. Psychophys. 16, 201-204.

Cooper, W. E. (1974c) Perceptuo-motor adaptation to a speech feature. Percept. Psychophys. 16, 229-234.

Cooper, W. E. (in press-a) Selective adaptation for acoustic cues of voicing in initial stops. J. Phonetics.

Cooper, W. E. (in press-b) Selective adaptation to speech. In Cognitive Theory, ed. by F. Restle, R. M. Shiffrin, J. N. Castellan, H. Lindman, and D. B. Pisoni. (Potomac, Md.: Erlbaum).

Cooper, W. E. and S. E. Blumstein. (1974) A "labial" feature analyzer in speech perception. Percept. Psychophys. 15, 591-600.

Crowder, R. G. (1971a) The sound of vowels and consonants in immediate memory. J. Verbal Learn. Verbal Behav. 10, 587-659.

Crowder, R. G. (1971b) Waiting for the stimulus suffix: Decay, delay, rhythm, and readout in immediate memory. Quart. J. Exp. Psychol. 23, 324-340.

Crowder, R. G. (1972) Visual and auditory memory. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Crowder, R. G. (1973) Precategorical acoustic storage for vowels of short and long duration. Percept. Psychophys. 13, 502-506.

Crowder, R. G. and J. Morton. (1969) Precategorical acoustic storage (PAS). Percept. Psychophys. 5, 365-373.

Cutting, J. E. (1973) Levels of processing in phonological fusion. Unpublished Doctoral dissertation, Yale University.

Cutting, J. E. (in press-a) Aspects of phonological fusion. Human Perception and Performance.

Cutting, J. E. (in press-b) Two left-hemisphere mechanisms in speech perception. Percept. Psychophys.

Cutting, J. E. and P. D. Eimas. (in press) Phonetic feature analyzers in the processing of speech by infants. In The Role of Speech in Language, ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press).

Cutting, J. E. and B. S. Rosner. (in press) Categories and boundaries in speech and music. Percept. Psychophys.

Darwin, C. J. (1969) Auditory perception and cerebral dominance. Unpublished Doctoral dissertation, University of Cambridge.

Darwin, C. J. (1971a) Dichotic backward masking of complex sounds. Quart. J. Exp. Psychol. 23, 386-392.

Darwin, C. J. (1971b) Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psychol. 23, 46-62.

Darwin, C. J. and A. D. Baddeley. (1974) Acoustic memory and the perception of speech. Cog. Psychol. 6, 41-60.

Day, R. S. (1968) Fusion in dichotic listening. Unpublished Doctoral dissertation, Stanford University.

Day, R. S. (1970a) Temporal order judgments in speech: Are individuals language-bound or stimulus-bound? Haskins Laboratories Status Report on Speech Research SR-21/22, 71-75.

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

Day, R. S. and J. M. Vigorito. (1973) A parallel between encodedness and the ear advantage: Evidence from a temporal-order judgment task. J. Acoust. Soc. Amer. 53, 358(A).

Day, R. S. and C. C. Wood. (1972) Mutual interference between two linguistic dimensions of the same stimuli. Paper presented at the 83rd meeting of the Acoustical Society of America, April 18-21, Buffalo, N. Y.

Delattre, P. C., A. M. Liberman, F. S. Cooper, and L. J. Gerstman. (1952) An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesized from spectrographic patterns. Word 8, 195-210.

Dorman, M. (1974) Discrimination of intensity differences on formant transitions in and out of syllable context. Percept. .'sychophys. 16, 84-86.

Dorman, M., D. Kewley-Port, S. Brady-Wood, and M. T. Turvey. (1973) Forward and backward masking of brief vowels. Haskins Laboratories Status Report on Speech Research SR-33, 93-100.

Dorman, M., D. Kewley-Port, S. Brady, and M. T. Turvey. (1974) Two processes in vowel perception: Inferences from studies of backward masking. Haskins Laboratories Status Report on Speech Research SR-37/38, 233-253.

Eimas, P. D. (1974) 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. M. Vigorito. (1971) Speech perception in infants. Science 171, 303-306.

Evans, E. F. and I. C. Whitfield. (1964) Classification of unit responses in the auditory cortex of the unanaesthetized and unrestrained cat. J. Physiol. 17, 476-493.

Fant, C. G. M. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).

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

Fant, C. G. M. (1966) A note on vocal tract size factors and nonuniform F-pattern scalings. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR-4.

Fant, C. G. M. (1968) Analysis and systhesis of speech processes. In Manual of Phonetics, ed. by B. Malmberg. (Amsterdam: North-Holland).

Fischer-Jørgensen, E. (1972) Perceptual studies of Danish stop consonants. Annual Report (Institute of Phonetics, University of Copenhagen) 6, 75-176.

Flanagan, J. L. (1972) Speech Analysis, Synthesis, and Perception, 2nd ed. (New York: Academic Press).

Fodor, J. A., T. G. Bever, and M. F. Garrett. (1974) The Psychology of Language. (New York: Graw Hill).

Foss, D. J. and D. A. Swinney. (1973) On the psychological reality of the phoneme: Perception, identification, and consciousness. J. Verbal Learn. Verbal Behav. 12, 246-257.

Fourcin, A. J. (1968) Speech source inference. IEEE Trans. Audio Electroacoust. AU-16, 65-67.

Fourcin, A. J. (1972) Perceptual mechanisms at the first level of speech processing. In Proceedings of the 7th International Congress of Phonetic Sciences. (The Hague: Mouton).

Frishkopf, L. and M. Goldstein. (1963) Responses to acoustic stimuli in the eighth nerve of the bullfrog. J. Acoust. Soc. Amer. 35, 1219-1228.

Fromkin, V. A. (1971) The nonanomalous nature of anomalous utterances. Language 47, 27-52.

Fromkin, V. A., S. Krashen, S. Curtiss, D. Rigler, and M. Rigler. (1974) The development of language in Genie: A case of language acquisition beyond the "Critical Period." Brain Lang. 1, 81-107.

Fujimura, O. and K. Ochiai. (1963) Vowel identification and phonetic contexts. J. Acoust. Soc. Amer. 35, 1889(A).

Fujisaki, H. and T. Kawashima. (1969) On the modes and mechanisms of speech perception. Annual Report of the Engineering Research Institute (University of Tokyo) 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 (University of Tokyo) 29, 207-214.

Fujisaki, H. and N. Nakamura. (1969) Normalization and recognition of vowels. Annual Report (Division of Electrical Engineering, Engineering Research Institute, University of Tokyo) 1.

Gerstman, L. J. (1957) Perceptual dimensions for the friction portion of certain speech sounds. Unpublished Doctoral dissertation, New York University.

Gerstman, L. J. (1968) Classification of self-normalized vowels. IEEE Trans. Audio Electroacoust. AU-16, 78-80.

Geschwind, N. and W. Levitsky. (1968) Human brain: Left-right asymmetries in temporal speech region. Science 161, 186-187.

Glanzman, D. L. and D. B. Pisoni. (1973) Decision processes in speech discrimination as revealed by confidence ratings. Paper presented at the 85th meeting of the Acoustical Society of America, April 10-13, Boston, Mass.

Godfrey, J. J. (1974) Perceptual difficulty and the right-ear advantage for vowels. Brain Lang. 4, 323-336.

Goldstein, L. M. and J. R. Lackner. (in press) Alterations of the phonetic coding of speech sounds during repetition. Cognition.

Greenberg, J. J. and J. J. Jenkins. (1964) Studies in the psychological correlates of the sound system of American English. Word 20, 157-177.

Haber, R. N. (1969) *Information-Processing Approaches to Visual Perception*. (New York: Holt, Rinehart, and Winston).

Hadding-Koch, K. and M. Studdert-Kennedy. (1964) An experimental study of some intonation contours. Phonetica 11, 175-185.

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

Haggard, M. P., S. Ambler, and M. Callow. (1970) Pitch as a voicing cue. J. Acoust. Soc. Amer. 47, 613-617.

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

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

Halwes, T. and J. J. Jenkins. (1971) Problem of serial order in behavior is not resolved by context-sensitive associative memory models. Psychol. Rev. 78, 122-129.

Hanson, G. (1967) Dimensions in speech sound perception: An experimental study of vowel perception. Ericsson Tech. 23, 3-175.

Harris, K. S. (1958) Cues for the discrimination of American English fricatives in spoken syllables. Lang. Speech 1, 1-17.

Harris, K. S., H. S. Hoffman, A. M. Liberman, P. C. Delattre, and F. S. Cooper. (1958) Effect of third-formant transitions on the perception of voiced stop consonants. J. Acoust. Soc. Amer. 30, 122-126.

Hirsch, H. V. B. and D. N. Spinelli. (1970) Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cat. Science 168, 869-871.

Hoffman, H. S. (1958) Study of some cues in the perception of the voiced stop consonants. J. Acoust. Soc. Amer. 30, 1035-1041.

House, A. S., K. N. Stevens, T. T. Sandel, and J. B. Arnold. (1962) On the learning of speech-like vocabularies. J. Verbal Learn. Verbal Behav. 1, 133-143.

Hubel, D. H. and T. N. Wiesel. (1962) Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. J. Physiol. 60, 106-154.

Jakobson, R., C. G. M. Fant, and M. Halle. (1963) Preliminaries to Speech Analysis. (Cambridge, Mass.: MIT Press).

Jakobson, R. and M. Halle. (1956) Fundamentals of Language. (The Hague: Mouton).

Jones, D. (1948) Differences between Spoken and Written Language. (London: Assn. Phonétique Internationale).

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

Julesz, B. (1971) Foundations of Cyclopean Perception. (Chicago: University of Chicago Press).

Kewley-Port, D. and M. S. Preston. (1974) Early apical stop production: A voice onset time analysis. J. Phonetics 3, 195-210.

Kimura, D. (1961a) Some effects of temporal lobe damage on auditory perception. Canad. J. Psychol. 15, 156-165.

Kimura, D. (1961b) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

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

Kimura, D. and M. Durnford. (1974) Normal studies on the function of the right hemisphere in vision. In Hemisphere Function in the Human Brain, ed. by S. J. Dimond and J. G. Beamont. [London: Paul Elek (Scientific Books)].

Kirman, J. H. (1973) Tactile communication of speech: A review and an analysis. Psychol. Bull. 80, 54-74.

Kirstein, E. (1971) Temporal factors in perception of dichotically presented stop consonants and vowels. Unpublished Doctoral dissertation, University of Connecticut.

Kirstein, E. (1973) The lag effect in dichotic speech perception. Haskins Laboratories Status Report on Speech Research SR-35/36, 81-106.

Klatt, D. H. and S. R. Shattuck. (1973) Perception of brief stimuli that resemble formant transitions. Paper presented at the 86th meeting of the Acoustical Society of America, October 30 - November 2, Los Angeles, Calif.

Köhler, W. and H. Wallach. (1944) Figural after-effects: An investigation of visual processes. Proc. Amer. Phil. Soc. 88, 269-357.

Konishi, M. (1965) The role of auditory feedback in the control of vocalization in the white-crowned sparrow. Z. f. Tierpsychol. 22, 770-783.

Kozhevnikov, V. A. and L. A. Chistovich. (1965) Rech' Artikuliatsia i vospriiatie. (Moscow-Leningrad). Transl. as Speech: Articulation and Perception. (Washington, D.C.: Clearinghouse for Federal Scientific and Technical Information) JPRS 30, 543.

Krashen, S. (1972) Language and the left hemisphere. Working Papers in Phonetics (University of California at Los Angeles, Phonetics Laboratory) 24.

Lackner, J. R. and L. M. Goldstein. (in press) The psychological representation of speech sounds. Cognition.

Ladefoged, P. (1967) Three Areas of Experimental Phonetics. (New York: Oxford University Press).

Ladefoged, P. (1971a) Preliminaries to Linguistic Phonetics. (Chicago: University of Chicago Press).

Ladefoged, P. (1971b) Phonological features and their phonetic correlates. Working Papers in Phonetics (University of California at Los Angeles) 21, 3-12.

Ladefoged, P. and D. E. Broadbent. (1957) Information conveyed by vowels. J. Acoust. Soc. Amer. 29, 98-104.

Ladefoged, P., J. DeClerk, M. Lindau, and G. Papçun. (1972) An auditory-motor theory of speech production. Working Papers in Phonetics (University of California at Los Angeles) 22, 48-75.

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

Lashley, K. S. (1951) The problem of serial order in behavior. In Cerebral Mechanisms in Behavior, ed. by L. A. Jeffress. (New York: Wiley) 112-136.

Lass, N. J., ed. (in press) Contemporary Issues in Experimental Phonetics. (Springfield, Ill.: C. C Thomas).

Lass, N. J. and R. M. Gasperini. (1973) The verbal transformation effect: A comparative study of the verbal transformations of phonetically trained and nonphonetically trained subjects. Brit. J. Psychol. 64, 183-192.

Lass, N. J. and S. S. Golden. (1971) The use of isolated vowels as auditory stimuli in eliciting the verbal transformation effect. Canad. J. Psychol. 25, 349-359.

Lass, N. J., L. K. West, and D. D. Taft. (1973) A non-verbal analogue to the verbal transformation effect. Canad. J. Psychol. 27, 272-279.

Lea, W. A. (1974) An algorithm for locating stressed syllables in continuous speech. J. Acoust. Soc. Amer. 55, 411(A).

Lenneberg, E. H. (1967) The Biological Foundations of Language. (New York: Wiley).

Lettvin, J. Y., H. R. Maturana, W. S. McCulloch, and W. H. Pitts. (1959) What the frog's eye tells the frog's brain. Proc. Inst. Rad. Engr. 47, 1940-1951.

Liberman, A. M. (1957) Some results of research on speech perception. J. Acoust. Soc. Amer. 29, 117-123.

Liberman, A. M. (1970) The grammars of speech and language. Cog. Psychol. 1, 301-323.

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., P. C. Delattre, and F. S. Cooper. (1952) The role of selected stimulus variables in the perception of the unvoiced stop consonants. Amer. J. Psychol. 65, 497-516.

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.

Liberman, A. M., P. C. Delattre, F. S. Cooper, and L. H. 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., K. S. Harris, J. Kinney, and H. Lane. (1961) The discrimination of relative onset time of the components of certain speech and non-speech patterns. J. Exp. Psychol. 61, 379-388.

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: Wiley) 307-334.

Licklider, J. C. R. and G. A. Miller. (1951) The perception of speech. In Handbook of Experimental Psychology, ed. by S. S. Stevens. (New York: Wiley) 1040-1074.

Lieberman, P. (1963) Some effects of semantic and grammatical context on the production and perception of speech. Lang. Speech 6, 172-179.

Lieberman, P. (1968) Primate vocalizations and human linguistic ability. J. Acoust. Soc. Amer. 44, 1574-1584.

Lieberman, P. (1970) Toward a unified phonetic theory. Ling. Inq. 1, 307-322.

Lieberman, P. (1972) The Speech of Primates. (The Hague: Mouton).

Lieberman, P. (1973) On the evolution of language: A unified view. Cognition 2, 59-94.

Lieberman, P. and S. Crelin. (1971) On the speech of Neanderthal man. Ling. Inq. 2, 203-222.

Lieberman, P., E. S. Crelin, and D. H. Klatt. (1972) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. Amer. Anthropol. *74*, 287-307.

Lieberman, P., K. S. Harris, P. Wolff, and L. H. Russell. (1971) Newborn infant cry and nonhuman primate vocalizations. J. Speech Hearing Res. *14*, 718-727.

Liljencrants, J. and B. Lindblom. (1972) Numerical simulation of vowel quality systems: The role of perceptual contrast. Language *48*, 839-862.

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. *35*, 1773-1781.

Lindblom, B. E. F. (1972) Phonetics and the description of language. In *Proceedings of the 7th International Congress of Phonetic Sciences*. (The Hague: Mouton) 63-97.

Lindblom, B. E. F. and M. Studdert-Kennedy. (1967) On the role of formant transitions in vowel recognition. J. Acoust. Soc. Amer. *42*, 830-843.

Lindblom, B. E. F. and J. Sundberg. (1971) Neurophysiological representation of speech sounds. Paper presented at the 15th World Congress of Logopedics and Phoniatrics, August 14-19, Buenos Aires, Argentina.

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.

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: Academia) 563-567.

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

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

Lotz, J., A. S. Abramson, L. H. Gerstman, F. Ingemann, and W. J. Nemser. (1960) The perception of English stops by speakers of English, Spanish, Hungarian, and Thai: A tape-cutting experiment. Lang. Speech. *3*, 71-77.

MacKay, D. G. (1970) Spoonerisms: The anatomy of errors in the serial order of speech. Neuropsychologia *8*, 323-350.

MacNeilage, P. F. (1970) Motor control of serial ordering of speech. Psychol. Rev. *77*, 182-196.

MacNeilage, P. F., T. P. Rootes, and R. A. Chase. (1967) Speech production and perception in a patient with severe impairment of somesthetic perception and motor control. J. Speech Hearing Res. *10*, 449-467.

Malmberg, B. (1955) The phonetic basis for syllable division. Studia Linguistica *9*, 80-87.

Marler, P. (1970) Bird song and speech development: Could there be parallels? Amer. Scient. *58*, 669-673.

Marler, P. (in press) On the origin of speech from animal sounds. In *The Role of Speech in Language*, ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press).

Marler, P. and P. Mundinger. (1971) Vocal learning in birds. In *Ontogeny of Vertebrate Behavior*, ed. by. H. Moltz. (New York: Academic Press) 380-450.

Marvilya, M. P. (1972) Spontaneous vocalization and babbling in hearing-impaired infants. In *Speech Communication Ability and Profound Deafness*, ed. by C. G. M. Fant. (Washington, D.C.: A. G. Bell Association for the Deaf).

Massaro, D. W. (1972) Preperceptual images, processing time, and perceptual units in auditory perception. Psychol. Rev. *79*, 124-145.

Mattingly, I. G. (1968) Synthesis by rule of General American English. Supplement to Haskins Laboratories Status Report on Speech Research, April.

Mattingly, I. G. (1971) Synthesis by rule as a tool for phonological research. Lang. Speech *14*, 47-56.

Mattingly, I. G. (1972) Speech cues and sign stimuli. Amer. Scient. 60, 327-337.

Mattingly, I. G. (1973) Phonetic prerequisites for first-language acquisition. Haskins Laboratories Status Report on Speech Research SR-34, 65-69.

Mattingly, I. G. (1974) Speech synthesis for phonetic and phonological models. In Current Trends in Linguistics, Vol. 12, ed. by T. A. Sebeok. (The Hague: Mouton).

Mattingly, I. G. (in press) The human aspects of speech. In The Role of Speech in Language, ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press).

Mattingly, I. G. and A. M. Liberman. (1969) The speech code and the physiology of language. In Information Processing in the Nervous System, ed. by K. N. Leibovic. (New York: Springer Verlag) 97-114.

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

McNeill, D. and L. Lindig. (1973) The perceptual reality of phonemes, syllables, words, and sentences. J. Verbal Learn. Verbal Behav. 12, 419-430.

McNeill, D. and B. Repp. (1973) Internal processes in speech perception. J. Acoust. Soc. Amer. 53, 1320-1326.

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, G. A., G. A. Heise, and W. Lichten. (1951) The intelligibility of speech as a function of the context of the test materials. J. Acoust. Soc. Amer. 41, 329-335.

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

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. Amer. 55, 390(A).

Miller, J. N., D. Sutton, B. Pfingst, A. Ryan, and R. Beaton. (1972) Single cell activity in the auditory cortex of rhesus monkeys: Behavioral dependency. Science 177, 449-451.

Mitchell, P. D. (1973) A test of differentiation of phonemic feature contrasts. Unpublished Doctoral dissertation, City University of New York.

Moffitt, A. R. (1971) Consonant cue perception by twenty- to twenty-four-week-old infants. Child Develop. 42, 717-731.

Molfese, D. L. (1972) Cerebral asymmetry in infants, children, and adults: Auditory evoked responses to speech and noise stimuli. Unpublished Doctoral dissertation, Pennsylvania State University.

Morse, P. A. (1972) The discrimination of speech and nonspeech stimuli in early infancy. J. Exp. Child Psychol. 14, 477-492.

Neisser, U. (1967) Cognitive Psychology. (New York: Appleton-Century-Crofts).

Nelson, P. G., S. D. Erulkar, and S. S. Bryan. (1966) Response units of the inferior colliculus to time-varying acoustic stimuli. J. Neurophysiol. 29, 834-860.

Nottebohm, F. (1971) Neural lateralization of vocal control in a passerine bird. I. Song. J. Exp. Zool. 177, 229-262.

Nottebohm, F. (1972) Neural lateralization of vocal control in a passerine bird. II. Subsong, calls, and theory of vocal learning. J. Exp. Zool. 179, 35-50.

Obusek, C. J. and R. M. Warren. (1973) A comparison of speech perception in senile and well-preserved aged by means of the verbal transformation effect. J. Gerontol. 28, 184-188.

Öhman, S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. J. Acoust. Soc. Amer. 39, 151-168.

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

Reed, S. K. (1973) Psychological Processes in Pattern Recognition. (New York: Academic Press).

Repp, B. H. (1973) Dichotic forward and backward masking of CV syllables. Unpublished Doctoral dissertation, University of Chicago.

Rosch, E. H. (1973) Natural categories. Cog. Psychol. 4, 328-350.

Sachs, R. M. (1969) Vowel identification and discrimination in isolation vs word context. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 93, 220-229.

Sales, B. D., R. A. Cole, and R. N. Haber. (1969) Mechanisms of aural encoding: V. Environmental effects of consonants on vowel encoding. Percept. Psychophys. 6, 361-365.

Savin, H. B. and T. B. Bever. (1970) The nonperceptual reality of the phoneme. J. Verbal Learn. Verbal Behav. 9, 295-302.

Schatz, C. (1954) The role of context in the perception of stops. Language 30, 47-56.

Scholes, R. J. (1968) Phonemic interference as a perceptual phenomenon. Lang. Speech 11, 86-103.

Shankweiler, D. P., W. Strange, and R. Verbrugge. (in press) Speech and the problem of perceptual constancy. In Perceiving, Acting, and Comprehending: Toward an Ecological Psychology, ed. by. R. Shaw and J. Bransford. (Potomac, Md.: Erlbaum Associates).

Shankweiler, D. P. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol. 19, 59-63.

Shearme, J. N. and J. N. Holmes. (1962) An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1 - formant 2 plane. In Proceedings of the 4th International Congress of Phonetic Sciences. (The Hague: Mouton) 234-240.

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) 67-113.

Singh, S. (1966) Cross-language study of perceptual confusions of plosive phonemes in two conditions of distortion. J. Acoust. Soc. Amer. 40, 635-656.

Singh, S. and D. Woods. (1970) Multidimensional scaling of 12 American English vowels. J. Acoust. Soc. Amer. 48, 104(A).

Sinnott, J. M. (1974) A comparison of speech sound discrimination in humans and monkeys. Unpublished Doctoral dissertation, University of Michigan.

Stetson, R. H. (1952) Motor Phonetics. (Amsterdam: North-Holland).

Stevens, K. N. (1960) Toward a model for speech recognition. J. Acoust. Soc. Amer. 32, 47-55.

Stevens, K. N. (1967) Acoustic correlates of certain consonantal features. Paper presented at Conference on Speech Communication and Processing, MIT, November 6-8, Cambridge, Mass.

Stevens, K. N. (1968a) Acoustic correlates of place of articulation for stop and fricative consonants. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 89, 199-205.

Stevens, K. N. (1968b) On the relations between speech movements and speech perception. Z. Phon., Sprachwiss. u. Komm. Fschg. 213, 102-106.

Stevens, K. N. (1972a) Segments, features, and analysis by synthesis. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press) 47-52.

Öhman, S. E. G. (1967) Numerical model of coarticulation. J. Acoust. Soc. Amer. 41, 310-320.

Papçun, G., S. Krashen, D. Terbeek, R. Remington, and R. Harshman. (1974) Is the left hemisphere specialized for speech, language, and/or something else? J. Acoust. Soc. Amer. 55, 319-327.

Parks, T., C. Wall, and J. Bastian. (1969) Intercategory and intracategory discrimination for one visual continuum. J. Exp. Psychol. 81, 241-245.

Perl, N. T. (1970) The application of the verbal transformation effect to the study of cerebral dominance. Neuropsychologia 8, 259-261.

Peterson, G. E. (1961) Parameters of vowel quality. J. Speech Hearing Res. 4, 10-29.

Peterson, G. E. and H. L. Barney. (1952) Control methods used in a study of vowels. J. Acoust. Soc. Amer. 25, 175-184.

Pettigrew, J. D. and R. D. Freeman. (1973) Visual experience without lines: Effects on development of cortical neurons. Science 182, 599-600.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Unpublished Doctoral dissertation, University of Michigan.

Pisoni, D. B. (1972) Perceptual processing time for consonants and vowels. Haskins Laboratories Status Report on Speech Research SR-31/32, 83-92.

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

Pisoni, D. B. (1973b) The role of auditory short-term memory in vowel perception. Haskins Laboratories Status Report on Speech Research SR-34, 89-118.

Pisoni, D. B. (in press) Dichotic listening and the processing of phonetic features. In Cognitive Theory, Vol. 1, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindam, and D. B. Pisoni. (Potomac, Md.: Erlbaum Associates).

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

Pisoni, D. B. and S. D. McNabb. (1974) Dichotic interactions of speech sounds and phonetic feature processing. Brain Lang. 4, 351-362.

Pisoni, D. B. and J. R. Sawusch. (in press) Category boundaries for speech and nonspeech sounds. Percept. Psychophys.

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

Pollack, I. and J. M. Pickett. (1963) The intelligibility of excerpts from conversation. Lang. Speech 6, 165-172.

Popper, R. D. (1972) Pair discrimination for a continuum of synthetic voiced stops with and without first and third formants. J. Psycholing. Res. 1, 205-219.

Porter, R. J. (1971) The effect of delayed channel on the perception of dichotically presented speech and nonspeech sounds. Unpublished Doctoral dissertation, University of Connecticut.

Posner, M. I., S. J. Boies, W. H. Eichelman, and R. L. Taylor. (1969) Retention of visual and name codes of single letters. J. Exp. Psychol. Monogr. 79, 1-16.

Potter, R. K., G. A. Kopp, and H. C. Green. (1947) Visible Speech. (New York: van Nostrand).

Rand, T. C. (1971) Vocal tract size normalization in the perception of stop consonants. Haskins Laboratories Status Report on Speech Research SR-25/26, 141-146.

Stevens, K. N. (1972b) The quantal nature of speech: Evidence from articulatory-acoustic data. In Human Communication: A Unified View, ed. by E. E. David and P. B. Denes. (New York: McGraw Hill) 51-66.

Stevens, K. N. (1973) Potential role of property detectors in the perception of consonants. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 110, 155-168.

Stevens, K. N. and M. Halle. (1967) Remarks on analysis by synthesis and distinctive features. In Models for the Perception of Speech and Visual Form, ed. by W. Wathen-Dunn. (Cambridge, Mass.: MIT Press) 88-102.

Stevens, K. N. and A. S. House. ·(1972) The perception of speech. In Foundations of Modern Auditory Theory, Vol. 2, ed. by J. Tobias. (New York: Academic Press) 3-62.

Stevens, K. N., A. S. House, and A. P. Paul. (1966) Acoustical description of syllabic nuclei. J. Acoust. Soc. Amer. 40, 123-132.

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.

Stevens, K. N., A. M. Liberman, M. Studdert-Kennedy, and S. E. G. Öhman. (1969) Cross-language study of vowel perception. Lang. Speech 12, 1-23.

Strange, W., R. Verbrugge, and D. P. Shankweiler. (1974) Consonantal environment specifies vowel identity. Paper presented at the 87th meeting of the Acoustical Society of America, April 23-26, New York City.

Studdert-Kennedy, M. (1974) The perception of speech. In Current Trends in Linguistics, ed. by T. A. Sebeok. (The Hague: Mouton). [Also in Haskins Laboratories Status Report on Speech Research SR-23 (1970) 15-48.]

Studdert-Kennedy, M. (in press) From continuous signal to discrete message: Syllable to phoneme. In The Role of Speech in Language, ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press).

Studdert-Kennedy, M. and K. Hadding. (1973) Auditory and linguistic processes in the perception of intonation contours. Lang. Speech 16, 293-313.

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

Studdert-Kennedy, M. and D. P. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Studdert-Kennedy, M., D. P. Shankweiler, and D. B. Pisoni. (1972) Auditory and phonetic processes in speech perception: Evidence from a dichotic study. Cog. Psychol. 2, 455-466.

Studdert-Kennedy, M., D. P. Shankweiler, and S. Schulman. (1970b) Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables. J. Acoust. Soc. Amer. 48, 599-602.

Summerfield, A. and M. Haggard. (1972) Perception of stop voicing. Speech Perception (Department of Psychology, The Queen's University of Belfast) Series 2, 1, 1-14.

Summerfield, A. and M. Haggard. (1973) Vocal tract normalization as demonstrated by reaction times. Speech Perception (Department of Psychology, The Queen's University of Belfast) 2.

Sussman, H. (1971) The laterality effect in lingual-auditory tracking. J. Acoust. Soc. Amer. 49, 1874-1880.

Sussman, H. M. and P. F. MacNeilage. (in press) Studies of hemispheric specialization for speech production. Brain Lang.

Sussman, H. M., P. F. MacNeilage, and J. Lumbley. (1974) Sensorimotor dominance and the right-ear advantage in mandibular-auditory tracking. J. Acoust. Soc. Amer. 56, 214-216.

Treon, M. A. (1970) Fricative and plosive perception-identification as a function of phonetic context in CVCVC utterances. Lang. Speech 13, 54-64.

Turvey, M. (1973) On peripheral and central processes in vision. Psychol. Rev. 80, 1-52.

Verbrugge, R., W. Strange, and D. P. Shankweiler. (1974) What information enables a listener to map a talker's vowel space? Paper presented at the 87th meeting of the Acoustical Society of America, April 23-26, New York City.

Vitz, P. C. and B. S. Winkler. (1973) Predicting the judged similarity of sound of English words. J. Verbal Learn. Verbal Behav. 12, 373-388.

Warren, R. M. (1968) Verbal transformation effect and auditory perceptual mechanisms. Psychol. Bull. 70, 261-270.

Warren, R. M. (1970) Perceptual restoration of missing speech sounds. Science 167, 392-393.

Warren, R. M. (1971) Identification times for phonemic components of graded complexity and for spelling of speech. Percept. Psychophys. 9, 345-349.

Warren, R. M. and R. L. Gregory. (1958) An auditory analogue of the visual reversible figure. Amer. J. Psychol. 71, 612-613.

Warren, R. M. and C. J. Obusek. (1971) Speech perception and phonemic restorations. Percept. Psychophys. 9 (3B), 358-362.

Weiss, M. and A. S. House. (1973) Perception of dichotically presented vowels. J. Acoust. Soc. Amer. 53, 51-58.

Werner, H. (1935) Studies on contour: I. Qualitative analyses. Amer. J. Psychol. 47, 40-64.

Whitfield, I. C. and E. F. Evans. (1965) Responses of auditory cortical neurons to stimuli of changing frequency. J. Neurophysiol. 28, 655-672.

Wickelgren, W. A. (1965) Distinctive features and errors in short-term memory for English vowels. J. Acoust. Soc. Amer. 38, 583-588.

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

Wickelgren, W. A. (1969) Context-sensitive coding, associative memory, and serial order in (speech) behavior. Psychol. Rev. 76, 1-15

Wollberg, Z. and J. D. Newman. (1972) Auditory cortex of squirrel monkey: Response patterns of single cells to species-specific vocalizations. Science 175, 212-213.

Wood, C. C. (1974) Parallel processing of auditory and phonetic information in speech perception. Percept. Psychophys. 15, 501-508.

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

Wood, C. C. and Day, R. S. (in press) Failure of selective attention to phonetic segments in consonant-vowel syllables. Percept. Psychophys.

Wood, C. C., W. R. Goff, and R. S. Day. (1971) Auditory evoked potentials during speech perception. Science 173, 1248-1251.

Woodworth, R. S. and H. Schlosberg. (1954) Experimental Psychology. (New York: Holt, Rinehart, and Winston).

Zlatin, M. A. (1974) Development of the voicing contrast: A psychoacoustic study of voice onset time. J. Acoust. Soc. Amer. 56, 981-994.

Speech Recognition Through Spectrogram Matching*

Frances Ingemann[+] and Paul Mermelstein
Haskins Laboratories, New Haven, Conn.

In order to assess human analysis of acoustic data before attempting such analysis by machine, we conducted a series of experiments in which subjects were asked to match spectrograms of continuous speech to reference spectrograms of the same words. Although error rates varied with sentence difficulty and size of vocabulary, comparison of the matches shows greater agreement in phoneme segments than other experiments have obtained in phonetic transcriptions of unknown utterances without semantic or syntactic processing. Accuracy in matching can be further improved by feedback in the form of spectrographic representation of a sequence of tentative matches spoken as if they made up the unknown utterance. Automatic matching of word- or syllable-sized acoustic patterns may provide a more accurate phonemic input to the syntactic-semantic component of a speech recognition system than other methods so far attempted.

The limited performance of speech recognition systems to date indicates to us that improved acoustic analysis as well as good semantic-syntactic analysis are prerequisites to better system performance. Human analysis of acoustic data without the use of nonacoustic information can be expected to assist the design of improved acoustic analysis systems.

The difficulty that people have in accurately identifying the phonetic content of spectrograms of unknown utterances has long been recognized by researchers in the field (see, for example, Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1968; also Fant, 1962). Until recently little experimentation had been undertaken since the early pioneering work at Bell Laboratories (Potter, Kopp, and Green, 1947). Within the past few years, interest in spectrogram reading has been renewed, at least partially in response to attempts at automatic speech recognition in the expectation that cues available to human spectrogram readers could be programmed into an automatic speech recognition system.

*A shorter version of this paper will be published in the Journal of the Acoustical Society of America.

[+] Visiting from University of Kansas, Lawrence.

Acknowledgment: We wish to acknowledge the contribution of F. S. Cooper and P. W. Nye to the initiation of these experiments. We are also grateful to the following people who not only served as subjects but also provided useful insights into the spectrogram-matching process and suggestions for further experimentation: L. Barton, R. Collier, F. S. Cooper, J. Gaitenby, L. Lisker, R. McGuire, P. W. Nye, and G. Sholes.

[HASKINS LABORATORIES: Status Report on Speech Research SR-39/40 (1974)]

Studies by Klatt and Stevens (1973), Lindblom and Svensson (1973), and Svensson (1974) have shown that subjects who are experienced in examining spectrograms can label phonetic segments correctly less than half the time when they are presented with spectrograms about which they have no additional information. These experiments have also shown that the addition of syntactic, semantic, and prosodic information can improve the performance significantly.

Our interest lay in finding out whether speech recognition could be improved without recourse to nonacoustic information. The technique we proposed was the matching of spectrograms of unknown utterances with reference spectrograms identified only by number so that success of the task depended almost entirely on the ability to match patterns visually. Klatt and Stevens (1973) also used spectrographic matching,but because the reference words were known, syntactic and semantic considerations entered into the selection of suitable matches. Our experiments were undertaken to evaluate human spectrogram-matching performance before attempting spectrogram matching by machine.

EXPERIMENT I

The first experiment was in the nature of a limited pilot study to determine whether subjects could match spectrograms at all. In this experiment, as in all the experiments described in this paper, spectrograms were based on the speech of a single female speaker (one of the authors), who recorded the samples in a sound-treated room using a clear (but not exaggerated or over-precise) reading style. Wide-band spectrograms were produced on a Voiceprint spectrograph using a frequency scale of 0-4800 Hz.

Subjects were asked to locate within spectrograms of five test sentences ten words given in reference spectrograms (see Table 1). The reference words were content words consisting of one, two, or four syllables spoken in the context "Say _____ again." Each reference word occurred once in the sentences, except that two monosyllabic words (grow and month) occurred in a suffixed form as well as in the uninflected form given as the reference word. Subjects were not told the meaning of either the reference words or the test sentences, but they were given the meaning of the reference frame.

TABLE 1: Sentences used in Experiment I.

1. Little children often chew bubble gum.
2. My friend's grandfather used to grow tomatoes.
3. He has too much month left at the end of his money.
4. There is a growing interest in Victorian houses.
5. Emergency regulations will be in effect for six months.

Underlined words were to be matched to reference spectrograms containing the same words.

Three subjects were used: one who had extensive experience examining spectrograms, one who had moderate experience, and one who had nc experience. All three subjects performed the task with few errors (75-83 percent correct). The subjects found particularly disconcerting the fact that they were told that two reference words occurred twice but they didn't know which. Six of the eight errors in matching are related to this aspect of the task. Table 2 lists the identifications made by each subject.

TABLE 2: Tabulation of responses on Experiment I.

			Responses		
			S1	S2	S3
Reference Words	left		-,E	-	-
	grow	1	0	0	0
		2	-	-	-
	month	1	-	-	-
		2	-	0	E
	friend's		-	-	-
	bubble		-	E	E
	children		-	-	-
	interest		-	-	-
	regulations		-	-	-
	Victorian		-	-	-
	emergency		-	-	-

Key: - correctly located
E incorrectly located
O not located

EXPERIMENT II

Since Experiment I had shown that spectrograms could be matched, a second experiment was devised to include all words in a randomly selected text to determine whether the task could be done as successfully with a larger set of reference words, some of which were unstressed.

The following passage, four sentences long, was chosen at random from a publication:

> When adults name things and persons for children they incidentally transmit the texture or grain of their reality. They do this by choosing for some referents names that categorize very broadly and, for some referents, names that categorize very narrowly. That is what this paper is about. It does not exhaust its subject if we understand its subject to be the function of names in tuning one consciousness to another (Brown, 1970).

These test sentences contain a total of 70 words, of which 51 are different.

Reference spectrograms were made of the 51 words in the context, "Say _____ again," in which again was given major sentence stress to prevent both contrastive stress on the reference word and a possible phrase boundary juncture between the reference word and again. In addition, a second version of some

TABLE 3: Responses to Experiment II.

	Responses by Subjects					
Words in Sentences to be Matched	S1	S2	S4	S5	S6	S7
when	-		-	-		
adults	-		? + -	-		
name	-		grain	-		
things	-		-	-		
and	in		in	in		
persons	-		-	-		
for	0		the	the		
children	-		-	-		
they	be		things	-		
incidentally	-		-	-		
transmit	- + the		-	-		
the	to		0	for		
texture	-		-	-		
or	-		for	0		
grain	-		-	-		
of	-		-	-		
their	-		-	-		
reality	-		-	-		
they					be	be
do					-	-
this					-	-
by					-	-
choosing					-	-
for					-	-
some					-	-
referents					-	-
names					things	-
that					to	-
categorize					-	-
very					-	or
broadly					-	referents
and					it	-
for					-	-
some					-	-
referents					-	-
names					things	-
that					to	we
categorize					-	-
very					-	-
narrowly					-	-
that	-		it			
is	-		-			
what	-		-			
this	-		its			
paper	-		-			
is	some		{some			
about	-					
it		0				
does		this				
not		reality				
exhaust		-				
its		in				
subject		-				
if		is				
we		-				
understand		-				
its		the				
subject		-				
to		0				
be		{paper				
the						
function		-				
of		0				
names		-				
in		-				
tuning		-				
one		when				
consciousness		-				
to		do				
another		reality				

- indicates correct response
0 indicates no response

monosyllabic function words in an unstressed context was provided. The second context was not identified for the subjects, who were told only that the word was in unstressed position in the second version.

Six subjects, all of whom had experience examining spectrograms, took part in the experiment. Each subject was given only one or two test sentences so that the refcrence set contained approximately twice as many words as a subject would find in his sentence.

The overall score of correct identifications was 67 percent. Most errors were made on monosyllabic words, particularly function words. A list of the errors by subject is given in Table 3 and a summary of the results by word type is shown in Table 4. Since some of the subjects took part in more than one experiment, the subjects are identified by number for the set of experiments, rather than separately for each experiment.

TABLE 4: Responses to Experiment II by word type.

	Number Attempted	Number Correct	Percent Correct
Total Matches	142	95	67%
Polysyllabic Matches	53	47	89%
Monosyllabic Matches	89	48	54%
Monosyllabic Content Word Matches	15	11	73%
Monosyllabic Function Word Matches	74	37	50%

EXPERIMENT III

Because monosyllabic words seemed to be more difficult to match than polysyllabic words, a third experiment consisting only of monosyllabic words was designed to examine this area more carefully. The difficulty of the task was increased by adding to the reference words other words that were phonetically similar.

The following test sentence consisting of ten monosyllables was used:

Ed will ask Ned to pay the bill for him.

The reference set consisted of the ten words in the test sentence plus the following 30, making a total of 40 words:

do
that
ill
add
A

win
wool
lass
last
mill

met
men
neck
beer
bid

bathe
bet
dub
dwell
took

Kate
coop
fin
full
thumb

shore
hen
as
them
her

Once again, for some of the monosyllabic function words a second unstressed variant in a context not known to the subject was provided.

The experiment used four subjects, all of whom had experience examining spectrograms. Words were identified correctly only 48 percent of the time. The responses of subjects are given in Table 5. In contrast to the previous experiment, content words were not more easily matched than function words. Content words were matched correctly only 45 percent of the time, while function words were matched 50 percent of the time.

TABLE 5: Responses to Experiment 3.

	Responses by Subjects			
Words to be Matched	S1	S2	S4	S8
Ed	0 + the	-	hen	neck
will	0	last	ill	for
ask	lass		lass	last
Ned	bet	-	-	-
to	do	took	-	-
pay	hen	-	-	-
the	-	that	do	do
bill	-	ill	-	mill
for	-	-	-	-
him	will	-	-	-

Key:
- correct response
0 no response

This experiment also pointed up the difficulty of locating word boundaries when the reference set includes words that can be confused. For example, ask was identified twice as last and twice as lass because the l of the preceding word will was assumed to be part of this word; to was once identified as took when it preceded pay.

A comparison of the string of phonemes in the test sentence with the string of phonemes in the matched reference words shows that the percent of phonemes correctly matched is considerably higher than the percent of words. Of the phonemes in the sentence, 72 percent were found to be correctly matched and 35 percent errors were made. The total of these exceeds 100 percent because two phonemes in the reference words were sometimes matched to a single phoneme in the sentence. The comparison of phonemes is given in Table 6.

When considered from the point of view of word recognition relative to phoneme recognition, the results correspond rather closely to the relationship found by Fletcher (1929) between syllable recognition and 'letter' recognition in testing noisy speech transmission systems. Fletcher's curves would predict 77 percent 'letter' [phoneme] recognition to accompany 48 percent syllable recognition. The predicted sentence intelligibility for human listeners under these conditions is 94 percent. These facts suggest that an automatic speech recognition system with performance on acoustic analysis comparable to the visual performance of our human subjects and with performance on the syntactic-semantic level comparable to that of human listeners can be expected to

TABLE 6: Comparison by phonemes of matches in Experiment III.

		Matches by Subjects			
		S1	S2	S4	S8
Phonemes in the Test Sentences	ɛ	0 + ðə	-	h-	n-
	d	m/n	-	n	k
	w	r	0	0	f
	I/ə	ʌ	0	-	ɔr
	l	-	-	-	-
	æ	-	-	l-	-
	s	-	-	-	-
	k	0	t	0	t
	n	b	-	-	-
	ɛ	-	-	-	-
	d	t	-	-	-
	t	d	-	-	-
	u/ʊ	-	-	-	-
	p	h	k-	-	-
	e	ɛn	-	-	-
	ð	-	-	d	d
	ə	-	-	u	u
	b	-	t	-	m
	I	-	-	-	-
	l	-	-	-	-
	f	-	-	-	-
	ɔr	-	-	-	-
	h	-	w	-	-
	I	-	-	-	-
	m	-	l	-	-

Key:
- correct response
0 no response

"understand" 94 percent of simple questions or instructions such as given in Fletcher's Intelligibility List.

EXPERIMENT IV

The number of errors in the preceding experiments led us to try spectrographic feedback as a means of improving performance. Experiment IV began, as did the previous two experiments, by asking the subject to match words in a sentence to reference words. The sentence to be matched was a truncated version of a sentence chosen at random from a paper:

> The way a speaker produces a string of phones will show a good deal of variability.

Two subjects were used: one who had participated in all three of the previous experiments and one who had participated in none. One subject (S2) was given 120 reference words, including all the words in Experiments II and III and 34 new ones. The other subject (S9) was given a subset of these, 78 in all, which included all words in Experiment III, 4 additional words from Experiment II, and the 34 new words. These words are listed in Table 7.

After the subject had tentatively matched the test sentence, the reference words he selected were read as a sentence with stress and intonation as close as possible to the original utterance. Spectrograms of this sequence of tentative matches were given to the subject to compare with the original sentence. He was then allowed to revise his list of matches and once again he was given spectrograms of the sequence of matches. This process was repeated until the subject indicated that he no longer wished to continue. Both subjects stopped with their third attempt. The subject was then asked to give a conference rating for each of his matches. The results are given in Table 8.

The subjects differed greatly in their matching ability, although spectrographic feedback improved both of their performances. Whereas S9 on the third try correctly matched all the words, S2 attained only 38 percent correct. Furthermore, only 50 percent of the matches in which S2 expressed high confidence were in fact correct. However, the ratings did have some validity in that none of the low-confidence matches were correct.

There are a number of possible explanations for the difference between the two subjects' performances. S9 had slightly less than two-thirds of the reference words that S2 had. In addition to having more opportunities to make an error, S2 also had a greater problem in handling the data physically: S2 had to sort through 146 reference spectrograms--26 of the reference words being represented by two spectrograms, one in the standard frame and one in unstressed position.

Another difference was that between trials 2 and 3, S9 requested and received spectrograms of a second reading of the original sentence so that he could get some indication of the variation to be expected in a rereading of the same sentence. S2 did not receive these spectrograms of the second reading.

A third difference was the amount of time spent on the task. Although no accurate record was kept, S9 estimated that he had spent about twice the time

TABLE 7: Reference words used in Experiment IV.

A
a*
able
(about)
add
(adults)
(and*)
(another)
as*
ask
astray
away
bathe
(be*)
beer
bet
bid
bill
bring
(broadly)
(by*)
(categorize)
(children)
(choosing)
(consciousness)
coop
could
deal
do*
(does*)
dub
dwell
Ed
eel
(exhaust)
fall
fin
foam
for*
(function)
good
(grain)
hairy
he*
hen
her*
him*
(if*)
ill
(in*)
(incidentally)
is*
(it*)
(its*)
Kate
lass
last
lit
love
lucky
men
met
mill
(name)
(names)
(narrowly)
neck
Ned
(not)
of*
(one)
(or*)
owe
(paper)
pay
peak
(persons)
phones
prod
produce
produces
(reality)
(referents)
says
shore
show
shower
(some)
speaker
stream
string
(subject)
swell
tea
(texture)
that*
the*
their*
them*
(they*)
(things)
(this)
thumb
to*
(transmit)
took
(tuning)
(understand)
variability
very
wag
way
(we*)
welsh
(what)
(when)
will*
win
wool
wringer

* Words followed by an asterisk were given in both stressed and unstressed forms.

() Words in parentheses were not given to S9.

TABLE 8: Responses to Experiment IV.

	Responses by Subjects							
	S2				S9			
Words to be Matched	First Attempt	Second Attempt	Third Attempt	Confidence Rating	First Attempt	Second Attempt	Third Attempt	Confidence Rating
the way	away	away	away	high	away	away	-	high
a	0	as	as + 0	mid	0	-	-	high
speaker	-	-	-	high	-	-	-	high
produces	-	-	-	high	-	-	-	high
a	as	as	as	mid	is	is	-	high
string	stream	stream	stream	low	-	-	-	high
of	dub	a	a	low	a	a	-	high
phones	for in	-	-	mid	-	-	-	high
will	swell	-	-	mid	a	a	-	high
show	-	shore	shore	low	-	-	-	high
a	0	0	0	none	0	-	-	high
good	-	-	-	mid	-	-	-	high
deal	ill	ill	ill	high	ill	- + ill*	-	high
of	if	if	if	high	-	-	-	high
variability	-	-	-	high	-	-	-	high

Key: - correct response
0 no response

*S9 by mistake submitted both his new match and his former match for this portion. He reported later that the resulting sentence was useful in putting the two words side by side and enabling him to confirm that his second choice was correct.

that S2 did. S9 reported that he used quite detailed coarticulation criteria whereas S2 attempted to judge matches on the basis of general configuration with variations as a function of stress and position of the segment within the word.

A fourth difference may lie in individual ability to perform the task. In the previous three experiments, S2 was usually at the lower end of the range of success. S9, on the other hand, was the only subject participating in the experiments who succeeded in reading more than occasional words in the sentence to be matched. He made only one mistake, identifying good as big. It should be noted, however, that most subjects did not make a serious effort at reading the spectrograms since that was given as a secondary instruction in Experiment I to S1 and S2 only and as an option "if you have time" in Experiment II. Since the few subjects who attempted to read spectrograms had not been very successful, the instruction was omitted in Experiments III and IV.

Although S2's word recognition score was only 38 percent on this experiment, a comparison of phonemes between the words he matched and the words in the original sentence gives a score of 82 percent correct and 24 percent error. When we compare this result with Experiment III, we see that although S2 made more word-matching errors than the average for Experiment III, he matched more phonemes.

ADDITIONAL OBSERVATIONS

Because one of our interests was in automatic speech recognition, we asked subjects to report informally on the procedures they used in matching. Most subjects categorized the spectrograms, with varying degrees of rigor, according to gross phonetic features; they also noted, even if not always consciously, stressed and unstressed syllables.

Some subjects began by categorizing the reference spectrograms according to length and whether they contained stop-like, s-like, or other fricative features, and whether the voiced segments contained readily identifiable rising, falling, or steady formant patterns. They then examined the test sentences seeking similar gross features as a clue to determining which category of reference spectrograms should be examined to find the closest match.

Some subjects began by searching among the reference words for something that could match a portion of the test sentences before or after a pause, since in that position at least one of the word boundaries was sure. After a match had been made, the adjacent portion was studied. Some subjects also scanned the test sentences to look for distinctive patterns of stops and fricatives and then searched for reference words that would fit. When several reference words had the same gross phonetic characteristics, more detailed comparisons were made involving frequency, duration, and manner cues that were not considered in the first categorization. These comparisons were made visually without making detailed measurements.

Most subjects eventually worked from the reference words, taking each in turn and visually scanning the test sentences to see if the reference word matched any portion of the sentence. This procedure was particularly effective for polysyllabic words, the patterns of which were easily recognized when embedded in sentences. However, this procedure would not be feasible for a large set of reference words.

Most subjects matched function words only after the prominent words were located and short intervening portions remained unidentified. Many subjects had difficulty distinguishing function words from transitional segments connecting two words.

CONCLUSIONS

Subjects can match spectrograms of unknown utterances with reference-word spectrograms better than they can transcribe the spectrograms directly in terms of a sequence of phonetic elements. Previous studies (Klatt and Stevens, 1973; Lindblom and Svensson, 1973; Svensson, 1974) have shown that use of syntactic, semantic, and prosodic information can materially enhance subjects' ability to transcribe. Our results indicate that subjects do not even make full use of the acoustic information present in the spectrograms unless they are given a means to assess the significant differences between the spectrographic manifestations of different words. One such means is comparison of spectrograms.

Even when the number of words correctly matched is low, the number of syllables in the utterance is for the most part preserved. Furthermore, the number of phonemes matched is higher than the number of correct identifications reported for other acoustic phoneme-recognition schemes. This suggests that the output of this analysis-by-matching technique could yield a more accurate input of the syntactic-semantic component to a speech recognition system than is now available.

A process that generates the sequence of matched words in a manner resembling as closely as possible that of the unknown utterance serves as a useful source of feedback to the subjects. Comparison of the newly produced form with the original reveals differences in patterns which suggest that new choices might be made at those points. However, since only two subjects took part in the experiment with feedback, the degree of improvement that might be generally obtained cannot be predicted.

Subjects' error rates in matching vary significantly with sentence difficulty, size of vocabulary, and general ability to predict the changes a word pattern may undergo when placed in an unknown context and spoken with a different prosody. At least for limited vocabularies, subjects are able to determine whether two spectrographic patterns do or do not correspond to different productions of the same word more reliably than they are able to assign phonetic labels to the speech stream as seen in the spectrograms. However, as the size of the vocabulary increases, the likelihood of selecting the correct word decreases. Since the analysis time and paper-handling difficulties also increase with vocabulary, manual execution of such tasks can rapidly become impractical. Subjects' ability to generalize the acoustic cues they observe so that they need not be presented with all spectrographic forms but only with a limited subset remains to be investigated.

Automation of this matching process would entail storage of a complete vocabulary in spectrographic form, an exorbitant requirement. Various parametric representations may be considered but the corresponding storage savings will have to be weighed against deterioration of performance as compared with full spectrogram matching. Although generalization of the appropriate acoustic information into a sufficient set of analysis rules remains the ultimate goal, studies of word matching provide useful comparisons with the performance of any other method.

We believe the results of these experiments warrant continuation of our studies with computer-assisted word retrieval as a means of developing automatic pattern-matching techniques that make best use of those cues found useful by humans in establishing reliable word matches.

REFERENCES

Brown, Roger. (1970) Psycholinguistics. (New York: Free Press) 3.

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

Fletcher, Harvey. (1929) Speech and Hearing. (New York: Van Nostrand).

Klatt, Dennis H. and Kenneth N. Stevens. (1973) On the automatic recognition of continuous speech: Implications from a spectrogram-reading experiment. IEEE Trans. Audio Electroacoust. AU-21, 210-217.

Liberman, A. M., F. S. Cooper, Donald P. Shankweiler, and Michael Studdert-Kennedy. (1968) Why are speech spectrograms hard to read? Amer. Ann. Deaf 113, 127-133.

Lindblom, Björn E. F. and Stig-Göran Svensson. (1973) Interaction between segmental and nonsegmental factors in speech recognition. IEEE Trans. Audio Electroacoust. AU-21, 536-545.

Potter, R. K., G. A. Kopp, and H. C. Green. (1947) Visible Speech. (Princeton, N. J.: Van Nostrand).

Svensson, Stig-Göran. (1974) Prosody and Grammar in Speech Production, Monograph 2 from the Institute of Linguistics, University of Stockholm.

Results of a VCV Spectrogram-Reading Experiment

G. M. Kuhn[+] and R. McI. McGuire[++]
Haskins Laboratories, New Haven, Conn.

ABSTRACT

We attempted to identify the consonant in 432 spectrograms of vowel-consonant-vowel (VCV) utterances. In five sessions of spectrogram reading, our overall identification rate was 83 percent. An error analysis of the results shows that:

1. During the course of the experiment, our identification rate improved from 75 to 90 percent.
2. Voicing, manner, and place errors occurred on the following percentages of the tokens:

Voicing	01%
Manner	07%
Place	16%

3. The greatest improvement in identification rate came in stops and fricatives, the two manner classes that were the most numerous.

We conclude that one can learn to do well at identifying consonants from spectrograms of utterances of this constrained phonetic type. In a further spectrogram-reading experiment we plan to prepare a checklist of the cues we have used to identify each consonant. We assume that the checklist will help us apply the cues more consistently, and that it will help indicate where further cues are needed.

INTRODUCTION

We wanted to see how well we could identify the consonant in spectrograms of the V'CV type (where ' indicates the presence of prominent stress on the following syllable). We were also interested in finding out whether our performance would improve over time and what kinds of errors we would make. We chose the V'CV frame for two reasons: first, it eliminated the significant problems

[+] Also University of Connecticut, Storrs.

[++] Also Yale University, New Haven.

[HASKINS LABORATORIES: Status Report on Speech Research SR-39/40 (1974)]

of locating and counting the consonants; and second, it is an environment that gave us a clear picture of the cues for the prestressed consonants.

The consonants we were interested in were the 24 phonemic consonants of "General American," and three important allophonic variants (the glottal stop, voiced h, and the apical flap). Our consonant inventory was then:

m n ŋ b p d t g k ʔ ǰ č ɾ v f ð θ z s ʒ ʃ ɦ h w l r y

We decided to embed each of the 27 consonants in the 16 environments formed by the inclusive combination of /i a u ɝ/, yielding a total of 432 utterances. We chose these vowels since they include extreme positions of the first three formants of the phonemic vowels of General American. The coarticulation of the same intervocalic consonant with these different vowels can produce significant changes in the acoustic patterns associated with the consonant (Liberman, Delattre, Cooper, and Gerstman, 1954; Öhman, 1966). Thus, though VCVs are simple in segmental structure, we could not assume that it would be a simple exercise to identify intervocalic consonants from spectrograms.

METHOD

Orthographic equivalents were chosen for the phonetic symbols (see Appendix 1). The use of orthographic equivalents permitted us to create and randomize the names of the VCVs using a character string editor on the Haskins Laboratories' DDP-224 computer. To avoid the possibility of identification by elimination, the names were randomize' only over the entire 432 positions. The randomized names were output from the computer as a printer listing (see Appendix 2). One of the experimenters read the listing of the randomization at a rate of one VCV per sec. The recording of this reading was used to make broad-band Voiceprint spectrograms of the frequency range 0-4.8 kHz. A linear frequency scale (1.2 kHz/ in) was selected.

The spectrogram-reading sessions were held both morning and afternoon of two consecutive days, and o. the afternoon of the third day. We read 40, 120, 100, 100, and finally, 72 spectrograms in the successive sessions. Each session proceeded as follows. The next 20 spectrograms were taken from the randomization. Each experimenter attempted to identify the 20 consonants, writing the name of the consonant and any comments and observations on his answer sheet. When both experimenters had finished with the set of 20, they compared answers and argued any differences, but did not change their answers. Then the computer listing was checked and the intended consonants determined. Errors were noted and rationalized. In each session, this process was repeated set by set, until we felt too tired to continue. No session lasted more than three hours.

After the experimental sessions, the audio recording was used in a control session where the experimenter who spoke the randomization attempted to identify the intended consonant by ear. The intended consonant was heard in all but three cases. In the following analysis the data were treated as though all consonants were heard as intended. Since the errant identifications are presented in Appendix 3, the reader can proceed with any data analysis he wishes. For our own part, we drew up identification matrices (see Appendix 4), and then we set up a feature distance metric for all the consonants. The following table of the consonants represents the voicing, manner, and place relationships as we—

somewhat arbitrarily--imposed them (compare with the scheme presented in Peterson and Shoup, 1966).

TABLE 1

	BILABIAL		UNI-LABIAL		LINGUA-DENT		ALVEOLAR		PALATAL		VELAR		GLOTTAL	
	+	−	+	−	+	−	+	−	+	−	+	−	+	−
NASAL	m						n				ŋ			
STOP	b	p					d	t			g	k		?
AFFRICATE									ǰ	č				
FLAP							ɾ							
FRICATIVE			v	f	ð	θ	z	s	ʒ	ʃ			ɦ	h
SONORANT	w						l		y		r			

From left to right, place of articulation moves back down the vocal tract, with voicing alternating within a place column. From top to bottom, the manner classes might be said to show a lesser duration of occlusion. We do not pretend that this is a perfect feature assignment, but only that it is one that could reasonably be used to find out whether the size of our identification errors decreased over time. In calculating feature distance we have assigned unit value to a minimal distinction in any dimension, and we have summed the distance along the three dimensions. Thus, if /m/ is identified as /r/, we have said that the feature distance is 5 in place, plus 5 in manner, plus 0 in voicing, for a total of 10. So defined, feature distance is the basis for the data in Figure 2.

RESULTS

Q1. How well were the consonants identified?

Both experimenters averaged 83 percent in consonant identification over all sessions. Given the fact that the phonetic context was chosen to increase the probability of success, we feel that this is not a particularly high rate of identification. In addition, the feature distance of many of the errors was so large that the average feature distance per error was 3.4.

Q2. Did the identification rate improve over time?

The results appear more positive when we look at identification and feature errors over time. Figure 1 shows that the 83 percent overall identification score hides the fact that the identification rate improved from 75 to 90 percent.

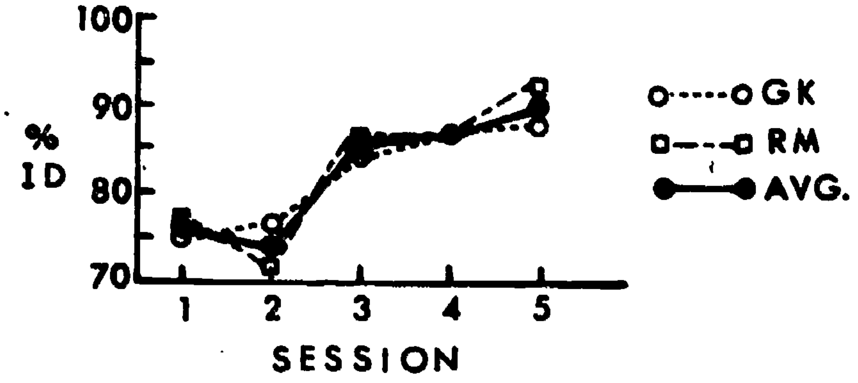


Figure 1: Identif!cation by session.

Figure 2 shows the average feature d.stance per token, an indicator of how wrong each identification was on the average. Here the results averaged over both experimenters show a continuing decrease from ses·'.on to session. It appears that the identification error rate fell by 60 percent, and that the feature distance per token fell by 66 percent. This latter result suggests that the errors grew not only less frequent, but also somewhat smaller.

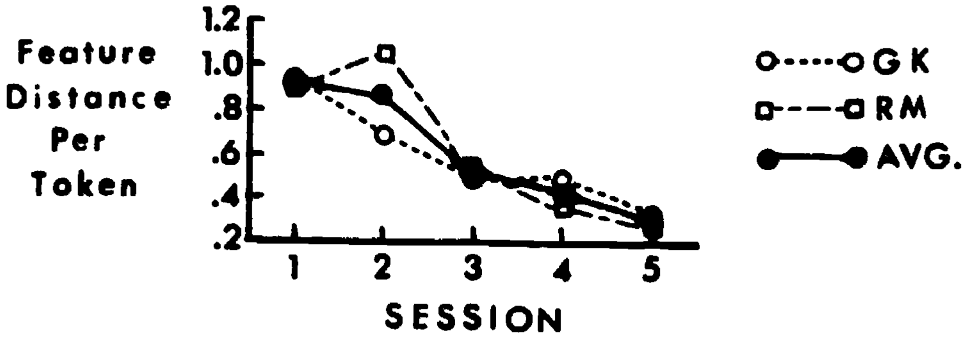


Figure 2: Average feature distance per token, by session.

Q3. What kinds of errors were made?

Overall, voicing, manner, and place errors occurred on the following percentages of the tokens:

Voicing	<01%
Manner	07%
Place	16%

In short, voicing errors were practically nonexistant, while manner errors occurred about half as often as errors of place. Figure 3 shows that manner and place identification errors both decreased over time. This figure also shows that the rate of manner errors dropped 66 percent to a final value of 5 percent. The rate of place errors, on the other hand, dropped only 50 percent, to a final value of 10 percent. Since the overall identification error rate and the place identification error rate both dropped to 10 percent, it follows that all manner errors in the last session were also errors of place. This result underscores the fact that while we have so far talked about manner and place independently, they are not truly independent. In other words, there are holes in the phonetic pattern, and an error of manner or place can force one to make an error in the other dimension.

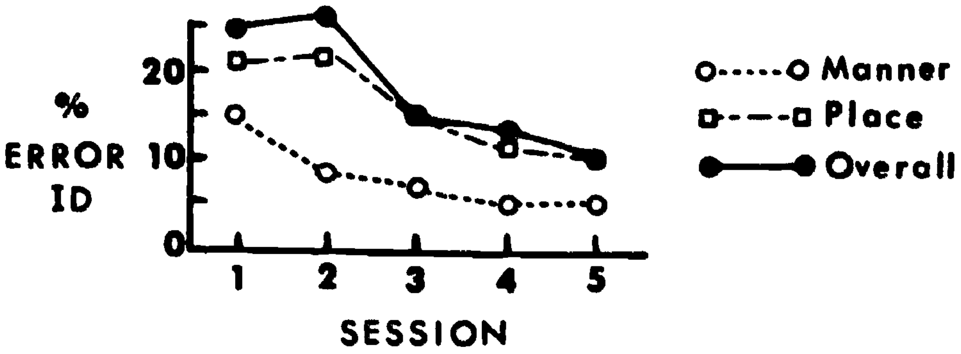


Figure 3: Manner and place identification errors, by session.

Figure 4 shows identification error rates by manner class. Nasals were incorrectly identified 32 percent of the time. Semivowels, stops, and fricatives were missed about 18 percent of the time. Affricates and flaps were missed less than 10 percent of the time. Figure 5 shows error rates for those manner classes that were represented by at least ten identifications in each of the five sessions. Nasals, stops, and fricatives show improvement, but semivowels do not. At the end of the experiment, nasals and semivowels have the highest error rates. The two experimenters do show some difference in their ability to identify stops and semivowels, but the overall pattern of their results is so similar that we feel it is reasonable to draw conclusions based on their averaged data.

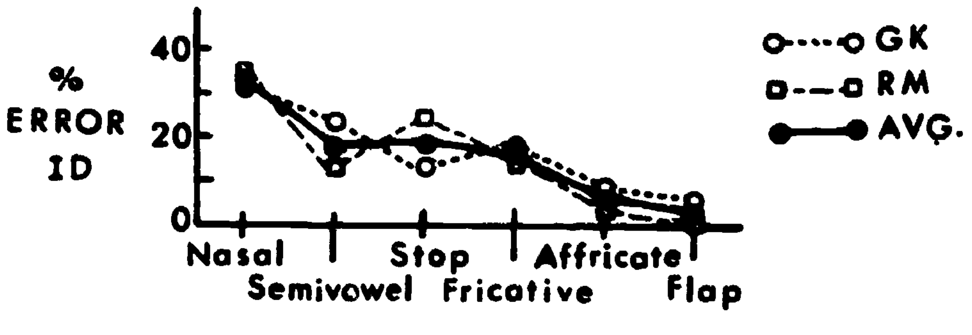


Figure 4: Identification errors by manner class.

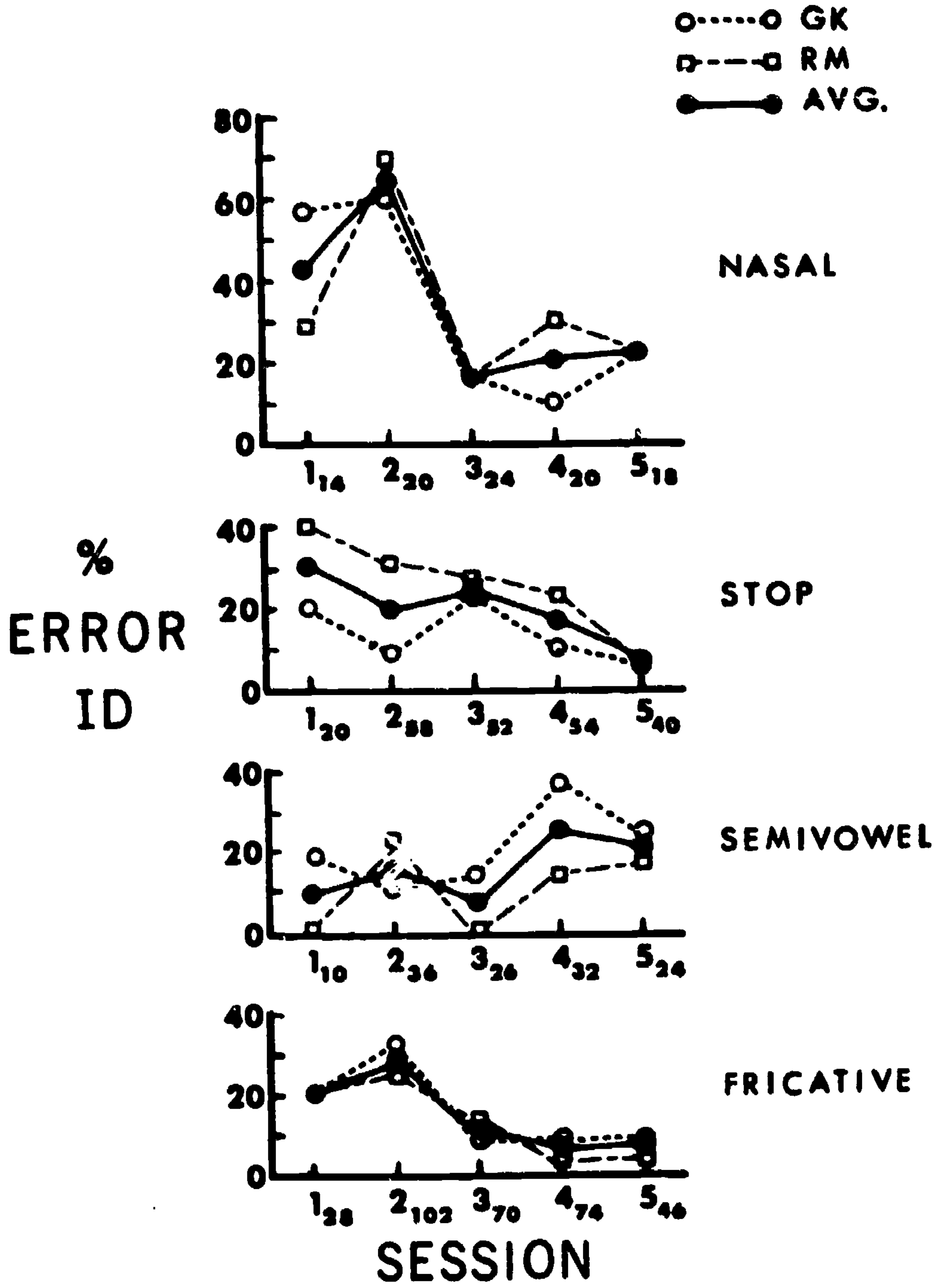


Figure 5: Identification errors per manner class, by session.

CONCLUSION

We interpret these results to show that one can learn to do well at identifying consonants from spectrograms of utterances of this phonetic type. Of course it should be born in mind that the VCV utterance is simple in structure, and that we set ourselves a very restricted task.

The greatest improvement in identification rate came on stops and fricatives, the two most numerous manner classes. As a consequence, we are tempted to assume that the error rates on nasals and semivowels would fall below their final value of 20 percent, if the number of tokens identified approached that for the stops and fricatives.

Our impression is that the cues we used are documented in the literature: the state of voicing, the shape of fricatives and bursts, the transitions of formants, to name just a few. All these cues are by now classic. We may indeed have made novel use of the cues, but since no explicit rules for identification were followed during the experiment, we have not presented tables of cues drawn up after the fact. Instead, we plan to test our cues on a second set of the same 432 VCVs. There, we will attempt to find out whether these explicit rules can help us to achieve more consistent resulcs, and whether they can bring the problems of place analysis into sharper focus.

REFERENCES

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 (8, Whole No. 379).

Öhman, S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. J. Acoust. Soc. Amer. 39, 151-168.

Peterson, G. E. and J. E. Shoup. (1966) A physiological theory of phonetics. J. Speech Hearing Res. 9, 5-67.

ORTHOGRAPHIC EQUIVALENT	PHONETIC SYMBOL
M	m
N	n
NG	ŋ
B	b
P	p
DD	d
T	t
G	g
K	k
?	ʔ
J	ǰ
CH	č
D	ɾ
V	v
F	f
DH	ð
TH	θ
ZH	ʒ
SH	ʃ
H	ɦ
HH	h
W	w
L	l
Y	y
R	r
EE	i
AH	ɑ
UW	u
ER	ɝ
' (apostrophe)	ˈ (stress)

Appendix 1: Orthographic equivalents.

```
VCV RANDOMIZATION. 15 JULY 1974.

ER'DUW    ER'NGUW   EE'ZEE    UW'SAH    UW'MAH    EE'SHEE   UW'NGER   AH'MER    ER'REE    UW'MUW
ER'TUW    UW'LEE    EE'ZHUW   EE'DAH    AH'GEE    AH'DEE    AH'DHER   EE'THEE   ER'BAH    AH'WAH
AH'?ER    UW'PAH    UW'KUW    UW'HER    EE'DHAH   UW'?UW    ER'GER    EE'SEE    ER'ZHAH   UW'MUW
ER'NAH    EE'ZHER   UW'YER    UW'DAH    UW'BUW    AH'TEE    ER'WEE    ER'HUW    EE'MER    EE'ZAH
AH'SUW    UW'VEE    AH'FER    UW'DDUW   EE'JEE    UW'VUW    EE'PER    ER'NER    ER'HAH    UW'SHEE
UW'BAH    UW'SHER   ER'FER    EE'YER    ER'NUW    AH'ZHAH   AH'NGUW   AH'THAH   UW'MEE    AH'SAH
EE'DHEE   ER'THER   EE'HEE    ER'NEE    ER'KAH    AH'DAH    AH'NGER   ER'KEE    UW'THAH   AH'RUW
EE'SHAH   EE'PEE    UW'DHUW   AH'ZER    ER'ZHUW   EE'RAH    EE'GER    AH'PUW    ER'KUW    AH'YER
AH'VEE    AH'THUW   AH'WUW    UW'FEE    AH'HHUW   ER'CHEE   AH'LAH    AH'FEE    AH'CHAH   EE'DDEE
UW'DHER   UW'ZUW    EE'RER    UW'HAH    ER'GAH    AH'SHAH   UW'SEE    UW'SER    ER'DDEE   ER'CHUW
EE'TER    EE'FEE    ER'ZEE    UW'RUW    UW'FAH    ER'BER    EE'DER    ER'WER    EE'GAH    EE'SAH
UW'HHUW   UW'?ER    UW'LAH    ER'BEE    EE'YAH    ER'NGAH   AH'MEE    UW'YUW    AH'THEE   EE'CHER
ER'VER    ER'?AH    AH'GUW    ER'LAH    ER'DAH    ER'YAH    UW'DUW    EE'WER    UW'WEE    ER'DHAH
UW'GER    ER'VUW    EE'DEE    UW'DDER   EE'DDAH   AH'HHEE   EE'HHER   ER'MEE    UW'GUW    AH'SER
AH'BUW    AH'SEE    EE'HHEE   AH'KUW    ER'FEE    AH'JAH    UW'NGUW   ER'FAH    ER'NEE    ER'ZHER
AH'HHAH   ER'DDER   ER'SAH    AH'KAH    EE'?UW    AH'TAH    ER'KUW    UW'TUW    EE'JAH    ER'FUW
ER'LEE    AH'PAH    ER'SER    EE'?UW    EE'YEE    ER'JER    AH'NER    AH'YUW    ER'NGER   AH'RER
UW'ZHER   ER'NGEE   AH'ZHEE   ER'LUW    EE'?AH    ER'THEE   UW'YEE    AH'HUW    UW'FUW    UW'JER
AH'PER    AH'WER    ER'SHER   ER'DHER   AH'LEE    EE'ZHAH   AH'SHER   ER'VAH    EE'DDER   ER'THUW
UW'VAH    AH'GER    ER'PER    UW'PEE    AH'CHUW   EE'NGEE   AH'DDER   UW'NUW    EE'SUW    UW'BEE
EE'UHUW   ER'?EE    AH'TUW    UW'ZAH    EE'RUW    EE'ZHEE   EE'FAH    UW'CHER   EE'KAH    ER'JAH
ER'?ER    EE'HER    AH'ZHUW   UW'JEE    EE'CHEE   UW'KER    ER'ZUW    UW'GEE    EE'VEE    ER'JUW
ER'MUW    UW'TAH    EE'WAH    UW'DHEE   EE'VER    UW'HHER   UW'REE    EE'HHAH   EE'NAH    UW'DER
EE'JUW    EE'GUW    EE'NEE    EE'PUW    AH'ZHER   AH'NEE    ER'CHAH   AH'HHER   ER'SHUW   AH'NGEE
ER'DHUW   UW'CHUW   AH'WEE    ER'PUW    ER'DHEE   ER'?UW    AH'KER    EE'NGUW   ER'ZER    EE'KER
UW'CHAH   UW'ZER    AH'CHER   EE'BAH    AH'HEE    UW'NER    UW'SHAH   EE'WEE    ER'GUW    EE'DDUW
ER'KAH    ER'PAH    UW'DDEE   AH'HER    UW'JUW    EE'HAH    UW'ZHEE   UW'VER    ER'ZAH    AH'YEE
AH'MAH    AH'DDAH   AH'SHEE   EE'?ER    EE'JEP    AH'GAH    AH'HAH    EE'KEE    UW'ZHAH   UW'HHAH
ER'LER    UW'PER    AH'NUW    EE'SER    UW'HHEE   EE'?EE    ER'SUW    ER'HHEF   AH'DUW    ER'YEE
ER'THAH   AH'ZAH    EE'REE    AH'LER    UW'LER    UW'YAH    EE'HHUW   AH'BAH    UW'THEE   EE'SHUW
UW'WAH    AH'FAH    UW'FER    AH'DER    UW'DHAH   EE'ZUW    ER'SHAH   ER'KER    EE'NUW    ER'HHAH
EE'HUW    ER'YER    ER'SHEE   EE'BUW    AH'JEE    UW'NAH    AH'FUW    AH'DDEE   ER'MER    UW'NEE
ER'GEE    ER'TAH    UW'KEE    ER'DDAH   EE'PAH    AH'BEE    EE'VUW    ER'DEE    AH'?UW    ER'DER
EE'KUW    EE'MAH    UW'NGEE   UW'?EE    EE'ZER    AH'MUW    ER'HHUW   UW'HEE    ER'MUW    AH'JER
ER'TEE    EE'THER   AH'DHUW   ER'RER    AH'PEE    EE'THUW   ER'ZHEE   AH'JUW    UW'TEE    EE'THAH
EE'WUW    UW'PUW    EE'LEE    UW'RAH    EE'BER    UW'WER    AH'VER    EE'CHAH   UW'DDAH   AH'NGAH
ER'DDUW   AH'SHUW   AH'BER    AH'VUW    ER'SEE    AH'DDUW   EE'TAH    ER'BUW    UW'GAH    ER'HHER
UW'DEE    EE'BEE    UW'SHUW   ER'HER    AH'NAH    UW'RER    ER'YUW    EE'VAH    EE'NGAH   UW'MER
EE'NGER   EE'NER    AH'KEE    AH'THER   EE'CHUW   EE'TEE    UW'NGAH   EE'LUW    AH'KAH    UW'BER
EE'FER    ER'JEE    EE'YUW    ER'TER    AH'LUW    UW'NUW    AH'ZUW    AH'TER    AH'YAH    EE'LAH
UW'SUW    ER'PEE    AH'?AH    EE'DHER   AH'?EE    EE'MUW    UW'ZHUW   ER'CHER   AH'VAH    AH'DHAH
ER'VEE    UW'JAH    ER'MAH    AH'DHEE   UW'?AH    UW'THUW   UW'CHEE   AH'CHEE   EE'LER    EE'SHER
UW'ZEE    EE'FUW    EE'GEE    AH'ZEE    UW'THER   EE'DUW    EE'MEE    UW'LUW    ER'WAH    AH'REE
UW'TER    UW'KAH
```

Appendix 2: Randomization of the VCV's.

APPENDIX 2

APPENDIX 3

IDENTIFICATION ERRORS: GK

No.	Sent	Recvd.	No.	Sent	Recvd.	No.	Sent	Recvd.
2	ER'NGUW	M	165	EE'YEE	NH	367	EE'TAH	CH
7	UW'NGER	M	166	ER'JER	DD	376	UW'RER	H
8	AH'MER	L	167	UW'YEE	DH	377	ER'YUW	Z
12	UW'LEE	DH	178	AH'HUW	V	385	EE'CHUW	J
17	AH'DHER	N	180	UW'JER	DD	387	UW'NGAH	M
22	UW'PAH	F	181	AH'PER	K	388	EE'LUW	V
25	EE'DHAH	D	193	ER'PER	K	410	AH'DHAH	L
26	UW'?UW	NH	194	UW'PEE	K	411	ER'VEE	DH
31	ER'NAH	NG	201	EE'DHUW	D	413	ER'MAH	NG
40	EE'ZAH	DH	209	EE'KAH	CH			
			218	UW'GEE	DH			
42	UW'VEE	DH	222	UW'TAH	CH			
43	AH'FER	?	228	EE'WHAH	SH			
46	UW'VUW	B	229	EE'NAH	NG			
48	ER'NER	NG	230	EE'DER	DH			
49	ER'NAH	NH	233	EE'NEE	NG			
53	ER'FER	NH						
54	EE'YER	D	266	EE'NAH	T			
55	ER'MUW	NG	274	EE'?ER	H			
56	AH'ZHAH	Y	293	EE'REE	V			
57	AH'NGUW	M	295	UW'LER	N			
58	ER'THAH	?	297	EE'WHUW	F			
60	AH'SAH	TH	301	UW'WAH	V			
64	ER'NEE	DH	320	UW'NEE	NG			
67	AH'NGER	N	323	UW'KEE	P			
80	AH'YER	ZH	338	UW'NEE	V			
85	AH'HHUW	?	339	ER'WUW	L			
90	EE'DDEE	B	352	UW'PUW	K			
92	UW'ZUW	S	353	EE'LEE	W			
96	AH'SHAH	TH	356	UW'WER	V			
119	AH'THEE	NH						
130	ER'DHAH	V						
136	AH'WHEE	TH						
145	ER'FEE	TH						
147	UW'NGUW	N						
149	ER'NEE	NG						
156	AH'TAH	K						
160	ER'FUW	NH						

IDENTIFICATION ERRORS: RM

No.	Sent	Recvd.	No.	Sent	Recvd.	No.	Sent	Recvd.
5	UW'MAH	L	145	ER'FEE	HH	368	ER'BUW	V
10	UW'MUW	L	147	UW'NGUW	M	375	AH'NAH	NG
17	AH'DHER	V	149	ER'NEE	NG	376	UW'RER	W
19	ER'BAH	DD	151	AH'WHAH	F	387	UW'NGAH	M
22	UW'PAH	TH	156	AH'TAH	K	388	EE'LUW	N
25	EE'DHAH	DD	160	ER'FUW	TH	411	ER'VEE	DH
26	UW'?UW	TH						
36	AH'TEE	CH	178	AH'HUW	V			
40	EE'ZAH	DH	181	AH'PER	F			
			188	ER'VAH	DH			
42	UW'VEE	DH	191	UW'VAH	B			
46	UW'VUW	?	194	UW'PEE	K			
48	ER'NER	NG	209	EE'KAH	T			
49	ER'HAH	NH	212	EE'HER	R			
53	ER'?ER	HH	216	UW'KER	P			
54	EE'YER	R	218	UW'GEE	DH			
57	AH'NGUW	M	221	ER'MUW	L			
64	ER'HEE	Z	222	UW'TAH	K			
67	AH'NGER	M	225	EE'VER	DH			
68	ER'KEE	CH	254	EE'BAH	V			
72	EE'PEE	TH	256	UW'WER	NG			
73	UW'DHUW	DD						
80	AH'YER	ZH	271	AH'MAH	NG			
86	ER'CHEE	J	272	AH'DDAH	G			
88	AH'FEE	NH	278	EE'KEE	CH			
91	UW'DHER	DD	293	EE'REE	V			
113	UW'LAH	N	303	UW'FER	TH			
114	ER'BEE	DD	320	UW'NEE	NG			
116	ER'NGAH	M	333	UW'NGEE	N			
117	AH'MEE	N	341	ER'TEE	CH			
121	ER'VER	B	342	EE'THER	NH			
122	ER'?AH	TH	345	AH'PEE	CH			
130	ER'DHAH	DD	349	UW'YEE	K			
131	UW'GER	B	352	UW'FUW	K			
132	ER'VUW	B	353	EE'LEE	R			
135	EE'DDAH	G						
139	UW'GUW	B						
141	AH'BUW	V						

Appendix 3: Identification errors.

RECEIVED

SENT	m	n	ŋ	b	p	d	t	g	k	ʔ	ǰ	č	ɭ	v	f	ð	θ	z	s	ž	ʃ	ɦ	h	w	l	y	r	ERRORS
m	14		1																						1			2
n		9	7																									7
ŋ	5	1	10																									6
b				16																								0
p					11				4						1													5
d				1		15																						1
t							13		1			2																3
g								15								1												1
k					1				14			1																2
ʔ										14												1	1					2
ǰ						2					14																	2
č											1	15																1
ɭ													15			1												1
v				1										13		2												3
f										1					12		1						2					4
ð		1											2	1		11									1			5
θ										1							14						1					2
z																1		14	1									2
s																	1		15									1
ž																				15						1		1
ʃ																	1				15							1
ɦ														2		1						11	1			1		5
h										1					1		1				1		12					4
w														2										13	1			3
l		1												1		1								1	12			4
y													1			1		:		1			1			11		5
r														1								1					14	2
RESPONSE TOTALS	19	12	18	18	12	17	13	15	19	17	15	18	18	20	14	19	18	15	16	16	16	13	18	14	15	13	14	

Appendix 4a: Identification matrix: GK

APPENDIX 4A

APPENDIX 4B

RECEIVED

SENT	m	n	ŋ	b	p	d	t	g	k	ʔ	ǰ	č	ɭ	v	f	ð	θ	z	s	ʒ	ʃ	ɦ	h	w	l	y	r	ERRORS
m	11	1	1																						3			5
n		11	5																									5
ŋ	5	1	10																									6
b				11		2								3														5
p					10				2			1			1		2											6
d						14		2																				2
t							11		3			2																5
g				2				13								1												3
k					1		1		12			2																4
ʔ										14							2											2
ǰ											16																	0
č											1	15																1
ɭ													16															0
v				1						1				8		4						2						8
f															11		2						3					5
ð						4								1		11												5
θ																	15						1					1
z																1		15										1
s																			16									0
ʒ																				16								0
ʃ																					16							0
ɦ														1				1				12	1				1	4
h															1								15					1
w																								16				0
l		1																				1			13		1	3
y																										15	1	1
r														1										1			14	2
RESPONSE TOTALS	16	14	16	14	11	20	12	15	17	15	17	20	16	14	13	17	21	16	16	16	16	15	20	17	16	15	17	

Appendix 4b: Identification matrix: RM

RECEIVED

SENT	m	n	ŋ	b	p	d	ṭ	g	k	ʔ	ǰ	č	ɪ	v	f	ð	θ	z	s	ž	ʃ	ɦ	h	w	l	y	r	ERRORS
m	25	1	2																						4			7
n		20	12																									12
ŋ	10	2	20																									12
b				27		2								3														5
p					21				6			1			2		2											11
d				1		29		2																				3
t							24		4			4																8
g				2				28								2												4
k					2		1		26			3																6
ʔ										28							2					1	1					4
ǰ						2					30																	2
č											2	30																2
ɪ													31			1												1
v				2						1				21		6						2						11
f										1					23		3						5					9
ð		1				4							2	2		22									1			10
θ										1							29						2					3
z																2		29	1									3
s																	1		31									1
ž																				31						1		1
ʃ																	1				31							1
ɦ														3		1		1				23	2			1	1	9
h										1					2		1				1		27					5
w														2										29	1			3
l		2												1		1						1		1	25		1	7
y													1			1		1		1			1			26	1	6
r														2								1		1			28	4
RESPONSE TOTALS	35	26	34	32	23	37	25	30	36	32	32	38	34	34	27	36	39	31	32	32	32	28	38	31	31	28	31	

Appendix 4c: Identification matrix: Total

APPENDIX 4C

Evidence for Spectral Fusion in Dichotic Release from Upward Spread of Masking

Terrance M. Nearey[+] and Andrea G. Levitt[++]
Haskins Laboratories, New Haven, Conn.

INTRODUCTION

Evidence from recent experiments conducted at Haskins Laboratories (Nye, Nearey, and Rand, 1974; Rand, 1974) indicates at least two important points. (1) The first formant (F1) of synthetic stimuli can mask higher formants (F2, F3, and the fricative components). In particular, it has been shown that this upward masking effect can result in a loss of information about the place of articulation for stop consonants.[1] (2) A release from masking on the order of 20 db can be obtained if the signal is spectrally divided and presented dichotically—F1 directed to one ear and the higher formants to the other ear.

One difficulty with the previous studies is that, in principle, it was possible for the listener to make the necessary discriminations by attending to the ear receiving the upper formant information alone. The fusion of the inputs to both ears and the perception of the combined sounds as speech were not necessary to produce correct responses. Nevertheless, anecdotal comments from the listeners appear to indicate that fusion normally took place. Further support is available from previous reports that listeners perceive spectrally divided signals as a single voice in a single location (Broadbent and Ladefoged, 1957; Cutting, 1973). Moreover, additional evidence (Carlson, Granström, and Fant, 1970) suggests that when spectral fusion of vowel formants occurs, phonetic information can be extracted from <u>both</u> dichotically presented channels.

Our present study sought the evidence that fusion occurs in conjunction with a release from masking. The basic technique was to create stimulus conditions where contributions from both channels were required to make the necessary phonetic judgments. Clear evidence for spectral fusion with a release from masking was found for vowels in one experiment. In a second experiment, the results suggest the operation of a similar fusion effect when listeners attempt to discriminate place versus voicing cues for stop consonants.

[+] Also University of Connecticut, Storrs.

[++] Also Yale University, New Haven, Conn.

[1] A similar information loss occurs in natural speech which has been low-pass filtered or mixed with Gaussian noise (Miller and Nicely, 1955).

[HASKINS LABORATORIES: Status Report on Speech Research SR-39/40 (1974)]

METHODS

Experiment 1

The essential strategy for testing a hypothesis of spectral fusion is to provide stimuli sets with no redundancy between channels for selected phonetic properties. In the first of our experiments, this was accomplished by the construction of a triplet of three-formant vowels, [ɛ], [æ], and [ɑ], with the following properties. (1) The nominal amplitude values of all the corresponding formants were the same, e.g., the amplitude of F1 of [ɛ] was equal to the amplitude of F1 of [ɑ]. In fact, small differences occur in the signals with different formant frequencies; however, these differences of less than 1 db are insignificant compared to the attenuation factors used in the experiment. (2) The durations of the vowels were identical. (3) The formant frequencies of the vowels were chosen so that [ɛ] and [æ] differed only in F1, while [æ] and [ɑ] differed only in F2. The frequency of F3 was the same for all three vowels. The practical consequence of this stimulus choice is that information from both F1 and F2 is required to keep all three vowels distinct. Loss of F1 implies the loss of the [ɛ]/[æ] distinction, while loss of F2 implies the loss of the [æ]/[ɑ] distinction.

Three formant transition burst patterns appropriate for [b], [d], and [g] were provided for each vowel stimulus, resulting in a final stimulus set of nine consonant-vowel (CV) syllables. The consonant portions were adjusted empirically for each of the nine stimuli with the restriction that vowels with identical steady-state F1's were provided with identical F1 transitions. The selection of F2,F3 values for the consonantal portions was restricted only by the requirement that they reach the steady-state vowel values by the time the F1 frequency transition had ended (see Figure 1).

Two tapes were prepared for the experiment. Each tape contained eight blocks of 27 randomly ordered stimuli (each of the nine stimuli appearing three times in each block). One complete tape was used for each condition, and tapes were alternated between conditions. There were six dichotic conditions, for a total of 1296 trials: (9 stimuli x 24 occurrences per tape x 2 ear/formant conditions x 3 attenuation levels--10, 20, and 30 db). In the two binaural conditions there were 432 trials: (9 stimuli x 24 occurrences per tape x 2 attenuation levels--10 and 30 db). The tapes were played at a baseline level of 85 db SPL.[2]

Twelve subjects, none of whom had any known hearing loss, participated in the experiment and were paid for their participation. They were divided into four equal groups; each group took part in a 2 1/2 hour session which included a 20 minute break. Presentation of the six dichotic conditions and the two binaural conditions was balanced across groups.

Initially, the subjects were told that they would be hearing nine CV syllables, [bɛ], [dɛ], [gɛ], [bæ], [dæ], [gæ], [bɑ], [dɑ], and [gɑ]. They were

[2] See Nye, Nearey, and Rand (1974) for a definition of the "baseline sound pressure level."

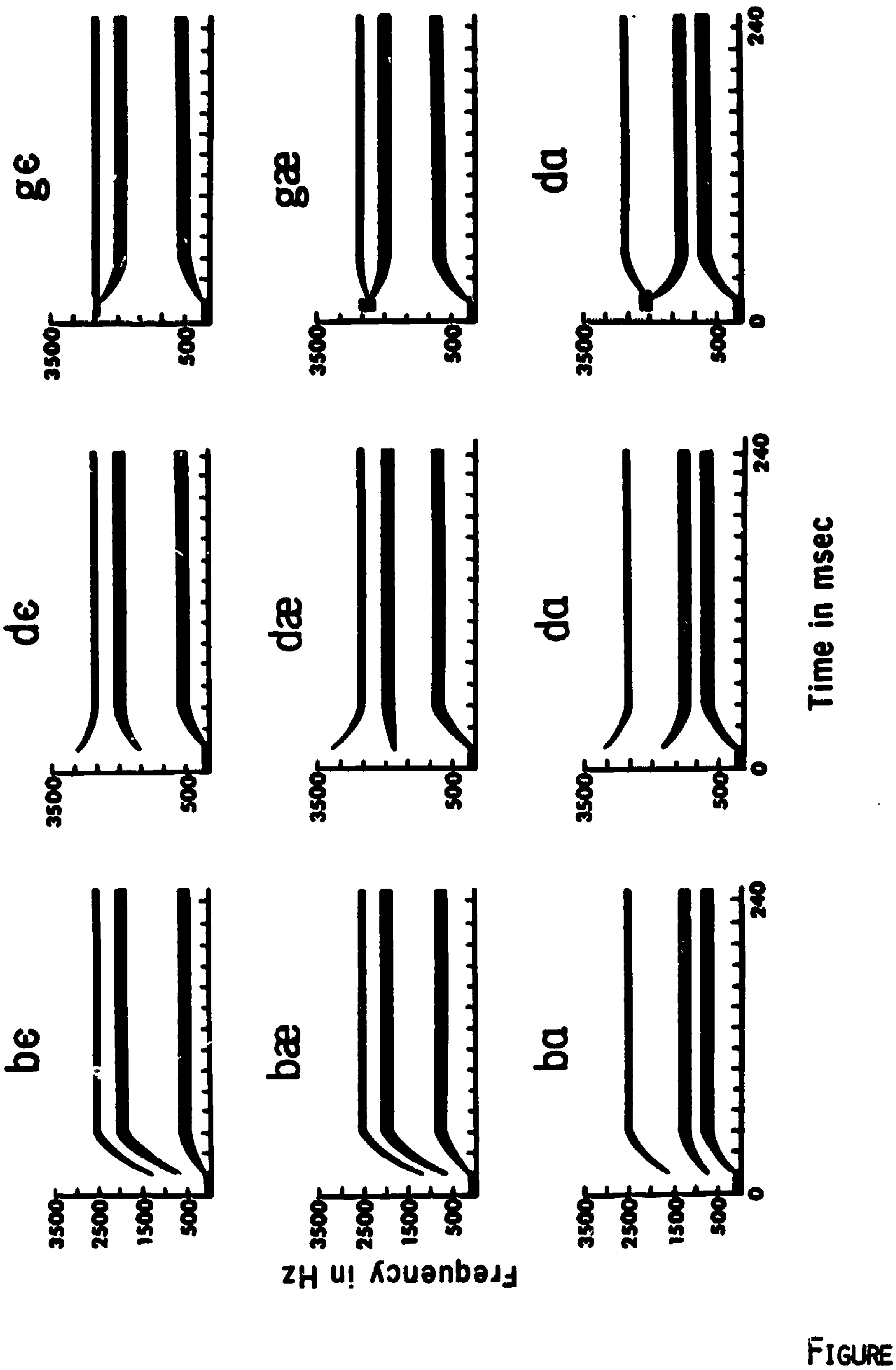

bɛ
dɛ
gɛ
bæ
dæ
gæ
bɑ
dɑ
dɑ
3500
2500
1500
500
0
240
Frequency in Hz
Time in msec

FIGURE 1

then instructed to write down either a phonemic transcription of what they heard or a corresponding English word, e.g., "bed" for [bε], "bad" for [bæ], etc. A chart of word responses was available for reference during the experiment. After a brief practice session the experiment began.

Experiment 2

For the second experiment a different set of nine stimuli was used: six consonant stimuli and three pseudo-consonants followed by the vowel [ɑ]. The six consonants were the voiced stops [b], [d], and [g] and the voiceless stops [p], [t], and [k], while the pseudo-consonants, or mixed stimuli, were half unvoiced-half voiceless sounds that have no counterparts in natural speech. The mixed stimuli were constructed from the F1 portion of the voiced stimuli and the F2,F3 portions of the voiceless stimuli. The important consideration here is that for all stimuli, place information was carried exclusively by F2 and F3, while in the case of the mixed stimuli there were conflicting voicing cues in F2 and F3 (voiceless) versus F1 (voiced) (see Figure 2). All stimuli were produced on the Haskins Laboratories' parallel formant resonance synthesizer.[3]

Six tapes, each containing 72 stimuli, were produced to give a total of 432 trials delivered at a baseline level of 85 db SPL: (9 stimuli x 2 repetitions x 3 presentation conditions x 2 noise conditions x 4 attenuation levels). Each tape was designed for monaural, dichotic, or binaural presentation, in one case without noise and in the other case with the addition of a Gaussian noise signal (signal/noise ratio +6 db). The F2 and F3 components of all the stimuli were attenuated 0, 10, 20, or 30 db. Furthermore, all the stimuli were randomized and balanced for ear of presentation on both the monaural and dichotic tapes.

Nine students, none of whom had any known hearing loss, were paid for their participation in the experiment, which lasted one hour. The subjects were divided into three groups. Although all of the groups encountered the three tapes without noise first, the order of presentation of the monaural, dichotic, and binaural tapes was balanced across groups.

RESULTS

The chief results of the first experiment are presented in Figure 3. In the dichotic condition there are essentially no vowel errors at any attenuation level. By contrast, in the binaural condition, the error rate for vowel identification rises to 32.7 percent under a 30 db attenuation of F2 and F3. An analysis of the errors reveals that this high rate is due almost entirely to [æ]/[ɑ] confusions, consistent with the masking of F2 and F3 by F1.[4] Identification of

[3] We thank Tim Rand who constructed the stimuli and prepared the tapes used in this experiment.

[4] This result is consistent with the findings of Ainsworth and Millar (1972). In experiments which varied vowel formant amplitude levels, vowel identification was basically unaffected until F2 amplitude reached a level of 28 db below F1. Beyond that level, they note that vowel errors occurred between vowels with the same F1 values.

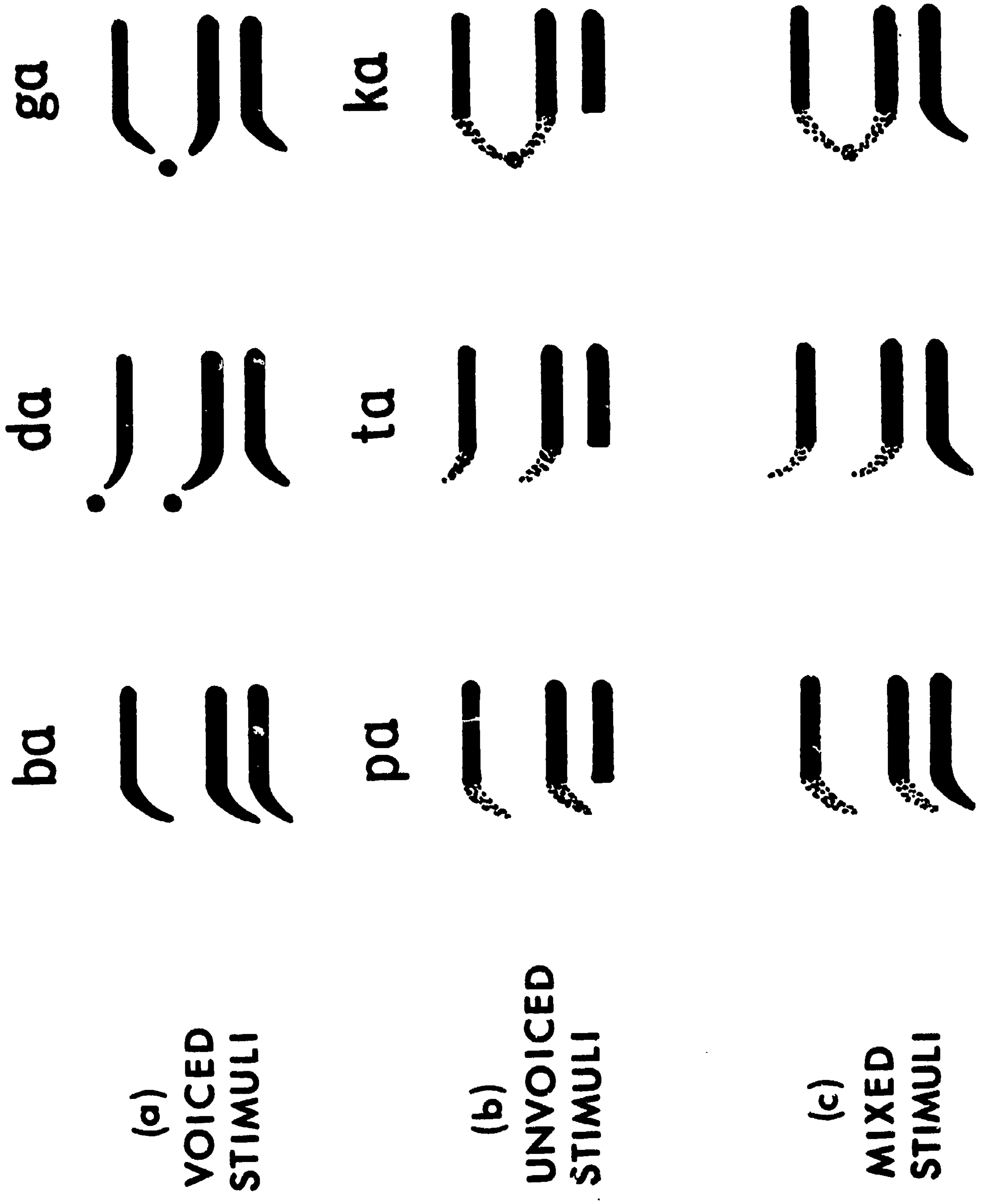

ga
ka
da
ta
ba
pa
(a) VOICED STIMULI
(b) UNVOICED STIMULI
(c) MIXED STIMULI

FIGURE 2

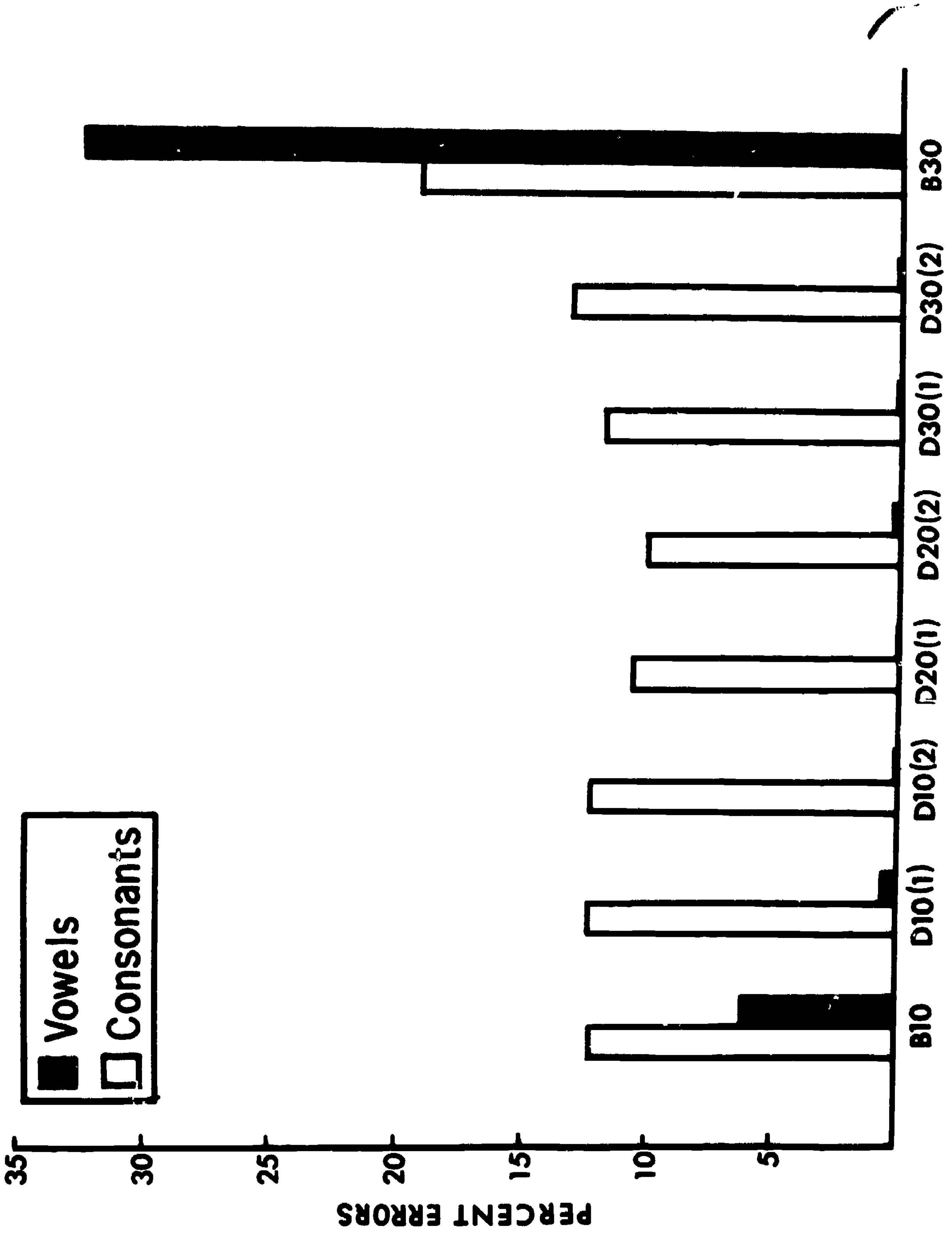

■ Vowels
□ Consonants
PERCENT ERRORS
35
30
25
20
15
10
5
B10
D10(1)
D10(2)
D20(1)
D20(2)
D30(1)
D30(2)
B30

FIGURE 3

the vowel [ɛ] remains essentially unaffected; the phonetic distinction between it and the other two vowels always involves an F1 difference. Even at the 10 db attenuation level, there are significantly more vowel errors in the binaural condition ($p \leq .01$), and once again the errors are heavily concentrated in [æ]/[ɑ] confusions. It should be pointed out that consonant errors are also significantly higher in the binaural 30 db attenuation condition than in the corresponding dichotic condition as indicated by a paired difference test ($p \leq .01$). There is considerable variability in the intelligibility of individual consonants, apparently due largely to differences in the energy relationships between the consonantal portions of individual tokens. One consonant, [g], performs extremely well in all conditions. Its resistance to masking is probably because in all three syllables containing [g] there is a simulated burst of relatively high F2 and F3 energy near the onset of each syllable where the amplitude of F1 is relatively low.

The results of the second experiment also show evidence of a release from masking as well as fusion of the spectrally divided signal. The results of this experiment were scored for number of correct responses for place and voicing in the cases of the voiced and voiceless stimuli, and for number of correct responses for place and proportion of responses "voiced" in the case of the mixed stimuli.

The findings for the voiced CV syllables show a highly significant release from masking for the dichotic no-noise condition when F2 and F3 are attenuated 30 db ($p \leq .01$) (see Figure 4). Overall performance is considerably lower for the tapes with noise, however, and no significant release from masking was found for dichotic presentation.

Under dichotic conditions an overwhelming proportion of the responses to the mixed stimuli are reported as "voiced." Given the nature of these unnatural stimuli and their randomized occurrence among pure voiced and pure voiceless stimuli, the fact that place identification is significantly better than random for the "voiced" responses provides some evidence that fusion is in fact taking place--because place information can only be extracted from the channel providing F2,F3 information, while voicing information can only be obtained from the fully voiced F1 transition supplied to the opposite ear. It should be pointed out, however, that the potency of aspiration of the higher formants as a cue to voicelessness may not be very great, and no formal control experiment with F2 and F3 presented alone was run to measure the strength of this feature.

DISCUSSION

The results of these two experiments confirm the evidence from previous studies (Nye, Nearey, and Rand, 1974; Rand, 1974) that release from masking occurs under the dichotic mode of presentation, in which F1 is sent to one ear and the higher formants are sent to the other ear. Further evidence indicates that fusion is in fact taking place under the dichotic presentation. This is seen most clearly for vowels in the first experiment in which fusion of the higher formants with F1 is found to occur in the dichotic condition. The mixed stimuli of the second experiment provide additional evidence for fusion, this time in the case of consonants. Although the experimental design in the second experiment did not incorporate a control condition which would have conclusively demonstrated that F2,F3 of the mixed stimuli are strong, voiceless cues when

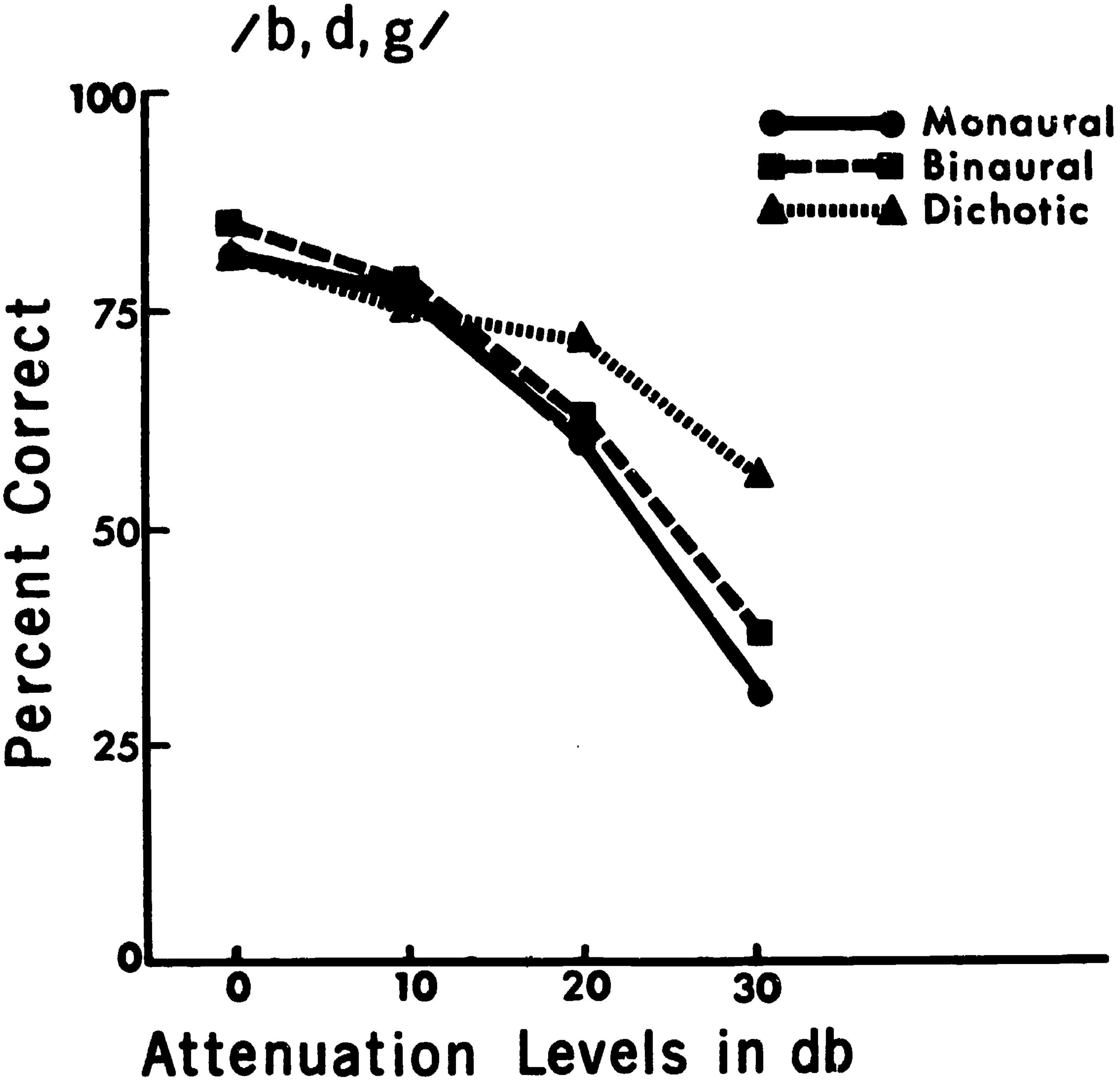

/b, d, g/
100
75
50
25
0
Percent Correct
0
10
20
30
Attenuation Levels in db
Monaural
Binaural
Dichotic

FIGURE 4

heard alone, subsequent testing with the mixed stimuli has shown that they are indeed heard as voiceless consonants. The extremely high incidence of voiced responses to these stimuli in the dichotic condition thus indicates subjective fusion of the voiced F1 component with the voiceless F2 and F3 components.

In the binaural conditicns of these experiments voiced F1 clearly acts as a strong mask on the attenuated higher formants, but the "cutback" F1 has no similar effect. However, whether the F1 transition alone is masking the higher formant transitions or whether there is backward masking of the steady-state F1 on the higher formant transitions is not clear. In order to provide further evidence about the type of masking that occurs, F1 can be temporally offset with respect to the higher formants so that it either precedes or follows F2 and F3 at equal intervals. This procedure should help to determine whether the F1 transition or the F1 steady-state portion of the vowel is the more effective masker of the higher formants, and whether backward masking in addition to simultaneous masking occurs in the binaural condition, and possibly in the dichotic condition. Research is currently underway to seek evidence of these possible masking effects.

REFERENCES

Ainsworth, W. A. and J. B. Millar. (1972) The effect of relative formant amplitudes on the perceived identity of synthetic vowels. Lang. Speech 15, 328-341.

Broadbent, D. E. and P. Ladefoged. (1957) On the fusion of sounds reaching different sense organs. J. Acoust. Soc. Amer. 29, 708-710.

Carlson, R., B. Granström, and G. Fant. (1970) Some studies concerning perception of isolated vowels. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) STL-QPSR 2-3, 19-35.

Cutting, J. E. (1973) Phonological fusion of synthetic stimuli in dichotic and binaural presentation modes. Haskins Laboratories Status Report on Speech Research SR-34, 55-56.

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

Nye, P., T. Nearey, and T. Rand. (1974) Dichotic release from masking: Further results from studies with synthetic speech stimuli. Haskins Laboratories Status Report on Speech Research SR-37/38, 123-137.

Rand, T. (1974) Dichotic release from masking for speech. J. Acoust. Soc. Amer. 55, 678-680. [Also in Haskins Laboratories Status Report on Speech Research SR-33 (1973), 47-55.]

The Tones of Central Thai: Some Perceptual Experiments*

Arthur S. Abramson+
Haskins Laboratories, New Haven, Conn.

INTRODUCTION

In recent years, research into prosodic features as part of the effort to understand the nature of speech communication has become an increasingly conspicuous aspect of experimental phonetics (Fry, 1968; Lieberman, 1974). Within prosody, such features as phonologically distinctive stress, segment length, and tone are clearly central to the sound pattern of a language in that they differentiate linguistic expressions. These are usually of more immediate interest to the linguist than are other types of prosodic features, and it is usually easier to design experiments testing hypotheses concerning them.[1] Prosodic research of this kind should make contributions to general phonology and, more narrowly, to our understanding of the phonology of a particular language, in this instance Thai.

The research reported here forms part of a larger project; other aspects of the project will be presented elsewhere. Taken together, these studies should furnish much information on the perceptually tolerable ranges of the tones of Thai. The work takes as its starting point the usual assumption that the major phonetic features of phonemic tone are found in the domain of pitch. The primary acoustic correlate of pitch is, of course, frequency. In instrumental analyses of tonal features, then, we measure the fundamental frequency of the voice as determined by the repetition rate of glottal pulsing.

*This is a slightly revised version of a chapter to be published in Studies in Tai Linguistics, ed. by Jimmy G. Harris and James Chamberlain. (Bangkok: Central Institute of English Language, in press).

+Also University of Connecticut, Storrs.

[1] This is not to assert that other realms of prosody, such as sentence intonation and rhythm, are irrelevant to the enterprise of phonetic research. Indeed, some important psycholinguistic questions have been raised in work on intonation; see, e.g., Lieberman (1967) and Studdert-Kennedy and Hadding (1973).

Acknowledgment: This research was conducted 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 the Faculty of Humanities, Ramkhamhaeng University, and the Central Institute of English Language, both in Bangkok.

[HASKINS LABORATORIES: Status Report on Speech Research SR-39/40 (1974)]

The language under analysis here is Central Thai (Siamese), the regional dialect that serves as the official language of Thailand. All native speakers used as subjects were from Bangkok or its close environs. The question of the tonal homogeneity of the central area of the kingdom has not been well explored[2] and thus cannot be categorically ruled out as a perturbing factor in some of the data presented; nevertheless, the speech of the test subjects gave an impression of general uniformity in tonal behavior.

One aim of this study was to determine how well the five tones of the language could be identified in isolation. It is at least conceivable that the identifiability of one or more of the tones would suffer without the benefit of an immediate context. Another aim was to reaffirm earlier work (Abramson, 1961, 1962) on the sufficiency of certain ideal fundamental frequency contours for the identification of the tones, using synthetic speech in which it would be possible to make frequency contours the only variable. In the expectation that less-than-perfect identification would be achieved with fundamental frequency alone, we also planned to learn whether the addition of variations in the amplitude of the speech signal would enhance the identifiability of the tones. Finally, we proposed to test the strong hypothesis that absolute fundamental frequency heights contribute nothing to the identification of the tones, while the shapes of the frequency contours carry all the information.

BACKGROUND WORK

Much of the present work is a continuation of earlier work done by the author (Abramson, 1962) with fewer informants and test subjects. That study showed (p. 128) that sets of tonally differentiated monosyllabic words, as produced by a single speaker, could be correctly identified nearly 100 percent of the time. In addition, fundamental frequency (f_0) measurements were taken from a large sampling of monosyllabic words with both short and long vowels, yielding average contours for the five tones (pp. 112-127).[3]

These average f_0 contours were then synthesized to see if Thai speakers could indeed identify each of the tones on the basis of f_0 alone. The synthesizer used was the Haskins Laboratories' Intonator (Abramson, 1962:20), which enabled the experimenter to analyze the spectrum of the speech signal and then resynthesize it on the machine's own "voice" source with new f_0 contours. Thus, most of the phonetic features of the original utterance are kept even when an f_0 contour is imposed. The five average tonal curves were thus imposed upon syllables that had originally carried the mid tone, namely /khaj/ 'dried sweat' and /loo/ 'unstable.'[4] The two perception tests prepared in this way exposed native Thai listeners to the tonal contours on both short and long vowels. The perceptual labeling of the randomized stimuli of these two tests showed clearly that the isolated f_0 contours provided sufficient cues for identifying the tones (pp. 131-132).

[2] The topic is being studied by Dr. Udom Warotamasikkhadit.

[3] It is gratifying to note that in data obtained some 14 years later (most of the tonal data for the 1962 publication were collected before 1960) Erickson (1974) provides general verification of the old contours while adding important information on the perturbing effects of initial consonants.

[4] In /loolee/.

In another experimental condition, five monosyllabic words minimally differentiated by tone were manipulated with the Intonator to yield 25 new syllables. That is, five new syllables were derived from each spoken syllable by removing the original tonal contour and replacing it with the synthetic contours, one by one, on this syllable carrier. This was done to see whether the curves carried enough information to override the effects of other features found in association with the pitch movements of the tones.[5] The results (pp. 131-134) included a small number of confusions apparently attributable to such concomitant features as variations in amplitude and duration, which would be likely to survive the analysis and resynthesis, although by and large the f_o curves were heard as intended. In general, then, the data supported the conclusion that the f_o contours isolated by means of acoustic analysis furnished sufficient cues by themselves for the identification of the five tones of Central Thai.

Experiment 1: Isolated Monosyllabic Words

Earlier work (Abramson, 1962:128) indicated that Thai listeners can identify the tones of monosyllabic utterances nearly perfectly. Although four sets of tonally differentiated words were used in that study, the productions of only one speaker furnished the stimuli. R. B. Noss (personal communication) has argued that generalizing from these results is, perhaps, not warranted and that normally the mid and low tones, at least, are difficult to identify unless they are embedded in a context. We thought that this objection might be handled by replicating the experiment with a somewhat larger number of speakers and listeners. In addition, data from responses to real speech were needed to furnish a standard for the later evaluation of results obtained with synthetic speech.

The following set of words was chosen for all the experiments to be described:

Tone	Thai Script	Transcription	Gloss
Mid	คา	/khaa/	'a grass (Imperata cylindrica)'
Low	ข่า	/khàa/	'galangal, a rhizome'
Falling	ข้า	/khâa/	'slave, servant'
High	ค้า	/kháa/	'to engage in trade'
Rising	ขา	/khǎa/	'leg'

The tone names are conventional but not fully descriptive. Ten native speakers of Central Thai recorded three or four randomizations each of the list with short pauses such that there were five tokens of each word in each randomization. The speakers, five men and five women, included nine university instructors and one clerk.

The tests were played to 25 native speakers of Central Thai through headphones in the language laboratory (Tandberg Teaching System IS 6 with Beyer

[5] This kind of information can preserve the tonal distinctions to a severely limited extent in whispered speech in which no f_o is present (Abramson, 1972).

DT98B Dynamic Headphones) of Ramkhamhaeng University, Bangkok, twice a week for a month. Only one test order was used for each of the ten speakers. Here and in later experiments the subjects were instructed to write the numerals 0, 1, 2, 3, or 4 for the mid, low, falling, high, and rising tones, respectively. These numbers are appropriate to the nomenclature of traditional Thai grammar and facilitated later scoring. Familiar as the subjects were with this convention, they were nevertheless provided with a sheet at each session showing the five words in Thai script with their numerical equivalents.[6]

The results are shown in Table 1, which is arranged as a confusion matrix with the tonal stimuli in the first column and the perceptual responses to them in the next five columns. The sixth column shows the total number of responses to each stimulus word. A perfect response to the tones as intended would yield 100 percent in each cell along the diagonal from the upper left to the lower right. The overall intelligibility of 98.6 percent, 85 errors out of 6100 responses, is high. Inspection of the responses to the mid and low tones indicates that some confusion between them accounts for most of the small number of errors.

TABLE 1: Confusion matrix of real-speech responses.

% Responses

	Labels:	Mid	Low	Falling	High	Rising	N
Stimuli	Mid	97.9	2.1				1220
	Low	3.4	96.6				1220
	Falling		0.2	99.1	0.4	0.3	1220
	High				100.		1220
	Rising		0.1	0.4		99.5	1220

Total = 6100
Subjects = 25
% Correct = 98.6

The data were also examined for individual differences among speakers and listeners. One of the ten speakers caused 45.9 percent of the errors, another speaker, 16.5 percent, and a third speaker, 12.9 percent. The recordings of only one speaker produced no errors at all. As for intersubject differences, the worst listener made 12.9 percent of the errors, followed by two who each made 8.2 percent of the errors. Three of the 25 subjects made no errors at all.

[6] 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.

This larger sampling of speakers and listeners supports the earlier conclusion that Thai words minimally differentiated by tones can nearly always be identified correctly even as isolated forms. Some speakers appear to provide minimal perceptual cues for words out of context, especially for the contrast between mid and low tones. Some listeners seem to require more than these minimal cues for the identification task. Finally, the data in Table 1 provide a baseline for the other experiments now to be discussed.

SYNTHETIC SPEECH

Experiment 2: Perception of f_0 Contours

To test once again the perceptual efficacy of the f_0 contours derived from speech measurements in the earlier study (Abramson, 1962:112-127), a different speech synthesizer, the Haskins Laboratories' formant synthesizer, was used under control of a computer. For the present experiment, the parameters specified were the frequency and amplitude values of the first three formants, the timing of source functions for voicing and voicelessness, the overall amplitude of the signal, and the fundamental frequency (f_0). Steady-state formant frequencies were chosen to yield a vowel acceptable as Thai /aa/; formant transitions (Liberman, Delattre, Cooper, and Gerstman, 1954) appropriate to the velar place of articulation were used, and voiceless aspiration was simulated by providing a long voicing lag (Lisker and Abramson, 1964, 1970) filled with turbulent noise in the regions of the formants. The overall amplitude was kept flat throughout the syllable except for a slight rise at the beginning and a slight fall at the end. These specifications yielded syllables of the type [kha:] which, it was hoped, with suitable f_0 contours would be heard as the five Thai words listed in the section on Isolated Monosyllabic Words. The five f_0 contours of the 1962 study (Abramson, 1962:127, Fig. 3.6) were retained with slight adaptations required by the nature of the computer program and imposed one-by-one upon this syllable. The contours, shown in the upper part of Figure 1, covered a range from 92 Hz to 152 Hz, which was reasonable for an adult male voice. Three tokens of each stimulus type thus produced were randomized into six test orders and played to 38 native speakers of Central Thai over a period of a month, together with other tests in the same sessions.

The results of these listening tests are shown in Table 2. Note that the tonal names in the first column are written with quotation marks. This is meant to convey that these f_0 contours were intended as those tones but can be so labeled only to the extent that the subjects accept them as such. In the same spirit, the word correct at the bottom of the table is also printed with quotation marks. Otherwise, the form of the confusion matrix is the same as that of Table 1.

Experiment 3: f_0 Plus Amplitude

Changes in the contraction of certain laryngeal muscles[7] and in subglottal air pressure can separately or together produce variations in the fundamental

[7] A University of Connecticut doctoral dissertation by Donna Erickson, soon to be completed, Laryngeal Mechanisms and Coarticulation Effects in the Tones of Thai, explores the role of intrinsic and extrinsic laryngeal muscles in the production of the tones of Thai.

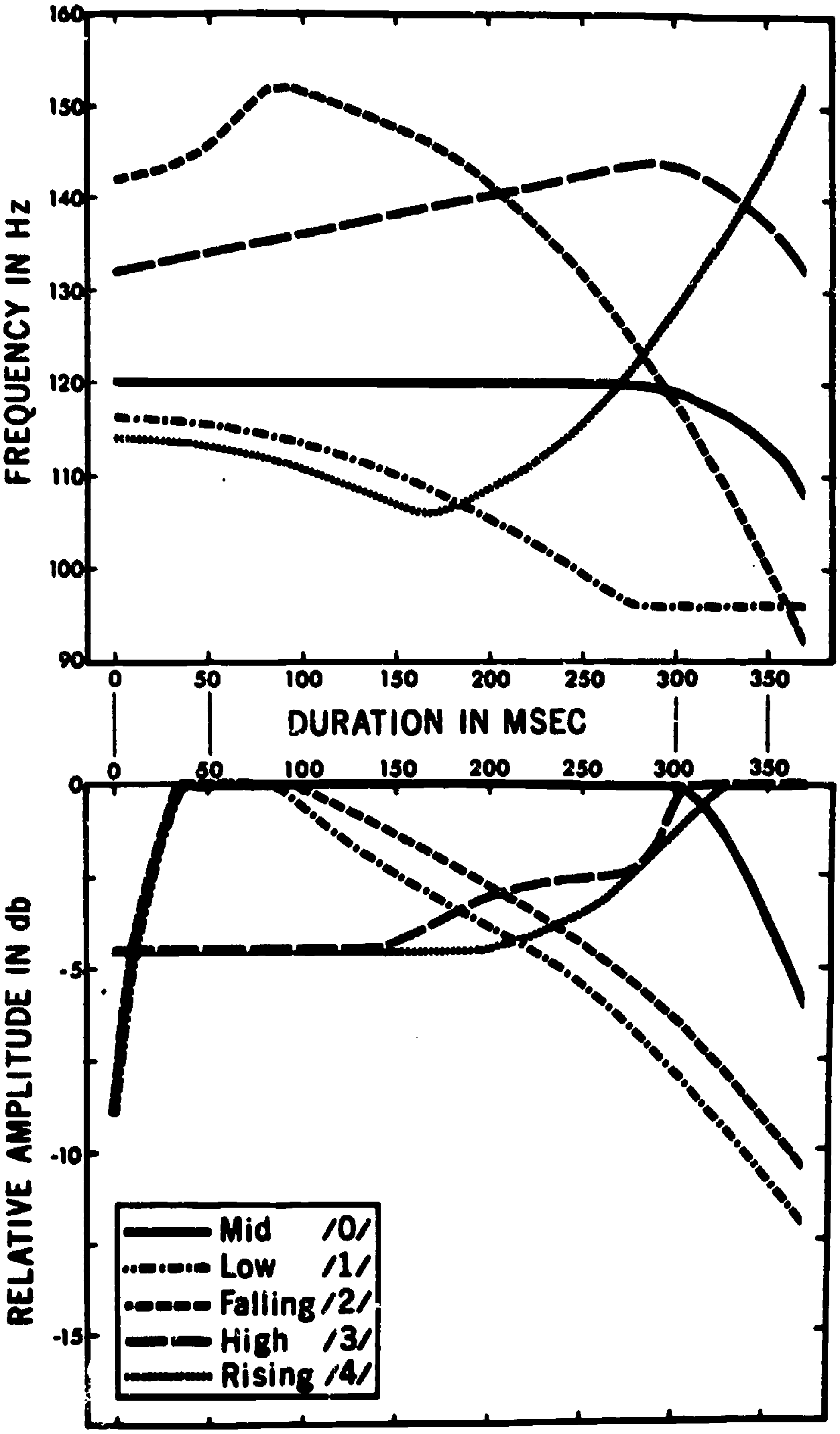


Figure 1: Upper part: f_o contours used in Experiment 2. Lower part: amplitude contours used in Experiment 3 in conjunction with the f_o contours.

TABLE 2: Synthetic speech: responses to five f_0 contours.

% Responses

	Labels:	Mid	Low	Falling	High	Rising	N
Stimuli	"Mid"	82.0	3.5	0.2	14.2		906
	"Low"	7.2	87.3	5.2	0.3		906
	"Falling"	0.4	0.7	97.8	0.6	0.6	906
	"High"	2.0		0.2	97.7	0.1	906
	"Rising"	0.2	0.2	0.3	0.1	99.1	906

Total = 4530
Subjects = 38
% "Correct" = 92.8

frequency of the voice. These mechanisms are also available for controlling the intensity of phonation and thus variations in the overall amplitude of the speech signal. To a certain extent, then, the two acoustic features, f_0 and amplitude, may covary. Since the major psychological correlate of amplitude or intensity is loudness, just as that of f_0 is pitch, it is not unreasonable to suppose that the ear may detect shifts in loudness in conjunction with large pitch excursions of one or more of the tones. If this is so, even though it has already been demonstrated with synthetic speech that certain ideal f_0 contours carry sufficient information for the identification of the five tones, the perceptual processing of some of the tones in real speech may include awareness of relative amplitude as a concomitant feature which, under certain conditions, may actually contribute to tonal identification. In the present study the question arises as to whether the discrepancies in intelligibility between Experiments 1 and 2 will be removed simply by the addition of appropriate amplitude contours.

To answer this question, we added the amplitude contours of the lower part of Figure 1 to the corresponding f_0 contours of the upper part. Amplitude is indicated in decibels (db) as decrements from the maximum output of the synthesizer at zero db. These amplitude contours are approximations derived from inspection of a small sample of amplitude displays made with sound spectrograms of Thai words in isolation. The new synthetic stimuli were randomized five times each into three test orders and played to 40 native speakers of Central Thai over a period of a month, together with other tests. The results are shown in the confusion matrix of Table 3.

Comparison of the Three Conditions

Inspection of Tables 1-3 shows that the overall identifiability of the stimuli moves from 98.6 percent for real speech through 92.8 percent for fundamental frequency alone to 96.1 percent for f_0 plus amplitude, thus suggesting that while f_0 alone is by and large a sufficient cue, its efficacy is enhanced

TABLE 3: Synthetic speech: responses to f_o + amplitude.

	Labels:	Mid	Low	Falling	High	Rising	N
Stimuli	"Mid"	95.7	1.0	0.2	3.0	0.1	1555
	"Low"	7.3	91.2	1.2	0.1	0.2	1555
	"Falling"	0.1	0.5	96.6	2.5	0.3	1555
	"High"	0.2	0.2	0.8	98.5	0.4	1555
	"Rising"	0.2	0.4	0.5	0.2	98.7	1555

Total = 7775
Subjects = 40
% "Correct" = 96.1

by the addition of amplitude information.[8] A comparison of the correct mean scores in the diagonals across the matrices is more to the point than just the overall scores. As we move from real speech (Table 1) to f_o alone (Table 2), the most striking changes are in the cells for the mid and low tones which lose, respectively, 15.9 percent and 9.3 percent. In Table 3, for f_o combined with amplitude information, the entries in the corresponding two cells move back in the direction of real speech, although the improvement is considerably greater for the mid tone.

Turning to the confusions in the matrices, we note a much greater scattering of errors in the two synthetic speech experiments as compared with real speech. Specifically, in Table 2 there is some confusion between the mid and high tones. The intended mid tone is called high 14.2 percent of the time. There is also some confusion in the other direction, high heard as mid, but only 2 percent of the time. Table 3 shows that the addition of amplitude information eliminates most of the confusion between these two tones. The only notable confusion in the real-speech test of Table 1 is between the mid and low tones. This confusion is even worse in Table 2. Under both conditions the hearing of the intended low tone as mid accounts for most of the confusion. This latter effect is not eliminated by the addition of amplitude information in Experiment 3, although the identification of the intended mid tone itself is now improved to the level of real speech or slightly better. The intended low tone is heard as falling 5.2 percent of the time in Table 2 but is nearly back to the level of real speech in Table 3. One small but puzzling distortion apparently caused by

[8] An experiment not performed as part of this research would be to try amplitude contours alone and then supplement them with f_o information. Previous research (Abramson, 1972) implied that amplitude alone would be not nearly sufficient for perception of the tones.

the addition of amplitude information is the hearing of the intended falling tone as high 2.5 percent of the time in Table 3.[9]

It seems safe to infer from the results of the preceding experiments that f_0 contours carry most of the information for tonal identification in Thai; that is, they carry sufficient information most of the time to identify words that are minimally distinguished by tone. A concomitant feature of some relevance for at least part of the tone system is contained in changes in the overall amplitude of the utterance. The confusions in the various matrices indicate the need for further information to improve the synthesis of Thai tones by rule. The improvements needed may be small refinements of the f_0 contours and better amplitude specification. In addition, simulating glottal tension in the voice source might make the high tone more natural and acceptable in utterance-final position.

Experiment 4: Perception of f_0 Levels

The five tones of Central Thai can be viewed as falling into two groups, the dynamic tones and the static tones (Abramson, 1962:9-11). In this scheme, the sharp upward and downward movements of the rising and falling tones place them in the dynamic category. Since the high, mid, and low tones often sound as if they simply occupy three levels, they are classified as static. Of course, the acoustical measurements, as reflected in the upper part of Figure 1, show that even the static tones undergo some f_0 movement. Kenneth Pike (1948:5) speaks of level tonemes and gliding tonemes: "a LEVEL toneme is one in which, within the limits of perception, the pitch of the syllable does not rise or fall during its production. A GLIDING toneme is one in which during the pronunciation of the syllable in which it occurs there is a perceptible rise or fall, or some combination of rise and fall, such as rising-falling or falling-rising." It may be the case, then, that speakers of the language, not to mention field phoneticians, hear the static tones of Thai as simple levels. A lengthy quotation from Pike (1948:4) is of considerable interest here:

> Tone languages have a major characteristic in common: It is the relative height of their tonemes, not their actual pitch, which is pertinent to their linguistic analysis. It is immaterial to know the number of vibrations per second of a certain syllable. The important feature is the relative height of a syllable in relation to preceding and following syllables. It is even immaterial, on this level of analysis (but not in the analysis of the linguistic expression of emotion), to know the height of a specific syllable in proportion to the general average pitch which the speaker uses. Rather, one must

[9] The effects set forth here seem obvious from simple inspection of the tables. Indeed, statistical analysis (t-tests for differences between correlated means) performed with the kind help of Dr. Lyle Bachman of the Central Institute of English Language and the Ford Foundation, Bangkok, shows them by and large to be significant at the 5 percent level of confidence or better. More refined statistical procedures might lead to further observations, but such effects would probably be so subtle as to be uninteresting for our understanding of the perception of tones.

> know the relationship of one specific syllable to the other syllables in the specific context in the particular utterance. A man and a woman may both use the same tonemes, even though they speak on different general levels of pitch. Either of them may retain the same tonemes while lowering or raising the voice in general, since it is the relative pitch of syllables within the immediate context that constitutes the essence of tonemic contrast.

Ilse Lehiste (1970:Chap. 3) provides a useful survey of these matters and related questions.[10]

How likely is it that absolute values of f_0 levels provide sufficient cues for the recognition of the tones of a given speaker of a tone language in a certain context? The relativity of the pitches of tones (Cook, 1972) is usually taken for granted. Indeed, the few studies that have yielded acoustic measurements of tones, e.g., for Mandarin Chinese (Howie, 1972) and Thai (Abramson, 1962; Erickson, 1974), show that the tones are characterized by at least some movement and, in some cases, much movement. That is, they tend not to be uttered with a flat, unchanging f_0. The one tone in Thai that appears most likely to have a flat f_0 in certain nonfinal environments is the mid tone. In an interesting experiment, Victor Zue (discussed in Klatt, 1973) demonstrated that Howie's Mandarin contours still showed rather high intelligibility even when the range of absolute f_0 is severely compressed, thus indicating that the pitch movement still available to the listeners carried sufficient information.

Of course, most of the foregoing observations are derived from tones in isolated words or at least in very short utterances. It seems likely that at least for some of the tones of Thai, presumably the high, mid, and low, the perturbations of their f_0 contours occasioned by the many coarticulations of running speech should produce flat variants here and there (Abramson, in preparation). If so, are such words understood by virtue of contextual redundancy or, to return to the question raised in the preceding paragraph, do the absolute levels furnish sufficient cues? That is, are "level" pitches assigned to tones only when they are the perceptual responses to small f_0 movements, or will true acoustic levels suffice? Experiment 4 was designed to provide some answers to the question.

As shown in Figure 1, the "voice" of the synthesizer in the experiments reported so far was set to range from 152 Hz down to 92 Hz. In this experiment the range was divided to produce 16 flat fundamental frequencies in steps of 4 Hz; the amplitudes were flat too, except for a slight rise at the beginning and a slight fall at the end. These variants were imposed upon the same basic syllable, randomized, and played to 37 native speakers of Thai for identification as members of the same set of five words as before. Table 4 reveals that the falling and rising tones, which are characterized by very dynamic movements, elicited practically no responses. Indeed, the 0.1 percent response to the 100 Hz level as the rising tone is so improbable as to suggest momentary inattention. From top to bottom, we have here a gradual crossover from the high

[10] An important background article for the phonological treatment of tone is Wang (1967).

TABLE 4: Synthetic speech: responses to 16 level f_0's.

% Responses

	Hz	Mid	Low	Falling	High	Rising	N
Stimuli	152	8.0	4.1	0.2	87.7		903
	148	7.9	4.2		87.6		903
	144	8.6	4.1	0.2	87.0		903
	140	12.1	4.3	0.1	83.5		903
	136	18.6	5.4	0.1	75.9		903
	132	29.2	5.4		65.3		903
	128	49.3	6.2		44.5		903
	124	65.3	5.5	0.1	29.0		903
	120	72.6	6.3	0.1	20.9		903
	116	73.0	12.2		14.7		903
	112	66.4	19.7		13.8		903
	108	42.1	45.7		12.2		903
	104	18.4	73.8		7.9		903
	100	11.0	81.5	0.3	7.1	0.1	903
	96	5.5	88.7	0.1	5.6		903
	92	4.8	90.1		5.1		903

Total = 14,448
Subjects = 37

tone through the mid tone to the low tone. Nowhere is 100 percent identification as a particular tone achieved. The closest is a peak of 90.1 percent for the low tone, which is comparable with the peak of 87.3 percent for the low tone in Experiment 2 (Table 2). The other two peaks here, 73 percent for the mid tone and 87.7 percent for the high tone, compare somewhat less well with their counterparts, 82 percent and 97.7 percent respectively, among the ideal contours of Experiment 2. Despite these peaks, it is important to note that all three tones persist in eliciting responses throughout their ranges. Most of this is accounted for not by the sporadic responses of all the subjects but rather the deviant response behavior of three of them. One subject agreed with the main group in calling the upper part of the range the high tone, but at 120 Hz she started crossing over to the low tone and remained there the rest of the way down with only scattered responses in the mid-tone column. The second subject of the three called the upper part of the range the mid tone and crossed over to the low tone at 108 Hz. The third subject deviated in the most surprising way from the performance of the main group: she assigned the upper part of the range to the low tone and, crossing over at about 120 Hz, the rest of the range to the high tone! (This subject's correct use of labeling conventions in other tests taken during the same sessions shows that she is not guilty of misuse of labels here.) Apparently, however, her psychological set shifts from time to time, because in some of the test sessions she assigned variations in the lower part of the range to the mid tone.

We may infer from the results of Experiment 4 that even in isolated monosyllabic words unchanging levels of fundamental frequency can carry considerable information as to the identity of the high, mid, and low tones. We suppose that in such a situation some accommodation to the speaker's pitch range is necessary. The subjects who took these tests were quite used to the "voice" of the synthesizer, and care was taken to confine the absolute levels of f_0 to the range already in use in other tonal experiments. At the same time, the fact that there is no gliding movement at all in the stimuli seems to cause a certain amount of confusion across the three categories that the subjects accepted. Indeed, for three of the 37 subjects this factor was perceptually very disrupting. They may represent a population of Thai speakers for whom the small gliding movements normally found in productions of the static tones in isolation are essential for correct identification. In the absence of any f_0 movement at all, it is not surprising that the falling and rising tones were not used as response categories.

Conclusion

The experiments described here lead to a few general conclusions about some aspects of the perception of the tones of Thai. First of all, it is more evident than heretofore that the intelligibility of tonally differentiated monosyllables presented in isolation is quite high. In addition, the average fundamental frequency contours obtained some years ago (Abramson, 1961, 1962), when applied as instructions to the parameters of a different synthesizer (and therefore presumably other synthesizers), still carry enough information for acceptable synthesis of Thai words. A slight reduction of the discrepancy in intelligibility between these contours and those of real speech can be effected by adding rough approximations of natural amplitude movements often found in correlation with shifts in fundamental frequency. To eliminate the small remaining discrepancy further work is needed. Finally, a continuum of level fundamental frequencies can be divided perceptually by native speakers of Thai into the high, mid, and low tones, but with very gradual transitions and with rather aberrant response patterns on the part of some test subjects; these data suggest that while levels, even in isolated syllables, can carry much information about these tones, there is still a fair amount of interference from the abnormal lack of change in fundamental frequency.

Given the paucity of similar perceptual data for other tone languages,[11] it is hard to say how we may generalize these findings beyond the Thai language. In fact, even within Thai there is little information about other major regional dialects.[12] Such phonetic features as "creaky voice" or "glottal tension" prominent in some tonal systems will probably require parameters other than simple control of fundamental frequency or amplitude for experimental investigation. Glottal tension is found in the high tone of Central Thai but appears to be unstable. Reports on more complicated manipulations of fundamental frequency to elucidate further the nature of tonal perception in Thai will be forthcoming.

REFERENCES

Abramson, A. S. (1961) Identification and discrimination of phonemic tones. J. Acoust. Soc. Amer. 33, 842(A).

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

Abramson, A. S. (1972) Tonal experiments with whispered Thai. In Papers in Linguistics and Phonetics to the Memory of Pierre Delattre, ed. by A. Valdman. (The Hague: Mouton) 31-44.

Abramson, A. S. (in preparation) The coarticulation of tones: An acoustic study of Thai. In Proceedings of the Eighth International Congress of Phonetic Sciences, Leeds, 1975.

Chan Yuen Yuen, Angela. (1971) A perceptual study of tones of Cantonese. Ph.D. dissertation, University College, London.

Cook, E-D. (1972) On the relativity of tones. Lingua 29, 30-37.

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. (in preparation) Laryngeal mechanisms and coarticulation effects in the tones of Thai. Ph.D. dissertation, University of Connecticut.

Fry, D. B. (1968) Prosodic phenomena. In Manual of Phonetics, ed. by B. Malmberg. (Amsterdam: North Holland) 365-410.

Howie, J. M. (1970) The vowels and tones of Mandarin Chinese: Acoustical measurements and experiments. Ph.D. dissertation, Indiana University.

Howie, J. M. (1972) Some experiments on the perception of Mandarin tones. In Proceedings of the Seventh International Congress of Phonetic Sciences, ed. by A. Rigault and R. Charbonneau. (The Hague: Mouton) 900-904.

[11] The most comparable study that comes to mind is Howie's (1972) work on Mandarin; for more detail see his doctoral dissertation (1970). For a study of Cantonese, see Chan (1971).

[12] Field work about to be started in Thailand by Jack Gandour of the University of California at Los Angeles should yield information along these lines.

Klatt, D. (1973) Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception. J. Acoust. Soc. Amer. 53, 8-16.

Lehiste, I. (1970) Suprasegmentals. (Cambridge, Mass.: MIT Press).

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, No. 8 (Whole No. 379).

Lieberman, P. (1967) Intonation, Perception, and Language. (Cambridge, Mass.: MIT Press).

Lieberman, P. (1974) A study of prosodic features. In Current Trends in Linguistics, Vol. 12, ed. by T. Sebeok et al. (The Hague: Mouton) 2419-2449.

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 Sixth International Congress of Phonetic Sciences. (Prague: Academia) 563-567.

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

Studdert-Kennedy, M. and K. Hadding. (1973) Auditory and linguistic processes in the perception of intonation contours. Lang. Speech 16, 293-313.

Wang, W. S-Y. (1967) Phonological features of tone. Int. J. Amer. Ling. 33, 93-105.

Phonetic Segmentation and Recoding in the Beginning Reader*

Isabelle Y. Liberman,[+] Donald Shankweiler,[++] Alvin M. Liberman,[++] Carol Fowler,[+] and F. William Fischer[+]

The beginning reader--the child of six or thereabouts--is an accomplished speaker-hearer of his language and has been for a year or more. Why, then, should he find it hard to read, as so many children do? Why does he not learn to read as naturally and inevitably as he learned to speak and listen? What other abilities, not required for mastery of speech, must he have if he is to cope with language in its written form?

If the beginning reader is to take greatest advantage of an alphabet and of the language processes he already has, he must convert print to speech or, more covertly, to the phonetic structure that, in some neurological form, must be presumed to underlie and control overt speech articulation. In the first part of the paper we will say why it might be hard to make the conversion properly--that is, so as to gain all the advantages that an alphabetic system offers. But the conversion from print to speech, whether properly made or not, may also be important to the child in reducing what is read to a meaningful message. This is so because of a basic characteristic of language: the meaning of the longer segments (for example, sentences) transcends the meaning of the shorter segments (for example, words) out of which they are formed. From that it follows that the shorter segments must be held in some short-term store until the meaning of the longer segments has been computed. In the second part of the paper we will consider the possibility that a phonetic representation is particularly suited to that requirement.

In referring to the conversion of print to speech, which is what much of this paper is about, we will not be especially concerned to make a distinction between overt speech and the covert neurological processes (isomorphic, presumably, with the phonetic representation) that govern its production and perception. We should only note that the beginning reader often converts to overt speech and the skilled reader to some more covert form. We should also note that the conversion to the covert form does not, of course, limit the reader to the relatively slow rates at which he can overtly articulate. We will also not be concerned with the distinction between the phonetic and the more abstract

*To appear in Reading: The CUNY Conference (tentative title), ed. by A. S. Reber and D. Scarborough (New York: Erlbaum Associates).

[+]University of Connecticut, Storrs.

[++]University of Connecticut, Storrs, and Haskins Laboratories, New Haven, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-39/40 (1974)]

phonological representations. Like many alphabetically written languages, English makes contact, not at the phonetic level, but at some more abstract remove, closer surely to the level of systematic phonologic structure (Chomsky, 1970; Klima, 1972) or, in the older terminology, to the phonemic and morphophonemic levels. That is an important consideration to students of the reading process, but it happens not to be especially relevant to our purposes in this paper. For convenience, then, we will speak of phonemes, phonetic segments, and phonetic structure without implying any differences in the abstractness of the units being referred to.

USING THE ALPHABET TO FULL ADVANTAGE

The need to segment phonetically

For the moment we will concern ourselves only with the first problem: what a child needs in order to read an alphabetic language properly. In that connection, let us look at the strategies the beginning reader might use to recover a phonetic representation of the written word. There are at least two possibilities: the child might work analytically, by first relating the orthographic components of the written word to the segmental structure of the spoken word, or he might do it holistically, as in the whole-word method, by simply associating the overall shape of the written word with the appropriate spoken word. In the whole-word strategy, the child not only does not need to analyze words into their phonetic components, but need not necessarily even be aware that such an analysis can be made. There are, however, many problems with this strategy. An obvious one, of course, is that it is self-limiting; it does not permit the child to take advantage of the fact that his language is written alphabetically. In the whole-word strategy, each new word must be learned as a unit, as if it were an ideographic character, before it can be read. Only if the child uses the more analytic strategy can he realize the important advantages of an alphabetically written language. Thus, given a word which he has heard or which is already in his lexicon, the child can read it without specific instruction, though he has never before seen it in print; or, given a new word which he has never before heard or seen, the child can closely approximate its spoken form and hold that until its meaning can be inferred from the context or discovered later by asking someone about it. In connection with the latter advantage, one might ask why the child cannot similarly hold the word in visual form. Perhaps he can. We know, however, that the spoken form can be retained quite easily and, indeed, that it can readily be called up. As to what can be done with a purely visual representation, we are not so sure. At all events, and as we will say at greater length later, spoken language, or its underlying and covert phonetic representation, seems particularly suited for storage of the short-term variety.

What special ability does the child need, then, if he is to employ the analytic strategy and thus take full advantage of the alphabetic way our language is written? In our view, it is the ability to make explicit the phonetic segmentation of his own speech. Consider, for example, what is involved in reading a simple word like *bag*. Let us assume that the child can identify the three letters of the word, and further, that he knows the individual letter-to-sound correspondences--the sound of *b* is /bʌ/, the sound of *a* is /æ/, and *g* is /gʌ/. If that is all he knows, however, he will sound out the word as *buhaguh*, a nonsense trisyllable containing five phonetic segments, and not as *bag*, a meaningful monosyllable with only three phonetic segments. If he is to map the printed, three-letter word bag onto the spoken word bag, which is already in his lexicon, he must know that the spoken syllable also has three segments.

The difficulties of making phonetic segmentation explicit

Given that the child must be able to make explicit the phonetic segmentation of the word, is there any reason to believe tha ne might encounter difficulties? There is, indeed, and it comes directly from research on acoustic cues for speech perception--the finding that there is, most commonly, no acoustic criterion by which the phonetic segmentation of a given word is dependably marked (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). Phoneme boundaries are not marked acoustically because the segments of the phcnetic message are often coarticulated, with the result, for example, that a consonant segment will, at the acoustic level, be encoded into--that is, merged with--the vowel. The word bag, for example, has three phonetic segments but only one acoustic segment. Thus, there is no acoustic criterion by which one can segment the word into its three constituent phonemes. Analyzing an utterance into syllables, on the other hand, may present a different and easier problem. We should expect that to be so because every syllable contains a vocalic nucleus and thus will have, in most cases, a distinctive peak of acoustic energy. These energy peaks provide audible cues that correspond approximately to the syllable centers (Fletcher, 1929). Though such auditory cues could not in themselves help a listener to define exact syllable boundaries, they ought to make it relatively easy for him to discover how many syllables there are and, in that sense, to do explicit syllable segmentation.

We should remark here that the analytic strategy we have been talking about does not mean reading letter by letter. Indeed, if the child is using the analytic strategy he most certainly cannot read that way. Sounding out the letters would produce nonsense, as in the example of buhaguh (for bag) offered above. Given the way phonetic segments are encoded or merged at the level of sound, the spoken form can be recovered only if, before making the conversion, the reader takes into account all the letters that represent the several phonetic segments to be encoded. In the example of bag, the coding unit is obviously the syllable. But coding influences sometimes extend across syllables, and in the case of prosody such influences may cover quite long stretches. We think, therefore, that the number of letters that must be apprehended before attempting to recover the spoken form may sometimes be quite large. In fact, we do not now know exactly how large these coding units are, only that they almost always exceed one letter in length. To identify such units is, in our view, a research undertaking of great importance and correspondingly great difficulty.

It should also be emphasized here that the child who finds it difficult to make explicit the phonetic segmentation of his speech need not have any problems at all in the regular course of speaking and listening. Children generally distinguish (or identify) words like bad or bag, which differ in only one phonetic segment. Indeed, there is evidence now that infants at one month of age discriminate ba from pa (and da from ta) and, moreover, that they make this discrimination categorically, just as adults do (Eimas, Siqueland, Jusczyk, and Vigorito, 1971). The child has no difficulty in speaking and listening to speech because there the segmentation of the largely continuous acoustic signal is done for him automatically by operations of which he is not conscious. In order to speak and listen, therefore, he need have no more conscious awareness of phonetic structure than he has of syntactic structure. We all know that the child can speak a grammatical sentence without being able to verbalize the rules he is using to form that sentence. Similarly, he can readily distinguish bad from bag without being able to analyze the phonetic structure underlying the distinction--that is, without an explicit understanding of the fact that each of these utterances consists of three segments and that the difference lies wholly in the third. But reading, unlike speech, does require an explicit analysis if the advantages of an alphabet are to be realized.

That explicit phonetic analysis might be difficult is suggested also by the history of writing (Gelb, 1963). In the very earliest systems the segment that the orthography represented was the word. Present-day approximations to that kind of writing are to be found in Chinese characters and in the very similar kanji that the Japanese use. Writing with meaningless units is a more recent development, the segment size represented in all the earliest forms being the syllable. An alphabet, representing the shortest meaningless segments (phones or phonemes), developed still later and apparently out of a syllabary. Moreover, all the other systems, whether comprising meaningful or meaningless units, and of whatever size, seem to have appeared independently in various places and at various times, but all the alphabets are considered to have been derived from a single original invention. It seems reasonable to suppose that the historical development of writing systems--from word, to syllable, to phoneme--might reflect the ease or difficulty of explicitly carrying out the particular type of segmentation that each of these orthographies requires. More to the point of our present concerns, one would suppose that for the child there might be the same order of difficulty, and, correspondingly, the same order of appearance in development.

Development of the ability to analyze speech into phonemes and syllables

We thus have reason to suppose that phonetic segmentation might be a difficult task, more difficult than syllabic segmentation, and that the ability to do it might, therefore, develop later. To test that supposition directly, we recently conducted an experiment. The point was to determine how well children in nursery school, kindergarten, and first grade (four-, five-, and six-year-olds) can identify the number of phonetic segments in spoken utterances and how this compares with their ability to deal similarly with syllables (Liberman, Shankweiler, Fischer, and Carter, 1974). The procedure was in the form of a game which required the child to indicate, by tapping a wooden dowel on the table, the number (from one to three) of segments (phonemes in the case of one group, syllables in the other) in a list of test words. To teach the child what was expected of him, the test list was preceded by a series of training trials in which the experimenter demonstrated how the child was to respond. The test itself consisted of 42 randomly assorted individual items of one, two, or three segments, presented without prior demonstration and corrected, as needed, immediately after the child's response. Testing was continued through all 42 items or until the child reached a criterion of tapping six consecutive trials correctly without demonstration. The children of each grade level were divided into two experimental groups, the one requiring phoneme segmentation and the other, syllable segmentation. Instructions given the two groups were identical, except that the training and test items required phoneme segmentation in one group and syllable segmentation in the other.

The results showed in more than one way that the test words were more readily segmented into syllables than into phonemes. At all grade levels, the number of children who were able to reach criterion was markedly greater in the group required to segment by syllable than in the group required to segment by phoneme. At age four, none of the children could segment by phoneme, whereas nearly half could segment by syllable. Ability to carry out phoneme segmentation successfully did not appear until age five, and then it was demonstrated by only 17 percent of the children. In contrast, almost half of the children at that age could segment syllabically. Even at age six, only 70 percent succeeded in phoneme segmentation, while 90 percent were successful in the syllable task.

The proportions of children at each age who reached criterion level in the minimum number of trials is another measure of the contrast in difficulty of the two tasks. For the children who worked at the syllable task, the percentage reaching criterion in the minimum time increased steadily over the three age levels: 7 percent at age four, 16 percent at age five, and 50 percent at age six. By contrast, we find in the phoneme group that no child at any grade level attained the criterion in the minimum time.

The data were also analyzed in terms of mean errors. In Figure 1 mean errors to passing or failing a criterion of six consecutive correct trials without demonstration are plotted by task and grade. Errors on ɔoth the syllable

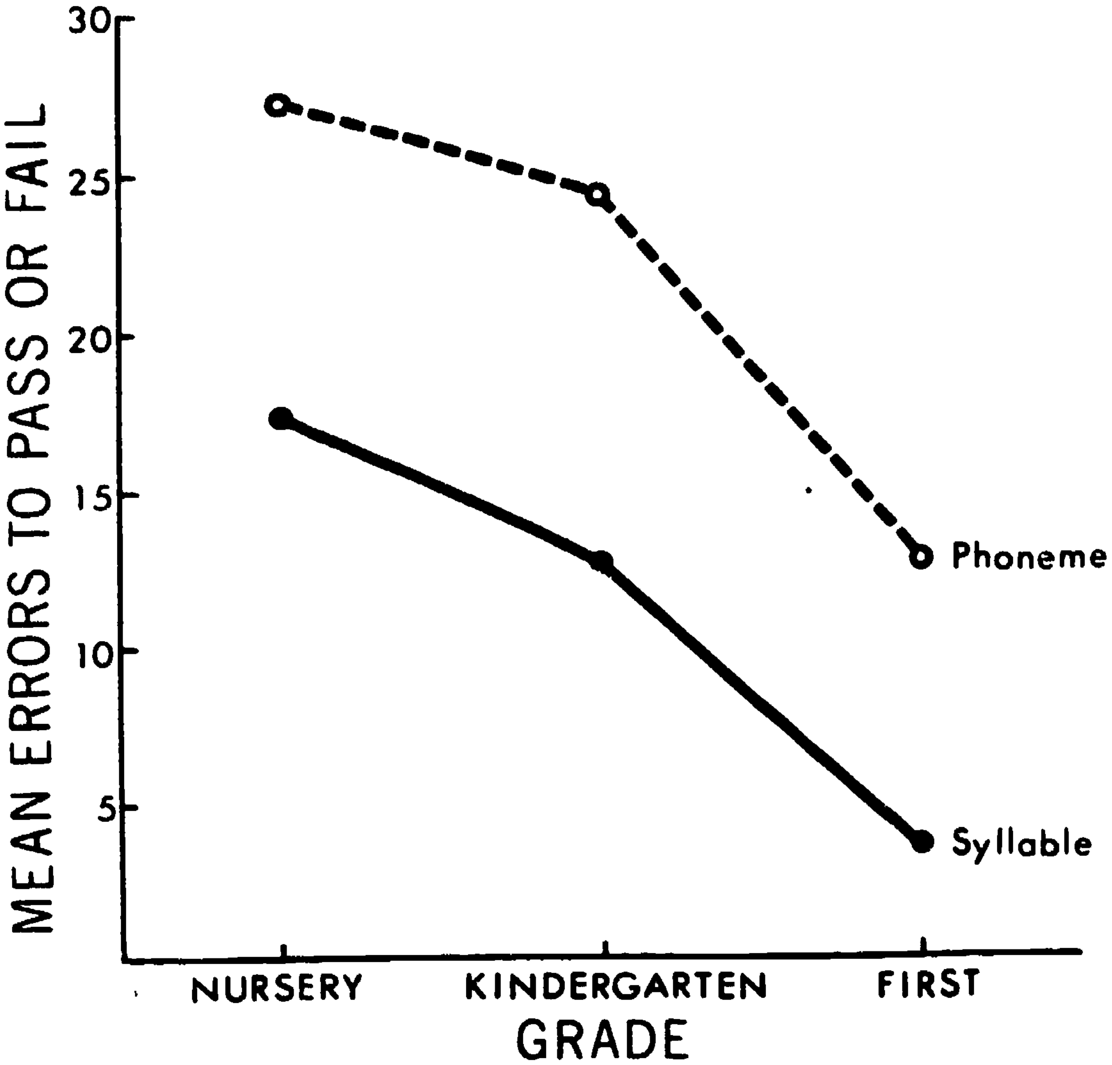


Figure 1: Mean number of errors to passing or failing a criterion of six consecutive trials without demonstration in phoneme and syllable segmentation.

and phoneme tasks decreased monotonically at successive grade levels, but the greater difficulty of phoneme segmentation at every level was again clearly demonstrated.

Segmentation and reading

The difficulty of phonetic segmentation has also been remarked by a number of other investigators (Rosner and Simon, 1970; Calfee, Chapman, and Venezky, 1972; Savin, 1972; Gleitman and Rozin, 1973; Elkonin, 1973; Gibson and Levin, in press). Their observations, together with ours described in the experiment above, also imply a connection between phonetic segmentation ability and early reading acquisition. This relationship is suggested in our experiment by the increase in number of children passing the phoneme-counting task, from only 17 percent at age five to 70 percent at age six. Unfortunately, the nature of the connection is in doubt. On the one hand, the increase in ability to segment phonetically might result from the reading instruction that begins between five and six. Alternatively, it might be a manifestation of some kind of intellectual maturation. The latter possibility might be tested by a developmental study of segmentation skills in a language community such as the Chinese, where the orthographic unit is the word and where reading instruction therefore does not demand the kind of phonetic analysis needed in an alphabetic system.[1]

In any event, since explicit phoneme segmentation is harder for the young child and develops later than syllable segmentation, one would expect that syllable-based writing systems would be easier to learn to read than those based on an alphabet. We may thus have an explanation for the assertion (Makita, 1968) that the Japanese kana, roughly a syllabary, is readily mastered by first-grade children. One might expect, furthermore, than an orthography which represents each word with a different character (as is the case in Chinese logographs and in the closely related Japanese kanji) would obviate the difficulties in initial learning that arise in mastering an alphabetic system. The relative ease with which reading-disabled children learn kanji-like representations of language while being unable to break the alphabetic code (Rozin, Poritsky, and Sotsky, 1971) may be cited here as evidence of the special burden imposed by an alphabetic script.

However, we need not go so far afield to collect indirect evidence that the difficulties of phoneme segmentation may be related to early reading acquisition. Such a relation can be inferred from the observation that children who are resistant to early reading instruction have problems even with spoken language when they are required to perform tasks demanding some rather explicit understanding of phonetic structure. Such children are reported (Monroe, 1932; Savin, 1972) to be deficient in rhyming, in recognizing that two different monosyllables may share the same first (or last) phoneme segment, and also in speaking Pig Latin, which demands a deliberate shift of the initial consonant segment of the word to initial position in a nonsense syllable added to the end of the word.

We, too, have explored directly, if in a preliminary way, the relation between ability to segment phonemes and reading. For that purpose we measured the reading achievement of the children who had taken part in our experiment on

[1] Unfortunately, a pure test will be hard to make. Children in the People's Republic of China are taught to read alphabetically before beginning their study of logographic characters.

phonetic segmentation, described above. Testing at the beginning of the second school year, we found that half of the children in the lowest third of the class in reading achievement (as measured by the word-recognition task of the Wide Range Achievement Test) had failed the phoneme segmentation task the previous June; on the other hand, there had been no failures in phoneme segmentation among the children who scored in the top third in reading ability (I. Liberman, 1973).

Data from the analysis of children's reading errors may also be cited as additional evidence for the view that explicit phoneme segmentation may be a serious roadblock to reading acquisition. If a chief source of reading difficulty is that the child cannot make explicit the phonetic structure of the language, he might be expected to show success with the initial letter--which requires no further analysis of the syllable--and relatively poor performance beyond that point. If he knows some letter-to-sound correspondences, and knows that he must scan in a left-to-right direction, he might simply search his lexicon for a word, any word, beginning with a phoneme that matches the initial letter. Thus, presented with the word bag, he might give the response butterfly. Such a response could not occur if the child were searching his lexicon for a word with three sound segments corresponding to the letter segments in the printed word. If, however, the child is unaware that words in his lexicon have a phonetic structure or if he has difficulty in determining what that structure is, then he will not be able to map the letters to the segments in these words. On these grounds, we would expect that in reading words such a child would make more errors on final consonants than on initial consonants. We have observed just this error pattern in a number of beginning and disabled readers aged seven to eleven (Shankweiler and Liberman, 1972; Liberman, 1973).

Further evidence comes from a recently completed study (Fowler, Liberman, and Shankweiler, in preparation) which showed that although the error rate in reading decreases markedly with grade level, the position effect (i.e., the discrepancy in error rate between initial and final consonants) is maintained as the child progresses through the early grades. The subjects in this study were second, third, and fourth grade children. The list of words to be read consisted of 38 monosyllables selected to give equal representation to the 19 consonant phonemes that can occur in both initial and final position in English words. Each phoneme was represented twice in the list in each position. The words were presented to the child singly to be read aloud to the best of his ability.

Analysis of the data shows final consonant errors to be at least twice as frequent as initial. At Grade 2, the rate of initial consonant errors (IC) was 8 percent as compared with 16 percent for final consonants (FC); at Grade 3, IC was 5 percent, FC 10 percent; at Grade 4, IC was 2 percent, FC 6 percent. It was clear that the FC-IC difference could not be accounted for by differences in the phonetic complexity of the consonants that tend to occur in initial and final position, because the consonant phonemes in the test list were controlled for frequency of occurrence and position in the word. But what about orthographic complexity? It was possible that the FC/IC difference might be due to the fact that a given phoneme occurring finally is spelled more complexly than that same phoneme in the initial position (g and j versus dge and ge). We therefore looked only at the errors on phonemes that are spelled simply (by a single letter) in both initial and final position (p,t,k,b,d,g,m,n,r). If the position effect had been due largely to orthographic complexity, it should have disappeared in this analysis. But it did not. Final consonants still produced more errors than initial.

It is clear that there is indeed a progression of difficulty with the position of the consonant segment in the word, the final consonants being more frequently misread than the initial. Similar findings have been reported by other investigators (Daniels and Diack, 1956; Weber, 1970) who examined error patterns in the reading of connected text. We found in an earlier study (Shankweiler and Liberman, 1972) that the initial-final difference cannot be a simple reflection of the error pattern in speech. There we presented, first for oral repetition and then for reading, a list of 204 monosyllables chosen to give equal representation to most of the consonants, consonant clusters, and vowels of English. The initial-final consonant error pattern was duplicated in reading, but in oral repetition the consonant errors were about equally distributed between initial and final position. Moreover, the initial-final error pattern in reading is also contrary to what would be expected in terms of sequential probabilities. If the child at the early stages of beginning to read were using the constraints built into the language, he would make fewer errors at the end than at the beginning of words, not more.

The contribution of orthographic complexity

In stressing the difficulty of phonemic segmentation, we do not intend to imply that no other problems are involved in reading an alphabetic language. For example, we realize that the mapping in English between spelling and language is sometimes complex and irregular.[2] Although that undoubtedly contributes to the difficulties of reading acquisition, we do not believe that the complexity of the orthography is the principal cause. Indeed, we know that it cannot be the only cause since many children continue to have problems even when the words are carefully chosen to include only those which map the sound in a consistent way and are part of the child's active vocabulary (Savin, 1972). Moveover, reading problems are known to occur in countries in which the writing system maps the language more directly than in English (Downing, 1973). In any event, the major irregularities of English spelling confronting the young child in the simple words he must read have to do mainly with the vowels.

Though we believe it to be of interest to examine the relation of orthographic complexity of the vowels to the problems of reading acquisition, and we are doing so (Fowler et al., in preparation), we suspect that getting the vowel exactly right may not be of critical importance in reading (though, of course, it is in spelling). If in the conversion to sound the child gets the phonetic structure correct except for errors in vowel color, he would not be too wide of mark, and many such errors would be rather easily corrected by context or by information obtained later.

[2] It is recognized that the "irregularities" of English spelling are more lawful than might appear, as in the spellings of "sign" and "signal," for example, which reflect morphological structure quite accurately (Chomsky, 1970). However, it must be said that this lawfulness can be appreciated only by the skilled reader and probably does not aid the beginner.

THE PHONETIC REPRESENTATION, SHORT-TERM MEMORY, AND READING

Phonetic recoding in reading as a way to tap primary language processes

Though beginning readers must surely recode phonetically if they are to cope with new words, we wonder what they do with words (and phrases) they have read many times. Do they, in those cases, construct a phonetic representation, using either of the two strategies we described earlier, or do they, as some believe (Bever and Bower, 1966), go directly from print to meaning?

One can think of at least two reasons why phonetic recoding might occur even with frequently read materials. A not very interesting reason is that, having adopted the phonetic strategy to gain advantages in the early stages of learning, the reader continues with the habit, though it may have ceased to be functional or may even have become, as some might think, a liability. There is a more interesting reason, however, and one we are inclined to take more seriously. It derives from the possibility that working from a phonetic base is natural and necessary if the reader (including even one who is highly practiced) is to take advantage of the primary language processes that are so deep in his experience and, indeed, in his biology. Consider, for example, that the normal processes for storing, indexing, and retrieving lexical entries may be carried out on a phonetic base. If so, it is hard to see why the reader should develop completely new processes, suited for the visual system, and less natural, presumably, for the linguistic purposes than the old ones. Or consider what we normally do in coping with syntax, an essential step in arriving at the meaning of a sentence. Though we do not know much about how we decode syntax, it is virtually certain that we are aided significantly by the prosody, which marks the syntactic boundaries. What, then, is the cost to our understanding of what we read if we do not recover the prosody, using for that purpose the marks of punctuation and such subtle cues as skillful writers may know how to provide (Bolinger, 1957)?

There are, of course, other natural language processes that the reader can exploit only by constructing a phonetic representation. Among them is short-term storage, and it is that process we will be concerned with in the remainder of this paper. As we pointed out earlier, it is characteristic of language that the meaning of longer segments (e.g., sentences) transcends the meaning of the shorter segments (words) from which they are formed. It follows, then, that the listener and reader must hold the shorter segments in some short-term store if the meaning of the longer segments is to be extracted from them. Given what we know about the characteristics of the phonetic representation, we might suppose that, as Liberman, Mattingly, and Turvey (1972) have suggested, it is uniquely suited to the short-term storage requirements of language. But apart from what we or they might suppose, there is relevant experimental evidence.

Phonetic representation of visually presented material in short-term memory

Some of the evidence comes from a class of experiments showing that when lists of letters or alphabetically written words are presented to be read and remembered, the confusions in short-term memory are phonetic rather than optical (Conrad, 1963, 1964, 1972; Sperling, 1963; Conrad and Hull, 1964; Conrad, Freeman, and Hull, 1965; Baddeley, 1966, 1968, 1970; Dornič, 1967; Hintzman, 1967; Kintsch and Buschke, 1969; Thomasson, 1970, reported in Conrad, 1972). From that finding it has been inferred that the stimulus items had been stored in phonetic rather than visual form. Indeed, the tendency to recode visually presented items into phonetic form is so strong that, as Conrad (1972) has emphasized, subjects consistently do so recode even in experimental situations in which it is clearly disadvantageous to do so.

A similar kind of experiment (Erickson, Mattingly, and Turvey, 1973) suggests that exactly the same kind of phonetic recoding occurs even when the linguistic stimuli are not presented in a form (alphabetic) that represents the phonetic structure. In that experiment the investigators used lists of kanji characters, which are essentially logographic, and Japanese subjects who were readers of kanji. As in the experiments with alphabetically spelled words, there was evidence that the stimulus items had been stored in short-term memory in phonetic rather than visual (or semantic) form.

There is also evidence that even nonlinguistic stimuli may, under some circumstances, be recoded into phonetic form and so stored in short-term memory. That evidence comes from work by Conrad (1972) who found that in short-term recall of pictures of common objects, confusions were clearly based on the phonetic forms of the names of the objects, rather than on their visual or semantic characteristics.

Though none of the experiments cited here dealt with natural reading situations, they are nevertheless relevant to the assumption that even skilled readers might recode phonetically, and that in so doing they might gain an advantage in short-term memory. It remains to be determined whether and to what extent readers rely on phonetic recoding for the short-term memory requirements of normal reading. Less generally, it remains to be determined also whether good and poor readers are distinguished by greater or lesser tendencies toward phonetic recoding. In the next section of this paper we will describe our first attempt to gain evidence on this question.

Phonetic recoding in good and poor beginning readers: an experiment

Given the short-term memory requirements of the reading task and evidence for the involvement of phonetic coding in short-term storage, we might expect to find that those beginning readers who are progressing well and those who are doing poorly will be further distinguished by the degree to which they rely on phonetic recoding. To our knowledge no one has investigated this possibility; consequently, we set out to do so. Our experiments will be described in detail elsewhere (Liberman, Shankweiler, Fowler, and Fischer, in preparation); here we will report briefly on only one experiment which is directly relevant to our present concerns.

In this experiment, we used a procedure similar to one devised by Conrad (1972) in which the subject's performance is compared on recall of phonetically confusable (rhyming) and nonconfusable (nonrhyming) letters. Our expectation was that phonetically similar items would maximize phonetic confusability and thus penalize recall in subjects who use the phonetic code in short-term memory. Sixteen strings of five upper-case letters were presented to the subjects by projector tachistoscope. Eight of the five-letter strings were composed of rhyming consonants (drawn from the set: B C D G P T V Z) and eight were composed of nonrhyming consonants (drawn from the set: H K L Q R S W Y). The two series of five-letter strings (confusable and nonconfusable) were randomly interleaved. An exposure time of 3 sec was adopted after preliminary studies had shown that even adult subjects require exposures in excess of 2 sec in order to report all five letters reliably. The test was given twice: first with immediate recall, then with delayed recall. In the first condition, recall was tested by having subjects print each letter string, in the order given, immediately after presentation. In order to make the task maximally sensitive to the recall strategy, we then imposed a 15-sec delay between tachistoscopic presentation and the response of writing down the string of letters.

As can be seen in Table 1, the subjects included three groups of school children who differed in level of attainment in reading as estimated by the word-recognition subtest of the Wide Range Achievement Test (WRAT). All were nearing completion of the sacond grade at the time the testr were conducted. There was no overlap in WRAT scores among the three groups. The first group, designated as the superior readers, comprised 17 children who were reading well ahead of their grade placement; they scored a mean grade equivalent of 4.9 on the WRAT. The second group, whom we call marginal readers, included 16 children who averaged slightly less than one half year of lag in reading achievement (grade 2.5). The third group, 13 children whom we call poor readers, obtained a mean WRAT equivalent of 2.0, indicating nearly a full year of retardation in reading. The three groups did not differ significantly in mean age. Their intelligence level, as measured by the Peabody Picture Vocabulary Test, was closely matched in the two extreme groups, the superior and poor readers. The difference in IQ level in the marginal group is apparently of no serious consequence since, as will be seen below, the performances of the marginal and poor groups on the experimental tasks were not appreciably different from each other.

TABLE 1: Estimated mean reading grade,* mean age, and IQ+ for second-grade school children grouped according to reading attainment.

Group	n	age	IQ	Reading Grade
Superior	17	8.0	113.9	4.9
Marginal	16	8.1	101.7	2.5
Poor	13	8.2	111.6	2.0

*Reading grade equivalent score on reading subtest of the Wide Range Achievement Test.

+Peabody Picture Vocabulary Test.

In Figure 2, which displays the data in terms of mean errors summed over all serial positions in the letter strings, the upper plot gives the results for superior readers, while the middle and lower plots show the results for the marginal and poor readers, respectively. We see at once that the main differences are between the superior readers and the other two groups. It was found, in fact, that the marginal and poor readers did not differ significantly in their overall performance. For this reason, we need not consider them separately here and will refer to them collectively instead as the "inferior" readers.

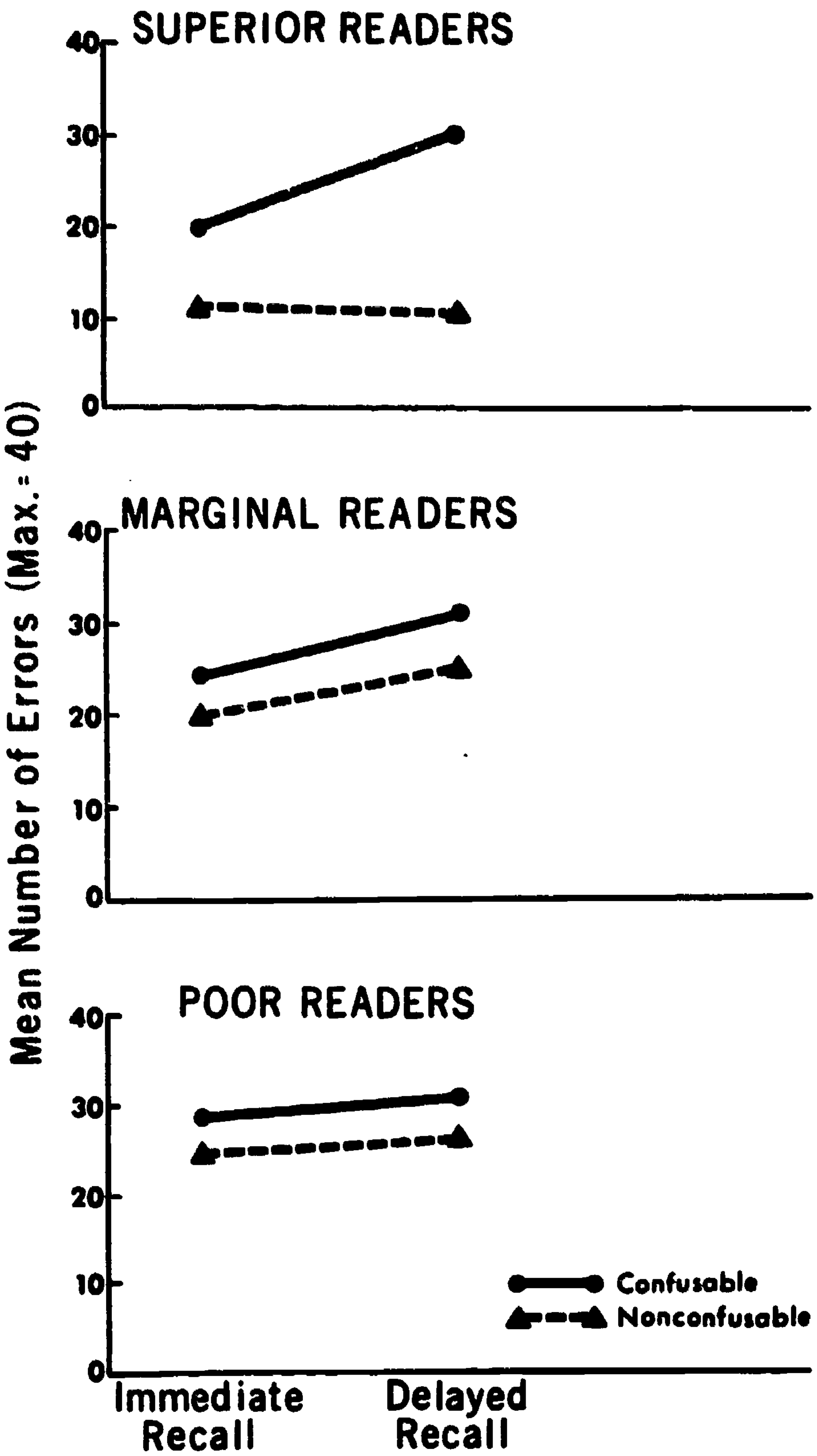


Figure 2: Mean recall errors summed over serial position.

It is immediately apparent that the superior group tends, overall, to make fewer errors in recall than the inferior readers. More notable, however, are the differences in the effects of phonetic similarity on the recall performance of the two reading groups. First, we see that though phonetic similarity caused some deterioration in immediate recall for all the children, the effect was much greater for the superior group than for the inferior readers. Second, the differential effect of phonetic similarity is even more marked in the delay condition. For the superior group, the interposition of a delay interval steeply increased errors of recall of the phonetically confusable strings but produced no effect on the recall of nonconfusable strings. We may suppose that in this group the phonetic similarity of the confusable strings caused interference with rehearsal during the delay interval. For the inferior readers, on the other hand, there is no such interaction; delay depressed their performances on both confusable and nonconfusable strings by nearly equal amounts.

The differential effect of phonetic similarity on the superior readers is again apparent in Figure 3, where the data are replotted as a function of serial position. An examination of the two graphs in the lower half of the figure shows that, after delay, the superior readers are sharply distinguished from the inferior groups in their better recall of nonconfusable strings, but are nearly indistinguishable from the others in their recall of confusable strings. Taken together, the two lower graphs make manifest the much greatei penal effect of phonetic confusability on the superior readers. The same differentially penal effect on this group is found also in the case of immediate recall, as seen in the upper graphs of Figure 3, but there the difference is less striking.[3]

In summary, then, the superior readers are strongly penalized by the phonetic similarity of the confusable strings of letters. The penalty is apparent in immediate recall and more marked in the delay condition. We conclude from these findings that the superior group is using a phonetic code in short-term memory. This is not to say, however, that the inferior readers are not recoding phonetically at all. Phonetic similarity does impair their performance somewhat, though the effect is clearly less than for the superior group. There may be several interpretations of this difference between the two reading groups in our study. One possibility is that the inferior readers rely less on phonetic recoding than the superior group and concurrently use other codes (visual codes, for example), which are unaffected by phonetic confusability. Another possibility, suggested by Crowder (personal communication), is that they may simply rehearse at a slower rate than the superior readers, thereby giving the confusable

[3] An analysis of variance performed on the data showed all main effects to be significant at $p < .001$ (Reading Group: $F_{2,\ 43} = 22.67$; Delay: $F_{1,\ 43} = 29.77$; Confusability: $F_{1,\ 43} = 73.00$). (The significance of the Reading Group factor is accounted for by the differences between the superior readers and the other two groups; the marginal and poor readers do not differ significantly from each other in recall.) The three-way interaction, Reading Group X Delay X Confusability, is statistically significant at $p < .001$ ($F_{2,\ 43} = 8.24$). Newman-Kuels post-hoc means tests reveal that for the superior readers, delay has a significantly greater effect on recall of confusable sequences than on recall of nonconfusable sequences. Among the marginal and poor readers, on the other hand, delay did not differentially affect performance on the two types of sequences.

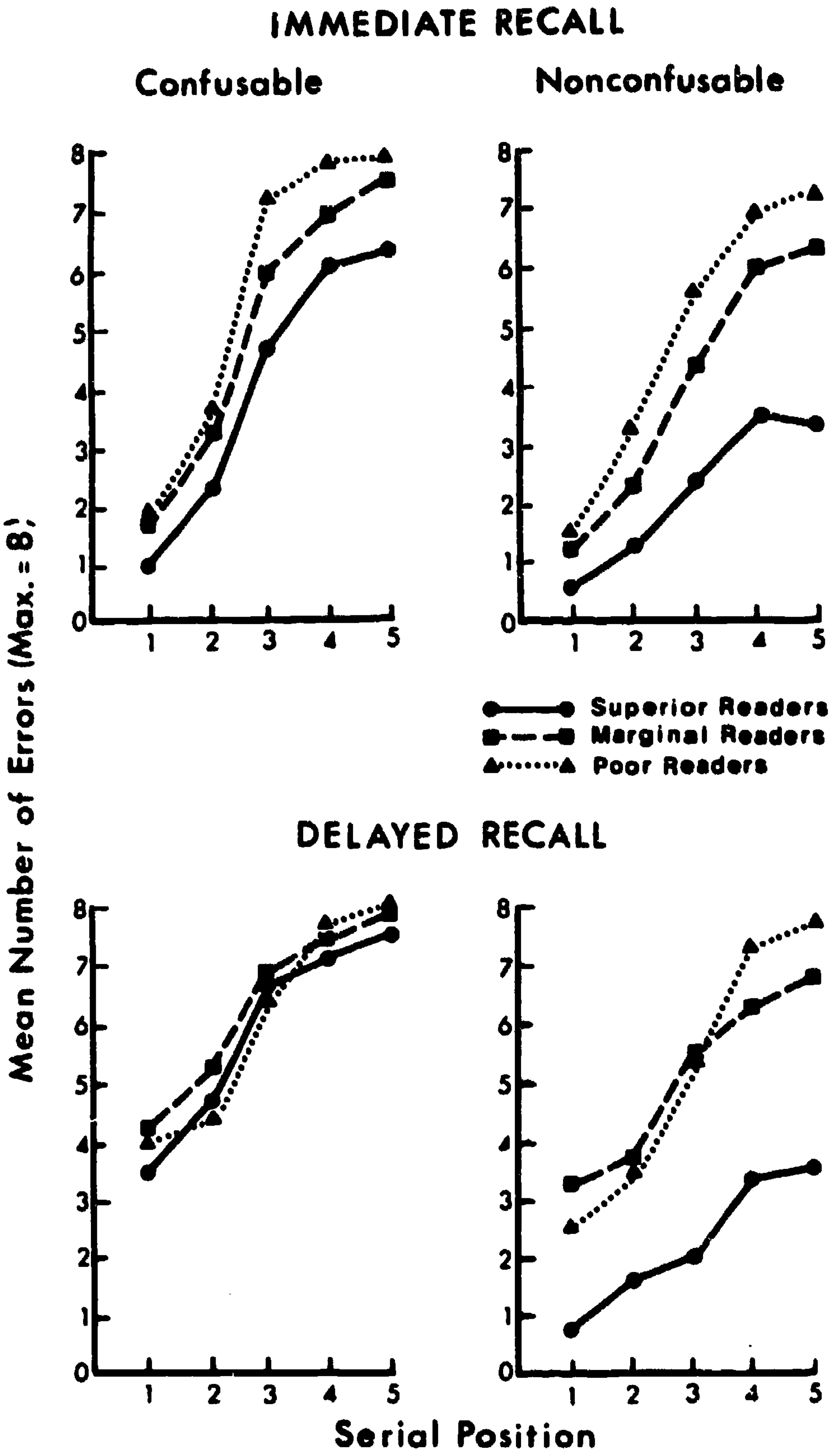


Figure 3: Recall data replotted as a function of serial position.

items less opportunity to interfere. Whatever interpretation is accepted (and the answers must await further investigation), we would emphasize that the failure of the superior readers to maintain their advantage over the inferior group in short-term memory when the items are phonetically confusable cannot be accounted for by assuming that the groups differ only with respect to a general memory capacity.

An auditory analog of our experiment would be one way to clarify the nature of the difference in short-term memory between the two groups of readers.[4] Since phonetic coding, as we said earlier, presumably cannot be avoided when the linguistic material arrives auditorily, auditory presentation should force the inferior reader into a phonetic mode. If an important component of his difficulty is that he is deficient in recoding visual symbolic material into phonetic form, then the phonetic similarity of auditorily presented stimuli should affect him as much (or as little) as it does the superior readers. While quantitative differences in memory capacity between the two groups may still show up in the general level of recall on the auditory presentation, the interaction of reading group and phonetic confusability should be diminished. If, on the other hand, the interpretation that implicates differences in rate of rehearsal between the groups is correct, the interaction should remain.

Obviously, many other refinements of the experimental task remain to be made. In particular, we hope in the future to use tasks that resemble more closely what happens in actual reading. At the very least, we should like to repeat the kind of experiment reported here, using words instead of letters. Only after that could we have a very high degree of confidence in the conclusion that seems to be suggested by the results of the present experiment--namely, that phonetic recoding is characteristic of skilled reading.

SUMMARY

By converting print to speech the beginning reader gains two advantages: he can read words he has never seen before, and he can, as he reads, fully exploit the primary language processes of which he is already master. If he is to realize the first advantage, he must make the conversion analytically, not by whole words. That analytic conversion requires, in particular, an explicit awareness that speech can be segmented into units of phonemic size. Given what we know about the relation of speech sounds to phonetic structure, we can see why explicit segmentation might be hard to achieve. Our recent research has shown that for young children such explicit segmentation is, in fact, difficult (more difficult in any case than segmentation into syllables) and that such difficulty may be related to success, or the lack of it, in the early stages of reading.

Among the primary language processes that the child can exploit by conversion to speech (either analytically or holistically) is the use of a phonetic representation to store smaller segments (words, for example) until the meaning of larger segments (phrases or sentences) can be extracted. Research on speech

[4] Since the auditory experiment would, of course, necessitate serial presentation, an additional visual condition, employing serial presentation, would be required to achieve comparability.

perception suggests that the phonetic representation may be uniquely suited to such storage. That the phonetic representation is, in fact, so suited is suggested by the outcome of many experiments on short-term memory. Now we have evidence from a similar experiment that, among second graders, good readers rely more on a phonetic representation than poor readers do.

REFERENCES

Baddeley, A. D. (1966) Short-term memory for word sequences as a function of acoustic, semantic, and formal similarity. Quart. J. Exp. Psychol. 18, 362-365.

Baddeley, A. D. (1968) How does acoustic similarity influence short-term memory? Quart. J. Exp. Psychol. 20, 249-264.

Baddeley, A. D. (1970) Effects of acoustic and semantic similarity on short-term paired associate learning. Brit. J. Psychol. 61, 335-343.

Bever, T. G. and T. G. Bower. (1966) How to read without listening. Project Literacy Reports No. 6, 13-25.

Bolinger, D. L. (1957) Maneuvering for accent and position. College Comp. Communic. 8, 234-238.

Calfee, R., R. Chapman, and R. Venezky. (1972) How a child needs to think to learn to read. In *Cognition in Learning and Memory*, ed. by L. W. Gregg. (New York: Wiley).

Chomsky, C. (1970) Reading, writing, and phonology. Harvard Educ. Rev. 40(2), 287-309.

Conrad, R. (1963) Acoustic confusions and memory span for words. Nature 197, 1029-1030.

Conrad, R. (1964) Acoustic confusions in immediate memory. Brit. J. Psychol. 55, 75-84.

Conrad, R. (1972) Speech and reading. In *Language by Ear and by Eye: The Relationships between Speech and Reading*, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Conrad, R., P. R. Freeman, and A. J. Hull. (1965) Acoustic factors versus language factors in short-term memory. Psychon. Sci. 3, 57-58.

Conrad, R. and A. J. Hull. (1964) Input modality, acoustic confusion, and memory span. Brit. J. Psychol. 55, 429-432.

Daniels, J. C. and H. Diack. (1956) *Progress in Reading*. (Nottingham: University of Nottingham Institute of Education).

Dornič, S. (1967) Effect of a specific noise on visual and auditory memory span. Scand. J. Psychol. 8, 155-160.

Downing, J. (1973) *Comparative Reading*. (New York: Macmillan).

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

Elkonin, D. B. (1973) [U.S.S.R.] In *Comparative Reading*, ed. by J. Downing. (New York: Macmillan).

Erickson, D., I. G. Mattingly, and M. T. Turvey. (1973) Phonetic activity in reading: An experiment with kanji. Haskins Laboratories Status Report on Speech Research SR-33, 137-156.

Fletcher, H. (1929) *Speech and Hearing*. (New York: Van Nostrand).

Fowler, C. A., I. Y. Liberman, and D. Shankweiler. (in preparation) Untitled manuscript.

Gelb, I. J. (1963) *A Study of Writing*. (Chicago: University of Chicago Press).

Gibson, E. J. and H. Levin. (in press) *The Psychology of Reading*. (Cambridge, Mass.: MIT Press).

Gleitman, L. R. and P. Rozin. (1973) Teaching reading by use of a syllabary. Read. Res. Quart. 8, 447-483.

Hintzman, D. L. (1967) Articulatory coding in short-term memory. J. Verbal Learn. Verbal Behav. 6, 312-316.

Kintsch, W. and H. Buschke. (1969) Homophones and synonyms in short-term memory. J. Exp. Psychol. 80, 403-407.

Klima, E. (1972) How alphabets might reflect language. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Liberman, A. M., F. S. Cooper, D. 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. (Washington, D.C.: Winston).

Liberman, I. Y. (1973) Segmentation of the spoken word and reading acquisition. Bull. Orton Soc. 23, 65-77.

Liberman, I. Y., D. Shankweiler, F. W. Fischer, and B. Carter. (1974) Reading and the awareness of linguistic segments. J. Exp. Child Psychol. 18, 201-212.

Liberman, I. Y., D. Shankweiler, C. Fowler, and F. Fischer. (in preparation) Phonetic recoding in good and poor readers.

Makita, K. (1968) Rarity of reading disability in Japanese children. Amer. J. Orthopsychiat. 38(4), 599-614.

Monroe, M. (1932) Children Who Cannot Read. (Chicago: University of Chicago Press),

Rosner, J. and D. P. Simon. (1970) The Auditory Analysis Test: An Initial Report. (Pittsburgh: University of Pittsburgh Learning Research and Development Center).

Rozin, P., S. Poritsky, and R. Sotsky. (1971) American children with reading problems can easily learn to read English represented by Chinese characters. Science 171, 1264-1267.

Savin, H. B. (1972) What the child knows about speech when he starts to learn to read. In Language by Ear and By Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Shankweiler, D. and I. Y. Liberman. (1972) Misreading: A search for causes. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Sperling, G. (1963) A model for visual memory tasks. Human Factors 5, 19-31.

Thomasson, A. J. W. M. (1970) On the Representation of Verbal Items in Short-Term Memory. (Nijmegen: Drukkerij Schippers). Cited by R. Conrad (1972).

Weber, R. (1970) A linguistic analysis of first-grade errors: A survey of the literature. Read. Res. Quart. 4, 96-119.

Word Recall in Aphasia*

Diane Kewley-Port+
Haskins Laboratories, New Haven, Conn.

ABSTRACT

Aphasic and normal subjects listened to lists of ten words in a probe recall paradigm. Memory function was assessed by estimating the probabilities of recalling a word from either long-term store or short-term store. When compared to the normal subjects, nine of the ten aphasic subjects showed deficient ability to recall a word from short-term store, and no capability to recall from long-term store. The memory functions of the remaining aphasic subject were anomalous: he showed no ability to recall words from short-term store, but an increased ability to recall from long-term store.

INTRODUCTION

Two common symptoms observed in patients diagnosed as aphasic are a greatly reduced vocabulary and difficulty in recalling strings of digits or words. To a psychologist these symptoms might indicate abnormal memory processes. Models of memory usually consist of the two components, long-term store (LTS) and short-term store (STS), where LTS is the permanent information store and STS retains briefly a small number of items (Waugh and Norman, 1965; Atkinson and Shiffrin, 1968). Accordingly, reduced vocabulary could be viewed as a problem in accessing and retrieving words from LTS, and shortened auditory retention span might reflect deficient STS function.

*A shorter version of this paper was presented at the 87th meeting of the Acoustical Society of America, New York, April 1974.

+Also City University of New York.

Acknowledgment: I thank Dr. Lesley Schwartz at the Northport Veterans Administration Hospital for permitting his patients to participate in this study and for the personal assistance he gave in effecting the research. I also thank the members of the Patrick J. Devany Post No. 964 of the Veterans of Foreign Wars for volunteering their time. This research was made possible through the facilities of the Communication Science Laboratory, Graduate Center, City University of New York, and was also supported in part by a grant from the National Institute of Dental Research to Haskins Laboratories.

Only a few experiments have been conducted to assess memory function in aphasic patients. This is not surprising since aphasia is defined as a language disorder, and although normal language communication must depend on normal memory function, the relationship between the two has rarely been discussed (cf. Norman, 1972; Aaronson, 1974). Nonetheless, in a study of patients with different neurologically based language disorders, Halpern, Darley, and Brown (1973) found that out of ten language functions, aphasics performed poorest on Adequacy (word-finding difficulties) and Auditory Retention Span.

Several studies have compared visual versus auditory STS functions with some aphasic patients. Luria, Sokolov, and Klimkowski (1967) studied two aphasics whose main symptom was the inability to repeat a series of auditorily presented words (acoustic-mnestic aphasia). They showed that difficulty in recalling a series of three to five words was specific to the auditory modality. They did not discuss a two-component memory model. Butters, Samuels, Goodglass, and Brody (1970) tested groups of brain-damaged patients, some of whom were aphasic, on recall of consonants presented visually or auditorily. A Peterson and Peterson (1959) paradigm was employed to test immediate and delayed recall for either single consonants or consonant trigrams. Patients with left-hemisphere, parietal brain damage (all eight were aphasic) had memory deficits in both visual and auditory tasks. Patients with left-hemisphere, frontal brain damage (seven aphasic, one nonaphasic) were thought to have no memory deficits, but rather an impairment in registration of the consonants. They concluded that "apparently, aphasic and memory disorders represent separate and independent processes" (p. 457). It is not clear that their own results fully support this generalization. In addition, they have not taken into account the fact that in normal subjects visually presented consonants are often encoded in a phonological form in immediate memory (Conrad, 1964; Wickelgren, 1965; Conrad, 1972), an effect which may have confounded their results (cf. Warrington and Shallice, 1972).

Two other studies have compared memory processes between aphasic and normal subjects. In one of those (Carson, Carson, and Tikofsky, 1968) only quantitative differences were found between aphasic and normal subjects in several learning tasks including a verbal serial learning task. In the other study (Swinney and Taylor, 1971) both quantitative and qualitative differences were observed in a nonverbal task examining the search process in STS.

A series of memory experiments for both STS and LTS have been conducted by Warrington and her colleagues using primarily one patient (KF) thought to have conduction aphasia, and whose main symptom was the inability to repeat a series of words (Warrington and Shallice, 1969; Warrington, Logue, and Pratt, 1971; Warrington and Shallice, 1972; Warrington and Weiskrantz, 1973). The main outcome of the research on KF is his selective impairment of auditory versus visual STS, and selective impairment of LTS versus STS. In another study of one conductive aphasic, Strub and Gardner (1974) accounted for the repetition difficulties primarily as a result of a linguistic-phonological deficit rather than a memory dysfunction.

The present study compared both long-term and short-term memory functions for groups of normal and aphasic subjects. We chose a probe-recall paradigm based on the work of Waugh and Norman (1965) in which separate functions for LTS and STS are derived from one set of data. The procedure was to present tape-recorded word lists, with each list followed by a tone and a word drawn from the list--the "probe word." The probe word occurred in various positions on the lists and the subject's task was to recall the word following the probe word. Previous experiments have reported right-ear superiority in recalling words presented monaurally (Bakker, 1969; Turvey, Pisoni, and Croog, 1972). Thus, the word lists were presented monaurally to test for possible ear differences in this experiment.

METHOD

Word Lists

In principle, the word lists were constructed to test the memory of an aphasic subject, not his difficulties in word usage. It has been reported that the variables of frequency of occurrence, part of speech, and abstractness (among others) do affect aphasic patients' use of words (cf. Halpern, 1972). With this in mind, the words chosen for this experiment were selected from the Thorndike and Lorge (1944) count of the 1000 most frequent words in English, excluding proper nouns and function words of three letters or less. Since high frequency words are easiest for aphasics to use (Schuell, Jenkins, and Landis, 1961) and many of the other variables correlate with word frequency, the words selected should present minimal difficulty for the subjects. Proper nouns and short function words were excluded for several reasons, one of which was that they tended to stand out in the word lists.

Thirty word lists of ten words each were constructed, words were selected in a quasi-random way, and only a few words were repeated across lists. Probe words were chosen so that positions 2, 4, and 6 were probed twice for each ear and positions 7, 8, and 9 were probed three times for each ear, following the procedures of Kintsch and Buschke (1969) and Turvey et al. (1972). Word lists were presented 15 times each to the left and right ears, with ear presentation alternated randomly.

The lists were read by the experimenter in a quiet room (IAC 1201) and recorded on a Uher 4200 tape recorder. The word lists were read at a rate of 3 sec/word, followed by a brief 1200 Hz tone and the probe word. Each list was read into only one channel of the tape recorder, the channel assignments alternating randomly. There was a 20 sec response interval between lists. The entire test tape lasted 30 minutes.

Subjects

Ten aphasic men (aphasics) and another group of eight men matched for age and education (normals) served as subjects. The aphasics were all patients in the Speech Pathology and Audiology Services clinic of Northport Veterans Administration Hospital, New York. The aphasics were judged to have mild to moderate aphasia as tested on the Short Examination for Aphasia (Schuell, 1957) and the Porch Index of Communicative Ability (Porch, 1967). Etiologies included both trauma and cerebral vascular accidents occurring from 6 months to 19 years prior to testing. In all cases there were symptoms indicating brain damage to the left hemisphere and in a few cases to both hemispheres. Their ages ranged from 26 to 56 years (mean = 47.6 years) and all were right handed. Education levels achieved included eighth grade (n=4), high school diploma (n=5), and college diploma (n=1). All patients had audiograms that were normal in both ears for their ages.

The normals were all veterans who volunteered their time for the experiment. They all appeared to have normal language function and had no known hearing difficulties. They ranged in age from 41 to 65 (mean = 60.6 years) and were all right handed. The education levels achieved included eighth grade (n=3), high school diploma (n=3), and college diploma (n=2).

Procedure

The procedure was the same for both groups. The subjects were verbally instructed to report the word on the list that followed the probe word. A few practice word lists were read until the experimenter was satisfied that the subject understood the instructions or he was eliminated from the experiment. The subjects were told that the task was very difficult, but that they were to think about each word as they heard it and not to go over previous words. The experimenter recorded the verbal response of the aphasics whereas the normals wrote down the responses themselves. No particular difficulty was encountered in understanding the responses spoken by this group of aphasic patients because of their moderate impairment.

The tape-recorded lists were played on a Uher 4200 through Grason-Stadler TDH-39 earphones. The playback channels were equated for equal intensity and presented to subjects at a comfortable listening level.

RESULTS

The results were first tallied by the number of correct responses for each ear for each subject. Unfortunately, no differences in recall were obtained from right- versus left-ear presentations for either normals or aphasics. Indeed, replication of the Turvey et al. (1972) probe recall experiment for normals now seems in doubt (Turvey, personal communication). Therefore, the data presented in this paper combine the results from both ears.

Examination of the overall pattern of correct response by probe position yielded similar functions for all subjects but one, aphasic CH. His data are reserved for later and the functions showing probability of recall for the remaining subjects are graphed in Figure 1. For both groups of subjects the probability of recall is low and constant for the words near the beginning of the list and increases rapidly for the last three items.

According to the Waugh and Norman (1965) model, the probability of a word entering secondary memory or long-term store, P(LTS), is constant over all probe positions if rehearsal of each item is constant (as the subjects were so instructed). On the other hand, the probability of an item being retained in primary memory or short-term store, $P(STS_i)$, is greatest for the most recently presented word and decreases monotonically for preceding words until it reaches zero when, presumably, the limited number of words stored in STS is exceeded. Assuming that the probabilities of a word being in LTS or STS are stochastically independent, then the probability of recalling a word at probe position i is: $P(r_i) = P(LTS) + P(STS_i) - P(STS_i)P(LTS)$. P(LTS) can be estimated over the constant portion of the recall functions by taking the mean of the recall probabilities at positions 2, 4, and 6. $P(STS_i)$ is then calculated from the equation above.

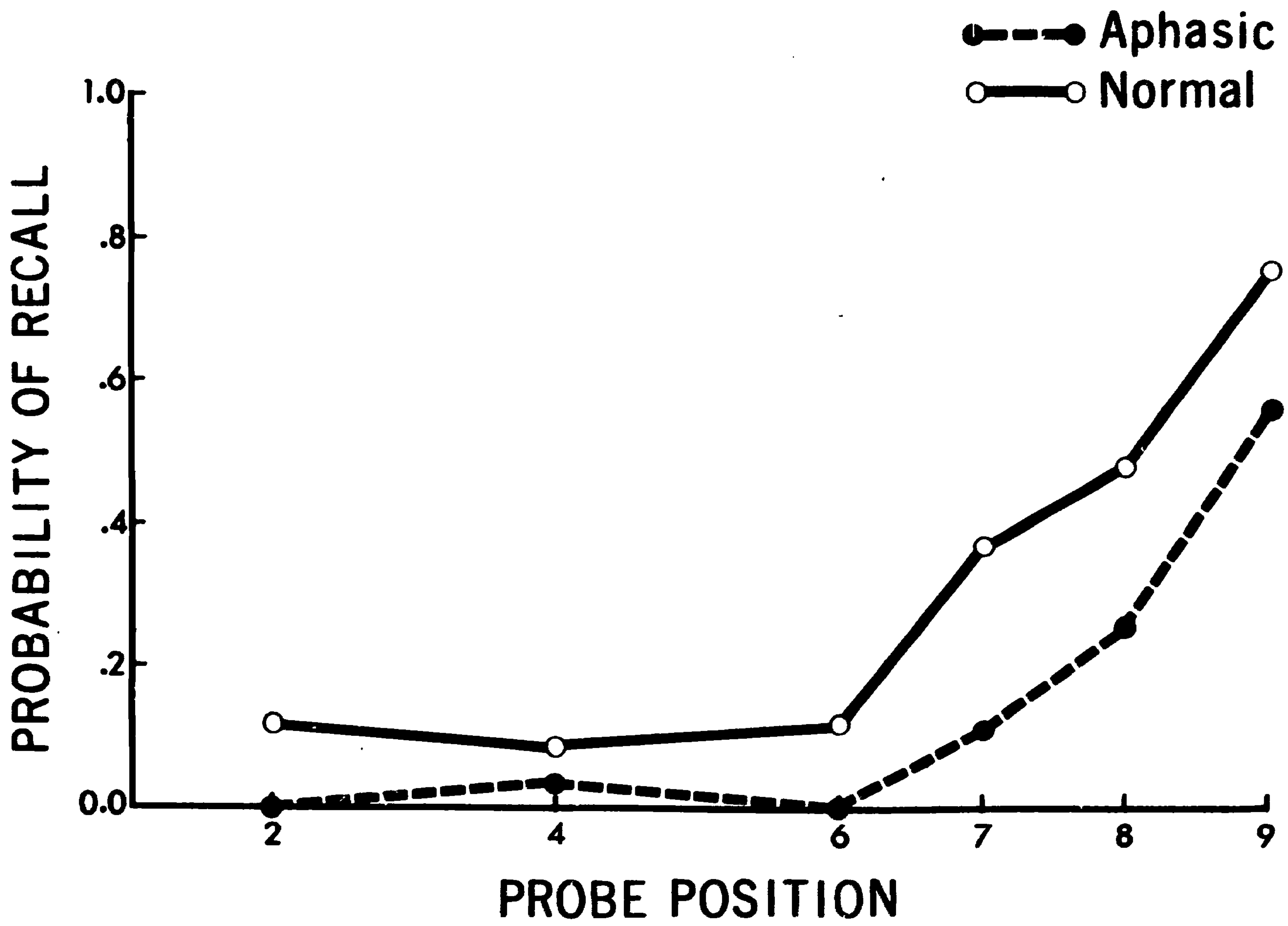


Figure 1: Functions of the number of correct responses divided by the number of trials (probability of recall) at each position probed in the word lists for nine aphasic and eight normal subjects.

FIGURE 1

Figure 2 displays the estimated probabilities for both groups of subjects. The most striking feature of Figure 2 is that the probability of recalling a word from LTS for the aphasics is zero. The P(LTS) for normals (.11) is in range of those reported in Waugh and Norman (1965). The estimated probabilities for STS produced similar recency components for normals and aphasics, although the P(STS_i) is depressed by about .17 at positions 7, 8, and 9 for the aphasics.

Figure 3 presents the recall function for aphasic subject CH. CH's function contrasts sharply with the normals and aphasics in Figure 1. The probability of recall at all probe positions is nearly a constant .5 with no recency effect in the final items. In terms of memory models, CH has a high probability of recalling an item from long-term store, P(LTS) = .44, and no evidence of recalling an item from short-term store.

DISCUSSION

A number of possible disorders of memory function are implied by the results of this experiment. The majority of the aphasics studied (nine of ten) retained the same number of words in short-term store as the normal subjects. However, the aphasics were unable to recall the words correctly from STS as often as the normals. It should be noted that these statements cannot be generalized since only mild to moderate aphasics participated in this experiment. For more severely impaired aphasics the instructions were too difficult to comprehend and presumably their memory processes might also be more impaired than those tested here. (Three moderately impaired aphasics were unable to comprehend the instructions.)

The results for STS agree with those obtained from the visual, serial search task of Swinney and Taylor (1971). Although some aphasics were unable to perform in their task, those who did were characterized as using a serial search process similar to normals, but more slowly and with more errors. These results are also in accord with the data shown by Carson et al. (1968:98). In the serial position curves obtained as an average of ten rote serial learning trials, the recency components for both aphasics and normals extend over the same number of items. However, the probability of recalling items is depressed for the aphasics.

In the probe recall experiment, nine out of ten aphasics were incapable of recalling words from long-term store. The conclusions of Carson et al. (1968: 110) that aphasics learned tasks slowly and "demonstrated limited retention and transfer of learning in general" might be related to deficient recall of material from LTS. The question arises, however, as to what extent the disorder observed is caused by the registration versus the retrieval of words from LTS. Analysis of the errors in this experiment showed that aphasics responded occasionally with words in early positions in the word lists (numbers 1 to 5) as well as with words from previous lists. Apparently, some words were registered in LTS. Clearly further research directed at the nature of retrieval of items from LTS should be undertaken.

We can conclude that the majority of aphasics demonstrated memory disorders characterized by reduced short-term store function and an absence of long-term store function. On the other hand, one aphasic subject apparently had a complementary disorder--no STS function and a heightened LTS function. I have no reason to believe that CH's results are due to any artifacts surrounding his testing

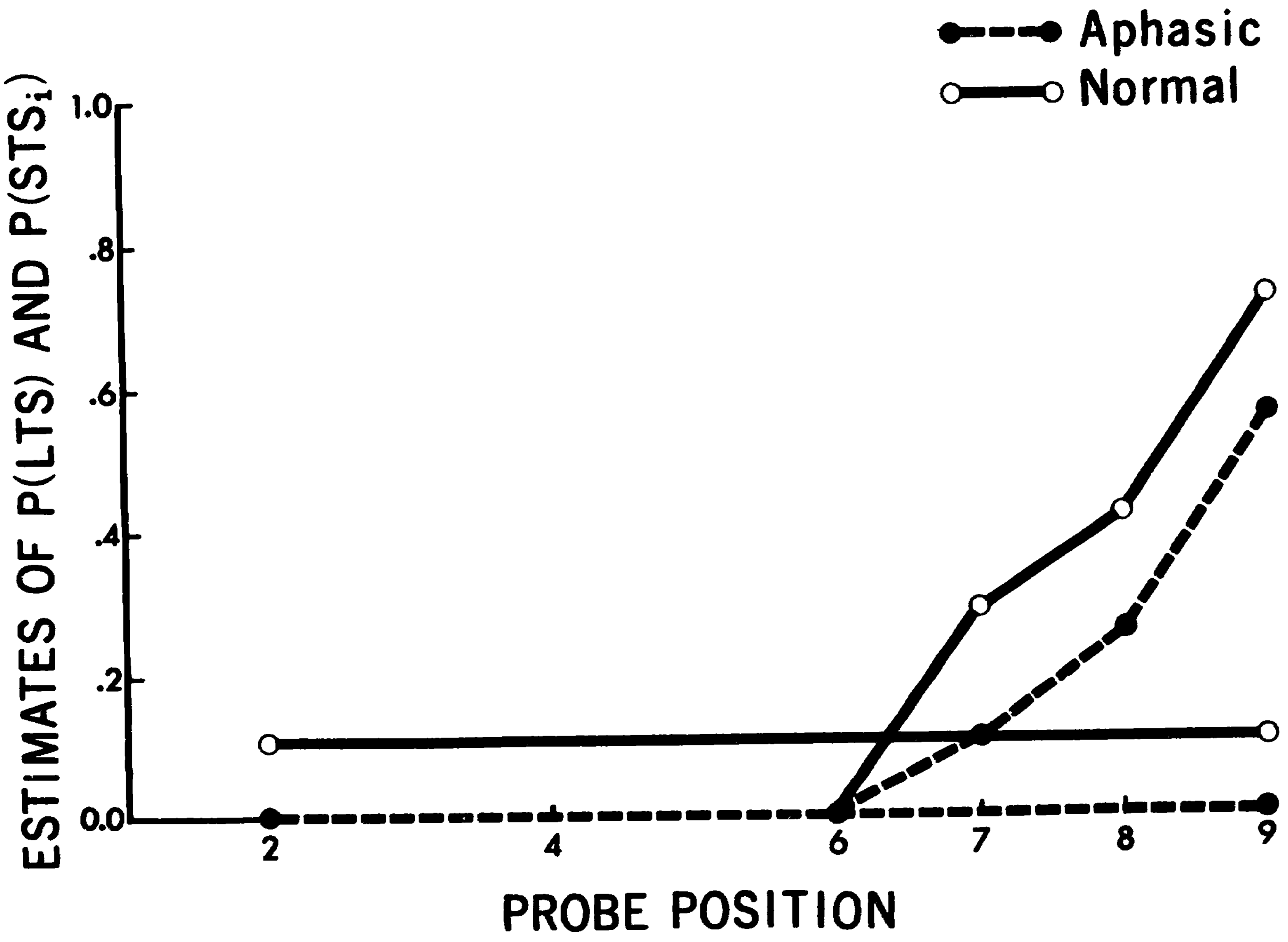


FIGURE 2

Figure 2: Estimates of the probability of words being recalled from long-term store or short-term store calculated from the functions presented on Figure 1 (see text).

FIGURE 3

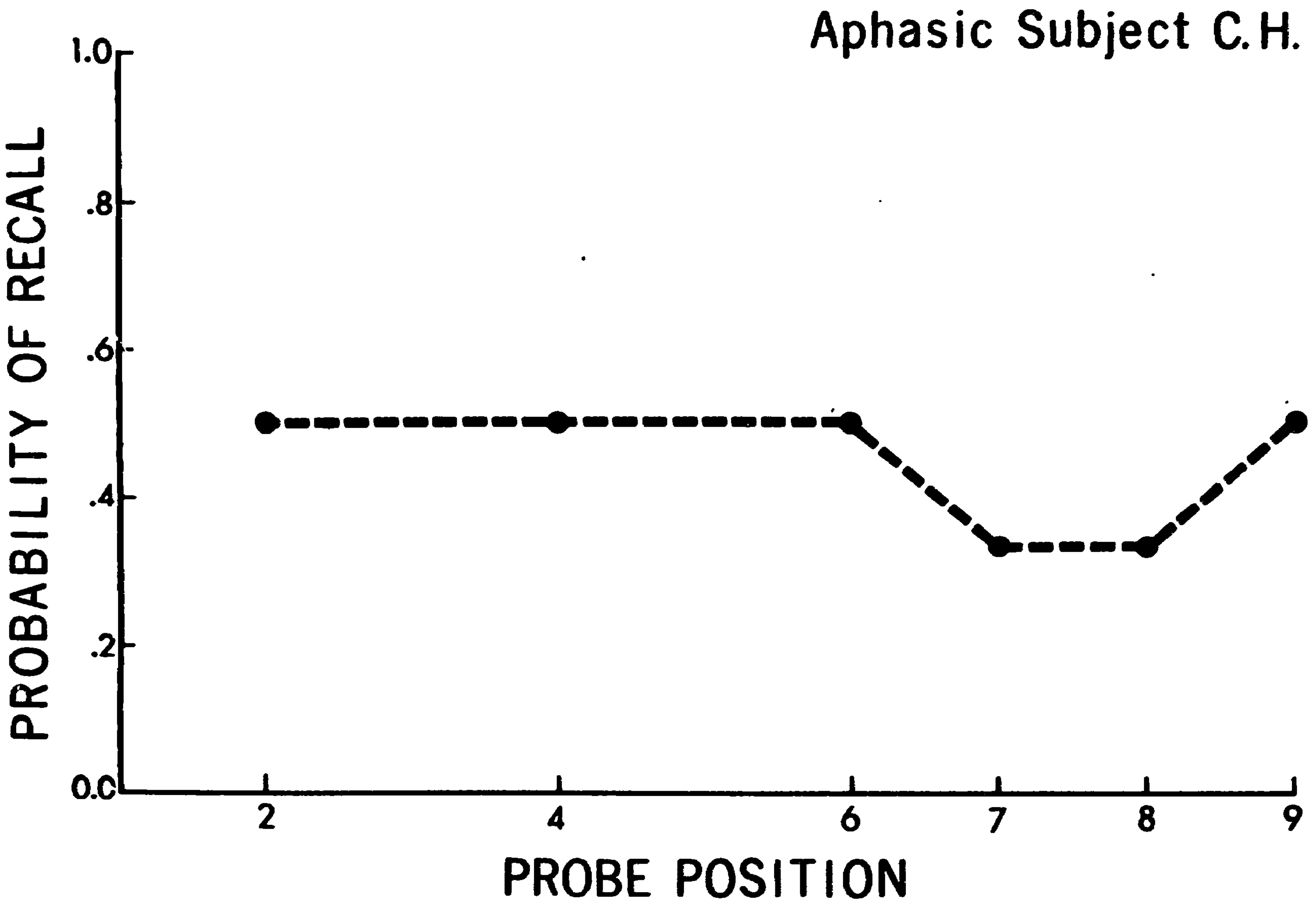


Figure 3: Number of correct responses divided by the number of trials (probablity of recall) at each position probed in the word lists for one aphasic subject.

or that he acted on different instructions than those given. (In informal testing with a graduate student I was unable to find a test strategy that could duplicate CH's results.)

We might be tempted to treat CH as an anomaly if it were not for the following studies. The study of two patients by Luria et al. (1967) is not directly comparable to the present study, but the patients did show strikingly different patterns of responses in auditory short-term memory tasks. Warrington et al. (1971) investigated three patients diagnosed as conductive aphasics. They concluded that these patients had relatively normal LTS function and severely impaired STS function for the auditory modality only. Strub and Gardner (1974) confirmed Warrington's results for another conductive aphasic. CH's results closely match those for the conductive aphasics except for the surprisingly high probability of recall from LTS. Unfortunately, CH stopped coming to the clinic and was not available for further testing.

We are thus left with the conclusion that aphasics differ from normals in both STS and LTS function, and further, that aphasics with different linguistic (and presumably neurological) deficits may have totally different memory disorders for auditorily presented words. We hope further research will clarify these statements and, in particular, incorporate the often observed memory differences for material presented in the auditory and visual modalities.

REFERENCES

Aaronson, D. (1974) Stimulus factors and listening strategies in auditory memory: A theoretical analysis. Cog. Psychol. 6, 108-132.

Atkinson, R. C. and R. M. Shiffrin. (1968) Human memory: A proposed system and its control processes. In The Psychology of Learning and Motivation; Advances in Research and Theory, Vol. II, ed. by K. Spence and J. Spence. (New York: Academic Press).

Bakker, D. J. (1969) Ear-asymmetry with monaural stimulation: Task influences. Cortex 5, 36-42.

Butters, N., I. Samuels, H. Goodglass, and E. Brody. (1970) Short-term visual and auditory memory disorders after parietal and frontal lobe damage. Cortex 6, 440-459.

Carson, D., P. Carson, and R. Tikofsky. (1968) On learning characteristics of the adult aphasic. Cortex 4, 92-112.

Conrad, R. (1964) Acoustic confusion in immediate memory. Brit. J. Psychol. 12, 1-6.

Conrad, R. (1972) Speech and reading. In Language by Ear and by Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Halpern, H. (1972) Adult Aphasia. (New York: Bobbs Merrill).

Halpern, H., F. L. Darley, and J. Brown. (1973) Differential language and neurological characteristics in cerebral involvement. J. Speech Hearing Dis. 38, 162-173.

Kintsch, W. and H. Buschke. (1969) Homophones and synonyms in short-term memory. J. Exp. Psychol. 80, 403-407.

Luria, A. R., E. N. Sokolov, and M. Klimkowski. (1967) Towards a neurodynamic analysis of memory disturbances with lesions of the left temporal lobe. Neuropsychologia 5, 1-10.

Norman, D. A. (1972) The role of memory in the understanding of language. In Language by Ear and by Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Peterson, L. R. and M. Peterson. (1959) Short-term retention of individual verbal items. J. Exp. Psychol. 58, 193-198.

Porch, B. (1967) The Porch Index of Communicative Ability. (Palo Alto, Calif.: Consulting Psychological Press).

Schuell, H., J. Jenkins, and L. Landis. (1961) Relationship between auditory comprehension and word frequency in aphasia. J. Speech Hearing Res. 4, 30-36.

Schuell, H. R. (1957) A short examination for aphasia. Neurology 7, 625-634.

Strub, R. L. and H. Gardner. (1974) The repetition defect in conduction aphasia: Mnestic or linguistic? Brain Lang. 1, 241-255.

Swinney, D. and O. Taylor. (1971) Short-term memory recognition search in aphasics. J. Speech Hearing Res. 14, 578-588.

Thorndike, E. L. and I. Lorge. (1944) The Teacher's Word Book of 30,000 Words. (New York: Bureau of Publication, Teachers College, Columbia University).

Turvey, M. T., D. Pisoni, and J. Croog. (1972) A right-ear advantage in the retention of words presented monaurally. Haskins Laboratories Status Report on Speech Research SR-31/32, 67-74.

Warrington, E. K., V. Logue, and R. T. C. Pratt. (1971) Localization of selective impairment of auditory short-term memory. Neuropsychologia 9, 377-387.

Warrington, E. K. and T. Shallice. (1969) The selective impairment of auditory verbal short-term memory. Brain 92, 885-896.

Warrington, E. K. and T. Shallice. (1972) Neuropsychological evidence of visual storage in short-term memory. Quart. J. Exp. Psychol. 24, 30-40.

Warrington, E. K. and L. Weiskrantz. (1973) Analysis of short-term and long-term memory defects in man. In The Physiological Basis of Memory, ed. by J. Deutsch. (New York: Academic Press) 365-393.

Waugh, N. C. and D. A. Norman. (1965) Primary memory. Psychol. Rev. 72, 89-104.

Wickelgren, W. A. (1965) Short-term memory for phonemically similar lists. Amer. J. Psychol. 78, 567-574

Linguistic and Nonlinguistic Stimulus Dimensions Interact in Audition but not in Vision

James E. Cutting[+]
Haskins Laboratories, New Haven, Conn.

During the processing of speech stimuli, irrelevant variation in fundamental frequency impedes the phonetic decision-making process, but irrelevant phonetic variation does not impede pitch decisions. No analogous asymmetry is found in vision: the processing of a letter of the alphabet is not impeded by irrelevant variation in type font, nor is the processing of type font impeded by irrelevant variation in letter. The differential results can be interpreted in terms of compulsory (in audition) versus optional (in vision) processing of linguistic features.

Garner (1974; Garner and Felfoldy, 1970) has employed two-choice speeded classification tasks to study patterns of interaction between stimulus dimensions. Often in card sorting tasks stimulus dimensions either mutually interfere or they do not interfere with one another during processing. Integral dimensions, such as value and chroma in Munsell color chips, produce interference when subjects are required to attend to orthogonal stimulus dimensions (Garner, Hake, and Eriksen, 1956). Separable dimensions, such as size and angle of dot arrays, do not interfere.

Day and Wood (1972a) and Wood (1973, 1974) found that linguistic and nonlinguistic dimensions of auditory stimuli were neither integral nor separable. There was some integrality (as seen in interference and in increased sorting time) but integrality was asymmetric. Likewise there was some separability (as seen in lack of interference and no increase in sorting time) but separability was also asymmetric. The linguistic dimension was place of articulation--[ba] versus [da] (Day and Wood, 1972a) or [bæ] versus [gæ] (Wood, 1973, 1974)--and the nonlinguistic dimension was fundamental frequency, or pitch--high (140 Hz) versus low (104 Hz), used in all three studies. Since it is not possible to mount auditory stimuli on cards these studies used a discrete reaction time paradigm.

Four tasks are of interest here. Decisions are made between (1) [ba] versus [da], for example, with no pitch variation within a task; (2) high pitch versus low pitch, with no phonetic variation within a task; (3) [ba] versus [da], with pitch varying randomly from trial to trial within a task; and (4) high versus low, with the phoneme varying randomly from trial to trial within a task. The

[+] Also Wesleyan University, Middletown, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-39/40 (1974)]

first two have been called control tasks, since only one dimension varies, whereas the last two have been called orthogonal tasks, since the nontarget dimension varies in an uncorrelated fashion. Reaction times are comparable for tasks 1, 2, and 3, but task 4 shows a marked increase in reaction time. That is, adding irrelevant phonetic information to the stimuli when making a pitch judgment has little or no effect on decision time, but adding irrelevant pitch information to the stimuli when making phonetic judgments increases reaction time by 8 percent (Day and Wood, 1972a), 12 percent (Wood, 1973), or even 14 percent (Wood, 1974). Such interactions appear to occur in audition only when one dimension is linguistic and the other is nonlinguistic. Two linguistic dimensions, such as consonants and vowels, yield integral or mutually interfering results (Day and Wood, 1972b), and two nonlinguistic dimensions, such as pitch and intensity, yield an integral pattern as well (Wood, 1973).

The present study was designed to determine if linguistic and nonlinguistic dimensions in visual stimuli would interact in a similar fashion. Lower-case letters *b* and *d* were used at different thicknesses.

Method

Eight decks of 32 cards each were prepared. Cards were made of cardboard, 2.5 X 3.5 inches with the upper left-hand corner clipped off in a manner similar to standard computer cards. Mounted on each card 1 inch down from the top and 1.25 inches from either side was the lower case *b* or *d*. Each was a 16 point Letraset Avant Garde Gothic press-on character in either a Medium or Bold font. The front surface of each card was then sealed in plastic.

Four decks were used in control tasks: Medium font *b* versus *d*, Bold font *b* versus *d*, Medium font versus Bold font *b*, and Medium font versus Bold font *d*. Every control deck contained 16 cards each of the two different categories to be sorted. The four remaining decks were used in orthogonal tasks. Each of these decks consisted of 8 Medium font *b* cards, 8 Bold font *b* cards, 8 Medium font *d* cards, and 8 Bold font *d* cards.

Twenty-four Yale University undergraduate and graduate students sorted each of the eight decks four times. Cards were sorted into two piles according to task instructions, either by letter or by thickness. The order in which subjects sorted the cards was determined by a balanced design.

Results

There were no asymmetries of sorting times, as shown in Table 1. Adding the irrelevant dimension of thickness increased sorting time by letter by only

TABLE 1: Mean sorting times, in seconds, for the four experimental conditions.

Stimulus Dimension	Condition	
	Control	Orthogonal
Letter	14.7	14.9
Thickness	14.4	14.8

1.4 percent, a nonsignificant increment. Adding the irrelevant dimension of letter increased sorting time according to thickness by only 2.7 percent, also nonsignificant. These results are similar to those found in a pilot study when sorting written versions of ba and da, printed in italics and standard type.[1]

Discussion

The letters b and d, when pronounced and transcribed, correspond to [bi] and [di]. These items differ only in vowel from the Day and Wood stimuli [ba] and [da]. The thicknesses of the letters, in Medium and Bold Avant Garde Gothic fonts, are roughly analogous to the dimension of pitch in the speech stimuli. Just as there can be no letters without thickness, there can be no speech stimuli without an excitation, normally the fundamental frequency (pitch). Thicknesses and pitches can vary within a wide range without decreasing identifiability of the letter or phoneme, but such variation does not change the linguistic message. Thus, the letters b and d in two different fonts would appear to be an appropriate analog to [ba] and [da] at two different pitches. Why then are the results not analogous?

While the paradigm in the present study differs from that of Day and Wood (1972a, 1972b) and Wood (1973, 1974), there is no logical reason for suspecting this difference to contribute to the different results. Both paradigms yield differences thought to reflect differential processing. Similarly, the lack of differences is thought to imply the lack of differential processing difficulty.

A more plausible explanation is that the difference between the visual and auditory results is caused by the nature of the stimulus dimensions. There is no question that for English speaking subjects the dimension of place of articulation, [ba] versus [da], is linguistic and that the dimension of pitch, high versus low, is nonlinguistic. In the visual stimuli of the present study, the dimension of thickness is certainly nonlinguistic, but perhaps the dimension of letter is not strictly linguistic. Perhaps subjects dismiss the pronunciations of [bi] and [di], and target for nonlinguistic form. Indeed many subjects in the present study volunteered that they merely looked for the loop at the lower end of the letter and sorted according to which side of the bar the loop was located. If indeed the letters were treated as different forms without reference to their pronunciations, the letter dimension is just as nonlinguistic as the dimension of thickness.

More broadly, then, a linguistic dimension in a visual pattern can be treated as language, as in the reading process, or it may be treated merely as nonlinguistic form. However, linguistic dimensions in an auditory pattern, at least those distinguishing stop consonants, cannot be dismissed as nonlinguistic form. It appears that linguistic analysis of speech is in some sense compulsory and dependent on prior--or parallel (Wood, 1974)--analysis of acoustic form.

REFERENCES

Day, R. S. and C. C. Wood. (1972a) Interactions between linguistic and nonlinguistic dimensions of the same stimuli. J. Acoust. Soc. Amer. 51, 79(A).

[1] J. Pomerantz, personal communication.

Day, R. S. and C. C. Wood. (1972b) Mutual interference between two linguistic dimensions of the same stimuli. J. Acoust. Soc. Amer. 52, 175(A).

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

Garner, W. R. and G. L. Felfoldy. (1970) Integrality of stimulus dimensions in various types of information processing. Cog. Psychol. 1, 225-241.

Garner, W. R., H. W. Hake, and C. W. Eriksen. (1956) Operationism and the concept of perception. Psychol. Rev. 63, 149-159.

Wood, C. C. (1973) Levels of processing in speech perception: Neurophysiological and information-processing analyses. Unpublished doctoral dissertation, Yale University (Psychology). (Issued as Supplement to Haskins Laboratories Status Report on Speech Research SR-35/36, S1-S68.)

Wood, C. C. (1974) Parallel processing of auditory and phonetic information in speech discrimination. Percept. Psychophys. 15, 501-508.

Laryngeal Muscle Activity, Subglottal Air Pressure, and the Control of Pitch in Speech

Rene Collier[+]
Haskins Laboratories, New Haven, Conn.

ABSTRACT

An experiment was performed to assess the degree to which laryngeal muscle activity and subglottal air pressure affect the rate of vocal cord vibration in speech. Attention was limited to those change: in the rate of vocal cord vibration that are associated with the articulatory implementation of prosodic categories such as intonation and prominence. Subglottal air pressure was measured directly through a catheter inserted between the cricoid and thyroid cartilages. Using hooked-wire electrodes, the electromyographic activity was recorded in the right and left cricothyroid muscles and in the sternohyoid, sternothyroid, and thyrohyoid muscles. The data were collected for one speaker of Dutch. The results of the experiment show that, in this speaker, (1) cricothyroid muscle activity bears the most direct relationship to all the major fundamental frequency (F_0) changes: contraction of that muscle raises F_0 while its relaxation has a F_0 lowering effect; (2) subglottal air pressure controls the gradually falling base line of the F_0 contour and gives support to a rapid F_0 drop if it occurs on the utterance-final syllable; and (3) the sternohyoid, sternothyroid, and thyrohyoid muscles have no systematic effect on F_0.

INTRODUCTION

The research reported in this paper concerns the general issue of pitch control in speech. More specifically, it is intended to clarify further the

[+] Visiting from the Belgian National Science Foundation (N.F.W.O.) and the University of Leuven, Belgium.

Acknowledgment: I wish to thank Dr. Franklin S. Cooper for his hospitality and my colleagues in Haskins Laboratories for their stimulating interest in my work. I am particularly indebted to T. Ushijima and H. Hirose who performed the experiment and whose comments, together with those of F. S. Cooper, K. S. Harris, G. Kuhn, and D. Kewley-Port, greatly improved the contents of this report. I also thank D. Zeichner and R. Sharkany for taking care of the engineering problems of the experiment, and R. McI. McGuire for his assistance with the F_0 measurements. Finally, I am grateful to A. McKeon who prepared the illustrations for this paper.

[HASKINS LABORATORIES: Status Report on Speech Research SR-39/40 (1974)]

relative importance of laryngeal and respiratory maneuvers in varying the rate of vocal cord vibration. We also focus attention upon the articulatory implementation of linguistic categories that are related to intonation and prominence.

The experiment included simultaneous recordings of subglottal air pressure and of electromyographic (EMG) activity in those laryngeal muscles that are usually assumed to participate in the control of fundamental frequency. The data were obtained for one subject, a native speaker of Dutch. Dutch was chosen because the intonational structure of the language is rather well understood, in both its acoustic and perceptual aspects. By including a large variety of pitch contours in the speech materials of the experiment we hoped to extend the range of observations beyond the simple dichotomy of "falling" versus "rising" pitch contours.

EXPERIMENTAL PROCEDURES

The perceptual experiments reported by Cohen and 't Hart (1967), Collier (1972), Collier and 't Hart (1972), and 't Hart and Cohen (1973) have resulted in a fairly complete inventory of the pitch contours that are acceptable in Dutch, together with a specification of their internal structure and of the degree of perceptual resemblance among them. Figure 1 presents a number of those contours in a stylized form. Each of them can be considered as a particular sequence of transitions between a low and a high pitch level. Those two reference levels can be approximated by parallel lines of gradually downward drifting pitch, the so-called high and low "declination line." The declination line effectively functions as a link between successive pitch movements, and is audible as such on all those syllables that do not carry a major change in pitch. The rate of declination may vary and individual pitch movements may overshoot or undershoot the declination line.

During the experiment all the contours of Figure 1 were spoken by the author, who is a native speaker of Dutch. He read lists of randomized utterances in which the word content was identical and the pitch contours were varied. Each contour type was repeated between 20 and 30 times. The utterances chosen were: "Heleen wil die kleren meenemen" [he.le.n wIl di kle:rə me.ne.mə] (Helen wants to take those clothes along); and "Heleen" [he.le.n] (Helen). The longer utterance was chosen according to the following criteria: meaningful Dutch, no open vowels (because jaw opening may influence the pattern of sternohyoid activity), a maximum of voiced segments (in order to obtain a continuous fundamental frequency curve), one voiceless segment (to facilitate the location of a line-up point for averaging the repetitions of each contour type), and at least three potentially prominent syllables.

Our intention was to sample EMG data for three intrinsic and three extrinsic laryngeal muscles and to record subglottal air pressure at the same time. The muscles chosen were: cricothyroid (CT), lateral cricoarytenoid (LCA), vocalis (VOC), sternohyoid (SH), sternothyroid (ST), and thyrohyoid (TH). Hooked-wire electrodes were inserted percutaneously into each of these muscles, following the techniques described by Hirose (1971). Subglottal air pressure (P_s) was measured directly: a plastic tube (0.035 inch inside diameter, 1 1/2 inches long) was placed around an 18-gauge steel needle and inserted through the cricothyroid membrane. When the needle was withdrawn the plastic tube was left

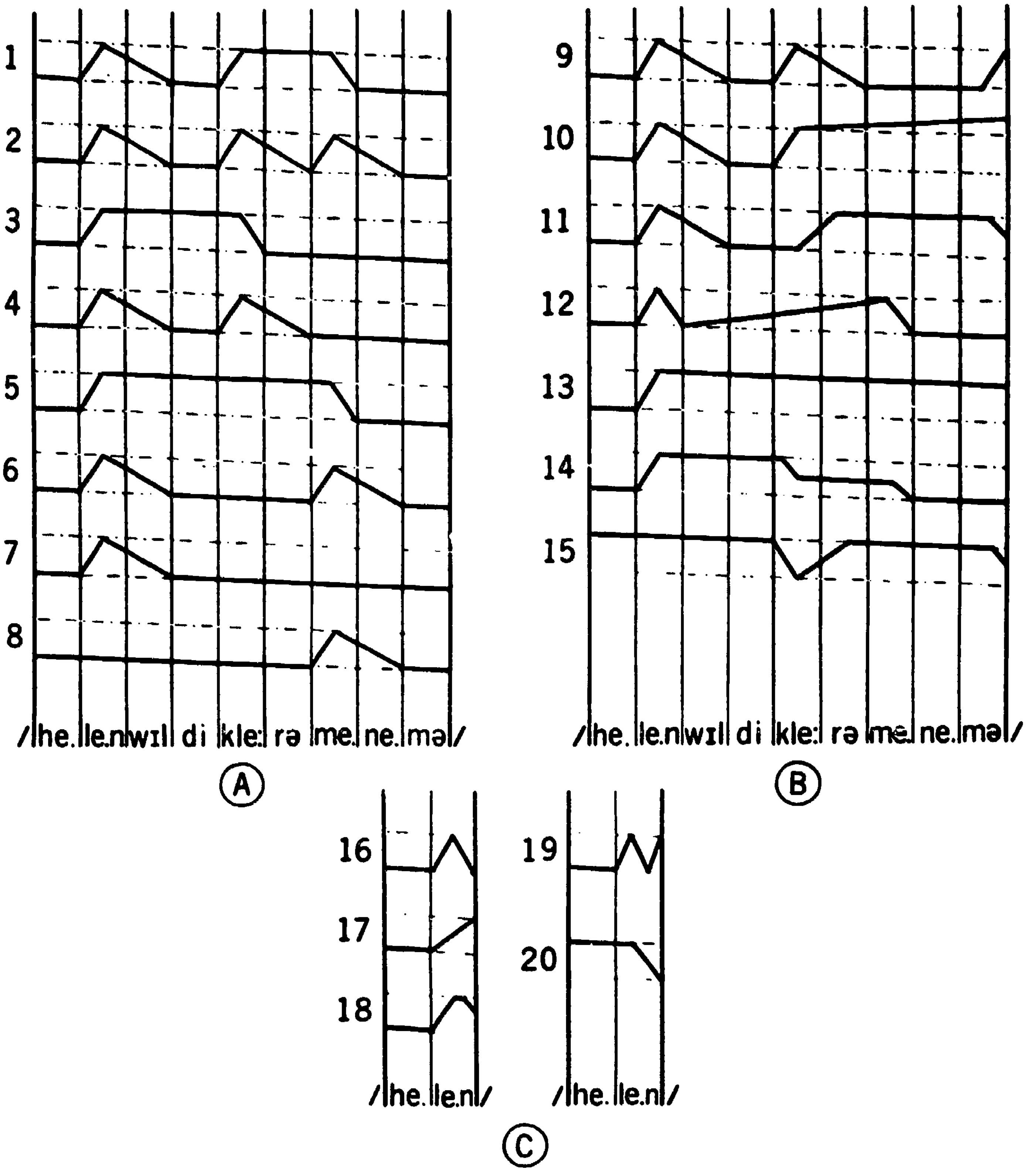

1
2
3
4
5
6
7
8
/he. le.n wɪl di kle: rə me. ne. məl/
A
9
10
11
12
13
14
15
/he. le.n wɪl di kle: rə me. ne. məl/
B
16
17
18
/he. le.n/
19
20
/he. le.n/
C

Figure 1: Stylized shape of the pitch contours used in the experiment.

in place. It was then coupled to a second tube (3/16 inch inside diameter, 8 inches long) and connected to a pressure transducer (Setra Systems, Model 236L).

The EMG and pressure signals were directed to differential amplifiers, then to distribution amplifiers. The physiological signals, together with the audio signal and timing pulses, were recorded on a 14-channel instrumentation recorder (Consolidated Electrodynamics VR-3300). The visual editing of the raw data and their computer processing were performed on the Haskins Laboratories' EMG data-processing system, following the procedures described by Port (1971) and Kewley-Port (1973, 1974). The P_s and EMG data were processed simultaneously. Before averaging, the physiological signals were optionally smoothed using an integration time constant of 75 msec. In order to increase the accuracy of comparative timing measurements, the data were also processed with a 20 msec time constant. F_0 was measured in selected tokens of each contour type using a computer-implemented adaptive autocorrelation method designed by Lukatela (1973).

RESULTS

Inspection of the processed data revealed that the electrode insertion into the VOC apparently had not reached the intended muscle but had been inserted instead into the CT, since the target muscle did not show the expected contraction for swallowing, coughing, or glottal stop production but was active for singing ascending pitch scales. Figure 2 shows the great similarity of the two CT channels. In presenting the data, only results obtained for the right-side CT will be considered. Figure 2 also shows that the SH and ST muscles have a rather similar pattern of activity. This similarity was also observed by Atkinson (1973). Therefore only the data on the SH will be presented. It is clear from Figure 3 that the TH muscle is not obviously related to F_0 changes: its pattern of activity varies little as a function of differences in the F_0 contours. TH will therefore not be considered in further presentations of the data. Since the insertion into the LCA deteriorated during the experiment, the data on this muscle could not be used.

All the relevant data of the experiment are grouped in the Appendix. They are displayed in the following order (from top to bottom in each illustration): fundamental frequency, cricothyroid, sternohyoid, and subglottal pressure. In each graph the thick line represents the physiological signal averaged over 20 to 30 repetitions of the same contour type. The thin line corresponds to the signal of the single token whose correlation with the averaged signal was found to be high on all channels (correlation coefficients ranging from $r_p = +.85$ to $+.98$). The F_0 curve at the top presents the pitch contour of that selected token. The longer utterances have been lined up with respect to the resumption of voicing after [k] in "Heleen wil die kleren meenemen;" the shorter utterances have been lined up with respect to the beginning of the second [e] in "Heleen." In all the illustrations the physiological data have been smoothed with an integration time constant of 75 msec.

In this section on Results we will examine the relationship between the various types of F_0 change and each of the physiological variables. We will point out particular EMG and P_s features that co-occur with the major F_0 changes, but will save the interpretation of the relative importance of these parameters for the Discussion section that follows.

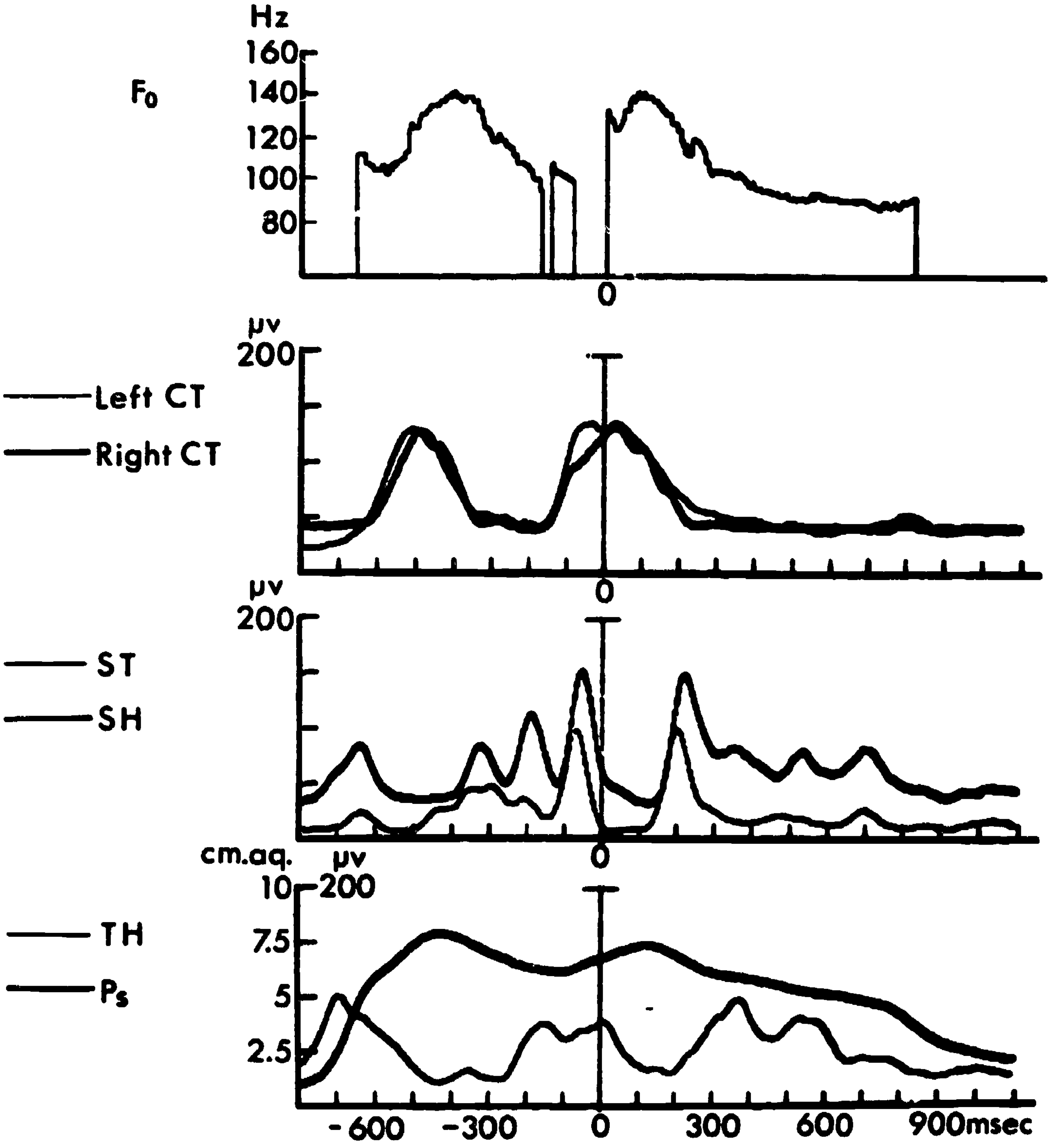


Figure 2 : The EMG activity in five laryngeal muscles and the subglottal air pressure variation, averaged over 26 repetitions of Contour 4.

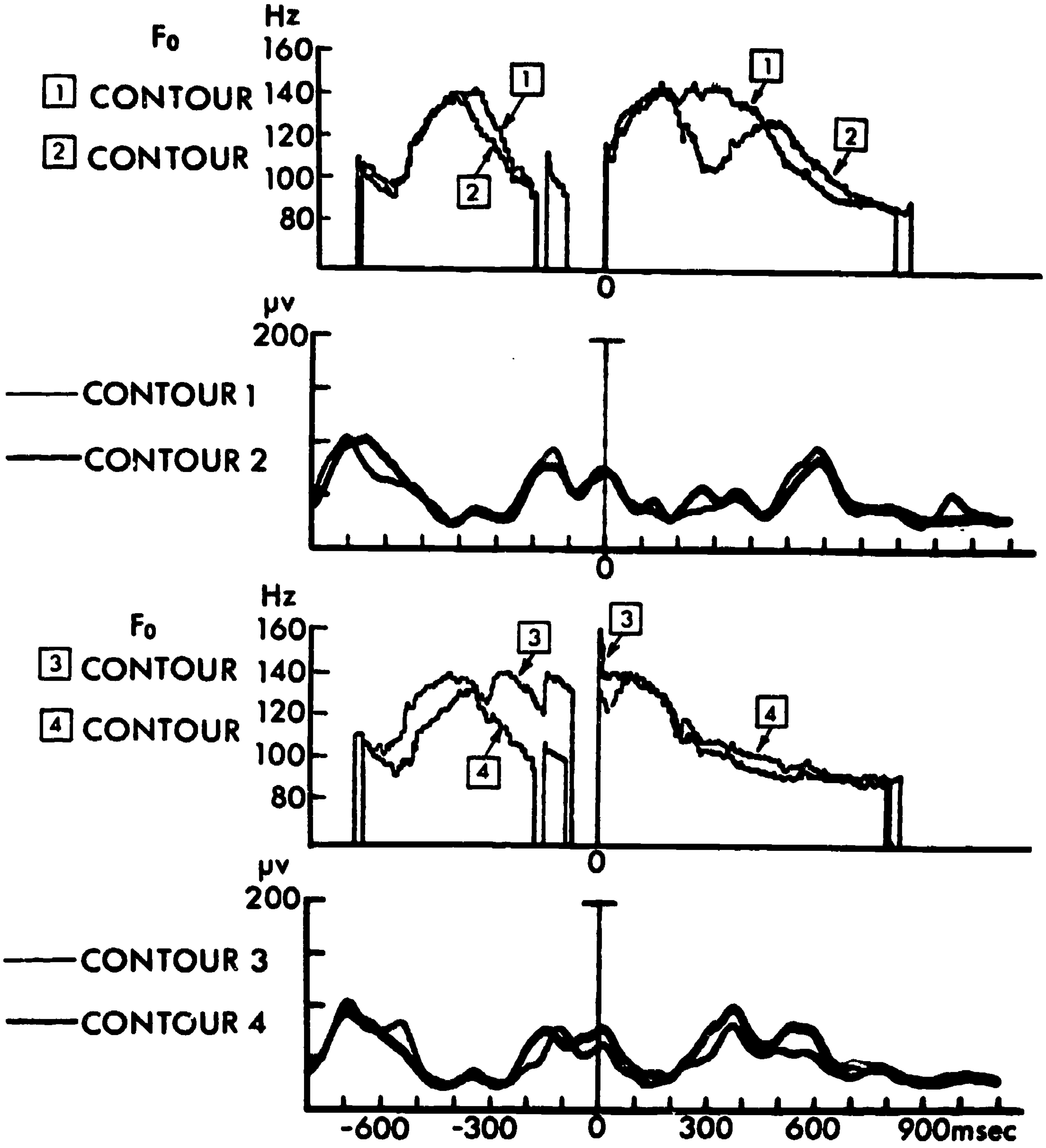


Figure 3: Thyrohyoid muscle activity in four diff^rent contourε.

Pitch Rises

All pitch rises appear to be preceded by increased CT activity. The time delay between a peak in CT activity and the corresponding F_0 maximum, as measured in 30 individual tokens, averages 94 msec (s = 32). The positive correlation between increasing F_0 and increasing CT activity appears to be good also when one examines the rate of increase and the magnitude of the F_0 rise. Thus, if F_0 rises abruptly, CT activity increases equally suddenly. Compare, for example, the first (sudden) and second (smooth) F_0 rises in Contour 12. Also, if two F_0 rises differ in magnitude, then the corresponding peaks in CT activity tend to differ accordingly. For example, compare the slightly different peak values of F_0 in the three rises of Contour 2, or the largely different F_0 peak values in Contours 10, 11, or 12. Contour 9, however, is different in this respect. The peak values of the second and third F_0 rises are almost identical and yet the corresponding CT peaks are different. Apparently, in this instance, where the third F_0 rise occurs on the utterance-final syllable, the CT contracts more strongly to overcome the pitch-lowering effect of the simultaneous, rapid P_s drop.

Usually, when the CT is contracting before a rise in F_0, the SH appears to show strongly reduced or suppressed activity. Examples of this relationship can be seen in Contours 2, 4, and 6. It is also apparent from the comparison of the second halves of Contours 9 and 10. However, there are cases where the SH does have a major peak during the CT contraction. This is particularly true of the SH peak preceding the line-up point in Contours 3, 5, and 13.

Most of the sudden F_0 rises are roughly correlated with increases in P_s, but in many of those instances the peaks in F_0 and P_s are not synchronous: the pressure peak precedes the F_0 peak by 80 msec in Contour 14; by 75 and 65 msec in Contour 6; and by 55, 40, and 35 msec in Contour 2. In other instances P_s and F_0 peaks are indeed synchronous, as in Contours 16, 18, and 19, all of which are short, bisyllabic utterances. In Contour 9 the first P_s peak leads the F_0 peak by some 50 msec, while the second P_s peak is synchronous with F_0. P_s does not increase if a F_0 peak occurs late in the utterance-final syllable, as in Contours 9 and 19. Also, gradual increases in F_0 are not invariably reflected in a smoothly rising P_s: in Contour 12 the correspondence is fairly good, but not in Contour 10. The increases in P_s, associated with rising F_0, range from 0.5 to 2.0 cm aq while the F_0 rises vary between 20 and 75 Hz.

Pitch Falls

Marked drops in F_0 appear to be preceded by relaxation of the CT in all utterances: e.g., in Contours 1, 2, 4, and 6. Apparently differences in the rate of F_0 decrease are also reflected in the rate of CT relaxation. In Contour 1 the first drop in F_0 is more rapid than the second, and so is the pattern of decreasing CT activity. In Contour 9 the first F_0 fall is more gradual than the second; the first F_0 fall in Contour 11 is less steep than the first in Contour 12. All these differences are reflected in the corresponding patterns of CT relaxation.

The onset of a F_0 fall coincides with the beginning of increased SH activity, so that the contraction of the SH reaches a peak by the time the F_0 fall is almost half completed. These temporal relationships are illustrated in Contours 2, 4, 9, and 16.

Most F_0 falls are accompanied by a decrease in P_s. This can be observed whether or not the F_0 fall occurs near the end of the utterance. Thus, if a F_0 fall occurs on a syllable that is not utterance-final (case A, for example in Contour 3), there is one decrease in P_s associated with this F_0 fall and a further one that is related to the end of phonation. If the final fall of a F_0 contour does occur on the utterance-final syllable (case B, for example in Contour 16), there is only one drop in P_s that marks both the F_0 fall and the end of phonation. In case A neither of the P_s drops excedes 2.5 cm aq. In case B the P_s decrease is at least 5 cm aq. In case A the last P_s decrease occurs 50-100 msec before the end of phonation, while in case B the only P_s drop begins 200-300 msec before this point.

High "Declination"

Stretches of high, nearly constant F_0 can be observed in Contours 3, 5, 13, and 14. During these portions the CT always shows continuing activity, while P_s is either constant or slowly falling. As long as F_0 is high, the SH shows partially reduced or suppressed activity.

Low "Declination"

Stretches of low, declining F_0 can be observed in Contours 3, 4, and 6, and especially in Contours 7 and 8. In all these utterance portions the CT is completely passive while the SH shows successive peaks of varying height. P_s is gradually falling and does so at a rate that matches the slowly falling F_0 rather well. In these cases a 5 Hz decrease in F_0 corresponds to a drop of 1 cm aq in P_s.

DISCUSSION

In the previous section we presented the experimental data, looked at the various types of F_0 change, and tried to indicate which physiological variation was associated systematically with the F_0 change. In the present section we approach the data from the opposite angle and examine each physiological parameter with respect to its effect on F_0 variation. The second part of the section deals with the relation between linguistic categories such as "breath group" and "prominence," and the articulatory mechanisms that implement them.

The Function of the Physiological Parameters

The cricothyroid muscle. The pitch-raising effect of CT contraction that we observe in our data has long been known. Ever since the early EMG studies on humans by Katsuki (1950) and Faaborg-Andersen (1957) it has been repeatedly confirmed that of all the intrinsic laryngeal muscles, the CT shows the most direct relationship to increasing F_0, both in singing and in speech. As to the F_0-lowering effect of CT relaxation, Lieberman (1970) observes: "There is no a priori reason to assume that all phonetic features must be implemented by tensing a particular muscle....It is therefore possible that abrupt falls in F_0 could be implemented by relaxing muscles that in their tensed state maintain higher F_0" (p. 199). Simada and Hirose (1970) and Sawashima, Kakita, and Hiki (1973) found that a steep decrease in CT activity preceded the F_0 drop associated with the accent kernel in Japanese. Atkinson (1973), too, has presented many instances where F_0 falls correlate with the decreasing activity of CT (and VOC and LCA).

Our data also indicate that whenever F_0 is high and level, the CT shows a pattern of continued activity. The amount of this activity may or may not be constant, so that prolonged CT contraction invariably correlates well with "high" F_0, but not always with "declining" F_0. Similarly, Sawashima et al. (1973) briefly mention the presence of high and gradually decreasing CT activity during a so-called plateau in two of their F_0 contours.

In our data, there is only one kind of F_0 change that bears no relationship to the CT: when F_0 is generally low and falling very gradually (at a rate of some 15 Hz/sec), the CT is not active and cannot be held responsible for the smoothly falling F_0. This can be seen in Contours 7 and 8. Apart from this exception, the CT always shows a very straightforward correlation with F_0: when one adjusts for the timing difference between the physiological and the acoustic event, the CT matches F_0 changes with respect to their direction (rising or falling), their magnitude (small or large), and their rate (gradual or sudden).

The sternohyoid muscle. Most researchers who have looked into the function of the SH in speech have argued that this muscle can participate in both segmental articulation and F_0 control. As far as segmental articulation is concerned, Ohala and Hirose (1970) and Simada and Hirose (1970) mention that SH contraction is associated with jaw opening, tongue lowering, and tongue retraction, all of which require lowering or fixation of the hyoid bone. Ohala (1970), Gårding, Fujimura, and Hirose (1970), and Atkinson (1973) also observe peaks in SH activity immediately before the onset of phonation and assume that the SH helps in preparing the larynx for the "speech mode." Atkinson further finds SH activity during voiceless consonant closure and relates this to the resumption of phonation after the consonant. Our test utterances do not contain open vowels, but most of the SH peaks can indeed be traced back to tongue retraction for the release of [l,n,d] or to tongue lowering for the release of [k]. Our data also show a peak in SH activity before the onset of phonation.

With respect to the participation of the SH in F_0 control, a number of researchers have pointed out that the SH is active during the transition from high to low F_0, as well as during low, level F_0 (Ohala, 1970; Atkinson, 1973; Sawashima et al., 1973). Ohala (1972) and Kakita and Hiki (1974) have attempted to account mechanically for the effect of SH contraction on F_0. In our own opinion, however, there remain reasons why the SH cannot be considered the primary effector of F_0 lowering. The first reason is that the lack of timing difference between the two variables makes a direct causal relationship unlikely. Since it takes time for a muscle to contract and become effective, one should expect SH activity to start well before the onset of the F_0 drop. In our data (and in those of Atkinson, 1973, and Sawashima et al., 1973) SH contraction coincides with the beginning of the F_0 fall. The second reason is that SH activity is the same for abrupt changes from high to low F_0 as for steady, low F_0. The third reason is the imperfect reciprocity between the patterns of CT and SH activity. If the two muscles were antagonists, SH would relax whenever CT contracts. However, our data show many exceptions to this tendency (for example, in Contours 3, 5, and 13). The fourth reason is that when one imitates a pitch contour by humming it (thus eliminating all segmental effects), there is almost no activity in the SH throughout the contour, not even for F_0 falls as large as 70 Hz.

Subglottal air pressure. Studies such as those discussed in Ohala (1970) indicate that variations in P_s can have an effect on the rate of vocal cord vibration. Our own data show that in certain pitch contour types the overall high and relatively steady P_s level is modulated by fluctuations that roughly correspond to the rises and falls in F_0. Ladefoged (1962) presented EMG data that show increased activity in the internal intercostal muscles immediately before accented syllables. Thus the momentary increases in P_s, associated with rising-falling F_0 on accented syllables, may well be the result of active, expiratory muscle control. Yet one must not overlook the possibility that the observed changes in P_s also reflect the variations in glottal resistance. Indeed, some of the intrinsic laryngeal muscles, especially the CT, but also the VOC and LCA, have a pitch-dependent pattern of activity. Thus it is conceivable that the synergetic, F_0-raising contraction of these muscles results in a gradually stronger resistance to the flow of air from the lungs, so that P_s passively increases in the vicinity of an accented syllable carrying a F_0 rise. The comparisons in Figure 4 show that the level of CT contraction may indeed explain to some extent the modulations in the P_s curve.

Whatever the origin of the P_s fluctuations may be, it appears that in our data they cannot be considered the prime cause of the major F_0 changes. For one thing, they appear to be too small to account for the full extent of the F_0 changes, unless one accepts as normal a $\Delta F_0/\Delta P_s$ ratio as high as 40/1. For another, the timing relationship between P_s and F_0 is highly variable and the expected synchrony of the peaks of the two variables is lacking in most of the cases. Finally, our data contain instances where P_s does not reflect the F_0 changes at all. This is particularly true of the F_0 rise in the second half of Contours 10 and 15, and of the F_0 falls in Contour 14.

Only two types of F_0 change in our data have a fairly good correlation with P_s. Subglottal pressure apparently controls the course of F_0 during those portions of the contour where no major rises or falls occur. In such cases the $\Delta F_0/\Delta P_s$ ratio is approximately 5/1. This interpretation corresponds to that of Atkinson (1973), who states: "The steady or slightly falling pressure...seems capable of controlling to a large extent any steady or slightly falling F_0 contour" (p. 117). In our description of F_0 falls we noted that P_s decreases more strongly and earlier in the utterance-final syllables that carry the terminal F_0 fall of the contour than in those that do not. Thus, considering its magnitude and timing, rapidly falling P_s may indeed control rapidly falling F_0 in utterance-final syllables. Here the $\Delta F_0/\Delta P_s$ ratio is 12/1. The difference between these two cases is that in the former, the CT is completely passive and cannot influence F_0, while in the latter, the F_0 fall can be related not only to falling P_s but also to decreasing CT activity.

The Articulatory Implementation of Prosodic Categories

The contours presented in Figure 1a are variants of the same intonation pattern (Cohen and 't Hart, 1967; 't Hart and Cohen, 1973). This pattern corresponds to the "unmarked breath group" [-BG] in the feature system of Lieberman (1967, 1970) and Lieberman, Sawashima, Harris, and Gay (1970). The contours in Figure 1b all differ from those in Figure 1a, and some, if not all, also differ from each other (Collier, 1972; Collier and 't Hart, 1972). In Figure 1b only Contours 9 and 10 (and its variant Contour 13) correspond to Lieberman's "marked breath group" [+BG]. In the feature system of Vanderslice and Ladefoged (1972) Contour 9 would be [+cadence, +endglide], while Contour 10 (or 13) would be [-cadence, +endglide]. Contours 11, 12, 14, and 15 cannot be described in terms of either feature system. Let us therefore limit our attention just to those contours that seem common to English and Dutch.

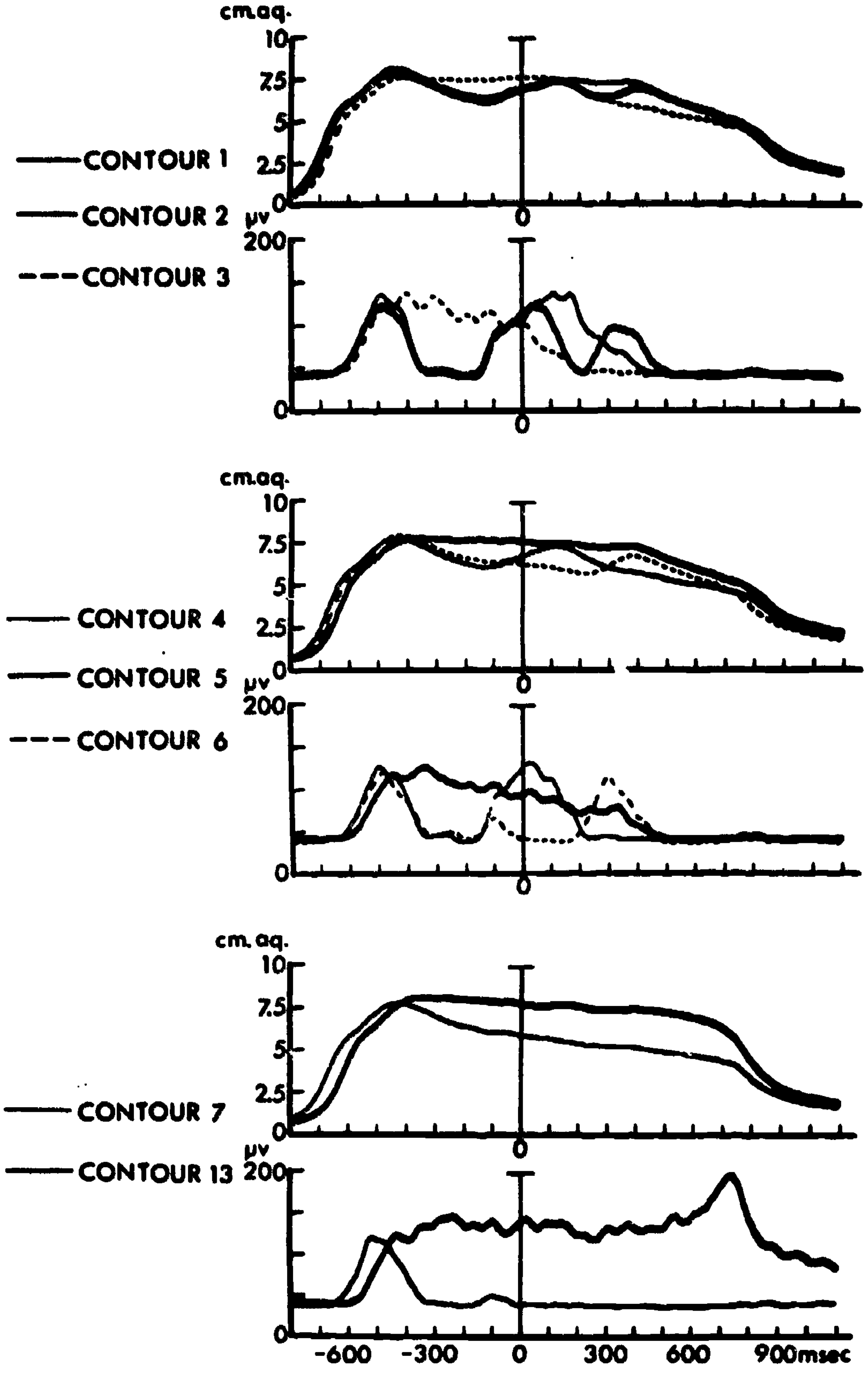


Figure 4: Cricothyroid muscle activity and subglottal air pressure variation in eight different contours.

Unmarked breath groups appear in Contours 1 to 8. They all show laryngeal activity associated with the F_0 rises and falls, irrespective of the location of the F_0 changes in the contour. Subglottal pressure also shows momentary rises and falls associated with the F_0 variations. Contours 9 and 10 (and 13) are examples of marked breath groups. They too show laryngeal activity all through the contour. Subglottal pressure also reflects the F_0 variation except for the F_0 rise on the utterance-final syllable, where P_s is falling.

If we look upon a F_0 contour as the product of both ± breath group and ± prominence specifications, there are at least two ways of sorting out the respective effects of these two prosodic categories. One way is to consider the breath group as the overall configuration of F_0 changes whose actual distribution over the contour is a function of the prominence specification of the syllables. The relationship between [BG] and [PROM] is then: all F_0 changes are the implementation of [BG], while some of them simultaneously implement [+PROM] by occurring on the prominent syllables. (Whether a F_0 change is a good cue for prominence depends mainly on its timing with respect to the syllable boundaries; see van Katwijk, 1974.) An alternative way is to consider the unmarked breath group to be the gradually declining base line of the overall F_0 contour, which further consists of major F_0 changes that implement [+PROM] and are superimposed on the breath group. The marked breath group is the same F_0 base line as the unmarked, but a rise in F_0 is added near the end. This rise, unlike some other major F_0 changes that precede it, does not implement [+PROM].

Evidently the major difference between these interpretations is that in the first view breath group is completely synonymous with F_0 contour, while in the second view breath group in fact equals declination line, with or without a terminal rise. The second interpretation is a possible paraphrase of the views expressed by Lieberman (1970) and Lieberman et al. (1970), while the first reflects our own position. It is clear that these differences of opinion are situated on the more abstract, linguistic level (where mental constructs such as breath group and prominence belong). Let us therefore examine separately what agreement there may be in specifying the articulatory correlates of these prosodic categories.

Lieberman (1970) and Lieberman et al. (1970) assume that the articulatory correlate of [-BG] is the pattern of respiratory muscle control that can be used to generate a relatively steady subglottal air pressure contour; [+BG] involves the participation of laryngeal muscles which produce the contour-final F_0 rise; [+PROM] syllables are characterized by momentary variations in both subglottal pressure and laryngeal tension. Lieberman further assumes that increased P_s on [+PROM] syllables is the "archetypal," primary correlate of this feature, while increased CT activity is a secondary characteristic. Our own view is that P_s controls the lower declination line only and that all major F_0 changes (whichever prosodic category they implement) are under laryngeal control. Thus we do not think that changing P_s can be the primary articulatory correlate of major F_0 variations, since the P_s variations in our data that are associated with accentual or intonational features are too small to account for the full extent of the F_0 change.

Lieberman (1970) appears to associate [+PROM] with rising-falling F_0. Whenever linguistic stress is manifested by unidirectional F_0 changes, i.e., by a simple rise or a simple fall, Lieberman considers these changes as the acoustic correlates of [+ACCENT UP], [+ACCENT DOWN], not of [+PROM]. He says that their articulatory implementation is under strictly laryngeal control (i.e., without accompanying P_s variation). By contrast, our own interpretation is that [+PROM] can be realized as rising-falling, rising, or falling F_0 (van Katwijk, 1974). This assumption is necessary to explain how pairs of contours such as 1 and 2, 3 and 4, 5 and 6 are considered as free variants by native speakers of Dutch ('t Hart and Cohen, 1973): both variants implement the same type of breath group with the same numbei of linguistic stresses on the same syllables. Thus it would not be plausible to consider the same syllable as [+PROM] in one variant and as [+ACCENT UP] or [+ACCENT DOWN] in the other. Apart from this difference in linguistic interpretation, we share the view with Lieberman that the articulatory correlate of these unidirectional F_0 changes is to be found in laryngeal maneuvers.

CONCLUSION

Our experiment suggests that the gradually falling baseline of a F_0 contour is controlled by the slowly decreasing subglottal air pressure, while the major deviations from this baseline (i.e., the rises and falls in F_0) are caused by the action of the cricothyroid muscle. The increasing activity of this muscle raises F_0, its continued contraction maintains high F_0, and its relaxation lowers F_0. Thus, the major differences between any two F_0 contours appear to be related systematically to differences in the activity of this single muscle. Other laryngeal or respiratory muscles may well assist in producing a change in F_0, but their effect is secondary.

REFERENCES

Atkinson, J. (1973) Aspects of intonation in speech: Implications from an experimental study of fundamental frequency. Ph.D. thesis, University of Connecticut.

Cohen, A. and J. 't Hart. (1967) On the anatomy of intonation. Lingua 19, 177-192.

Collier, R. (1972) From pitch to intonation. Ph.D. thesis, University of Leuven, Belgium.

Collier, R. and J. 't Hart. (1972) Perceptual experiments on Dutch intonation. In Proceedings of the Seventh International Congress of Phonetic Sciences. (The Hague: Mouton) 880-884.

Faaborg-Andersen, K. (1957) Electromyographic investigation of intrinsic laryngeal muscles in humans. Acta Physiol. Scand. 41, Suppl. 140.

Gårding, E., O. Fujimura, and H. Hirose. (1970) Laryngeal control of Swedish word tones. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) 4, 45-54.

Hirose, H. (1971) Electromyography of the articulatory muscles: Current instrumentation and techniques. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.

Kakita, Y. and S. Hiki. (1974) A study of laryngeal control for voice pitch based on anatomical model. Preprints of the Eighth International Congress on Acoustics, London, July, 222.

Katsuki, Y. (1950) The function of the phonatory muscles. Jap. J. Physiol. 1, 29-36.

Kewley-Port, D. (1973) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research SR-33, 173-183.

Kewley-Port, D. (1974) An experimental evaluation of the EMG data processing system: Time constant choice for digital integration. Haskins Laboratories Status Report on Speech Research SR-37/38, 65-72.

Ladefoged, P. (1962) Subglottal activity during speech. In Proceedings of the Fourth International Congress of Phonetic Sciences. (The Hague: Mouton) 73-91.

Lieberman, P. (1967) Intonation, Perception, and Language. (Cambridge, Mass.: MIT Press).

Lieberman, P. (1970) A study of prosodic features. Haskins Laboratories Status Report on Speech Research SR-23. 179-208.

Lieberman, P., M. Sawashima, K. S. Harris, and T. Gay. (1970) The articulatory implementation of breath group and prominence: Cricothyroid muscular activity in intonation. Language 46, 312-327.

Lukatela, G. (1973) Pitch determination by adaptive autocorrelation method. Haskins Laboratories Status Report on Speech Research SR-33, 185-193.

Ohala, J. (1970) Aspects of the control and production of speech. U.C.L.A. Working Papers in Phonetics 15, 1-192.

Ohala, J. (1972) How is pitch lowered? J. Acoust. Soc. Amer. 52, 124(A).

Ohala, J. and H. Hirose. (1970) The function of the sternohyoid muscle in speech. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) 4, 41-44.

Port, D. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research SR-25/26, 67-72.

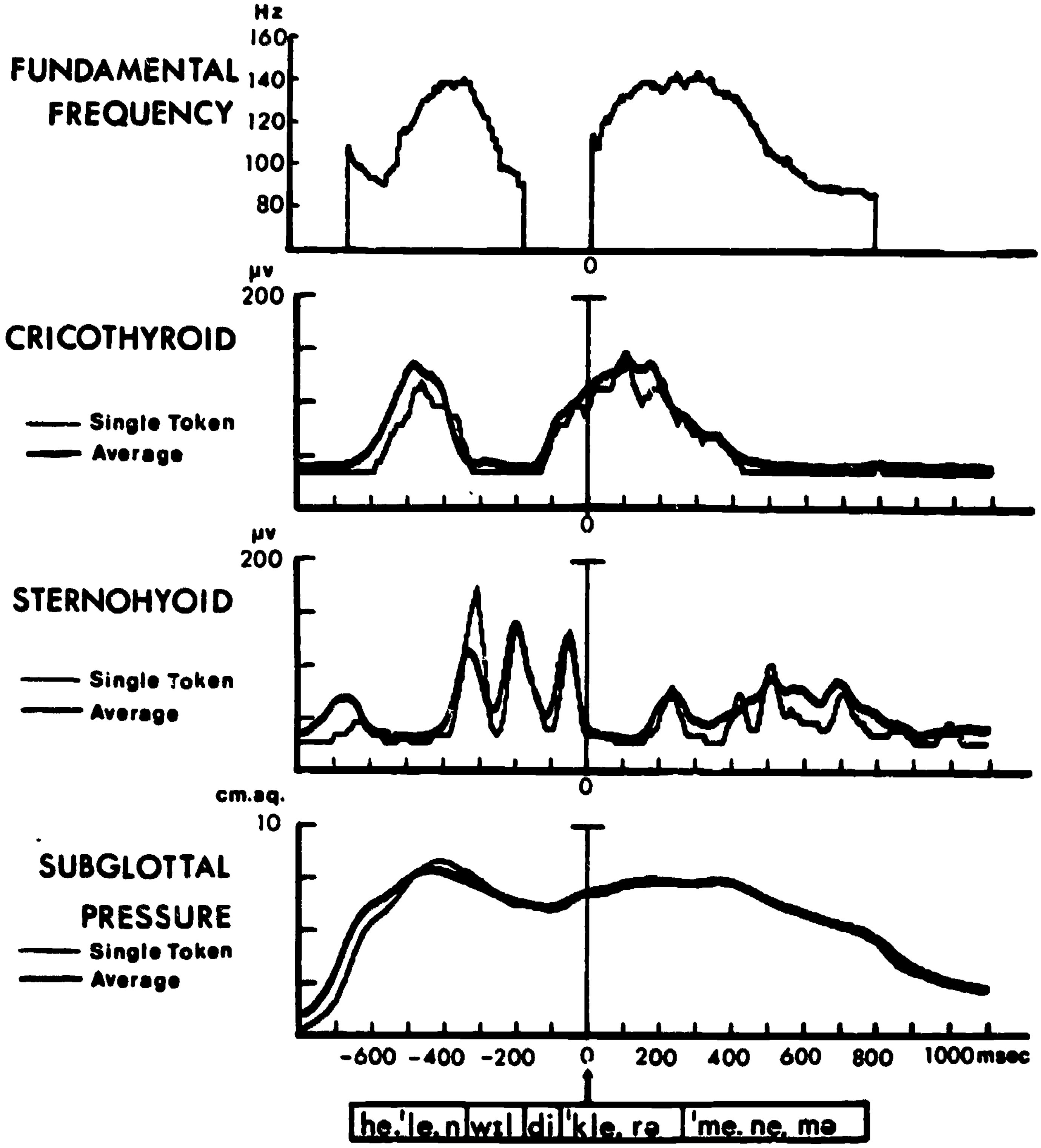
't Hart, J. and A. Cohen. (1973) Intonation by rule: A perceptual quest. J. Phonetics 1, 309-327.

Vanderslice, R. and P. Ladefoged. (1972) Binary suprasegmental features and transformational word-accentuation rules. Language 48, 819-838.

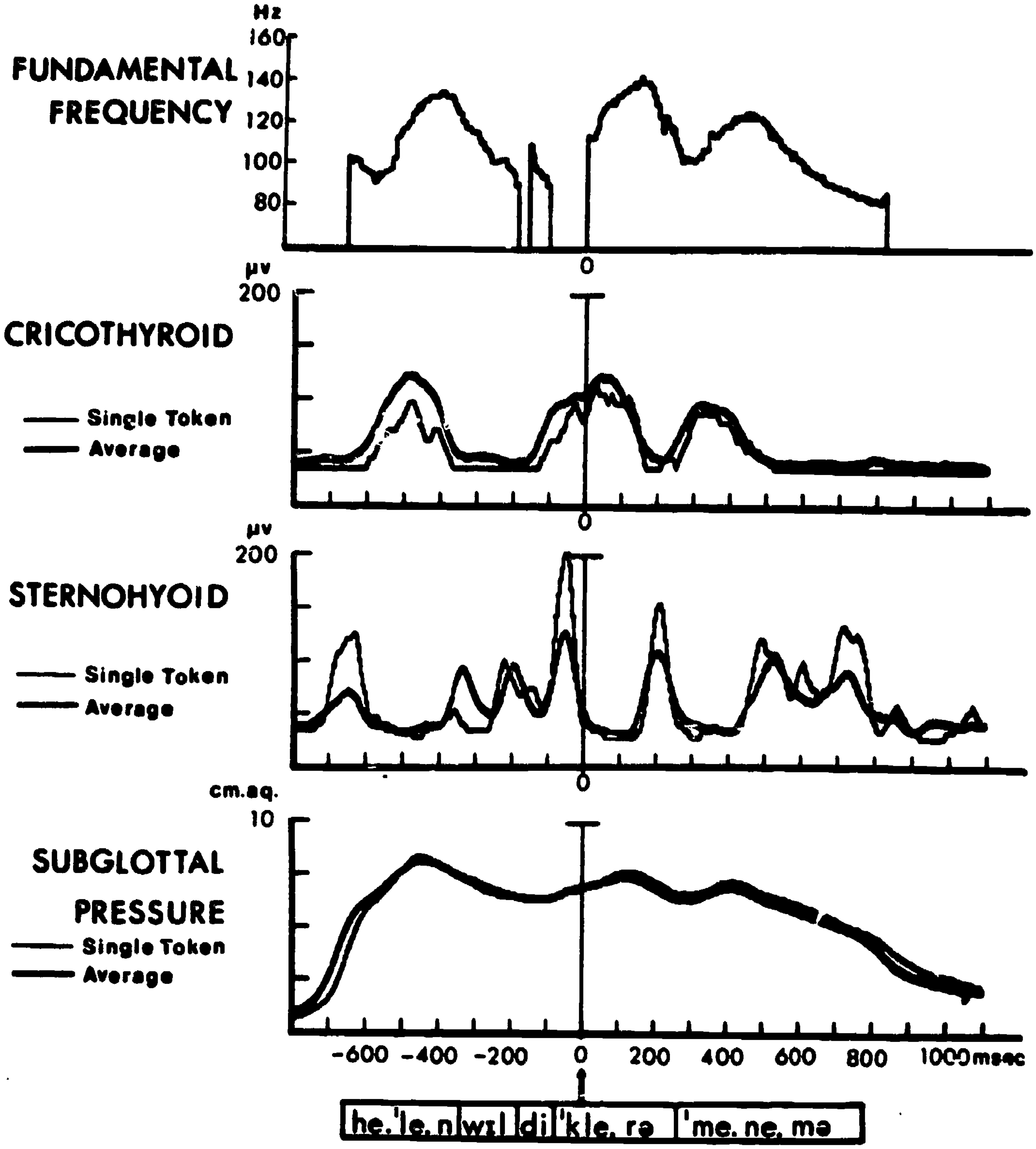
van Katwijk, A. (1974) Accentuation in Dutch. Ph.D. thesis, University of Utrecht, The Netherlands.

Sawashima, M., Y. Kakita, and S. Hiki. (1973) Activity of the extrinsic laryngeal muscles in relation to Japanese word accent. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) 7, 19-25.

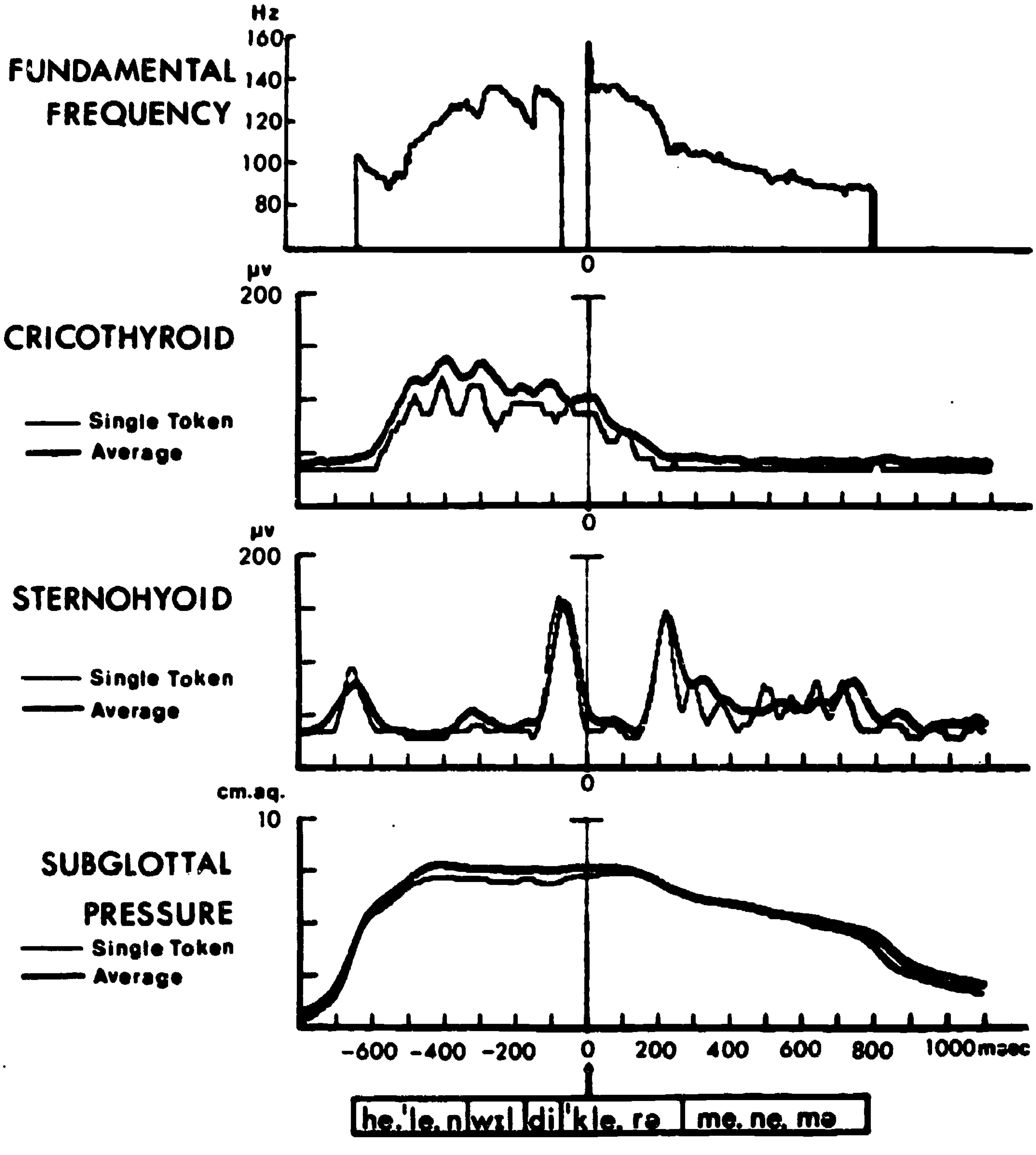
Simada, Z. and H. Hirose. (1970) The function of laryngeal muscles in respect to word accent distinction. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) 4, 27-40.


Hz
160
140
120
100
80
FUNDAMENTAL FREQUENCY
0
μv
200
CRICOTHYROID
Single Token
Average
0
μv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL PRESSURE
Single Token
Average
-600
-400
-200
0
200
400
600
800
1000 msec
he.'le.n
wɪl
di
'kle.rə
'me.ne.mə

CONTOUR 1

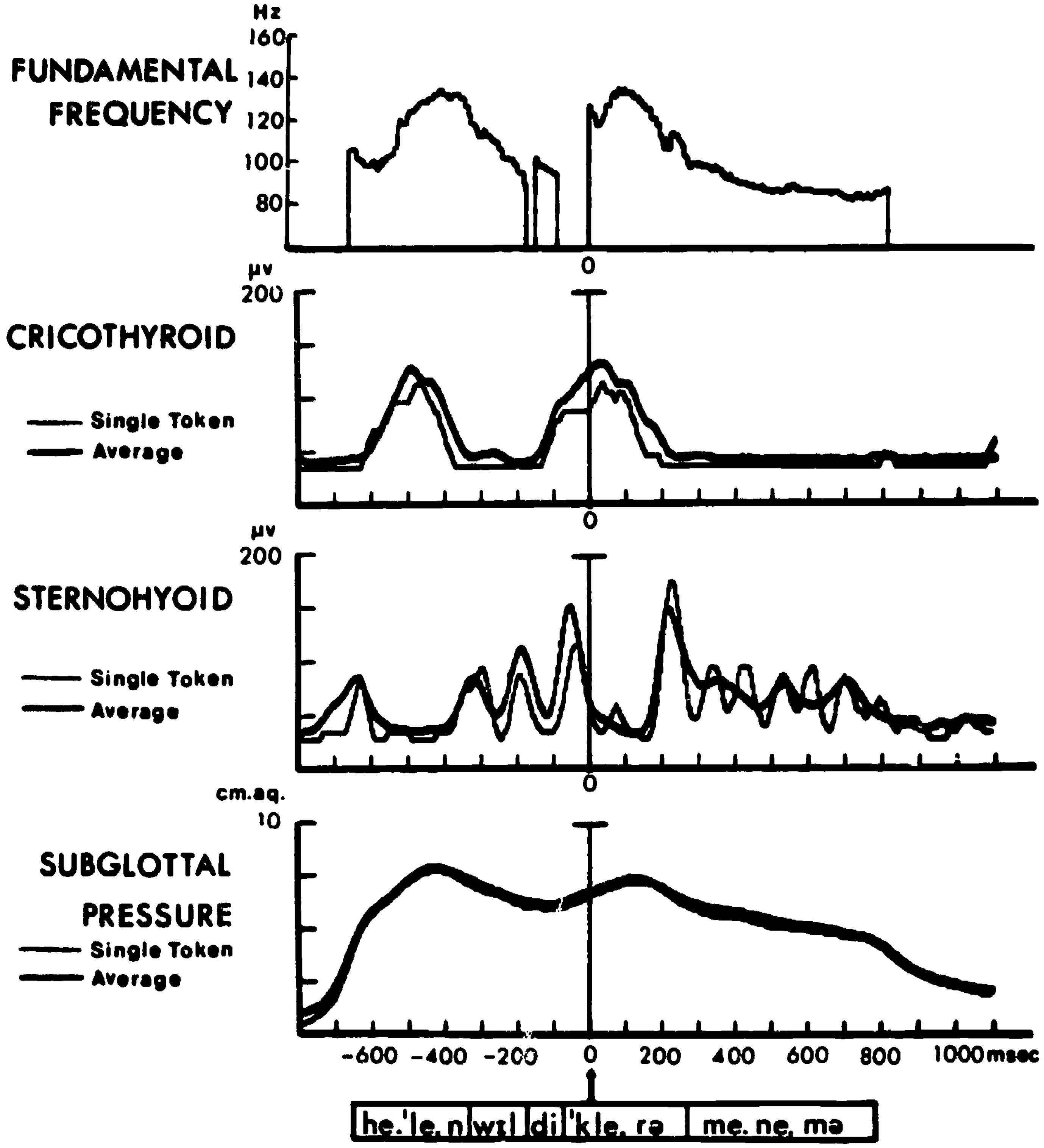

Hz
160
140
120
100
80
FUNDAMENTAL FREQUENCY
0
µv
200
CRICOTHYROID
Single Token
Average
0
µv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL PRESSURE
Single Token
Average
-600 -400 -200 0 200 400 600 800 1000msec
he.'le.n
wɪl
di
'kle.ra
'me.ne.mə

CONTOUR 2

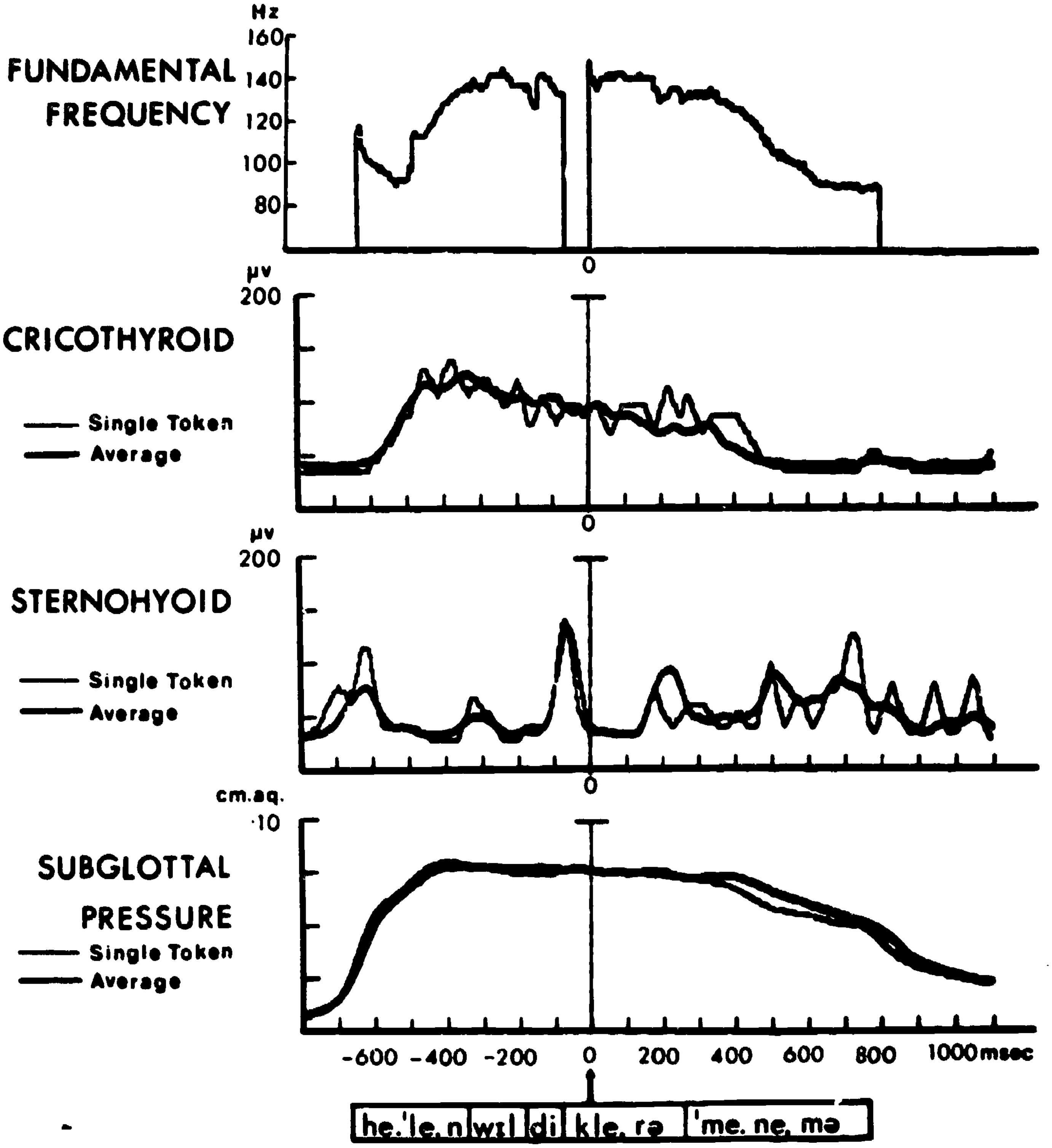

Hz
160
140
120
100
80
FUNDAMENTAL FREQUENCY
0
µv
200
CRICOTHYROID
Single Token
Average
0
µv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL PRESSURE
Single Token
Average
-600 -400 -200 0 200 400 600 800 1000msec
he.'le.n wɪl di 'kle.rə me.ne.mə

CONTOUR 3

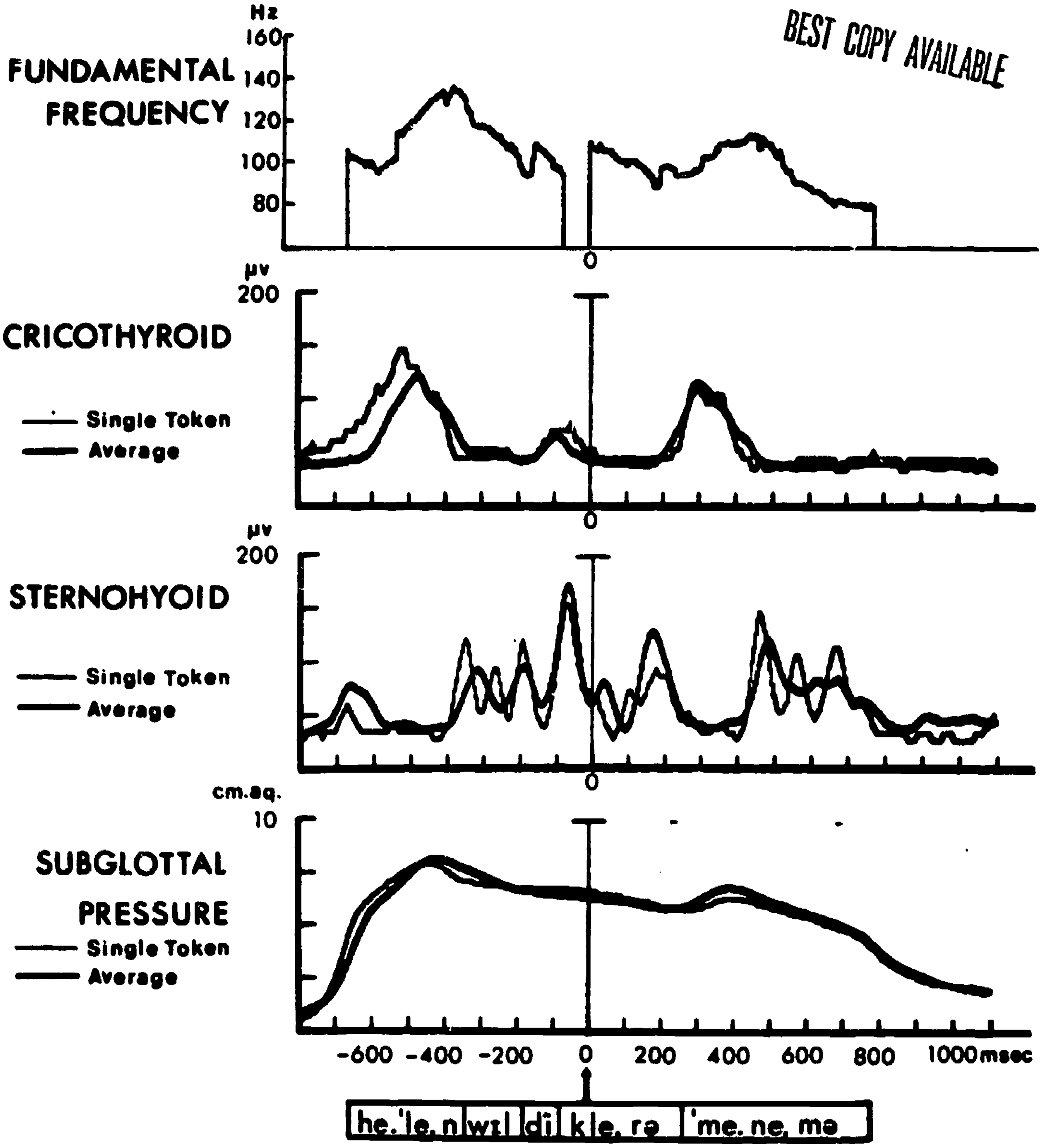
BEST COPY AVAILABLE



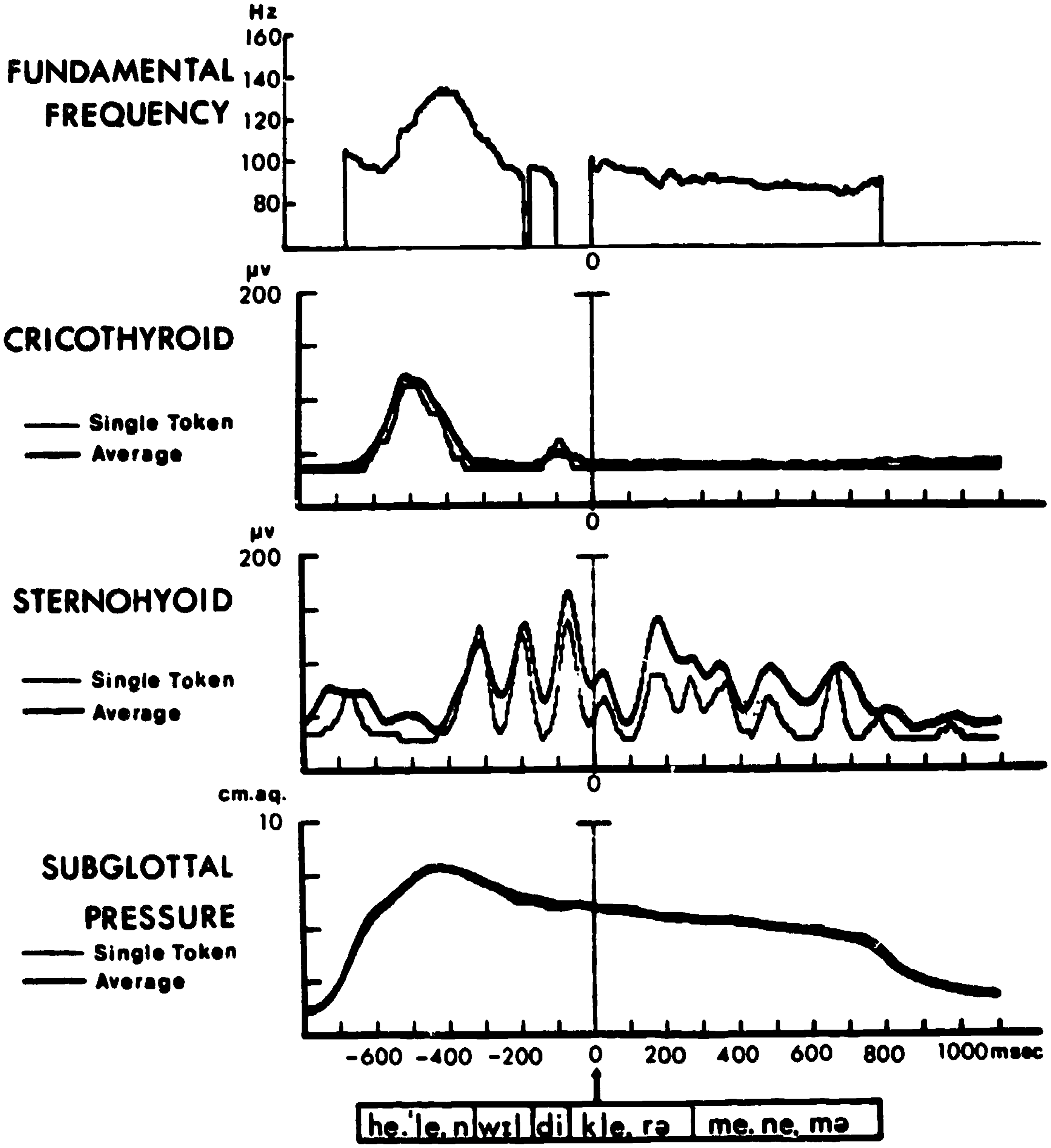
CONTOUR 4


Hz
160
140
120
100
80
FUNDAMENTAL FREQUENCY
0
μv
200
CRICOTHYROID
Single Token
Average
0
μv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL PRESSURE
Single Token
Average
-600
-400
-200
0
200
400
600
800
1000msec
he.ˈle.n
wɪl
di
kle.rə
ˈme.ne.mə

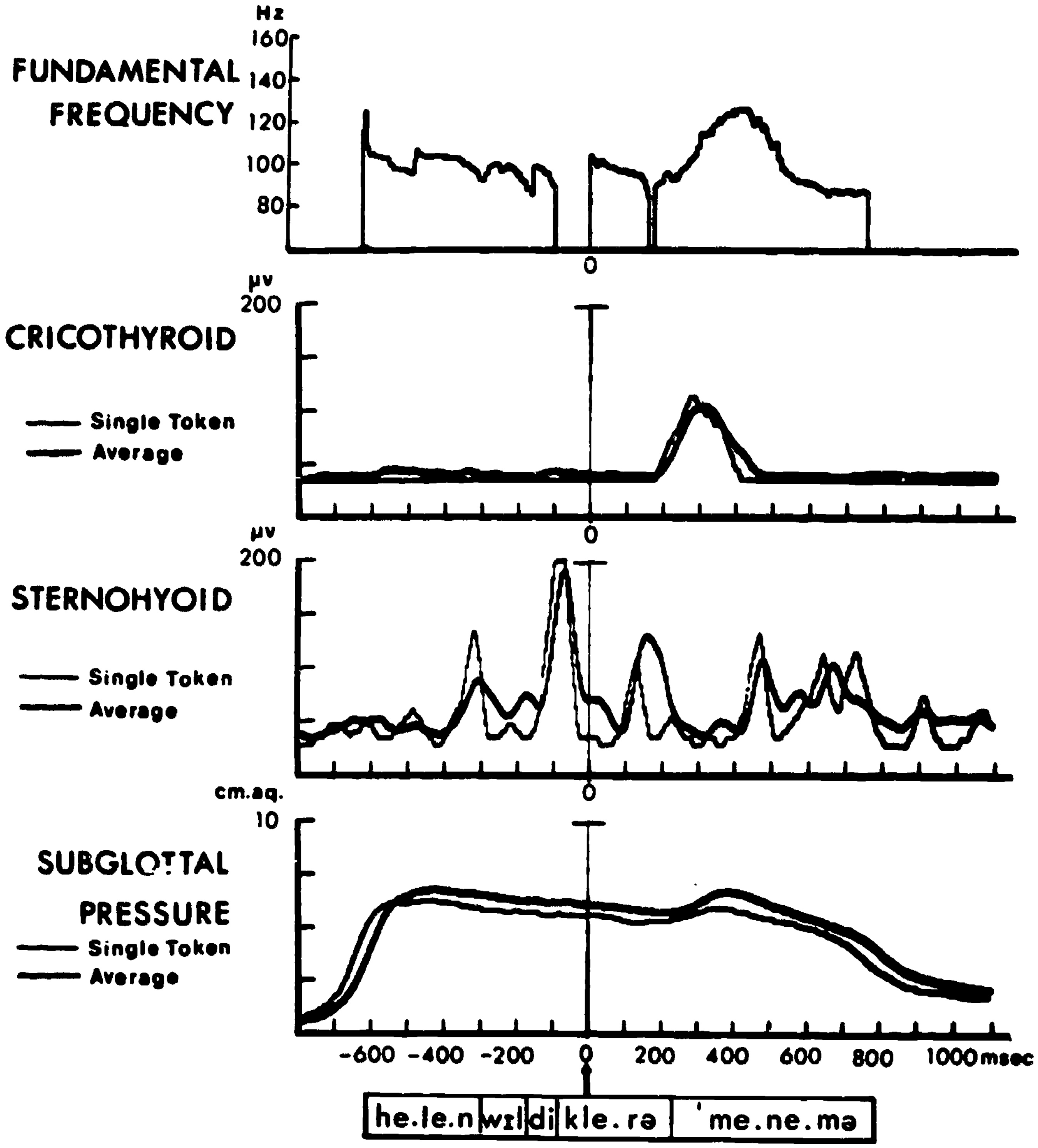
CONTOUR 5


FUNDAMENTAL FREQUENCY
Hz
160
140
120
100
80
0
CRICOTHYROID
μv
200
Single Token
Average
0
STERNOHYOID
μv
200
Single Token
Average
0
SUBGLOTTAL PRESSURE
cm.aq.
10
Single Token
Average
-600 -400 -200 0 200 400 600 800 1000msec
he.'le.n wɪ dĩ k e. rə 'me. ne. mə

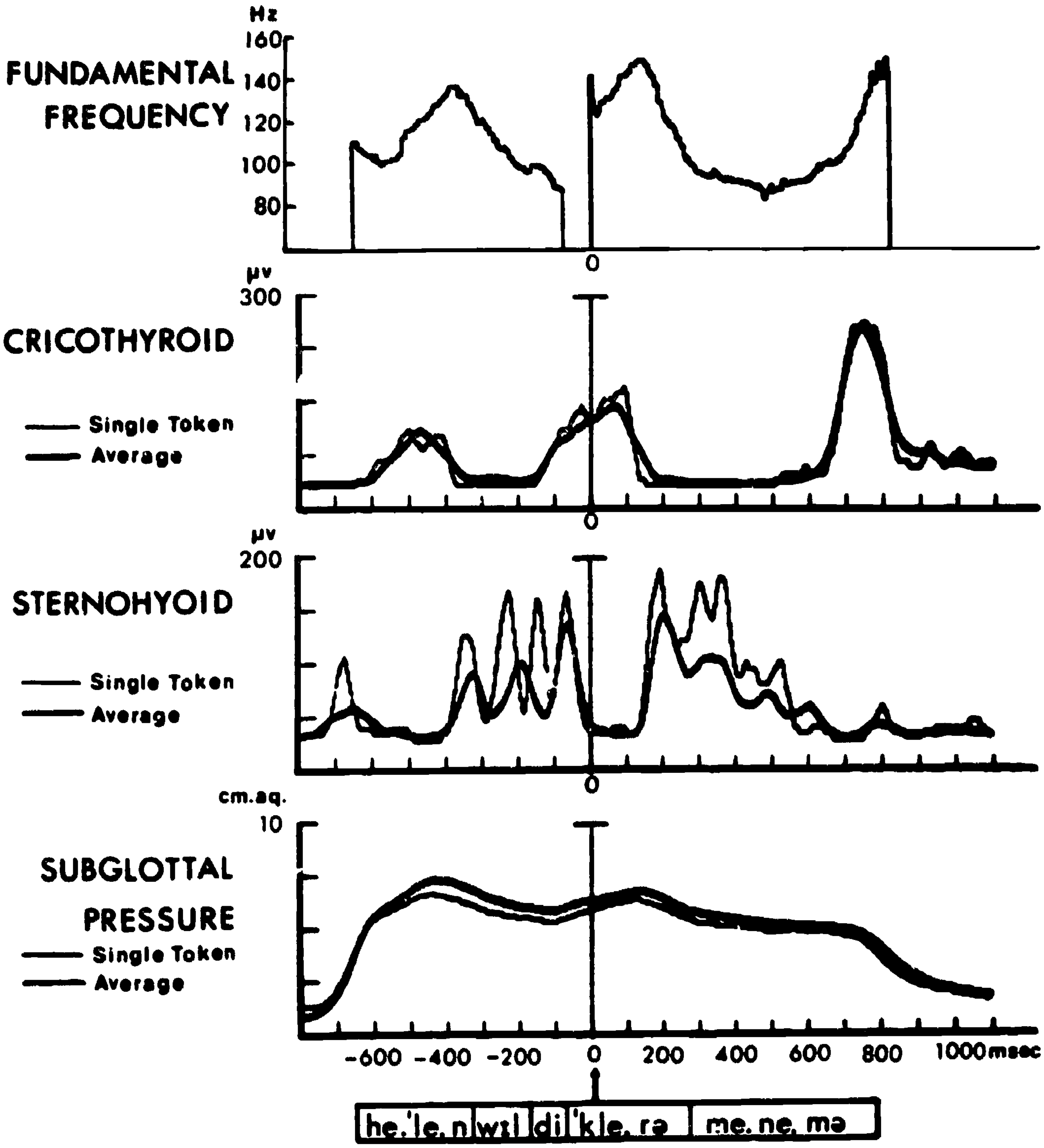
CONTOUR 6


Hz
160
140
120
100
80
FUNDAMENTAL FREQUENCY
0
µv
200
CRICOTHYROID
Single Token
Average
0
µv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL PRESSURE
Single Token
Average
-600 -400 -200 0 200 400 600 800 1000msec
he.'le.n
wɪl
di
kle.rə
me.ne.mə

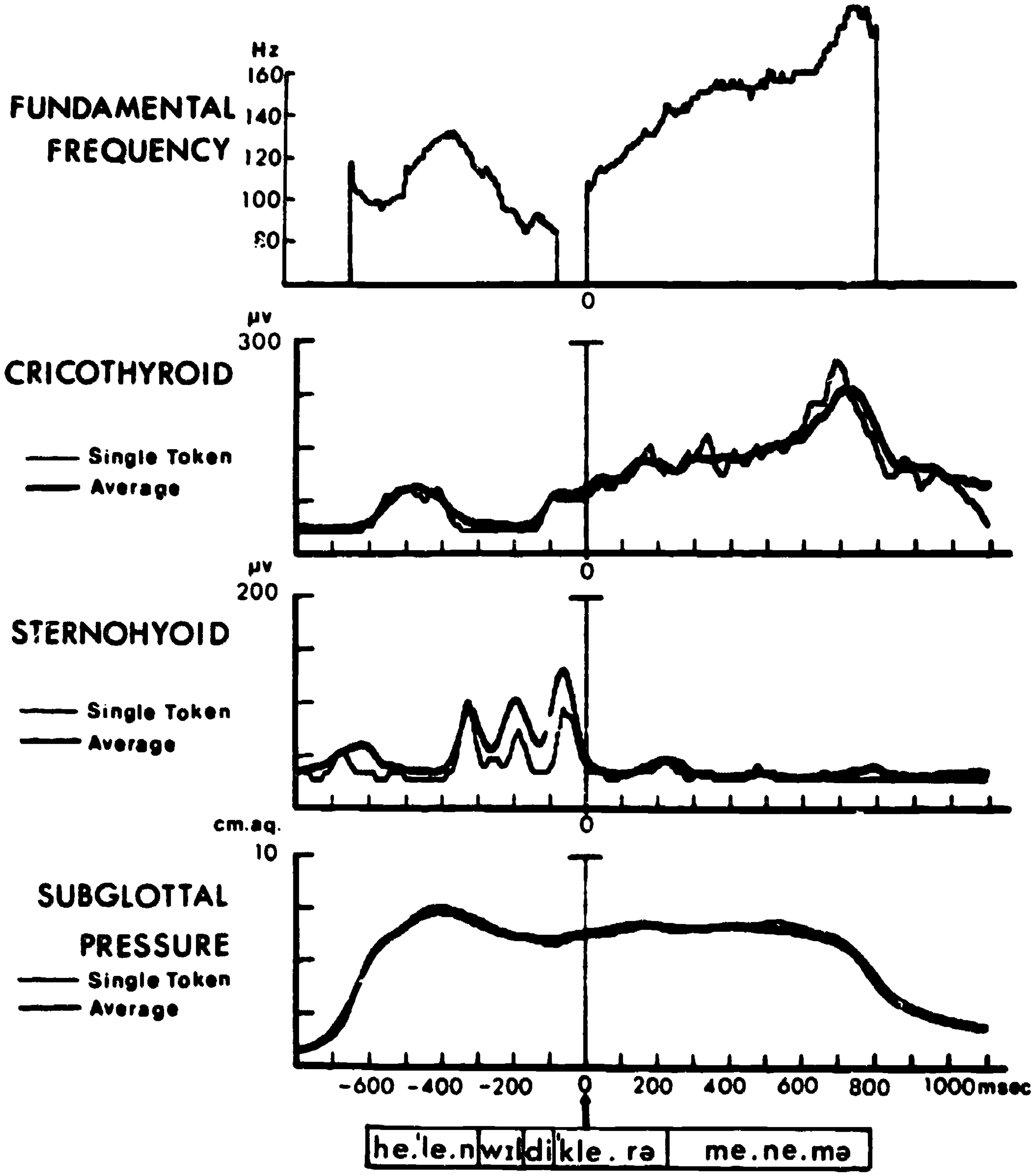
CONTOUR 7


Hz
160
140
120
100
80
FUNDAMENTAL
FREQUENCY
0
μv
200
CRICOTHYROID
Single Token
Average
0
μv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL
PRESSURE
Single Token
Average
-600
-400
-200
0
200
400
600
800
1000 msec
he.le.n
wɪl
di
kle.rə
'me.ne.ma

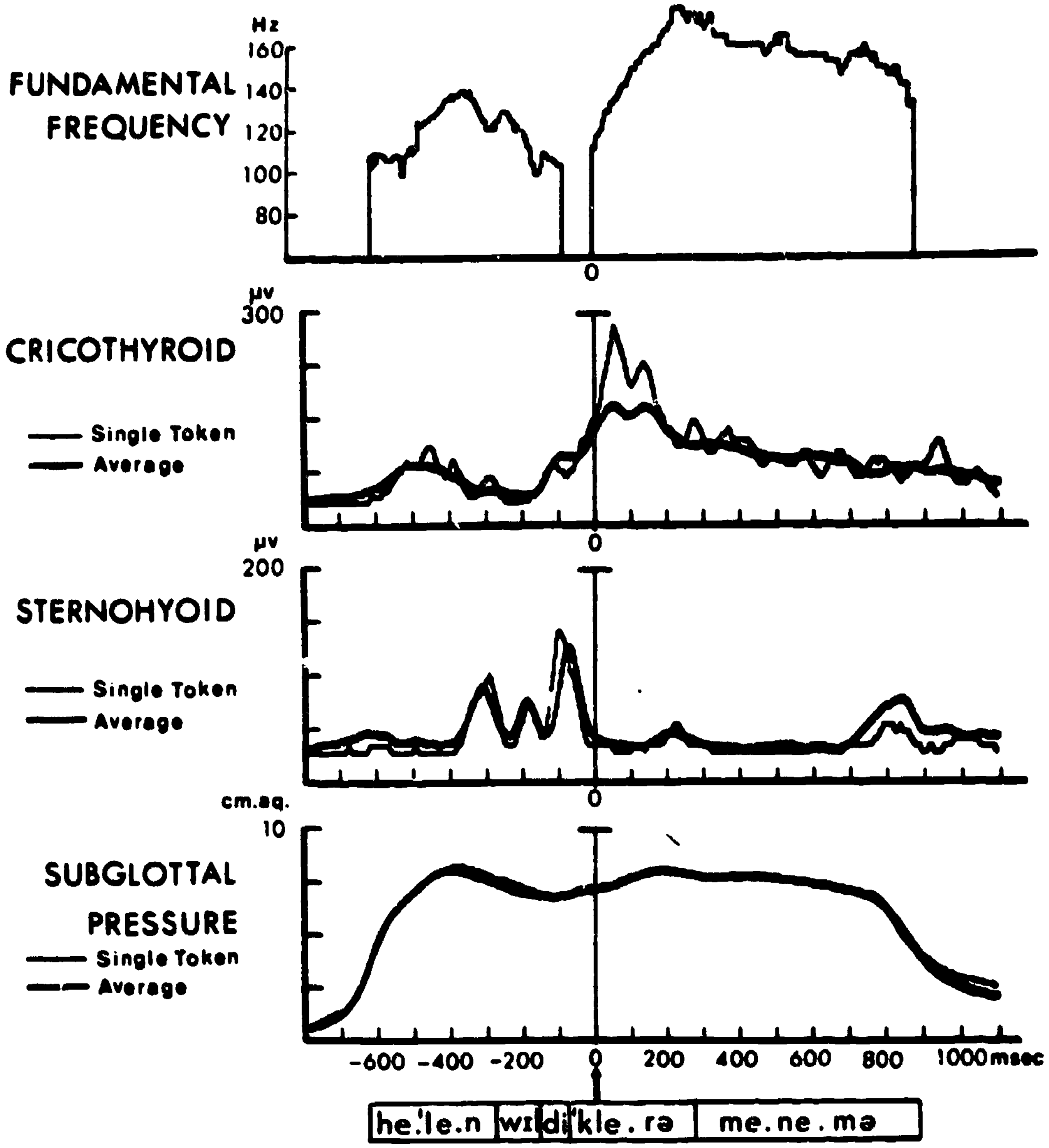
CONTOUR 8


Hz
160
140
120
100
80
FUNDAMENTAL FREQUENCY
0
µv
300
CRICOTHYROID
Single Token
Average
0
µv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL PRESSURE
Single Token
Average
-600 -400 -200 0 200 400 600 800 1000msec
he.'le.n wɪl di 'kle.rə me.ne.mə

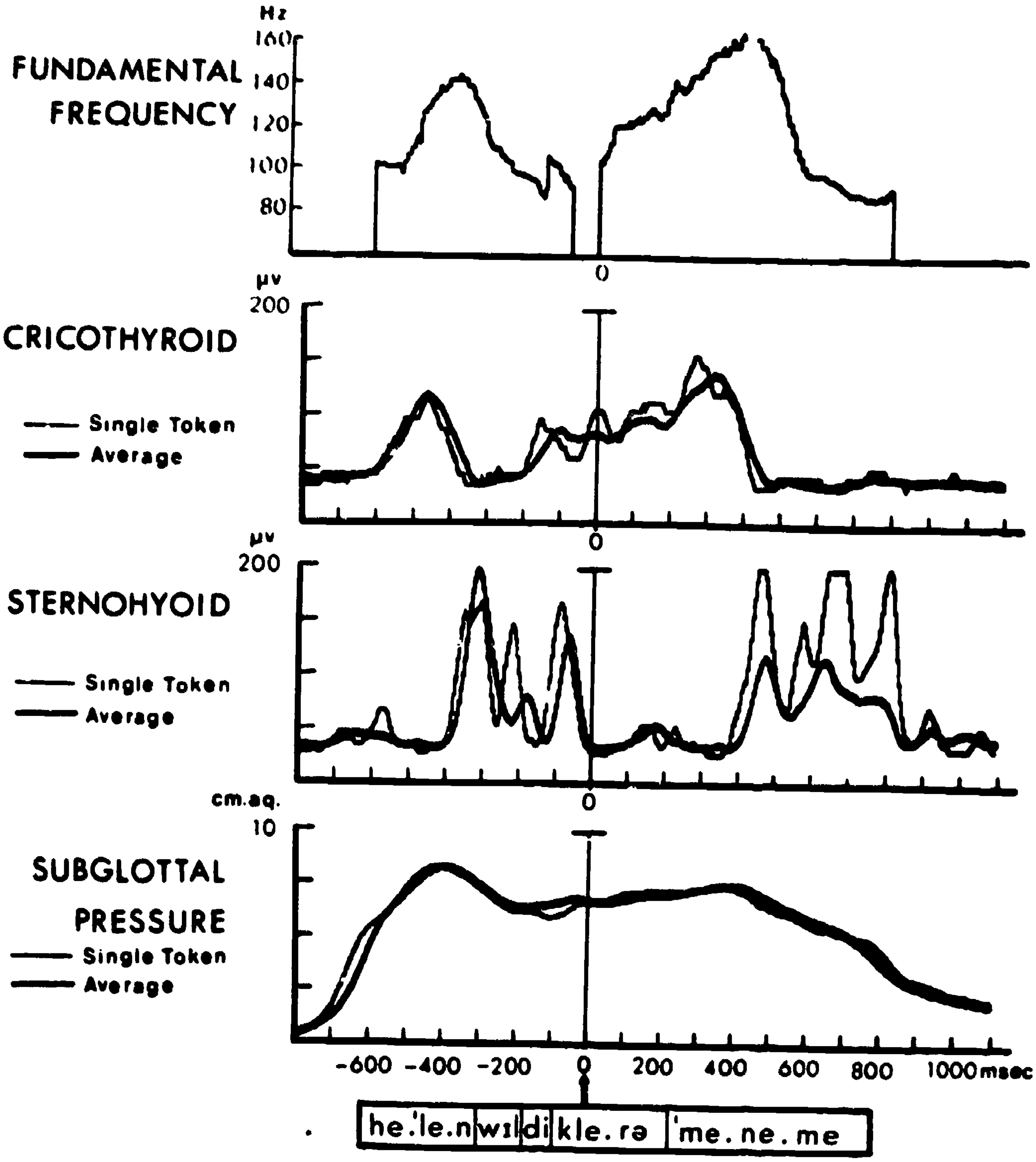
CONTOUR 9


Hz
160
140
120
100
80
FUNDAMENTAL FREQUENCY
0
μv
300
CRICOTHYROID
Single Token
Average
0
μv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL PRESSURE
Single Token
Average
-600 -400 -200 0 200 400 600 800 1000msec
he.'le.n wɪ di 'kle.rə me.ne.mə

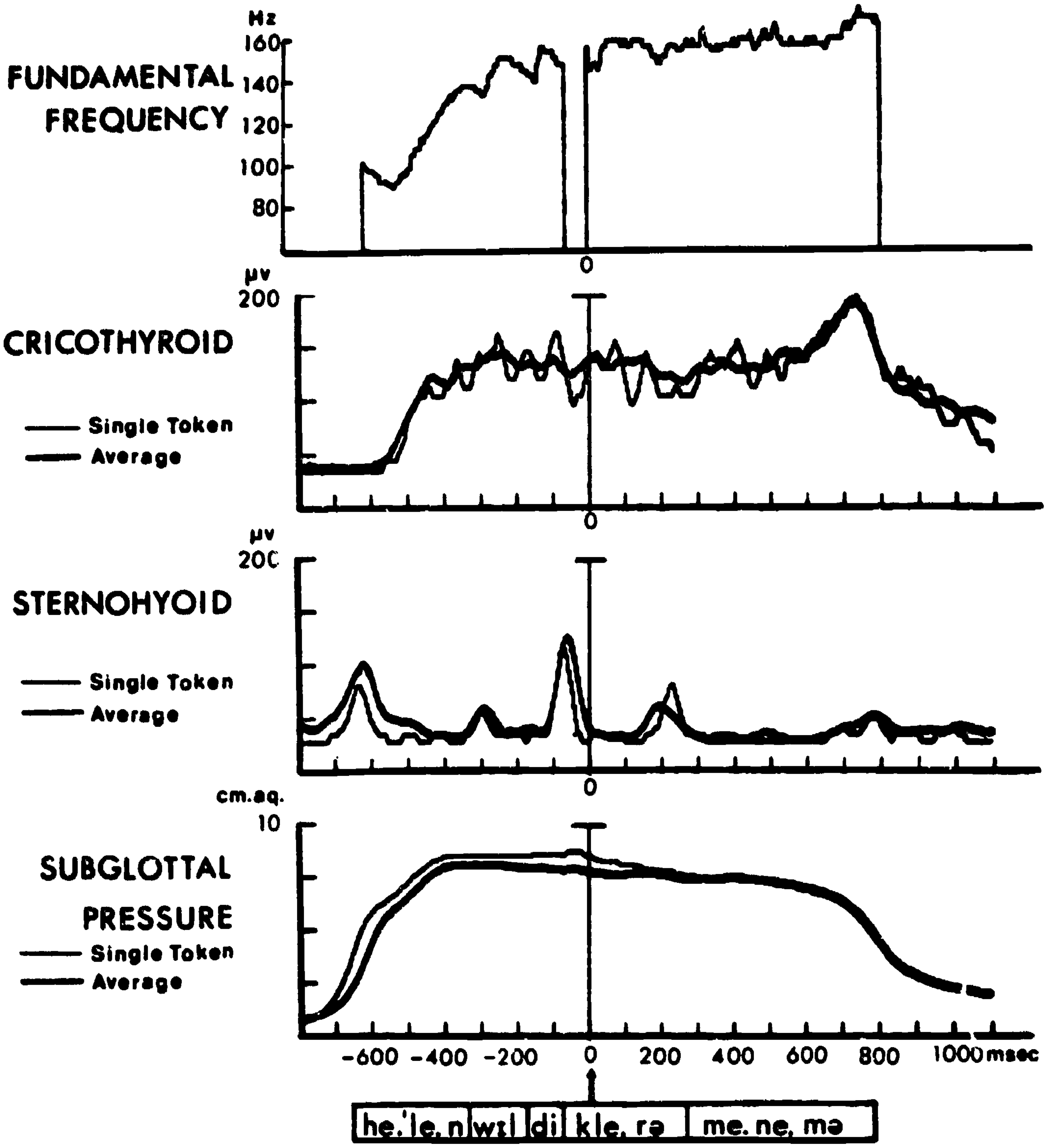
CONTOUR 10


Hz
160
140
120
100
80
FUNDAMENTAL FREQUENCY
0
μv
300
CRICOTHYROID
Single Token
Average
0
μv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL PRESSURE
Single Token
Average
-600 -400 -200 0 200 400 600 800 1000msec
he.'le.n
wɪ
di
'kle.rə
me.ne.mə

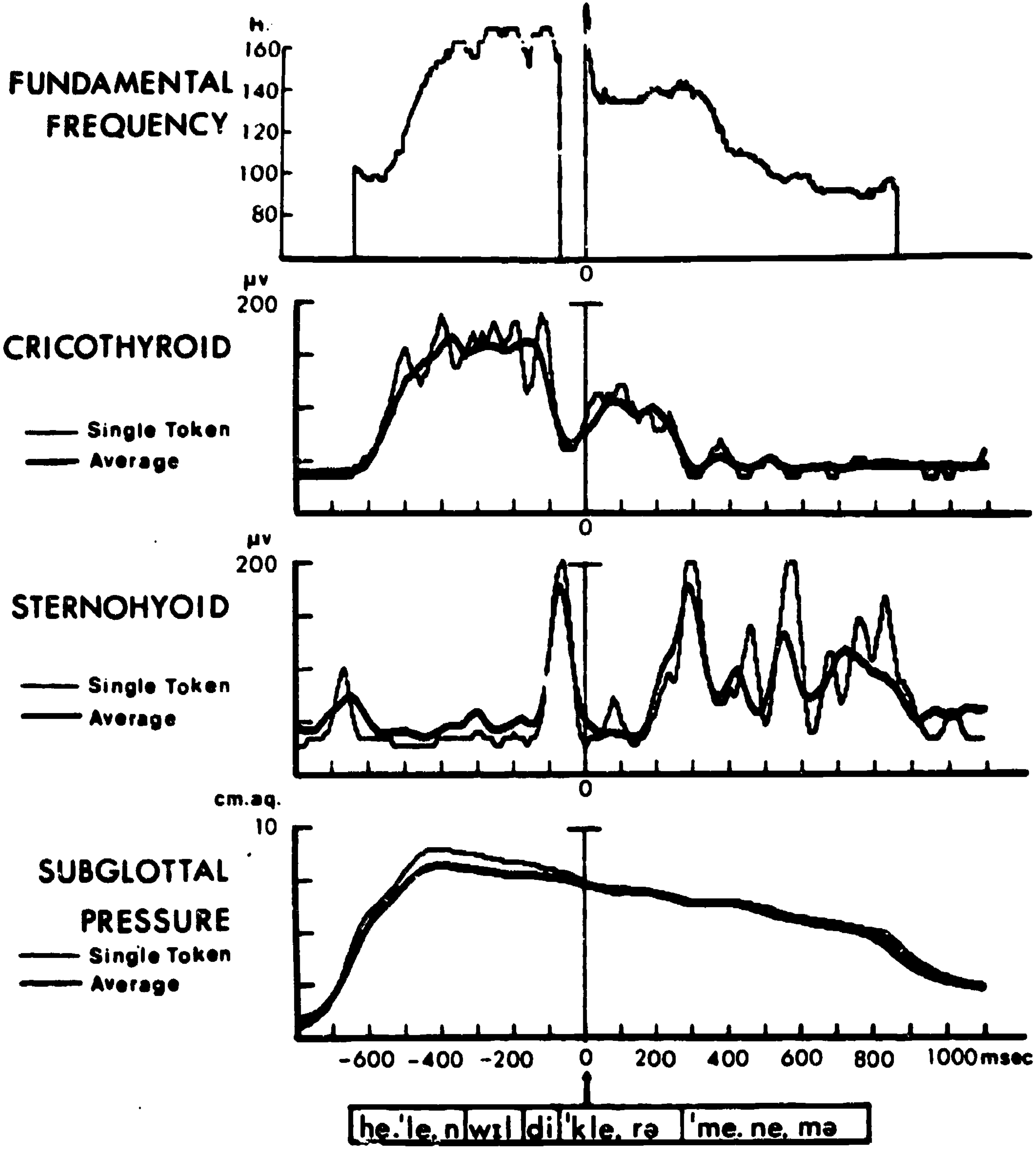
CONTOUR 11


FUNDAMENTAL FREQUENCY
Hz
160
140
120
100
80
0
CRICOTHYROID
μv
200
Single Token
Average
0
STERNOHYOID
μv
200
Single Token
Average
0
SUBGLOTTAL PRESSURE
cm.aq.
10
Single Token
Average
-600 -400 -200 0 200 400 600 800 1000msec
he.ˈle.n wɪl di kle.rə ˈme.ne.me

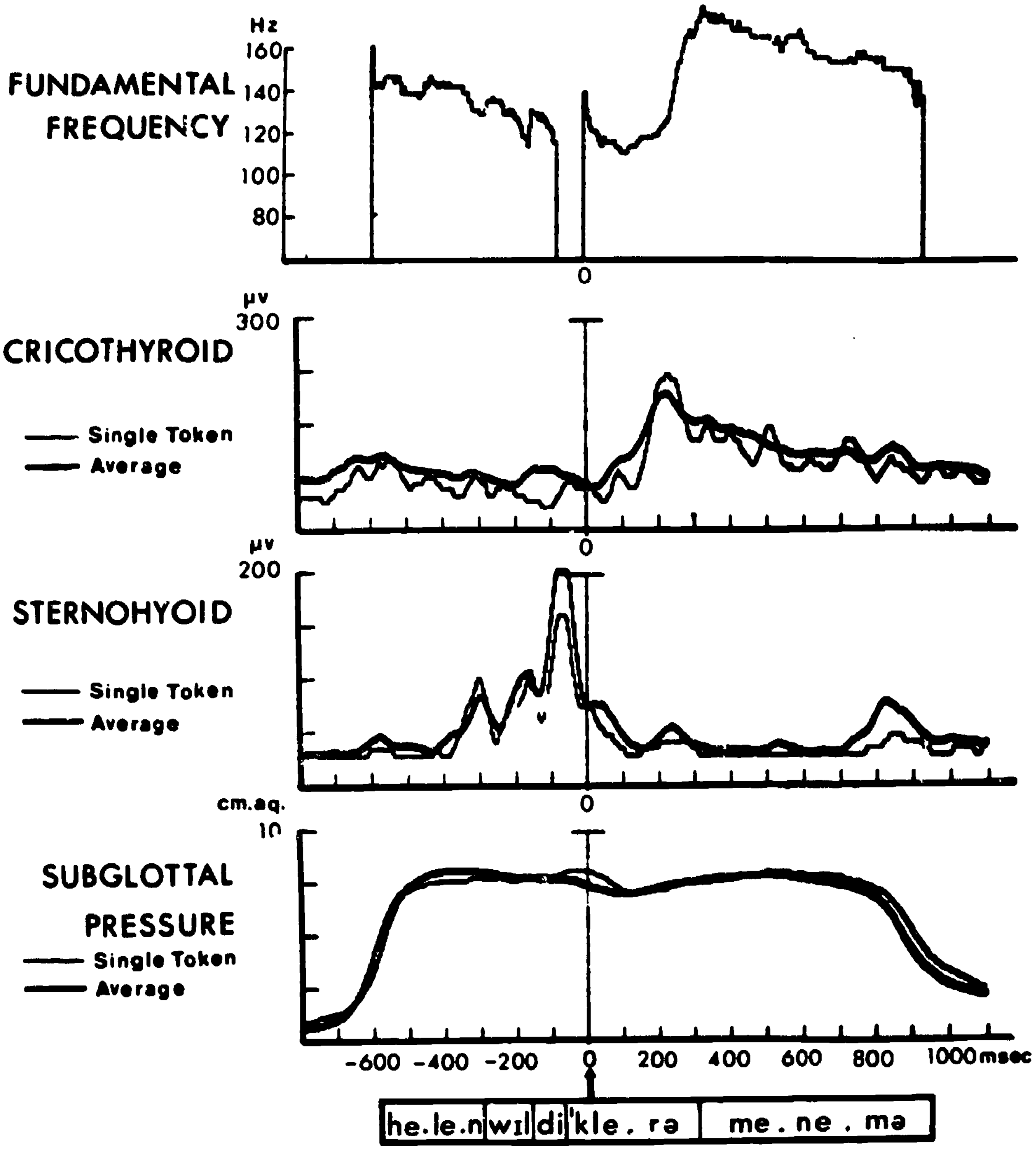
CONTOUR 12


Hz
160
140
120
100
80
FUNDAMENTAL
FREQUENCY
0
µv
200
CRICOTHYROID
Single Token
Average
0
µv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL
PRESSURE
Single Token
Average
-600
-400
-200
0
200
400
600
800
1000 msec
he.ˈle.n
wɪl
di
kle.rə
me.ne.mə

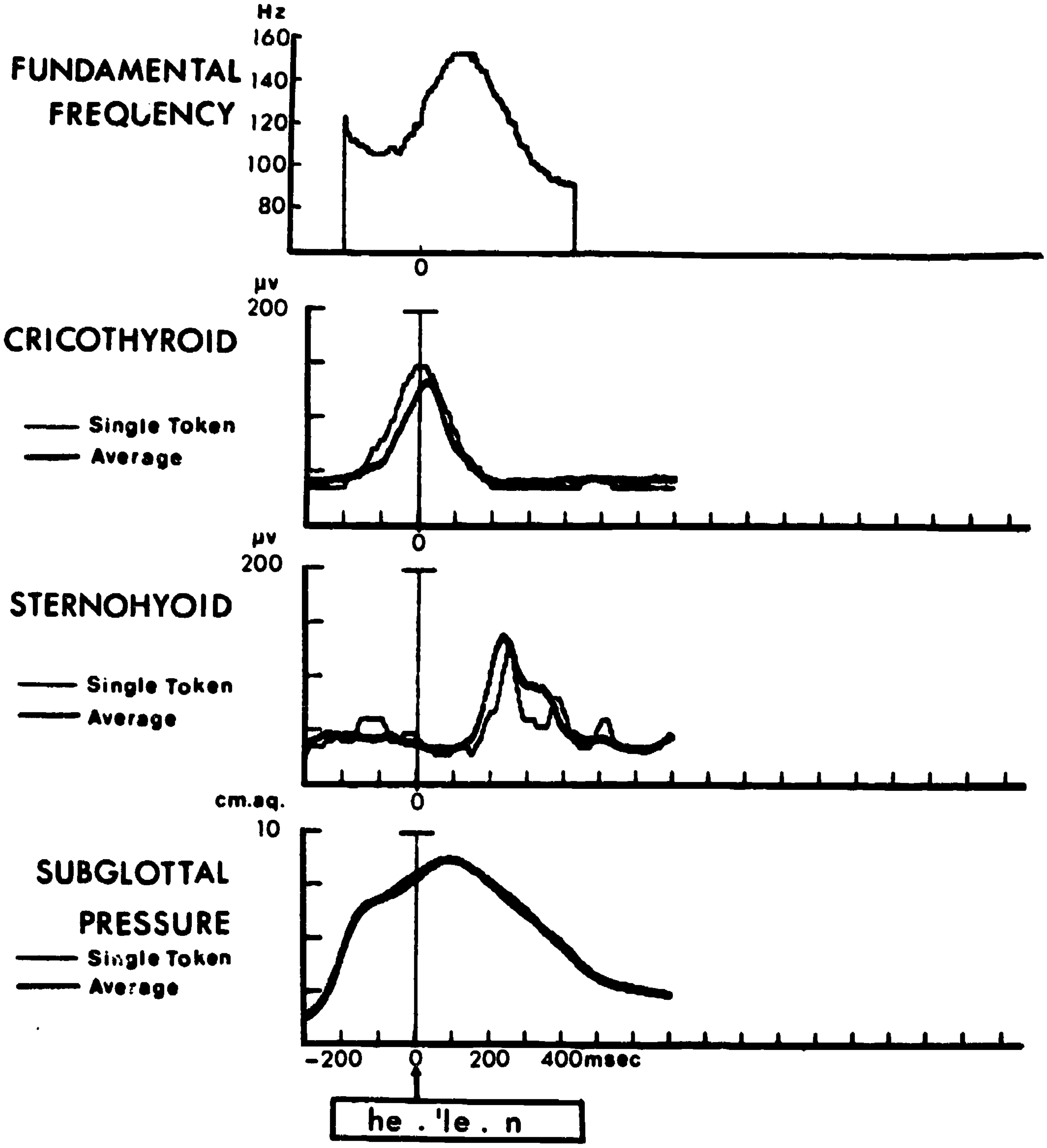
CONTOUR 13


Hz
160
140
120
100
80
FUNDAMENTAL
FREQUENCY
0
μv
200
CRICOTHYROID
Single Token
Average
0
μv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL
PRESSURE
Single Token
Average
-600
-400
-200
0
200
400
600
800
1000msec
he.'le.n wɪl di 'kle. rə 'me. ne. mə

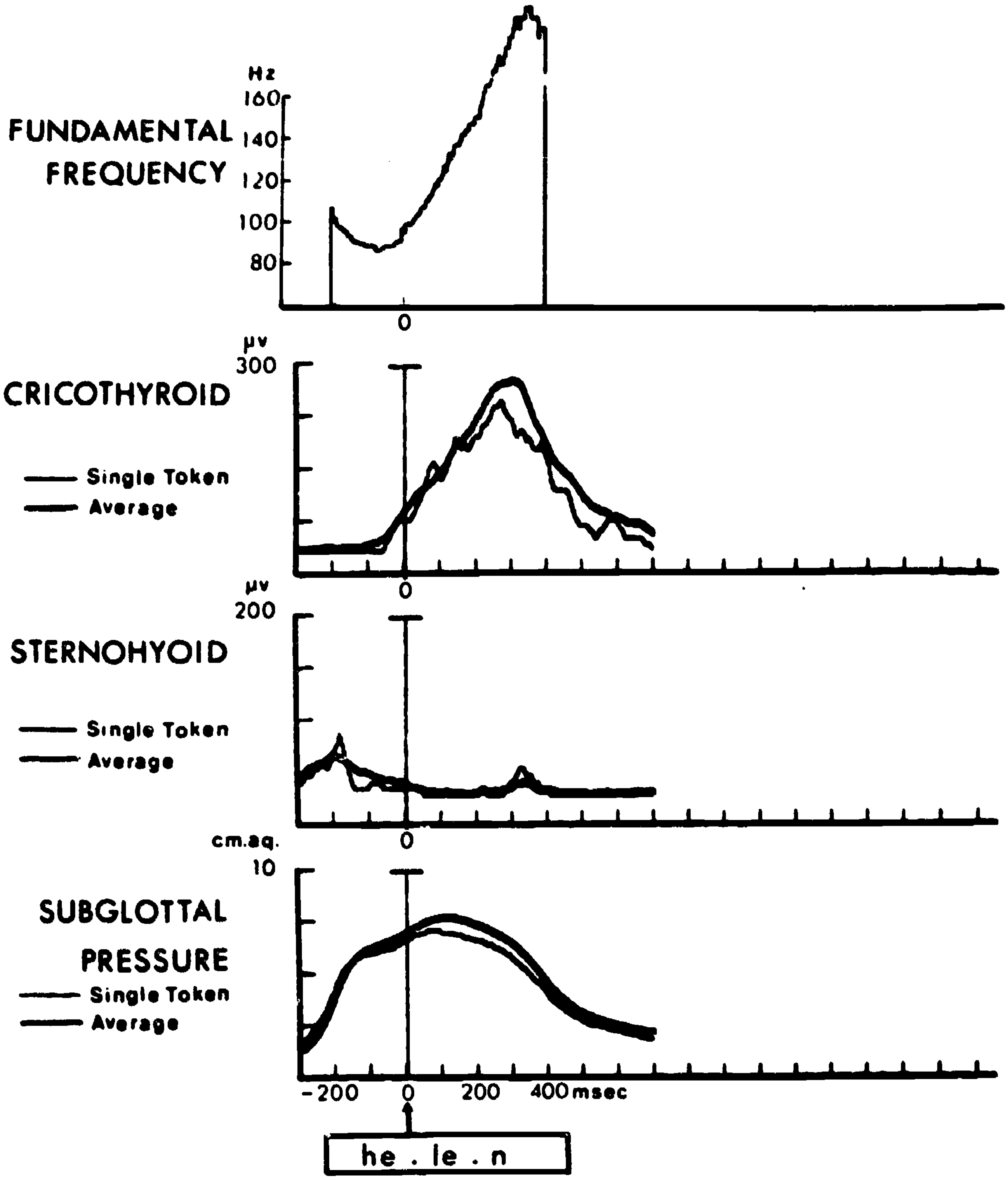
CONTOUR 14


FUNDAMENTAL FREQUENCY
Hz
160
140
120
100
80
0
CRICOTHYROID
µv
300
— Single Token
— Average
0
STERNOHYOID
µv
200
— Single Token
— Average
0
SUBGLOTTAL PRESSURE
cm.aq.
10
— Single Token
— Average
-600 -400 -200 0 200 400 600 800 1000msec
he.le.n wɪl di ˈkle . rə me . ne . mə

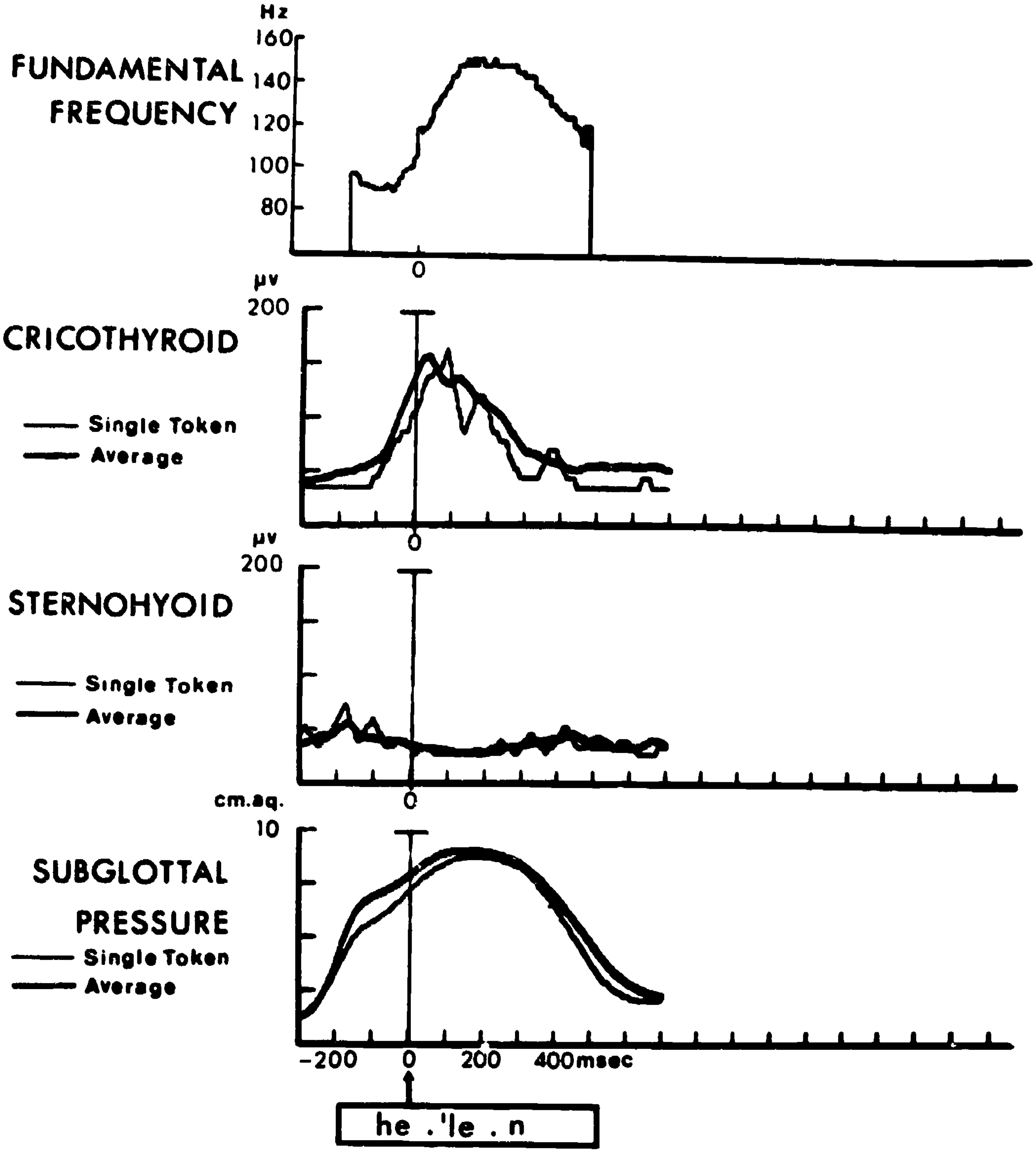
CONTOUR 15


Hz
160
140
120
100
80
FUNDAMENTAL
FREQUENCY
0
µv
200
CRICOTHYROID
Single Token
Average
0
µv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL
PRESSURE
Single Token
Average
-200
0
200
400msec
he . 'le . n

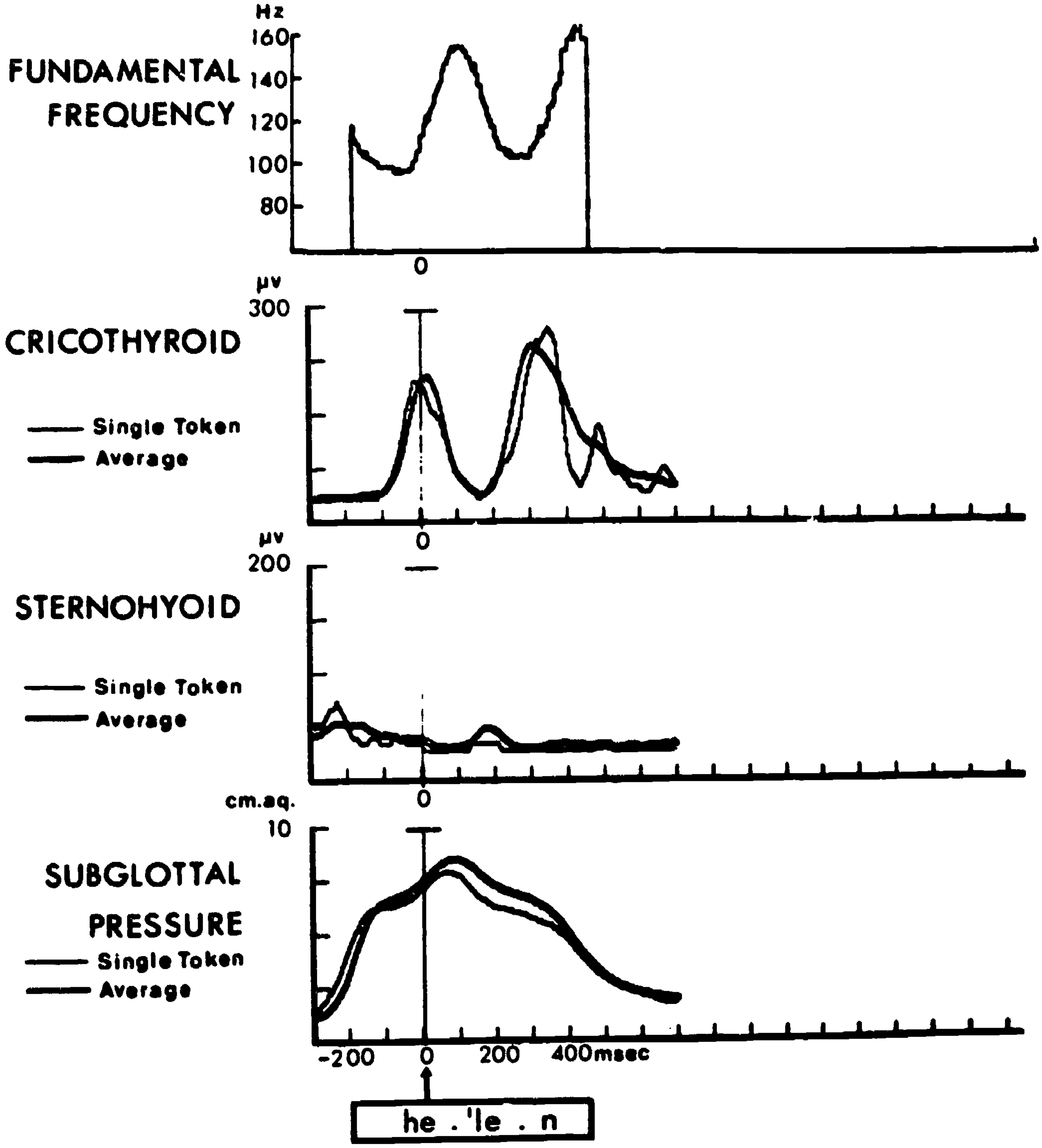
CONTOUR 16


Hz
160
140
120
100
80
FUNDAMENTAL
FREQUENCY
0
µv
300
CRICOTHYROID
Single Token
Average
0
µv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL
PRESSURE
Single Token
Average
-200
0
200
400 msec
he . le . n

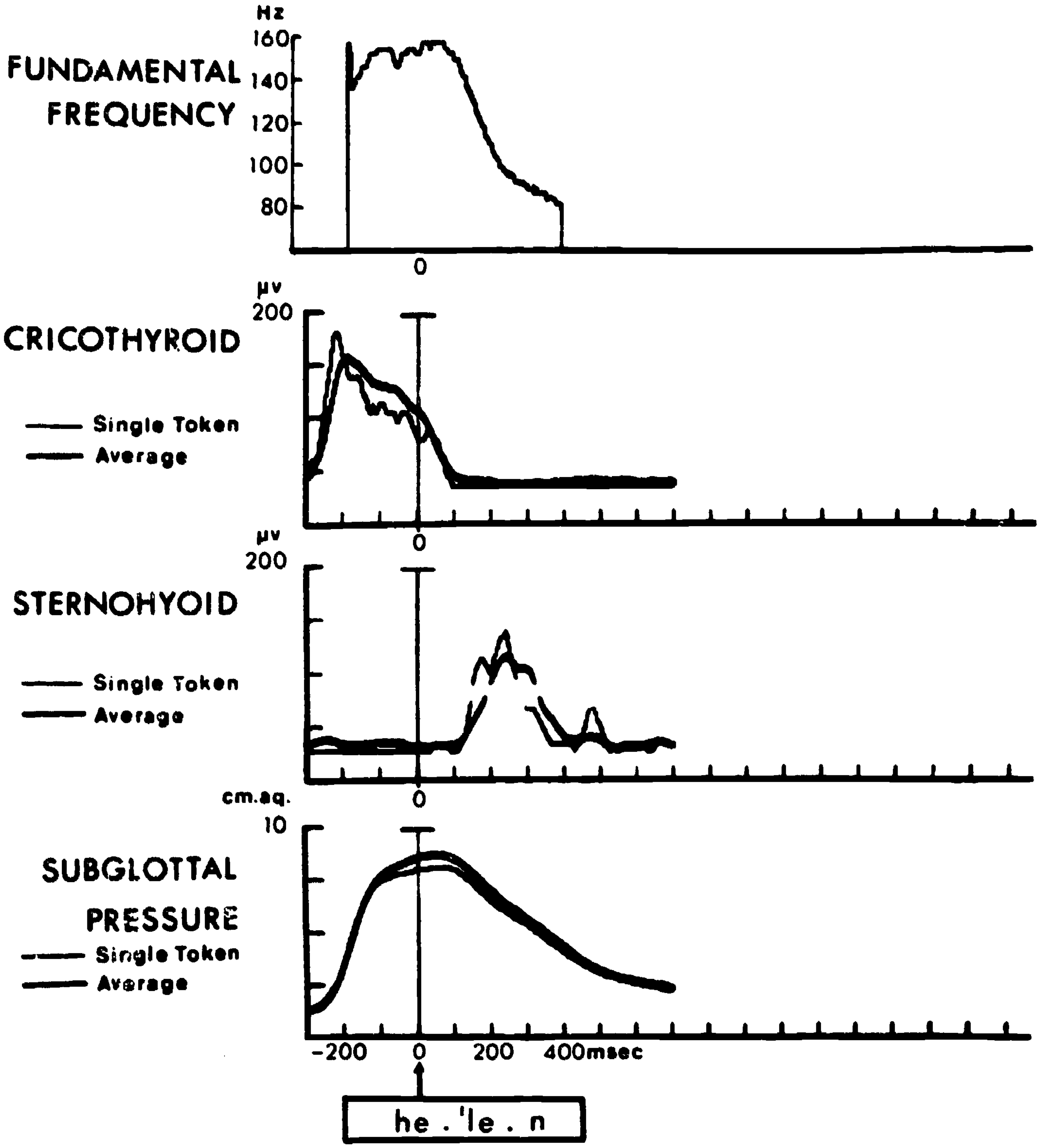
CONTOUR 17


Hz
160
140
120
100
80
FUNDAMENTAL FREQUENCY
0
µv
200
CRICOTHYROID
Single Token
Average
0
µv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL PRESSURE
Single Token
Average
-200
0
200
400msec
he .'le . n

CONTOUR 18


Hz
160
140
120
100
80
FUNDAMENTAL
FREQUENCY
0
µv
300
CRICOTHYROID
Single Token
Average
µv
0
200
STERNOHYOID
Single Token
Average
cm.aq.
0
10
SUBGLOTTAL
PRESSURE
Single Token
Average
-200
0
200
400msec
he . 'le . n

CONTOUR 19


Hz
160
140
120
100
80
FUNDAMENTAL
FREQUENCY
0
μv
200
CRICOTHYROID
Single Token
Average
0
μv
200
STERNOHYOID
Single Token
Average
0
cm.aq.
10
SUBGLOTTAL
PRESSURE
Single Token
Average
-200
0
200
400msec
he . 'le . n

CONTOUR 20

A Cinefluc:ographic Study of Vowel Production*

Thomas Gay[+]
Haskins Laboratories, New Haven, Conn.

The purpose of this experiment was to study the effects of changes in both phonetic context and speaking rate on the movements toward and attainment of target positions for the vowels /i/, /a/, and /u/. Two subjects read lists of nonsense words containing these vowels in vowel-consonant-vowel (VCV) combination with the consonants /p/, /t/, and /k/, at both slow and fast speaking rates. Lateral-view X-ray films were recorded along with the acoustic signal. Results showed that during slow speech the target positions of both /i/ and /u/ remain highly stable across changes in both the preceding and following consonant and vowel. The production of /a/, although not subject to right-to-left effects beyond the following consonant, is sensitive to changes in the consonant, as well as in the vowel preceding the consonant. These coarticulation effects, however, are not reflected as such in the acoustical measurements. The production of all three vowels during fast speech is characterized by articulatory undershoot and an upward shift in the frequencies of both the first and second formants. These results are discussed in terms of a target-based description of vowels.

Coarticulation has been the subject of considerable interest in recent physiological speech research, yet one that is still little understood. Although the variability in the production of a phone at all levels of the peripheral production process is well documented, there is little data on the exact nature and extent of most coarticulatory phenomena.

One good example is vowel production. MacNeilage and DeClerk (1969) and Harris (1971) have shown that different motor command strategies can be used for

*Also in Journal of Phonetics (1974) 2, 255-266.

[+]Also University of Connecticut Health Center, Farmington.

Acknowledgment: I wish to thank Dr. J. Daniel Subtelny, head of the Department of Orthodontics, Eastman Dental Center, Rochester, N. Y., for the use of his cinefluorographic facilities. This research was supported, in part, by a grant from the National Institute of Neurological Diseases and Stroke (NS-10424), and by a National Institute of Dental Research grant to Haskins Laboratories.

a vowel depending upon the phonetic context in which it is placed. However, it is not clear whether these different strategies are reflected in differences in the target positions of the vowel. Indeed, it might be argued that coarticulation at the motor command level simply reflects a strategy to attain a quasi-invariant articulatory target position (MacNeilage, 1970). Unfortunately, however, the available data that bear on this point are somewhat contradictory. For example, the physiological data of Houde (1967), MacNeilage and DeClerk (1969), and Gay, Ushijima, Hirose, and Cooper (1974) suggest that vowel stability is more the rule than the exception, while the X-ray data of Kuehn (1973) suggest a good deal of positional variability for the vowel target (specifically /a/). Target variability is also evident for faster speech (Gay et al., 1974; Kuehn, 1973) and destressed speech (Lindblom, 1963; Kent and Netsell, 1971).

The purpose of the experiment reported here was to examine more closely a number of aspects of vowel production. The most important of these concerns the nature of a vowel target and whether it can be defined in terms of a three-dimensional articulatory coordinate system (MacNeilage, 1970). We used cinefluorgraphy to study the effects of changes in both phonetic context and rate of speech on the movements of the tongue and jaw during the production of selected vowels. This was designed to provide a descriptive account of the movements toward and attainment of vowel target positions under the constraints of a variety of linguistic demands known to be sources of articulatory variability. In addition, we used acoustical analysis to determine whether any variability evident at the articulatory level is reflected in the formant structure of the vowel.

METHOD

Subjects and Speech Material

Subjects were two adult males, FSC and TG, both native speakers of American English. The speech material consisted of the consonants /p,t,k/ and the vowels /i,a,u/ in a trisyllable nonsense word of the form, /pV_1CV_2pə/, where V_1 and V_2 were all possible combinations of /i,a,u/ and C was either /p/, /t/, or /k/. The 27 utterance types were randomly ordered into a master list. Each utterance, preceded by the carrier phrase, "It's a....," was produced at two speaking rates: slow (or normal) and fast. Each rate was based on the subject's own appraisal of comfortable slow and fast rates. A brief practice session preceded the filming session. The subjects were also instructed to say the first two syllables of the utterance with equal stress, and the final syllable unstressed.

Data Recording

Lateral-view X-ray films were recorded with a 16 mm cine camera at a speed of 64 fps. The X-ray generator delivered 1 msec pulses to a 9 in image intensifier tube. Two lead pellets ([illegible].5 mm diameter) were attached to the surface of the tongue along the midline. The pellets were located on the dorsum at points approximately 2 and 3 in from the tip. Cyanoacrylate was used as the adhesive. A barium sulfate paste was also used as a contrast medium on the tongue, and tantalum was applied along the midline of the nose and lips to outline those structures. The acoustic signal was recorded on magnetic tape and synchronized with the film record by means of camera-generated synchronization pulses.

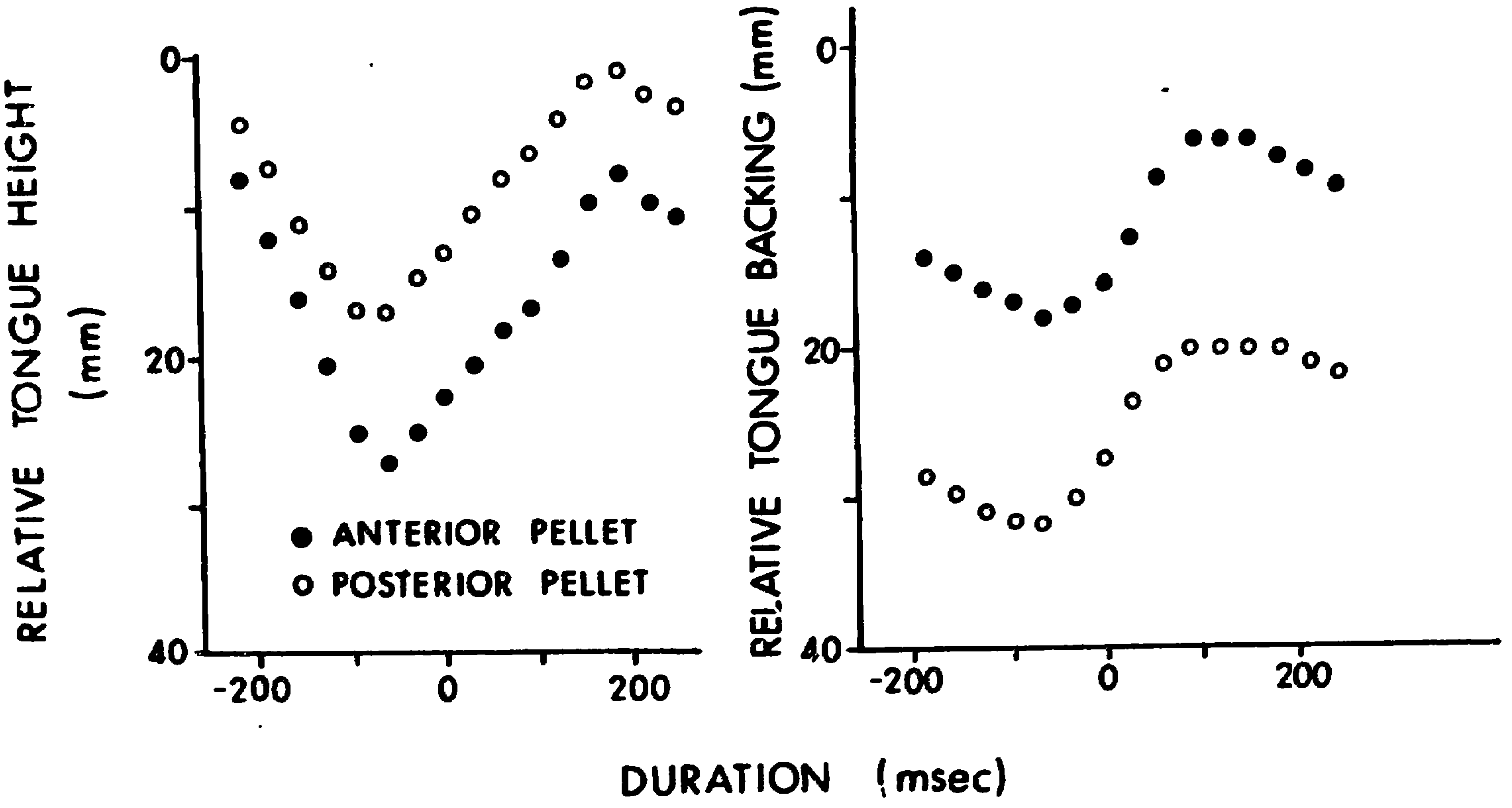


FIGURE 1

Figure 1: Typical pellet measurements of relative tongue height and relative tongue backing. The utterance is /api/, Subject FSC. O on the abcissa equals time of closure for /p/.

BEST COPY AVAILABLE

Data Analysis

The X-ray films were analyzed frame-by-frame, using a Perceptoscope f.lm analyzer. The film was projected life size to a writing surface via an overhead mirror system. Each of the two pellets was tracked in a coordinate system that used fixed landmarks as reference points. These points, along with an outline of the hard palate and upper central incisors, were drawn on a master template. Photocopies of the master were then used as templates for the measurements of each film frame. Measurements were made from the time of /k/ release to the time of closure for the final /p/. Rechecks of a number of measurements revealed a pellet measurement error of no more than 1 mm.

Each of the pellets was tracked in two dimensions: tongue height versus time and tongue backing versus time. Examples of these graphs for FSC are shown in Figure 1. The relative tongue backing measurements provided little in the way of useful data and will not be presented in this form. As can be seen in Figure 1, the ballistic patterns for both pellets are essentially the same; the only real di'ference is a greater amplitude of movement for the anterior pellet during the production of the open vowel /a/. For no other reason, this pellet will be used to illustrate the data in the Results section.

Jaw movement in the vertical plane was also tracked frame-by-frame. This was done by measuring the vertical distance between the upper and lower central incisors.

Besides tracking the dynamics of tongue movement, pellet positions can also be used to construct vowel target positions in the traditional articulatory sense. Figure 2 shows a typical configuration for the vowels /i,a,u/; the pellet positions appear as the three points in the two-dimensional articulatory triangle.

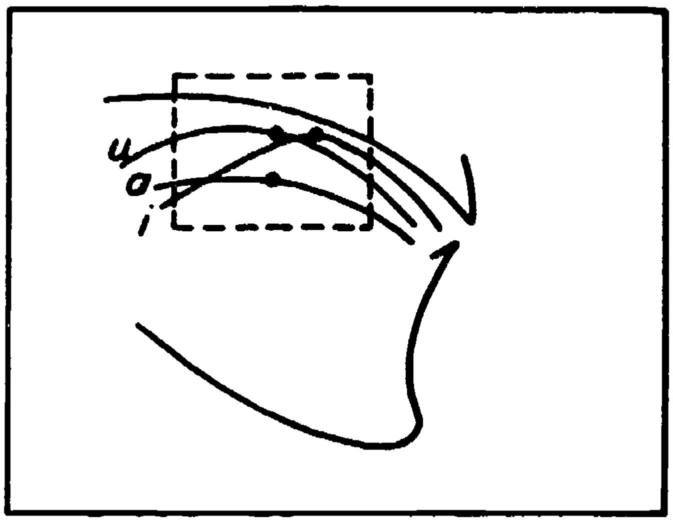


Figure 2: Typical pellet positions at the target positions for /i/, /a/, and /u/.

Wide-band spectrograms of all the utterances produced during the X-ray run were made from the accompanying magnetic tape recording. Duration measurements were also made from the spectrograms. Durations from the time of /k/ release to the time of /p/ closure averaged 560 msec and 390 msec for FSC and 510 msec and 370 msec for TG, for the normal and fast speaking rates, respectively.

RESULTS

This section is divided into three parts: the effect of phonetic context on vowel target position, the effect of speaking rate on vowel target position, and the acoustical consequences of these articulatory effects.

Phonetic Context Effects

In this section the effect of the intervocalic consonant on the target positions of the first and second vowels, the effect of the second vowel on the target position of the first vowel, and the effect of the first vowel on the target position of the second vowel will be described.

Figures 3 and 4 summarize the effect of the intervocalic consonant on the target positions of the first and second vowels. These figures show the relative positions in two dimensions of the anterior pellet at the vowel target (point of farthest articulator displacement). For both subjects, the target positions for /i/ and /u/ in both pre- and postconsonantal positions are quite stable across changes in the consonant. Generally speaking, target variability for /i/ and /u/ rarely exceeded 2 mm, and never exceeded 3 mm. For /a/, however, individual differences appear. While the positions for TG remain stable, the targets for FSC show a rather strong consonant effect, primarily in the height dimension. These differences occur for both the first and second vowels and span a distance of almost 8 mm. Displacement for both the first and second vowels is least when the consonant is /t/ and greatest when the consonant is /p/.

The reason for these differences becomes apparent in the movement tracking measurements. Figure 5 shows the measurements for tongue height (anterior pellet) and jaw opening for the entire VCV utterance, for both subjects. While the data for TG show essentially identical movement patterns and target positions throughout the utterance, the data for FSC show variability for all three phonetic segments. This variability appears in both the displacement and velocity components of the curves. The tongue not only extends farther for the vowel, but also moves more quickly (steeper slope) toward its target when the consonant is /p/.

Apparently, displacement for the vowel is greatest when the consonant is /p/ because the tongue and jaw are least involved in the production of this consonant. Both /t/ and /k/, on the other hand, are characterized by greater degrees of jaw closure; this probably acts to constrain the degree of opening for the adjacent vowels. Although the displacement differences for the tongue cannot be accounted for entirely by differences in jaw opening, the correlation between the two measures is obviously quite high. This is shown by the fact that the curves for the tongue body closely shadow those for the jaw.

In addition to affecting the displacement of its neighboring vowels, the consonant also conditions the timing of the movement toward the vowel. Figure 6 shows the tongue height measurements for the utterances /ipa/, /ita/, and /ika/.

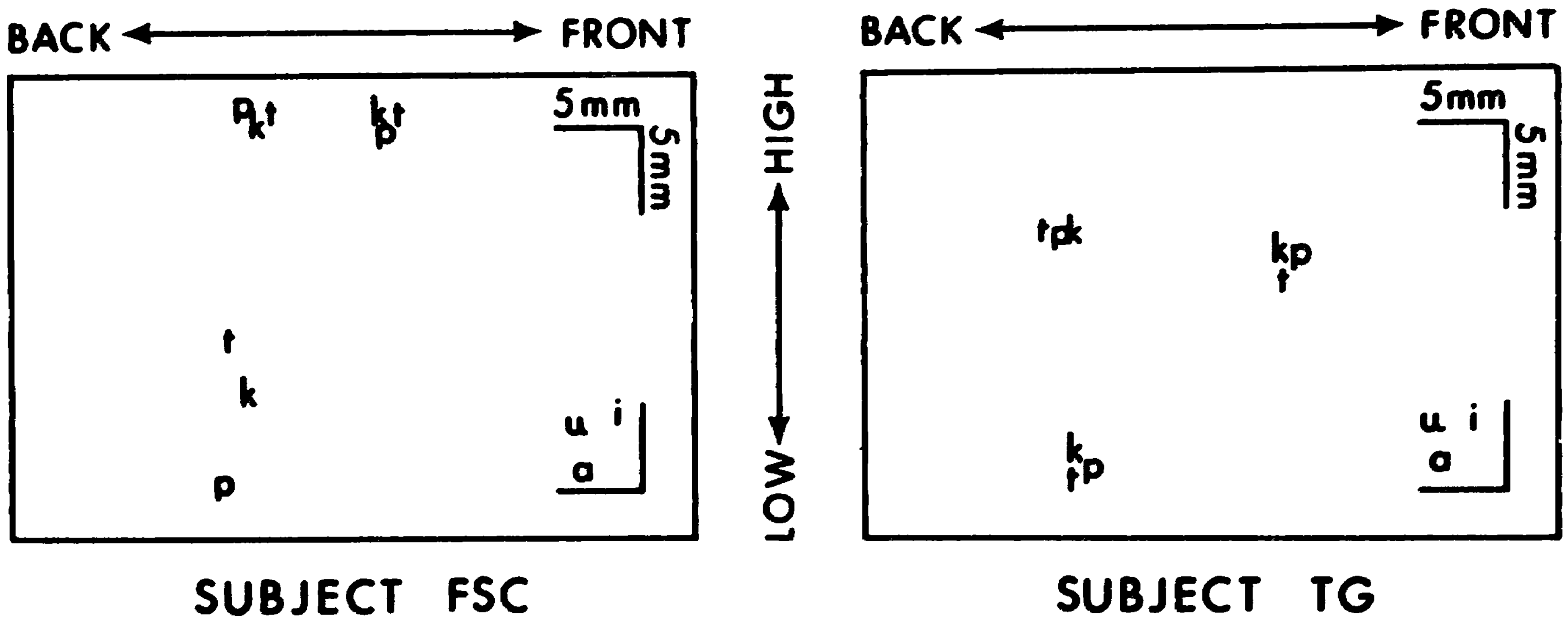


Figure 3: Effect of the consonant on the target positions of the first vowel. The second vowel is /a/ (/ipa, ita, ika, apa, etc./).

BEST COPY AVAILABLE

FIGURE 3

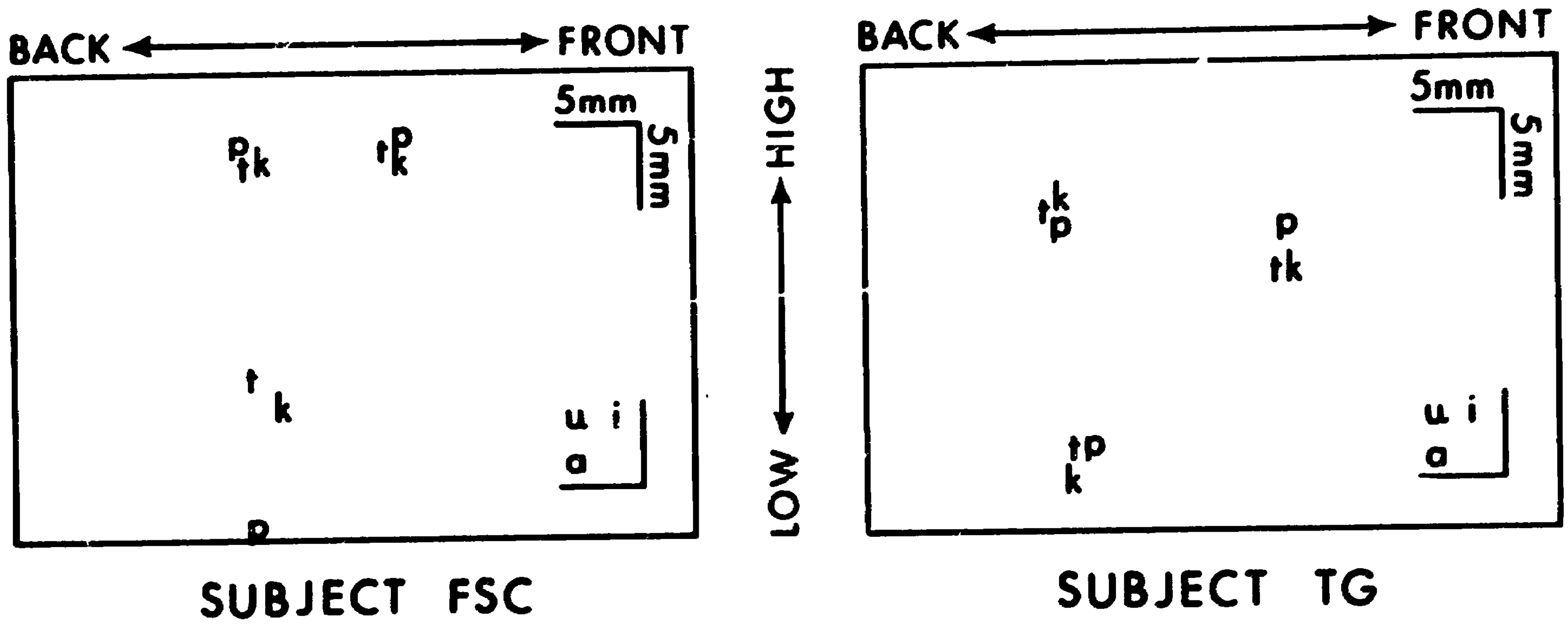


FIGURE 4

Figure 4: Effect of the consonant on the target positions of the second vowel. The first vowel is /a/ (/api, ati, aki, apa, etc./).

FIGURE 5

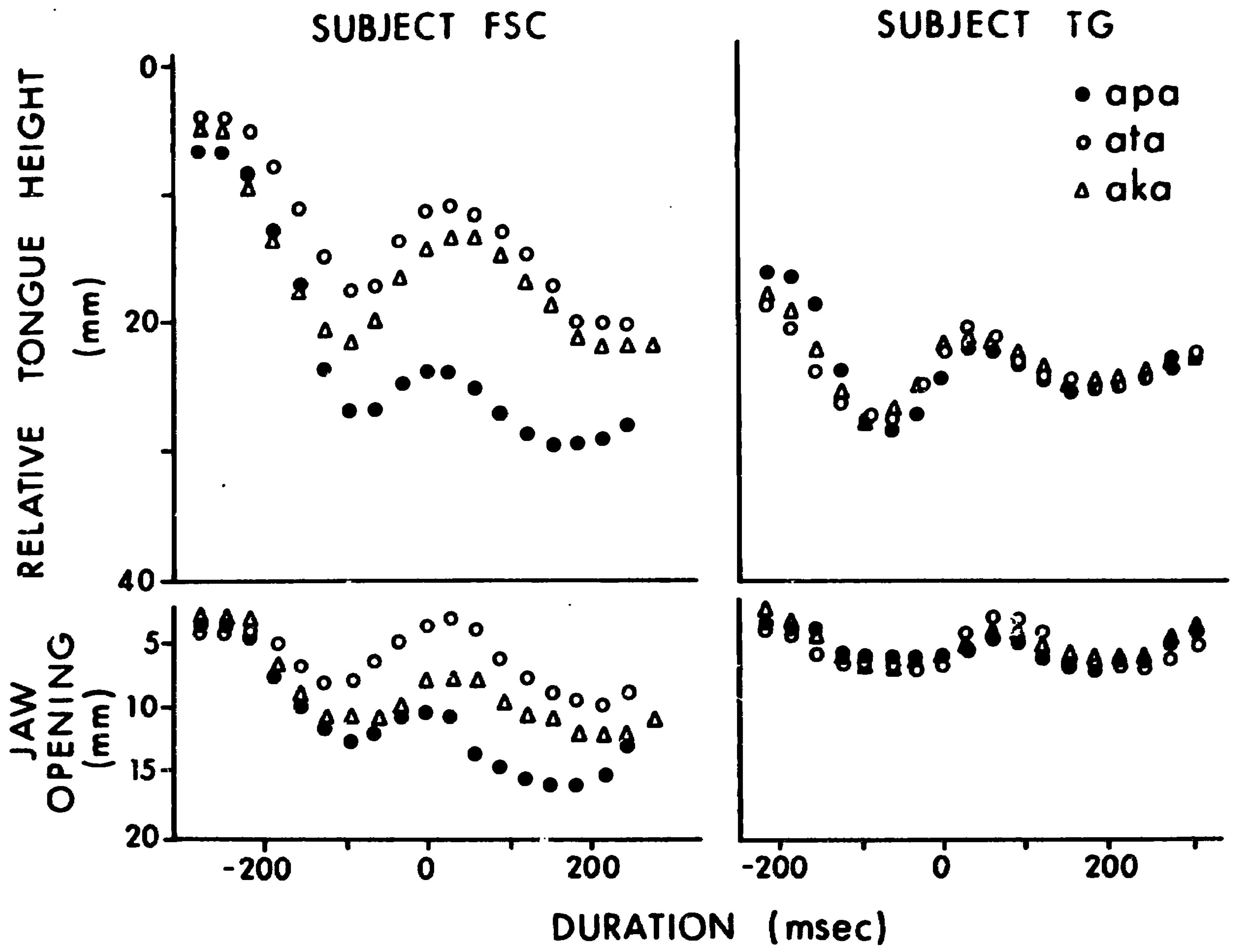


Figure 5: Effect of the consonant on tongue height and jaw opening for both the first and second vcwels. 0 on the abcissa equals time of closure for the consonant.

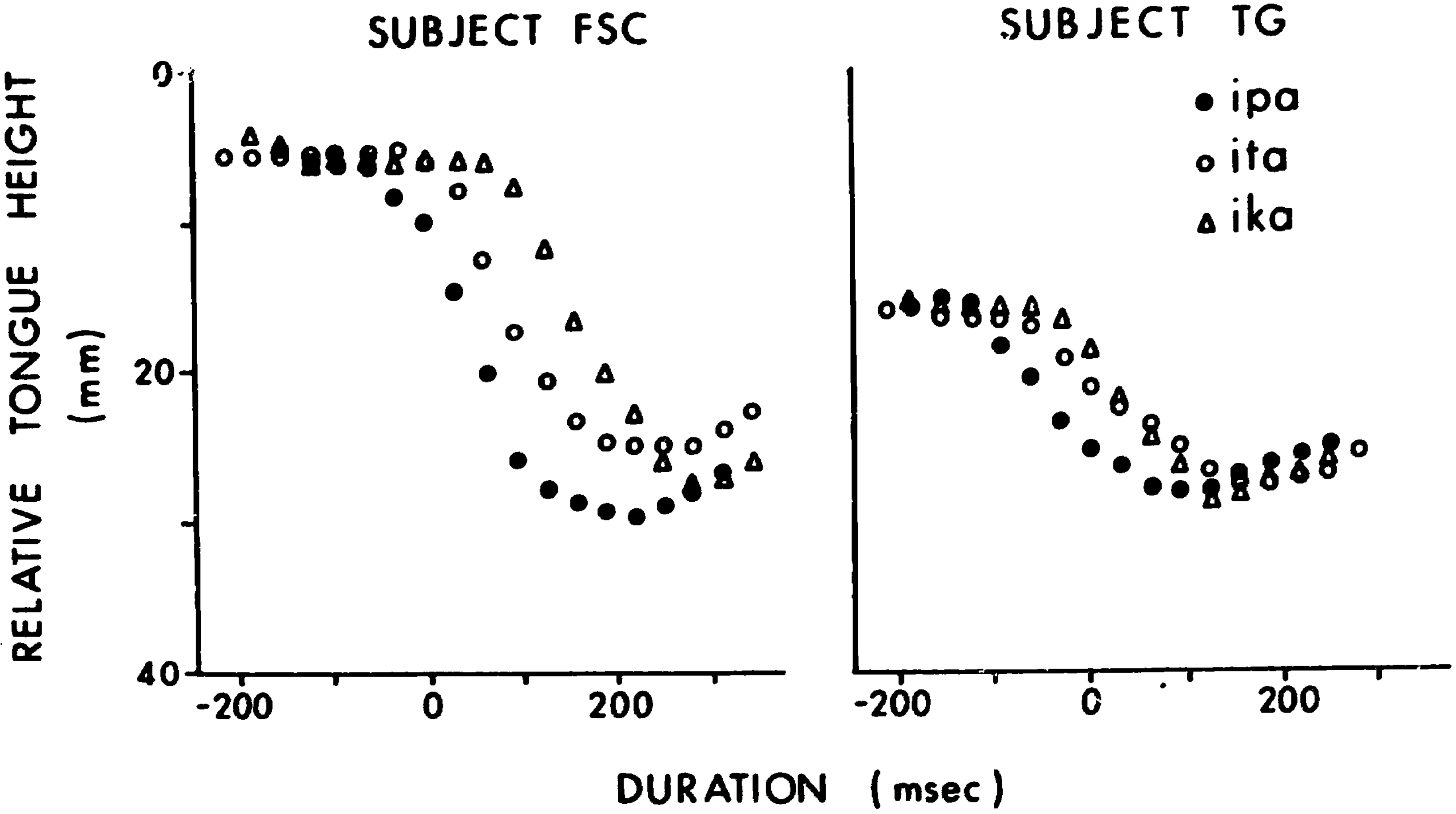


Figure 6: Effect of the consonant on the timing of tongue movements toward the second vowel. 0 on the abcissa equals time of closure for the consonant.

FIGURE 6

For both subjects, movement toward the second vowel occurs earliest when the consonant is /p/. This effect occurs whenever the tongue moves from /i/ or /u/ to /a/.[1] Again, these differences are apparently related to the independence of the tongue during the articulation of /p/.

Individual differences in displacement also appear for /a/. While tongue displacement for the vowel remains stable in all consonant contexts for TG, consonant effects are evident for FSC (greater displacement for the second vowel when preceded by /p/, /k/, /t/, in that order). Interestingly, these differences appear to be due solely to differences in the timing of the movement from the consonant; in contrast, the effects described earlier were apparently caused by differences in the degree of displacement for the consonant.

Figure 7 summarizes the effect of the second vowel on the target positions of the first vowel. This figure shows the target positions of the first vowel as a function of different second vowels in the /p/ consonant context. For both subjects, the target positions of all three first vowels are stable across changes in the second vowel (again, generally within a range of 2 mm). This stability is also evident when the consonant is /t/ and /k/. Apparently, right-to-left effects do not extend across the consonant to the preceding vowel.

Although the first vowel in the VCV utterance is not sensitive to any right-to-left effects beyond the consonant, the second vowel is subject to some left-to-right, or carryover, vowel effects; these effects, however, are fairly complicated and linked to the consonant.

When the consonant is /p/ the first vowel has no real effect on the target position of the second vowel (Figure 8). All three vowels maintain positional stability. However, when the intervocalic consonant is either /t/ or /k/, left-to-right effects appear. Although the targets for both /i/ and /u/ (in the second vowel position) remain stable, the first vowel exerts a strong effect on the target position of /a/, this time for both subjects. These effects are illustrated in Figure 9 (two-dimension measurements) and Figure 10 (tongue height versus time measurements) for the /t/ consonant environment. These figures show less opening for /a/ when the first vowel is /a/ than when the first vowel is either /i/ or /u/.

At first glance these effects are quite surprising. It would seem intuitively more likely that greater degrees of opening for the second vowel would be caused by a more open first vowel. However, closer inspection of Figure 10 can explain these effects. At the time of closure for the consonant (0 on the abcissa), both the tongue body and jaw are in approximately the same position for each of the three first vowels. Up until this point, however, the tongue is closing toward this position from /a/, whereas it is opening toward this position from both /i/ and /u/. Thus, the tongue is moving in different directions at this point, and, in effect, has a head start towards the second vowel when the first vowel is close.

[1] This effect was also observed in the front-back dimension when the tongue moved from /i/ to /u/, and vice versa.

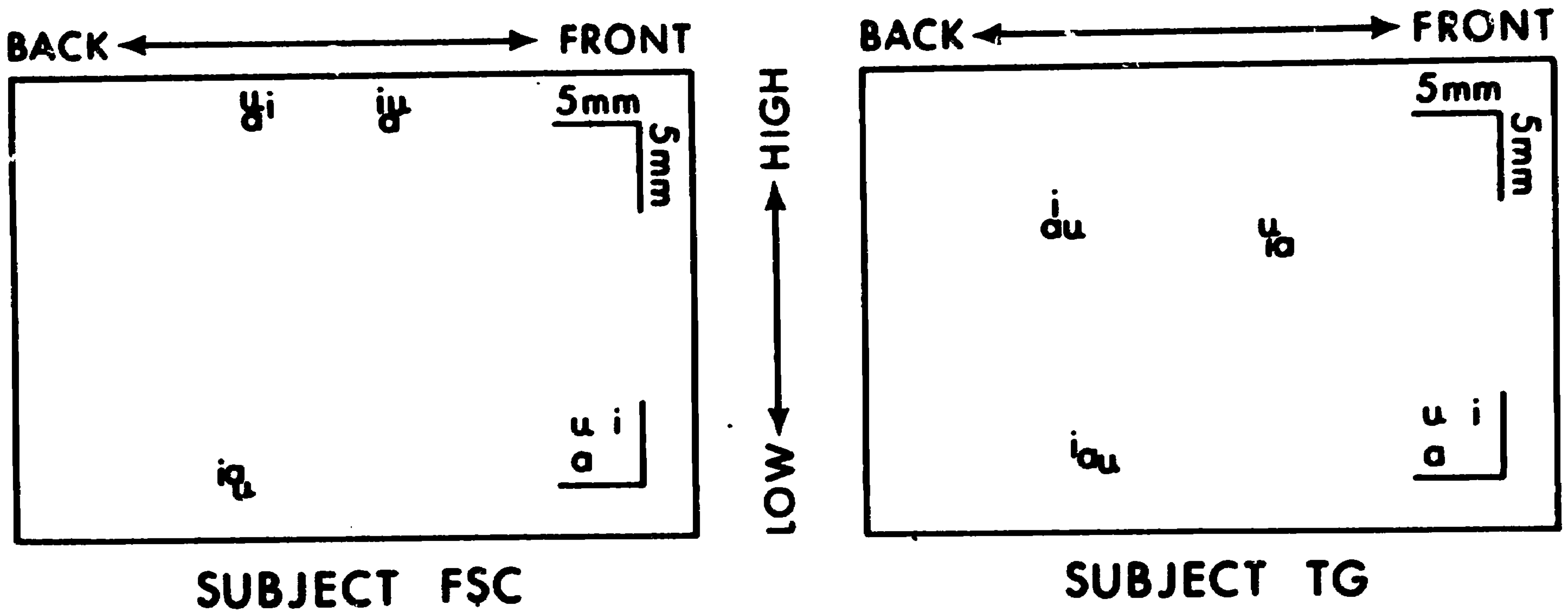


Figure 7: Effect of the second vowel on the target positions of the first vowel. The consonant is /p/ (/ipi, ipa, ipu, api, etc./).

FIGURE 7

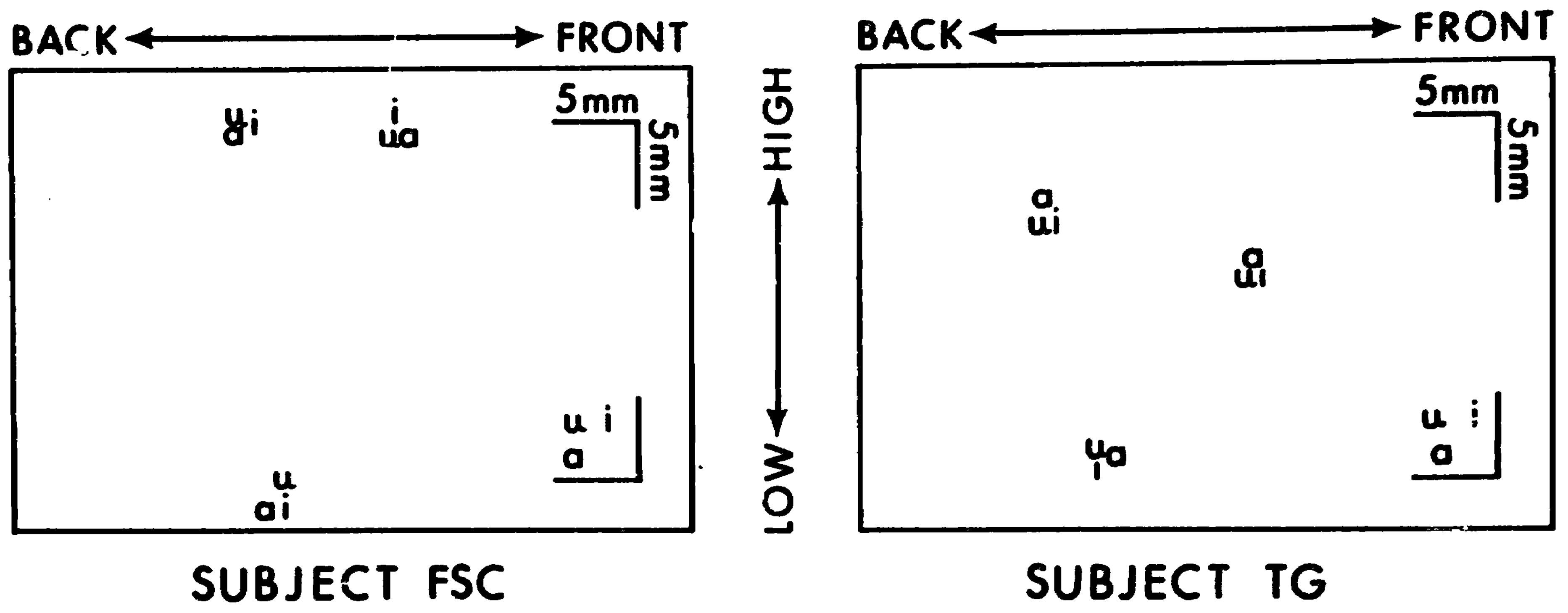


Figure 8: Effect of the first vowel on the target positions of the second vowel for the consonant /p/ (/ipi, api, upi, ipa, etc./).

FIGURE 8

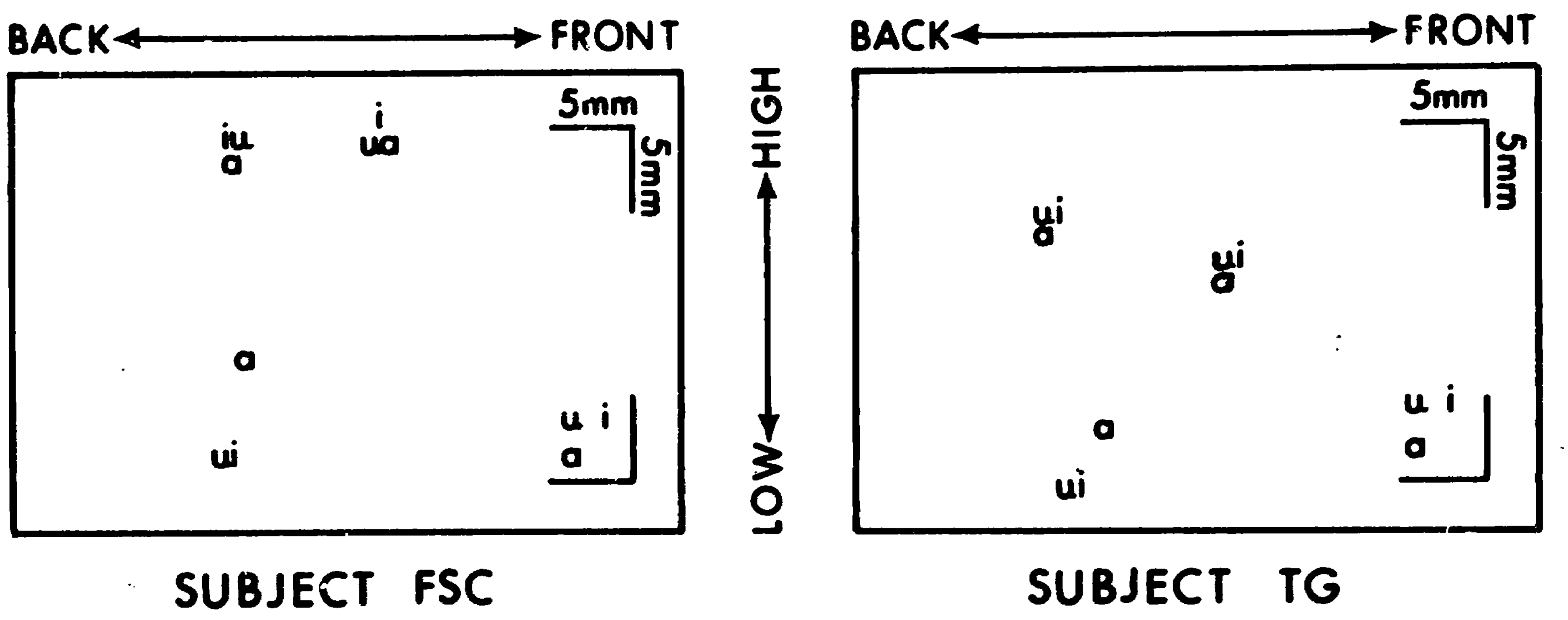


FIGURE 9

Figure 9: Effect of the first vowel on the target positions of the second vowel for the consonant /t/ (/iti, ati, uti, ita, etc./).

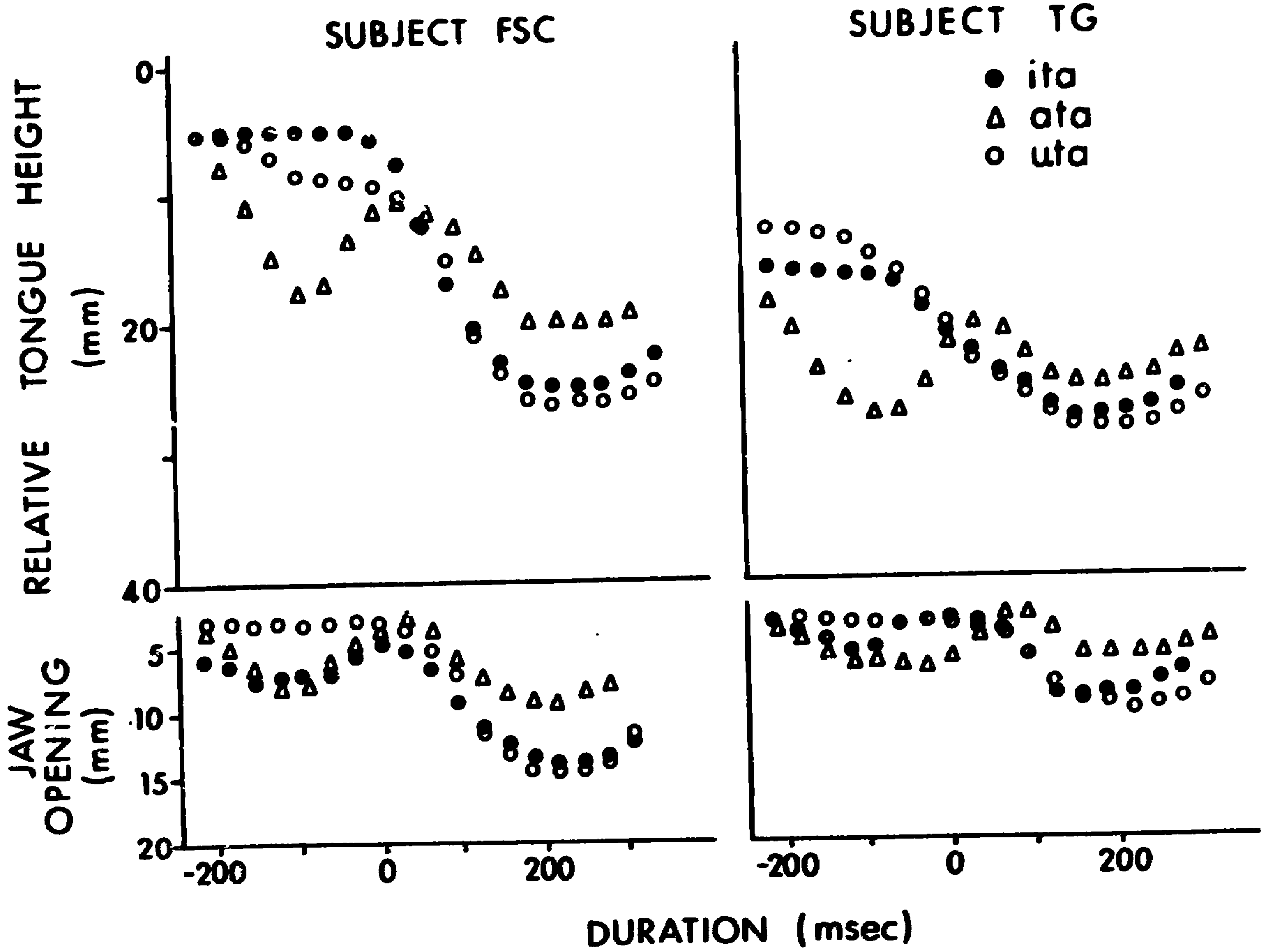


Figure 10: Effect of the first vowel on tongue height and jaw opening for the second vowel. 0 on the abcissa equals time of closure for the consonant.

FIGURE 10

To determine the extent to which jaw opening controls tongue height for an open vowel, the tongue measurements were plotted against the jaw measurements (tongue - jaw) to obtain the net movement curves for the tongue, i.e., independent from the jaw. These data are shown for both subjects in Figure 11. Since the three vowels in this figure are characterized by almost equal degrees of displacement at and near the target, it is apparent that the differences in opening for the vowel are controlled by the jaw (Lindblom and Sundberg, 1971).

The jaw opening data are interesting from another point of view: /u/ is the only vowel characterized by a closed jaw position. Both /a/ and /i/ are produced with a more open jaw--/a/ for obvious reasons, and /i/, probably to make room for the bunching of the tongue. Although the degree of jaw opening for /i/ shown in this figure approaches that for /a/, in most cases, jaw opening for /i/ is somewhat less than this, usually one-half to, at most, two-thirds that for /a/.

The results of this section can be summarized as follows. The vowel targets of both /i/ and /u/ are highly stable across changes in either the consonant or the vowel. The targets for /a/ are more variable, especially for one subject. Target position variability, when it does appear, is conditioned by both the consonant (left-to-right and right-to-left effects) and the first vowel (left-to-right effects). Right-to-left effects of the second vowel on the first vowel were virtually nonexistent.

Speaking Rate Effects

An increase in speaking rate generally resulted in a decrease in articulatory displacement for the vowel. Although target undershoot was usually present in the speech of both speakers, in a number of instances an increase in speaking rate had no appreciable effect on the displacement of the vowel. These occurrences were more frequent for TG. Undershoot occurred both more often and to a greater degree for /a/, and averaged 3-5 mm for FSC and 1-3 mm for TG.

The context effects that appeared at the slow speaking rate were generally absent at the fast speaking rate. This is probably because jaw movement was more restricted during fast speech; thus, the jaw-dependent contextual effects tended to disappear.

Perhaps the most interesting fast-speech effect occurred for the /apa/ sequence (illustrated in Figure 12). For both subjects, tongue movement from the initial /k/ to the second vowel occurs as a single articulatory gesture through both the first vowel and the consonant; there is no articulatory target evident for the first vowel! In this instance also, the tongue moves somewhat independently from the jaw, with the jaw either holding steady (FSC) or closing (TG) for the intervocalic consonant while the tongue continues moving downward for the second vowel.

Acoustical Consequences

Wide-band spectrograms were made of all the utterances spoken by both speakers. However, because of the noise produced by both the X-ray generator and cine camera, a number of tokens (/u/, in particular) could not be analyzed. The noise also obscured the first formant frequencies of both /i/ and /u/ to a point where most of these measures must be considered dubious. Nonetheless, the

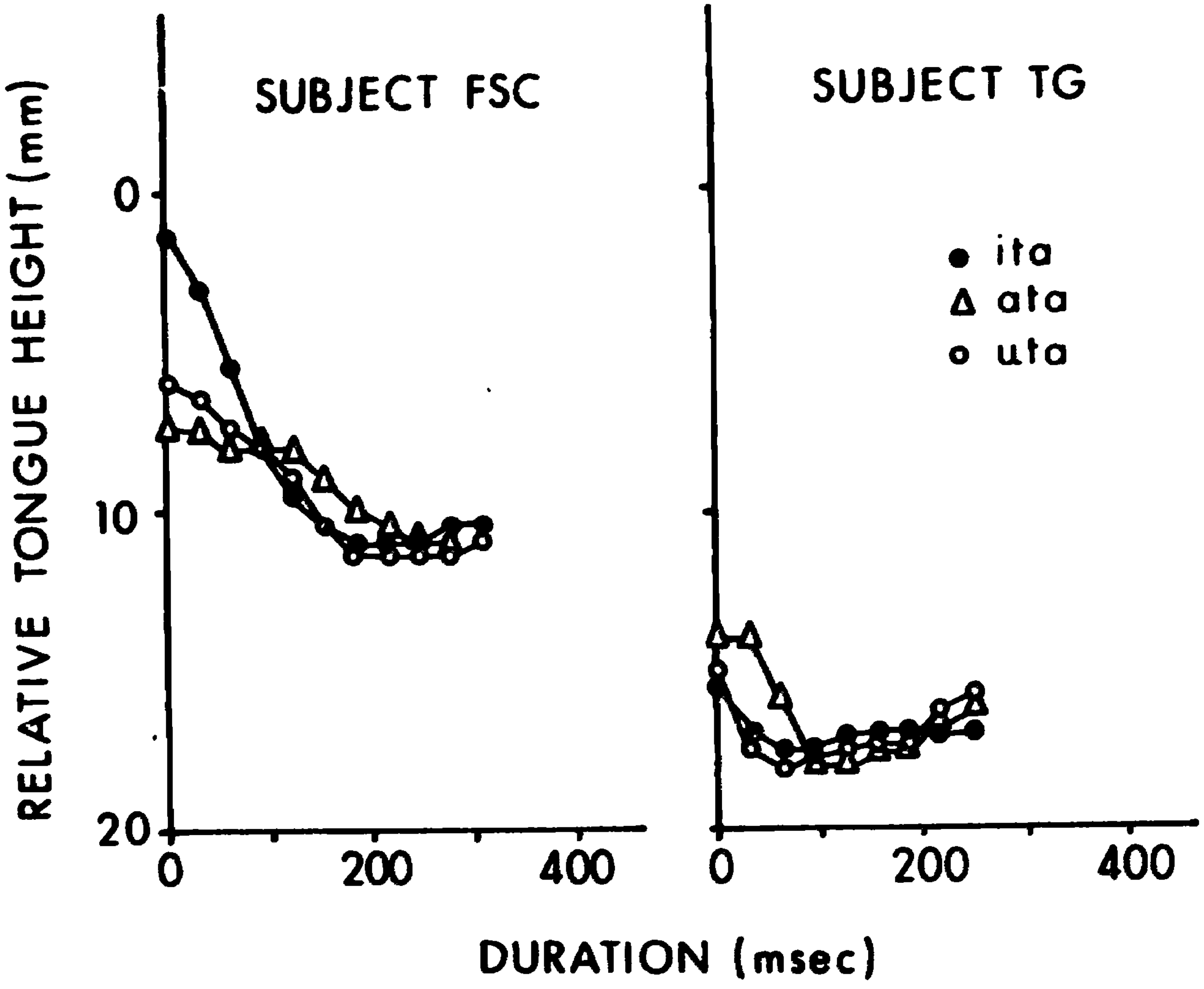


Figure 11: Net tongue movement for /a/. Each data point represents the relative position of the jaw subtracted from the relative position (height) of the tongue. 0 on the abcissa equals time of closure for the consonant.

FIGURE 11

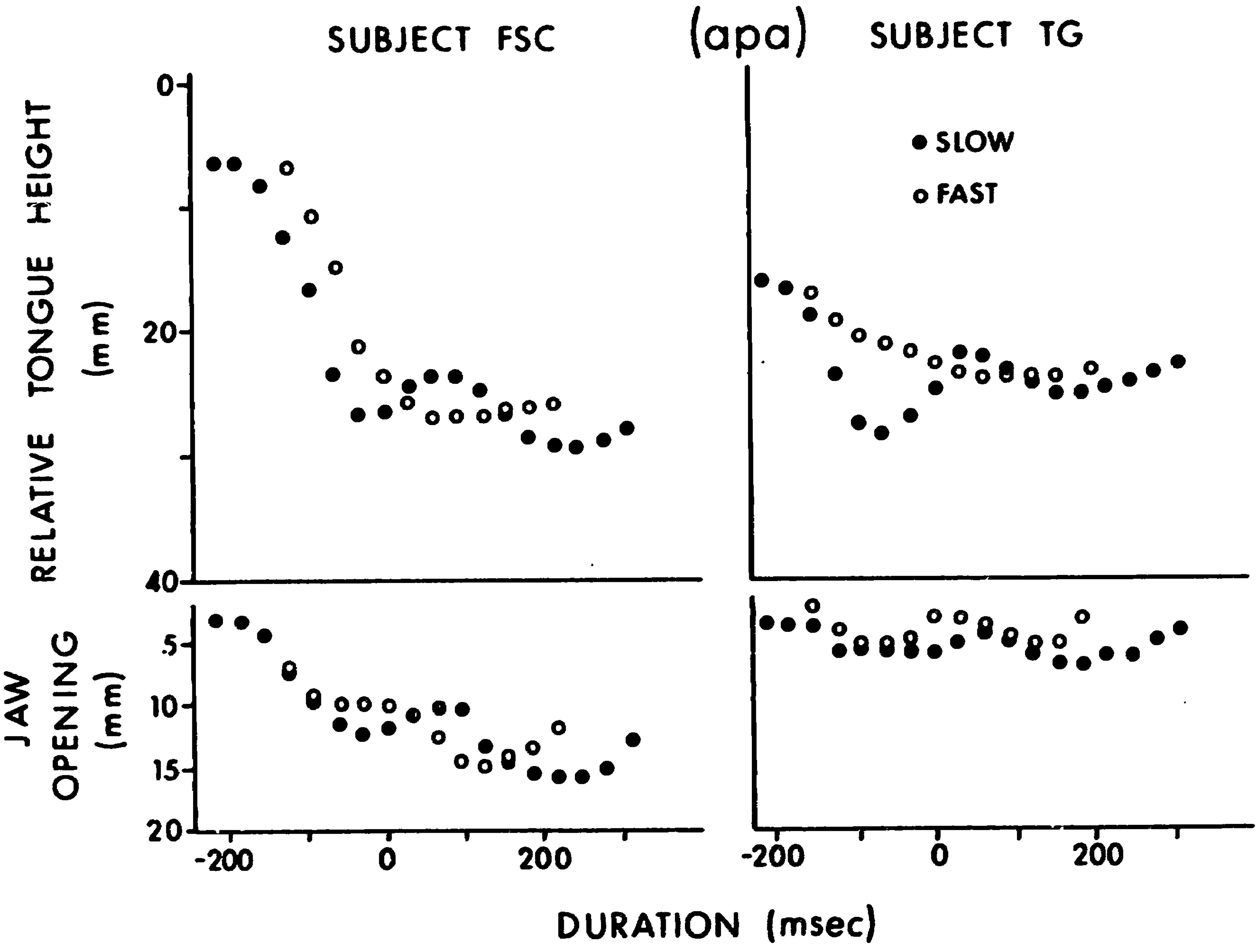


FIGURE 12

Figure 12: Effect of speaking rate on tongue height and jaw opening for /apa/. O on the abcissa equals time of closure for the consonant.

acoustical measures still provided adequate information to shed some light on the two most important questions at issue: (1) whether the context-dependent articulatory variability resulted in corresponding acoustic variability, and (2) whether the undershoot effects evident during fast speech reduced the acoustic vowel triangle towards the neutral schwa.

Although the spectrograms showed the presence of considerable acoustic variability, this variability could not be attributed to any of the coarticulation effects described earlier; acoustic variability occurred almost at random. In fact, the only consistent acoustic effect occurred for TG where the first and second formant frequencies of /a/ showed a consonant effect. First and second formant frequencies for /a/ averaged 775 Hz and 1300 HZ when the intervocalic consonant was /p/ and 825 Hz and 1425 Hz when the intervocalic consonant was either /t/ or /k/. We should remember, however, that the coarticulation effects of the consonant occurred only for FSC and not TG! The articulatory targets of TG were stable for these utterances. The ranges for first and second formant frequencies across all consonants are shown in Table 1.

TABLE 1: Ranges of formant frequencies for all occurrences of /i/, /a/, and /u/ during the slow speaking rate condition. Values are rounded to the nearest 25 Hz.

	Subject FSC		Subject TG	
	F1	F2	F1	F2
/i/	350 - 450	1850 - 2125	475 - 575	2150 - 2375
/a/	700 - 850	1150 - 1375	750 - 850	1275 - 1475
/u/	450 - 525	875 - 1025	575 - 625	950 - 1075

The effect of an increase in speaking rate on the formant frequencies of all three vowels is shown, for both subjects, in Figure 13. These graphs show the F1-F2 coordinate positions for all occurrences (where measurable) of /i,a,u/ at both the slow (s) and fast (f) speaking rates.

For both subjects, an increase in speaking rate is accompanied by an increase in the frequency levels of both the first and second formants. The increases are generally greater for FSC than for TG. The formant frequency measurements for both subjects show the same range of variation during fast speech as during slow speech. Some overlap of coordinate positions is also evident for the two speaking rates. The increases in formant frequencies for /i/ and /u/ might be explained by the more open vocal tract observed for these vowels during fast speech (articulatory undershoot for /i/ and /u/ results in a greater degree of openness). This explanation, however, could not apply to the formant frequency shift observed for /a/. The most important aspect of these measurements, however, is that the acoustic triangle is not reduced towards the neutral schwa during fast speech, i.e., articulatory undershoot during fast speech does not produce the same acoustic result as articulatory undershoot during destressed speech (Lindblom, 1963).[2] These different acoustic effects are probably related

[2] The notion of vowel neutralization during fast speech does not hold up even if the somewhat dubious measures of F1 for /i/ and /u/ are discounted; the upward shift of F2 for /i/ precludes this possibility.

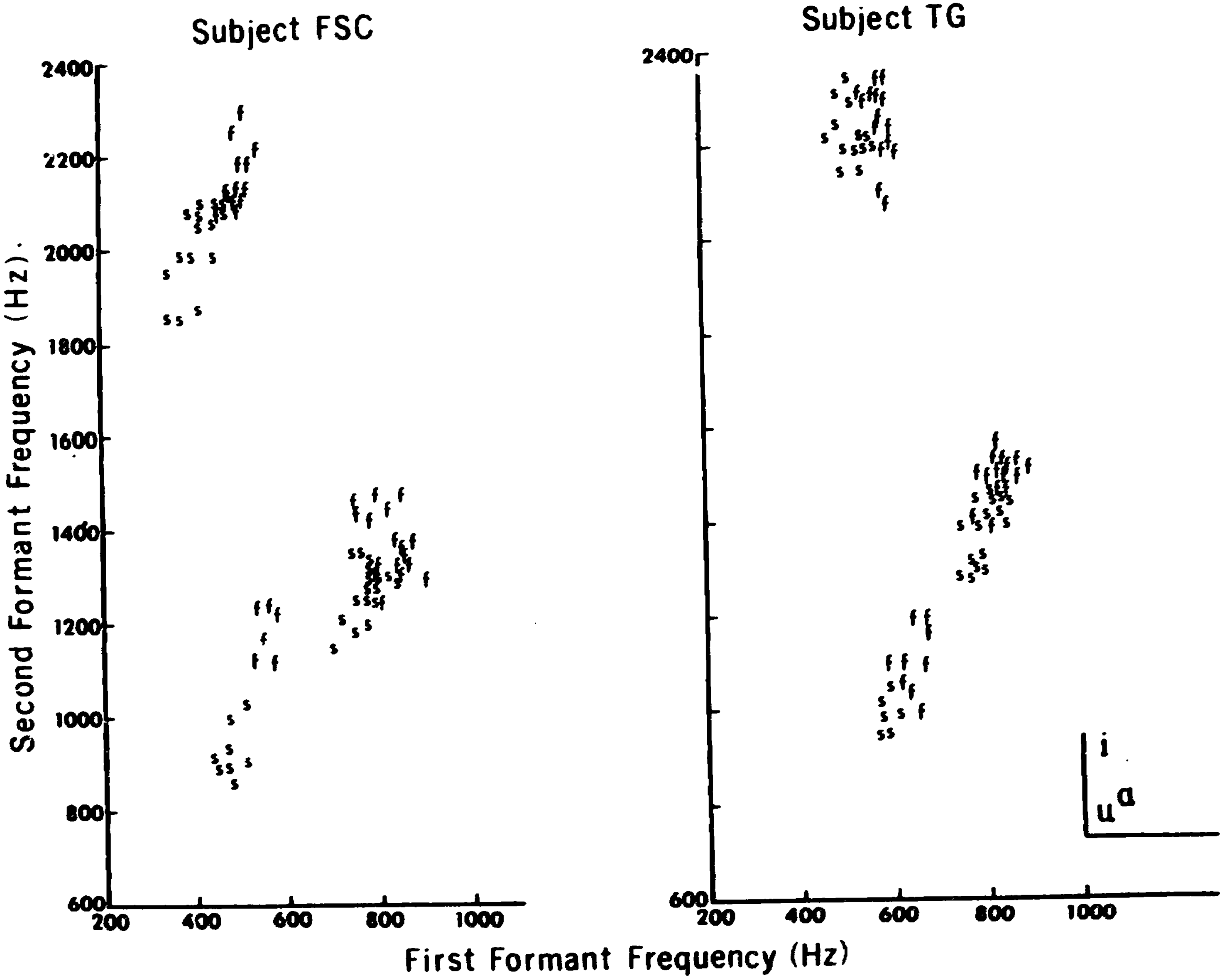


FIGURE 13

Figure 13: First and second formant frequency measurements for all occurrences of /i/, /a/, and /u/ for the slow (s) and fast (f) speaking rates. All measurements are to the nearest 25 Hz.

to differences in the magnitude of articulatory undershoot characterizing the vowel during fast and destressed speech.

DISCUSSION

The major findings of this experiment can be summarized as follows. During slow speech, the target positions of both /i/ and /u/ remain relatively stable across changes in both the preceding and following consonant and vowel. The production of /a/, although not subject to right-to-left effects beyond the following consonant, is sensitive to changes in the consonant, as well as in the vowel preceding the consonant. These coarticulation effects, however, are not reflected, as such, in the acoustical measurements. The production of /i/, /a/, and /u/ during fast speech is characterized by articulatory undershoot and an upward shift in the frequencies of the first and second formants.

These results tend to be somewhat perplexing; there is, on the one hand, a strong tendency for target position stability for /i/ and /u/, and on the other hand, almost as strong a tendency for variability for /a/. The extent of these differences in variability is considerable: maximally, 3 mm for /i/ and /u/, and 8 mm for /a/. Yet acoustic variability for /a/ is no greater than that for /i/ and /u/. This seems to indicate that the acoustical properties of /i/ and /u/ are more sensitive to articulatory variability than those for /a/, at least for the parameters measured in this study. However, Stevens' (1972) view that opening for an open vowel, for example, can be perturbed considerably without any change occurring in the acoustic output is accommodated by these data only if the acoustic variability observed were the result of an articulatory perturbation not measured in this experiment (pharyngeal cavity size, in particular).

Perhaps it is the relative acoustic insensitivity to articulatory variability that allows the speech production mechanism a certain degree of latitude in the production of /a/; on the other hand, the effects might be primarily inertial. While /i/ and /u/ targets are attained primarily by movements of the tongue, opening for /a/ is controlled by the jaw. It is conceivable that because of its greater mass and the nature of its suspension system, the jaw cannot be moved about with the same degree of accuracy as the tongue.

An increase in speaking rate is also accompanied by articulatory variability; undershoot for the vowel target is more the rule than the exception. However, reduction towards schwa is not evident in the acoustic measures. This, of course, is not necessarily an unexpected result. If vowels produced during faster speech were neutralized, fast speech would be characterized by unintelligible strings of consonants and schwas. Even though undershoot for the vowel is evident for both fast and destressed speech, it is obvious that the two features are controlled by two different strategies. The differences in the acoustic effect are probably due to the magnitude of articulatory differences.

Unlike stress effects, speaking rate effects cannot be attributed solely to articulatory sluggishness. The data of both Gay et al. (1974) and Gay and Ushijima (in press) show (for the same two subjects used in this experiment) that vowels produced during fast speech are characterized by a *decrease* in the activity level of the muscle; in other words, undershoot is *programed* into the gesture. This means, in effect, that a gesture towards a vowel is not directed toward one specific target position. The gesture can be modified to some degree without the loss of whatever perceptually significant acoustic feature limits the vowel. Although the acoustic data during fast speech show an upward shift in both the first and second formants for both subjects, it is not known whether these shifts occur simply within the field of each individual vowel or represent a generalized upward shift of the entire triangle.

Because variability is built into the production of a phone at a level higher than the peripheral speech mechanism, a vowel target cannot be internalized, much less operationally defined, as an invariant event. Nonetheless, MacNeilage's (1970) three-dimensional coordinate system still seems to be the best basis for describing a vowel. However, such a specification would have to be expanded to include a spatial field, the boundaries of which are defined by the acoustic limits of the vowel.

Although the data of this study easily fit into a field-specified vowel system, the entire schema is by no means complete. First, this experiment studied only the tongue-jaw system; the entire pharyngeal cavity remains unspecified. Second, the point vowels, although delimiting, and perhaps normalizing the vowel space for a given individual, cannot serve to specify all the vowels of a language (Lieberman, in press). Indeed, before a general model of vowel production can be posited, the intermediate vowels must likewise be specified.

In summary, this experiment produced two major findings. First, articulatory variability in terms of vowel target position exists, but not to a degree that correlates to the existing acoustic variability. Second, articulatory variability also occurs with an increase in speaking rate; speaking rate effects, however, unlike stress effects, are accompanied by an upward shift in formant frequencies.

REFERENCES

Gay, T. and T. Ushijima. (in press) Effect of speaking rate on stop consonant-vowel articulation. In Proceedings of the Speech Communication Seminar, Stockholm, 1974. (Stockholm: Almquist and Wiksell).

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

Harris, K. S. (1971) Action of the extrinsic musculature in the control of tongue position: A preliminary report. Haskins Laboratories Status Report on Speech Research SR-25/26, 87-96.

Houde, R. A. (1967) A study of tongue body motion during selected speech sounds. Doctoral dissertation, University of Michigan.

Kent, R. D. and R. Netsell. (1971) Effects of stress contrasts on certain articulatory parameters. Phonetica 24, 23-44.

Kuehn, D. P. (1973) A cinefluorographic investigation of articulatory velocities. Doctoral dissertation, University of Iowa.

Lieberman, P. (in press) On the Origins of Language: An Introduction to the Evolution of Human Speech, MacMillan Series on Physical Anthropology. (New York: MacMillan).

Lindblom, B. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. 34, 1773-1781.

Lindblom, B. and J. Sundberg. (1971) Acoustical consequences of lip, tongue, jaw, and larynx movements. J. Acoust. Soc. Amer. 50, 1166-1179.

MacNeilage, P. F. (1970) Motor control of serial ordering of speech. Psychol. Rev. 77, 182-196.

MacNeilage, P. F. and J. L. DeClerk. (1969) On the motor control of coarticulation in CVC monosyllables. J. Acoust. Soc. Amer. 45, 1217-1233.

Stevens, K. N. (1972) The quantal nature of speech. In Human Communication: A Unified View, ed. by E. E. David and P. B. Denes. (New York: McGraw-Hill).

Mechanisms of Duration Change*

K. S. Harris+
Haskins Laboratories, New Haven, Conn.

ABSTRACT

Changes in stress, speaking rate, and terminal consonant are known to modify the duration of vowels in spoken utterances (Lehiste, 1970). Electromyographic investigation permits a detailed examination of the mechanisms underlying these observed effects; in particular, it is possible to test Lindblom's (1963) hypothesis that context-dependent vowel color alterations result from a change in the timing of the signals to the articulators, rather than from a reorganization of the articulatory process. Results suggest that the articulatory process itself is reorganized, and that reorganization is different for the three types of change.

INTRODUCTION

In 1963, Lindblom wrote a classic paper on the effects of variations in word stress on the target formant position of vowels. The object of his experiments was to show that the so-called "vowel neutralization phenomenon" could be derived from a very simple model of upper vocal tract control. The phenomenon itself is well-known; briefly, as a syllable is destressed, its vowels will tend to be more neutralized, as well as shorter in duration. Lindblom suggested that the neutralization is a consequence of the shortening. He made a number of spectrographic measurements, showing a regular relationship between duration of vowels and their target formant positions. The relationship is consonant with a model in which the signals sent to the articulators are determined by a stored template for each vowel, independent of its stress position; if signals are sent to the articulators at rates greater than some critical value, target position is not attained before new signals arrive; thus, the shorter the vowel, the greater the target undershoot. In his original paper, Lindblom (1963) suggests that the same model can be applied to the effects of changes in speaking rate

*Paper presented at the Speech Communication Seminar, Stockholm, Sweden, 1-3 August 1974, and to be published in the conference proceedings (Stockholm: Almqvist and Wiksell; and New York: Wiley).

+Also the Graduate School of the City University of New York.

Acknowledgment: The series of experiments reported in part here were performed in close association with L. Raphael, T. Gay, and T. Ushijima.

and changes in stress on vowel color. The model presumably could be extended to apply to any context effect that might be expected to cause changes in vowel duration, such as the well-known effect of the voicing status of the final consonant on the vowel. Lindblom did not suggest the level at which constant signals are presumed to be sent to the articulators. The simplest suggestion is that the signals to the muscles might be constant. If this were true, we might expect electromyographic (EMG) signals to the muscles to be of equal size, under conditions of varying stress, speaking rate, and voicing status of the final consonant. A secondary result of Lindblom's model is that the timing relationship between consonant and vowel signals may be expected to change, as the duration of the vowel changes.

The experiment described here was the latest in a series of tests of Lindblom's hypothesis; the results will be compared with a number of related experiments.

EXPERIMENT

Procedure

A single speaker recorded the four-syllable nonsense utterances /apipipa/ and /apipiba/ under several conditions: stress was placed on either the first or the second /i/; there were two speaking rates; and thus, there was a total of eight utterance types. Utterances were arranged in random lists of 28, with slow and fast lists alternated. After the removal of faulty utterances, there were between 24 and 33 utterances of each type for averaging.

Hooked-wire electrodes were inserted bilaterally into the anterior portion of the genioglossus muscle (GGA). A single electrode was inserted into the posterior portion of the genioglossus muscle (GGP). Results from other placements for this experiment will not be discussed here. Electrode construction and placement are discussed in Hirose (1971).

EMG data were amplified and recorded on 16-channel instrumentation tape, together with the acoustic signal and code pulses for later computer analysis. After inspection for artifacts, EMG signals were processed and averaged by techniques previously described by Port (1971) and Kewley-Port (1973).

Wide-band spectrograms were made of 56 utterances, half from near the beginning and half from near the end of the recording session. Thus, acoustic records were available of seven utterances of each type. On the spectrograms, measurements were made of the duration of the two syllables containing /i/, from the release of closure to closure or the end of voicing. The second and third formants were measured in each syllable at their peak frequency.

Results

Averaged EMG curves for one of the GGA leads are shown in Figure 1. The point on the time line marked "zero" is the averaging lineup point for each utterance, and is at the /p/ closure of the second syllable. The duration of the acoustic events is indicated above each figure. Clearly the EMG signal associated with each stressed syllable has a somewhat larger peak height, and a somewhat longer duration, than its unstressed counterpart. This is in accord with earlier results (Harris, 1973). Further, there is a systematic tendency

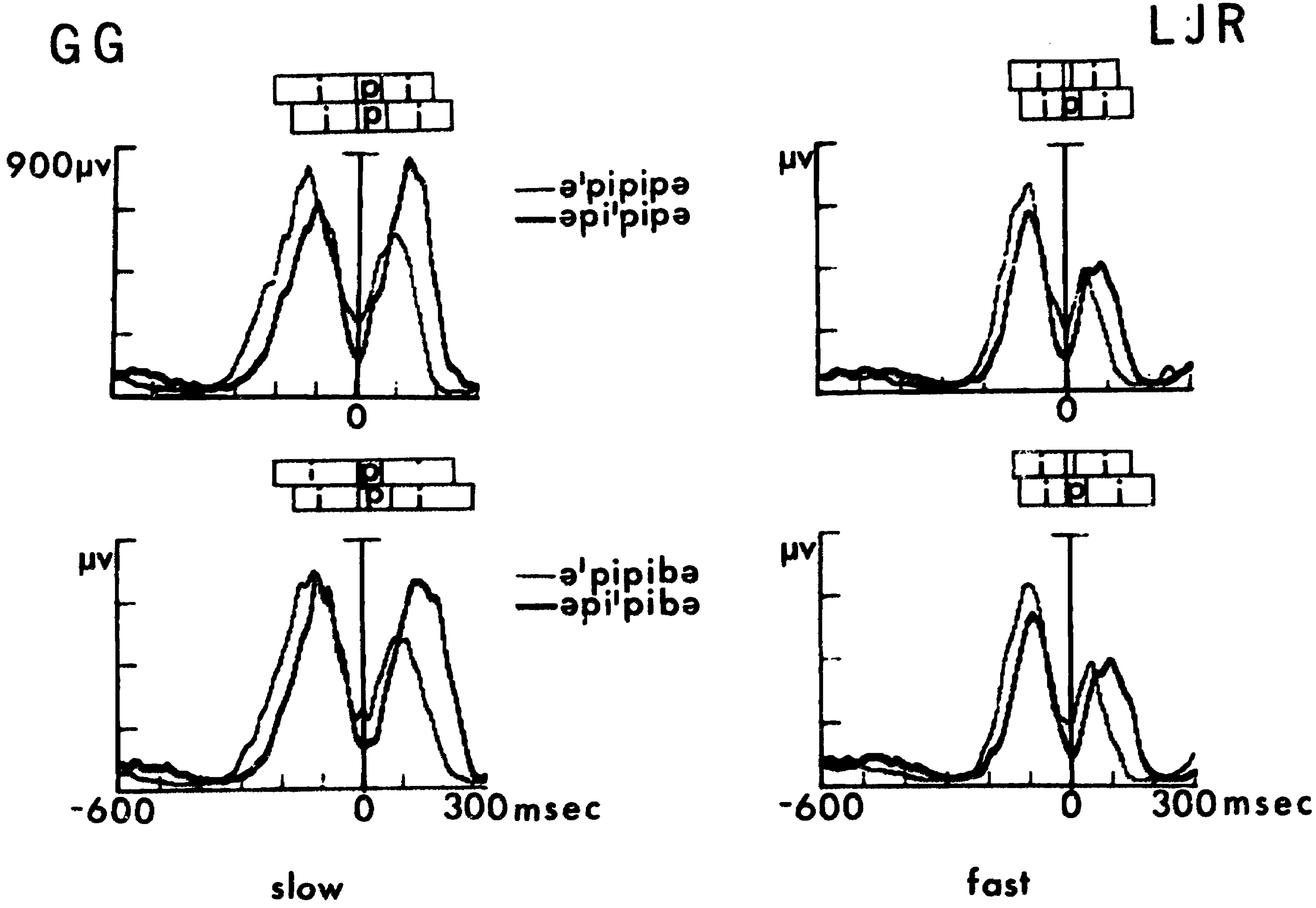

GG
LJR
900μv
μv
μv
μv
—əˈpipipə
—əpiˈpipə
—əˈpipibə
—əpiˈpibə
-600
0
300msec
-600
0
300msec
slow
fast

FIGURE 1

for the fast speaking condition to show smaller peak heights than the slow condition (Gay, Ushijima, Hirose, and Cooper, 1974). Although it is less obvious, there is no overall systematic trend for vowel peaks associated with terminal /p/ and /b/.

Correlation coefficients were calculated between various measures. Results are shown in Table 1. Correlations are quite high, and are uniformly significant at the .05 level or greater. The correlations between the formant levels and duration are essentially a reiteration of Lindblom's results, in a somewhat different form--that is, formants reach a more extreme value as duration lengthens. The correlations between peak height and duration, and peak height and formant value, however, are a contradiction of Lindblom's hypothesis.

TABLE 1

	F_2	F_3	GGA_R	GGA_L	GGP
duration	.89	.81	.57	.60	.56
F_2			.48	.57	.54
F_3			.55	.66	.69
GGA_R				.93	.90
GGA_L					.97

While the overall correlations are rather high, they are somewhat misleading, since the effects of the terminal /b/ on half the utterances are masked. More detailed results, for the second syllable only, are presented in Table 2.

TABLE 2

	p				b			
	Slow		Fast		Slow		Fast	
	$\acute{S}_2$	S_2	$\acute{S}_2$	S_2	$\acute{S}_2$	S_2	$\acute{S}_2$	S_2
duration in msec	167	145	131	117	201	179	147	133
F_2 in Hz	1945	1907	1857	1807	1936	1949	1879	1816
F_3 in Hz	2320	2232	2303	2191	2345	2300	2250	2177
GGA_R in μV	876	616	471	460	745	553	460	454
GGA_L in μV	311	263	227	189	283	247	229	191
GGP in μV	343	271	242	186	299	246	215	172

An inspection of the table shows the following results:

1) The expected effects of stress, speaking rate, and voicing on the duration of the second syllable are obtained.

2) Values of F_2 and F_3 are more extreme for slow speech and for stressed production. However, there is no systematic tendency for the values to differ for terminal /p/ and /b/. The result is not surprising in view of the classic literature. So far as I know, it has never been suggested that vowels are more neutral before voiceless consonants.

3) Stress and speaking rate affect the peak values of muscular activity. In ten of twelve possible comparisons, peaks are higher for terminal /p/ than for terminal /b/. The result suggests a systematic trend. However, Raphael (1974) has examined the peak heights associated with vowel production in a large number of utterances in which the terminal consonant is a voiced or voiceless stop or fricative. His results show a prolongation of the vowel signal before voiced consonant, but no systematic differences in amplitude. Obviously, this result requires more systematic examination.

DISCUSSION

It is interesting to consider Lindblom's explanation of the stress effect in light of the picture it gives of the organization of running speech. In his model, pushed to an extreme, a series of signals are sent to the articulators, which depend for their identity on the phonetic specification of the segments. Changes of stress or speaking rate will affect the timing of the arrival of these signals (as will certain segmental characteristics of the sequence, by extension) but not their relative size. The resulting acoustic output will vary, not because of variation in the signal size, but because of changes in the relative timing of, for example, successive vowel and consonant signals. Furthermore, differences between contexts vary along the single dimension of time.

The results described above suggest that the real picture is substantially more complex. Signal size for vowels varies systematically with duration for changes in stress and speaking rate, and does not (apparently) vary with the duration changes conditioned by the voicing status of the terminal consonant. Let us examine the stress and speaking rate variations first, since Lindblom's model is intended to apply only to them. Is the target position observed due entirely to the size difference observed, or may the result be due in part to Lindblom's proposed mechanism? In Lindblom's model, any duration change automatically generates a change in target position, unless it is counteracted by some other adjustment. Therefore, if vowels are longer preceding /b/, then signals should be smaller, if the acoustic target for the formants is to be the same. As noted above, in the present experiment, there is a trend in this direction, although the same trend is not seen in other experiments. This point must be examined further.

In the present experiment, the effects of stress and speaking rate are at least qualitatively homogeneous. However, there is some evidence that consonant and vowel signals do not behave in parallel ways under the two manipulations.

Some years ago, we found that consonants associated with heavily stressed syllables will be produced with stronger associated articulation (Harris, Gay, Sholes, and Lieberman, 1968). Gay et al. (1974) have found that faster speaking rates are associated with higher consonant peak heights. These effects are in opposite directions to the associated vowel articulations. Our information is, however, concerned with the somewhat special circumstance of the labial consonant surrounding the vowel, and should be examined in more varied environments.

Lindblom's model is in one respect similar to a much earlier formulation proposed by Cooper, Liberman, Harris, and Grubb (1958). In both models, constant signals yield a variable output, due to variability in timing. In all the experiments we have performed in manipulating stress, speaking rate, and context, we seem to get the opposite result. The relative timing of consonant and vowel gestures to different articulators seems to be very closely time locked, while the amplitude and duration of gestures for particular segments vary substantially. We will be interested to see what happens in those conditions where the same articulator is involved in both consonant and vowel gestures.

REFERENCES

Cooper, F. S., A. M. Liberman, K. S. Harris, and P. M. Grubb. (1958) Some input-output relations observed in experiments on the perception of speech. In Second International Congress on Cybernetics. (Namur, Belgium: International Association of Cybernetics).

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

Harris, K. S. (1973) Stress and syllable duration change. Haskins Laboratories Status Report on Speech Research SR-35/36, 31-38.

Harris, K. S., T. Gay, G. Sholes, and P. Lieberman. (1968) Electromyographic measures of consonant articulations. Haskins Laboratories Status Report on Speech Research SR-13/14, 137-152.

Hirose, H. (1971) Electromyography of the articulatory muscles: Current instrumentation and technique. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.

Kewley-Port, D. (1973) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research SR-33, 173-183.

Lehiste, I. (1970) Suprasegmentals. (Cambridge, Mass.: MIT Press).

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. 35, 1773-1781.

Port, D. K. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research SR-25/26, 67-72.

Raphael, L. J. (1974) The physiological control of durational differences between vowels preceding voiced and voiceless consonants in English. Haskins Laboratories Status Report on Speech Research SR-39/40 (this issue).

The Physiological Control of Durational Differences between Vowels Preceding Voiced and Voiceless Consonants in English*

Lawrence J. Raphael[+]
Haskins Laboratories, New Haven, Conn.

ABSTRACT

A series of two electromyographic experiments was designed to determine the nature of the muscular activity underlying the articulation of consonant-vowel-consonant (CVC) syllables in which identical vowels differed in duration because of the voicing characteristic of the consonant that followed them. Results indicate that the most reasonable hypothesis to explain the durational differences posits a sustention of muscular activity in the articulatory gesture of the vowel preceding voiced consonants, relative to the gesture for vowels preceding voiceless consonants. It is noted that the acoustically determined differences between vowels, the differences between the durations of the muscular-articulatory gestures for the vowels, and the temporal displacement of the final consonant peaks generally show remarkably similar values.

The differences between the durations of vowels preceding voiced and voiceless consonants in English is well documented in the phonetic literature (Locke and Heffner, 1940; Kenyon, 1951; Peterson and Lehiste, 1960; House, 1961; Gimson, 1962). Investigators have described and/or commented on the perceptual consequences of these differences (Denes, 1955; Noll, 1960; Jakobson and Halle, 1967; Raphael, 1972), and have theorized as to whether the variation in vowel duration is a physiologically mandated behavior, one that is learned, or, to some extent, both (Zimmerman and Sapon, 1958; Peterson and Lehiste, 1960; House, 1961; Delattre, 1962; Elert, 1964; Chen, 1970).

Little, however, has been discovered or written about the physiological activity that must underlie durational differences, no matter what their cause. The present study was undertaken to specify the muscular activity governing the articulatory gestures for vowels preceding both voiced and voiceless consonants. The studies referred to above, whether based on impressionistic evidence or on the analysis of acoustic records, suggest three hypotheses for the mechanism that renders vowels longer in duration before voiced consonants than before

*To be published in Journal of Phonetics (1975) 3, 25-33.

[+]Also Herbert H. Lehman College of the City University of New York.

[HASKINS LABORATORIES: Status Report on Speech Research SR-39/40 (1974)]

voiceless consonants. The first, and perhaps the simplest, posits a greater duration of muscular activity for a vowel preceding a voiced consonant than for one preceding a voiceless consonant. Under this hypothesis final consonant articulations of either voicing type would be more or less identical, with the same time of onset relative to the offset of the preceding vowel.

The second hypothesis posits muscular activity of the same duration for vowels in both the voiced and voiceless environments. The durational difference would then be effected by a difference in the timing of the onset of muscular activity of the following consonants in relation to the offset of preceding vowel activity: relatively earlier in the voiceless case and relatively later in the voiced case. Such differences have been found in lip and jaw movements for stops [Kozhevnikov and Chistovich, 1965; Ohala, Hiki, Hubler, and Harshman, 1968; Chen, 1970; Kim and MacNeilage, 1972 (cited in MacNeilage, 1972); Leanderson and Lindblom, 1972], although the magnitude of these differences does not appear to be great enough to account for the durational difference between English vowels (MacNeilage, 1972).

The third hypothesis merges the first two and posits differences both in the duration of muscular activity for vowels and in the relative timing of the onset of the muscular activity for the following consonants.

EXPERIMENT I

Procedure

We constructed a series of real-word, minimal pair, CVC test utterances in which the articulation of the initial consonants and vowels would be essentially controlled by a muscle or set of muscles different from and independent of the muscles controlling the articulations of the final consonants. For example, in the minimal pair leaf-leave, it was assumed that the initial consonant and the vowel would be controlled by lingual muscles, whereas the final consonant would be under the control of labial muscles. (This proved to be the case for a pair such as leaf-leave, but for other pairs, namely those containing back, lip-rounded vowels, the separation of muscle function was not as clear as had been desired. For pairs such as bought-bawd and moat-mowed there was genioglossus activity for the vowel as well as for the final consonant. A more anterior electrode placement might reduce or eliminate this activity. Thus, some of the information concerning onset of muscle activity for the final consonant was obscured, although by no means completely.) Of the six minimal pairs used, three were of a labial-to-lingual configuration (mowed-moat, bawd-bought, moos-moose), and three of a lingual-to-labial configuration (leave-leaf, thieve-thief, lab-lap).

Two subjects took part in the experiment. Both read the words in isolation from a series of ten randomized lists. At least 12 and as many as 19 tokens of each type were used to produce the averaged electromyographic (EMG) curves.

The muscles explored for labial articulation were the orbicularis oris and the depressor anguli oris. Lingual articulation was investigated by recording EMG signals from the genioglossus muscle.

Concentric needle electrodes of standard Disa type were used for insertions into the muscles. Both the EMG output and a voice trace were recorded on magnetic tape for subsequent computer processing. The onset of voicing was used as a reference line-up (zero) point in the data manipulation.

Results and Discussion

The only effect consistently found was that of greater duration of muscular activity in the articulation of vowels preceding voiced consonants. Figure 1 shows the most common manifestation of this effect:[1] the peaks associated with the vowel articulation occur almost simultaneously in both the voiced and voiceless cases; there is a sustention of muscular activity in the voiced case relative to the voiceless case; the onsets of the muscular activity for the following consonants occur at approximately the same time relative to the offset of muscular activity for the preceding vowel; the onset durations and slopes for the muscular activity associated with the following consonants are generally equivalent. Certainly there is no durational difference between the onsets of consonant activity (relative to the preceding vowel) on the order of the durational differences between vowels as determined from acoustic records. For the utterances and subject of Figure 1, the average vowel duration for thief was 150 msec, and that for thieve 360 msec. The durational difference between the muscular activity underlying the vowel articulations, with reference to the time each EMG curve reaches its base line, is on the order of 220 msec, quite close to the 210 msec difference between the durations of the vowels in the acoustic measurement.

This sustention of muscular activity following the peak for the vowel was found for both subjects in most cases. Figure 2 shows typical examples of the averaged durational differences between the EMG signals caused by the sustention. Figure 3 displays the temporal displacement of the terminal consonant peaks associated with the vowels of Figure 2. Note that these final-consonant peaks were displaced from each other by time values approximately equal to both the durational differences between the EMG signals for the preceding vowels and their acoustically determined durational differences. The data for the vowel duration differences (both acoustic and EMG) and the temporal displacement of the EMG peaks of the final consonants are summarized in Table 1.

One other articulatory strategy can be found underlying the durational differences between vowels: in two cases there is a delay in the onset and peaking of muscular activity for the vowel preceding the voiced consonant (Figures 2, 3, bottom graphs). This, in turn, causes a delay in the peak of muscular activity

[1] In order to maximize, visually, the temporal relationships between EMG traces, the peaks for all vowels and consonants have been equated in height, regardless of their actual microvolt values. However, these values are recorded: in Figure 1, the microvolt values on the left-hand ordinate are those of the vowel (genioglossus) gesture, and the microvolt values on the right-hand ordinate are those of the final consonant (orbicularis oris) gesture. In Figures 2-5, the microvolt values for the vowel preceding the voiceless consonant and for the voiceless consonant itself are shown on the left-hand ordinate; those for the vowel preceding the voiced consonant and for the voiced consonant itself are shown on the right-hand ordinate. The actual values shown are peak values for vowel or consonant.

FIGURE 1

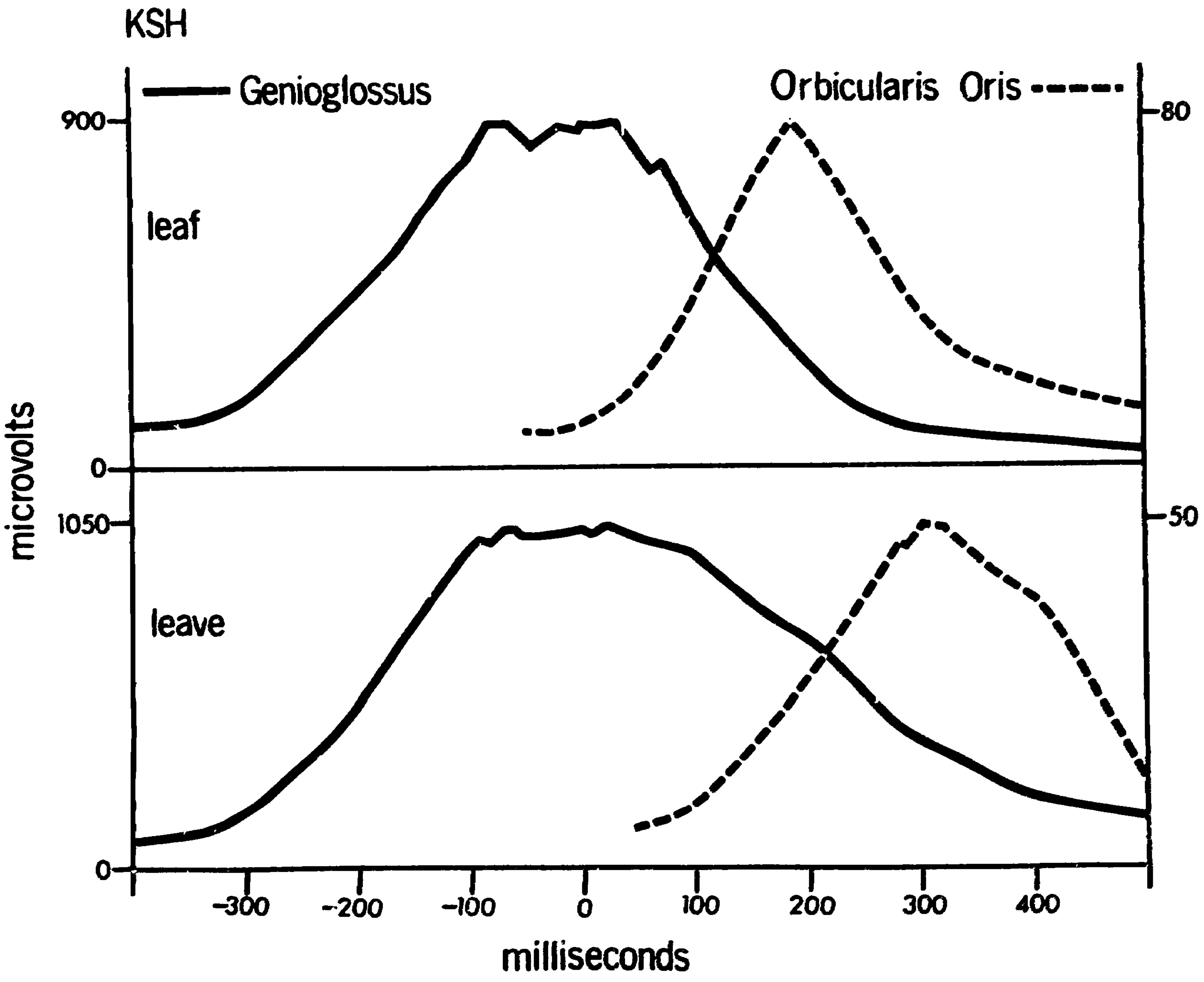


Figure 1: Typical EMG activity for two syllables contrasting in the voicing characteristic of the final consonant.

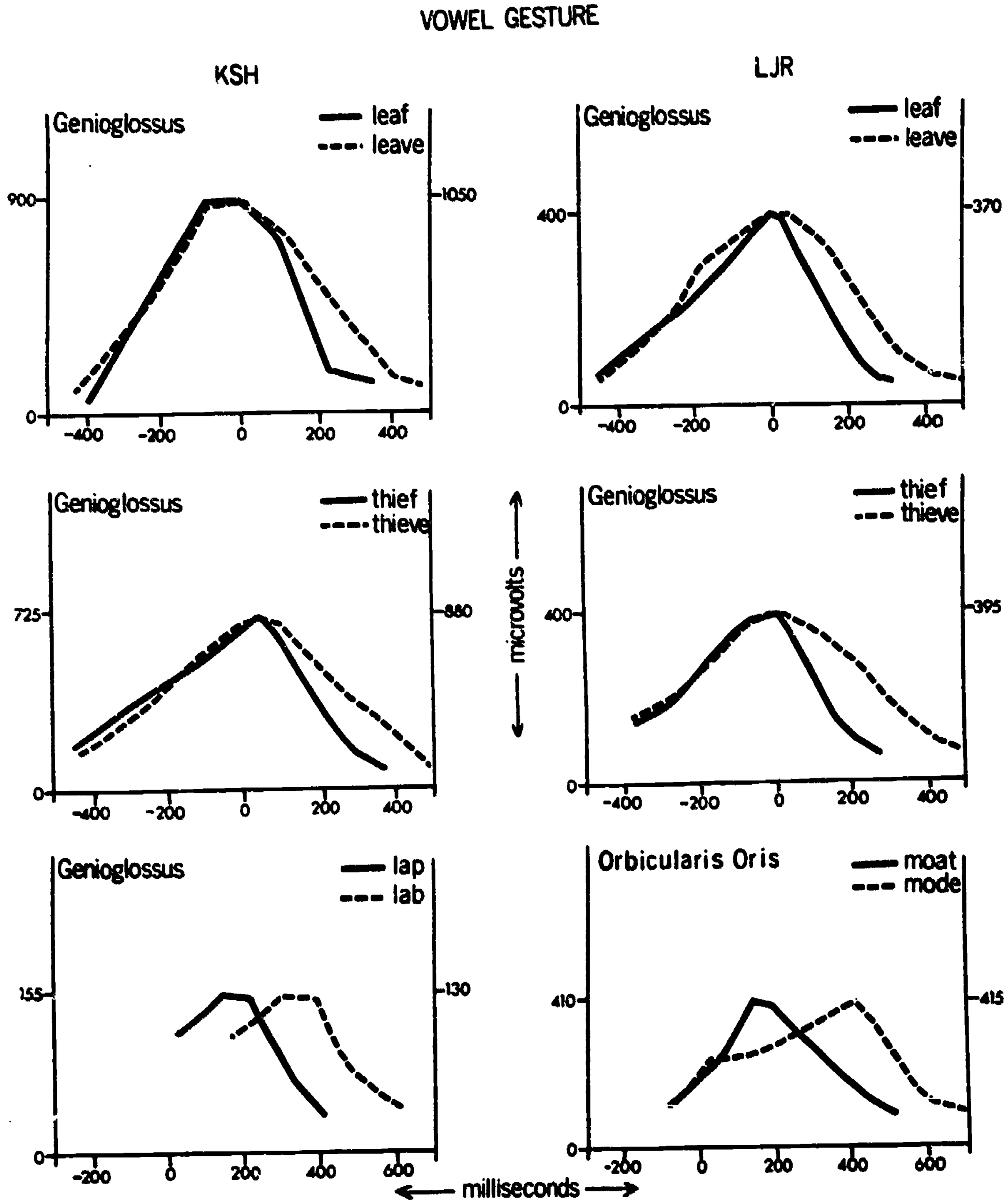


Figure 2: Paired EMG signals for identical vowels, one preceding a voiced and the other a voiceless, syllable-final consonant.

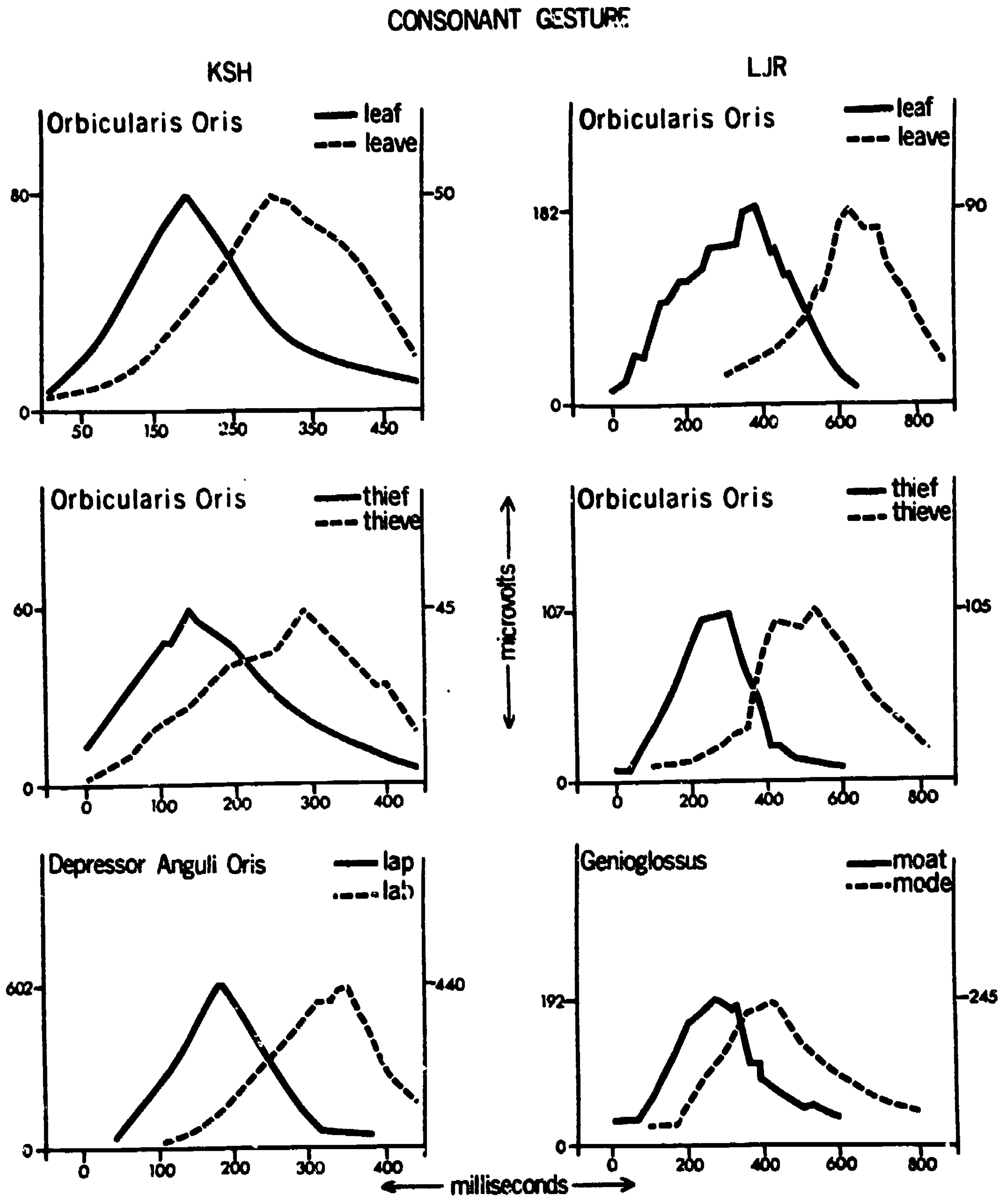


Figure 3: Paired EMG signals for voiced/voiceless syllable-final consonants.

TABLE 1: Comparison of vowel duration differences as determined spectrographically and by EMG measurements with temporal displacement of final consonant EMG peaks.

	Duration Differences (msec)		Consonant Peak Displacement (msec)
	Vowel (Acoustic)	Vowel (EMG)	(EMG)
Subject LJR			
leaf-leave	260	230	210
bought-bawd	155	110	115
moose-moos	180	160	80
moat-mowed	190	185	135
thief-thieve	300	290	170
lap-lab	175	180	165
Subject KSH			
leaf-leave	175	185	250
bought-bawd	90	100	90
moose-moos	100	120	110
moat-mowed	125	160	150
thief-thieve	210	220	200
lap-lab	105	85	130

for the vowel, so that even though the slopes of the offsets of EMG activity are virtually identical in both voiced and voiceless cases, the separation of the final voiced and voiceless consonant peaks in this case is still the result of the difference in timing and duration of vowel articulation. It may be possible, since the initial consonants in these utterances are semivocalic in nature, that part of the durational difference usually carried by the vowel is absorbed by the preceding consonant. Although the acoustic records do not reveal any consistent differences between the durations of these initial consonants, the EMG signals for the initial /l/ and /m/ in the voiced syllables show a slower onset and later peak than do those in the voiceless syllables. Thus the data do not provide a consistent explanation of this effect. Further, there remains the question of why one subject shows the effect for /l/ and not for /m/, and the other for /m/, but not for /l/.

EXPERIMENT II

Procedure

The minimal-pair utterances of the second experiment were disyllables beginning with schwa, followed by /p/. The interconsonantal vowel was variously /i I e ɛ æ a ʌ ɔ o ʊ u/. The final consonant was either /p/ or /b/. Utterances of these types provided negligible coarticulation effects, for the muscles investigated, between consonant and vowel, or between the vowel and the final consonant. One of the two subjects who provided data in this experiment had also participated in Experiment I. A third subject (who had also provided data in Experiment I) read an alternate list of utterances ending in /k/ or /g/ from which only final-consonant EMG data were obtained.

Data were obtained from the orbicularis oris muscle for the labial stop consonants and from the mylohyoid muscle for the alternate utterances ending in velar stops. Vowel data were obtained from the genioglossus and inferior longitudinal muscles. Electrode insertions were made as described by Hirose (1971) for the orbicularis oris, mylohyoid, and genioglossus muscles. The insertion into the inferior longitudinal muscle was made at the lower tongue surface near the back of the anterior third, approximately 1 cm from the lateral margin and roughly parallel to the lower surface of the tongue at a depth of approximately 5 mm.

The EMG signals were stored on magnetic tape for subsequent data processing. The onset of voicing of the interconsonantal vowel was used as a reference line-up (zero) point for the data manipulation and displays.

Results and Discussion

As in Experiment I, there was a greater duration of muscular activity in the articulations of vowels preceding voiced consonants than in those preceding voiceless consonants. Both subjects showed this main effect for all muscles and for all vowels for which data were obtained in this experiment. Figure 4 (a, b) displays the genioglossus data for the two subjects. The similarity of the vowel sustention effects, as shown in the EMG curves, among subject LJR in Experiment I (Figure 2) and in Experiment II, and subject FBB in Experiment II is readily apparent. The similarity is further reflected in the EMG curve for the inferior longitudinal muscle for subject LJR (Figure 4c).

The displacement of the final consonant peaks is illustrated in Figure 5 for one of the subjects of this experiment and for the subject who read the alternate list of utterances ending in /k/ or /g/. As in the first experiment, the acoustic durational differences as determined from spectrograms, the EMG durational differences, and the temporal displacement values of the EMG peaks for the final consonants show remarkably similar values (Tables 2 and 3).

CONCLUSION

The data presented here provide strong confirmation for the first hypothesis presented above. That is, the acoustically measured durational differences long observed between vowels preceding voiced and voiceless consonants are primarily controlled physiologically by motor commands to the muscles governing the articulators that are active in the formation of vowels. The timing of these commands is generally such that after the peak of the articulatory-muscular activity has been reached, the articulators are maintained (although not statically) in shapes and positions appropriate for vowels somewhat longer when they precede voiced consonants.

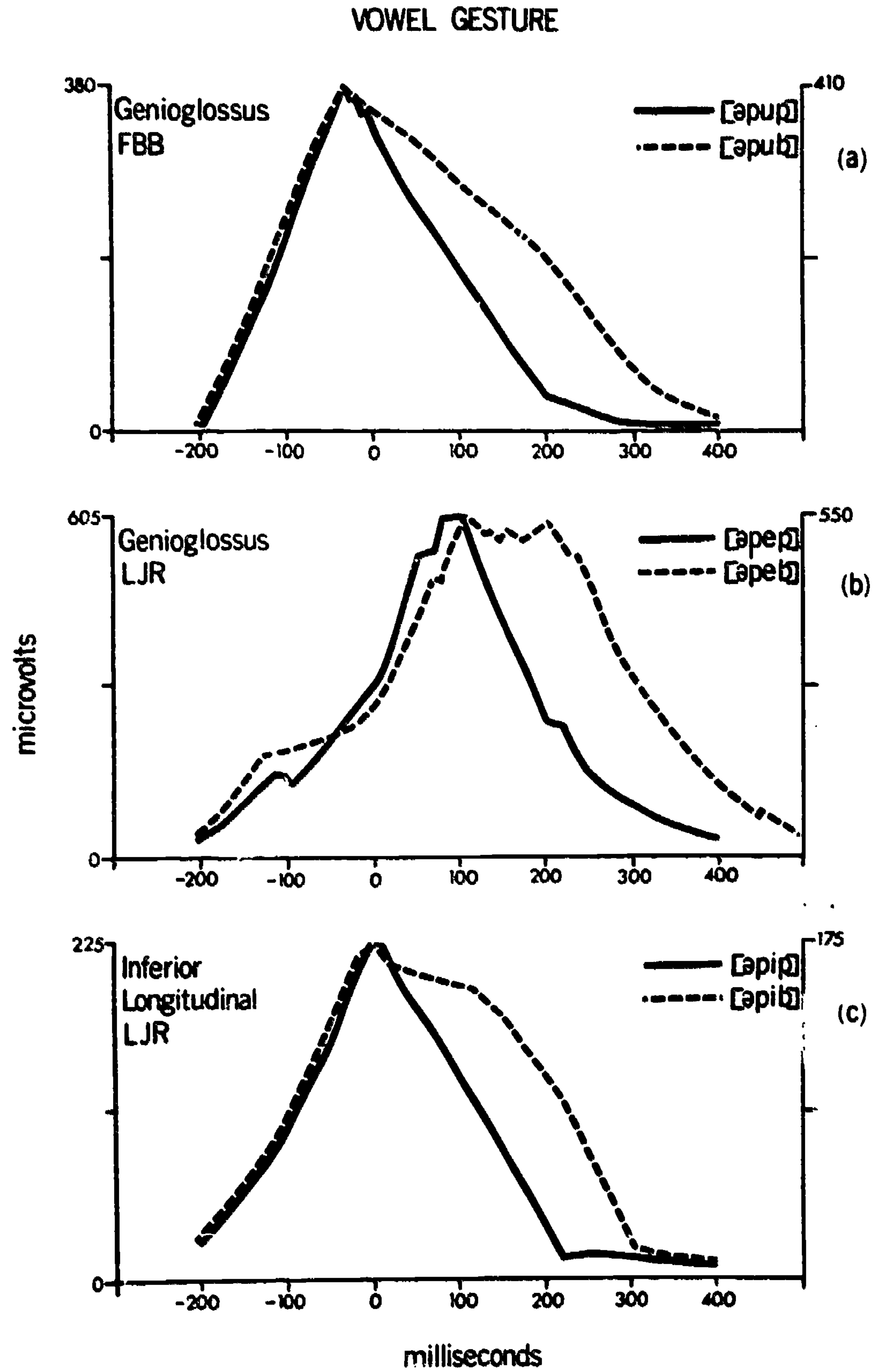


Figure 4: Paired EMG signals for identical vowels, one preceding a voiced and the other a voiceless, syllable-final consonant.

FINAL CONSONANT GESTURE

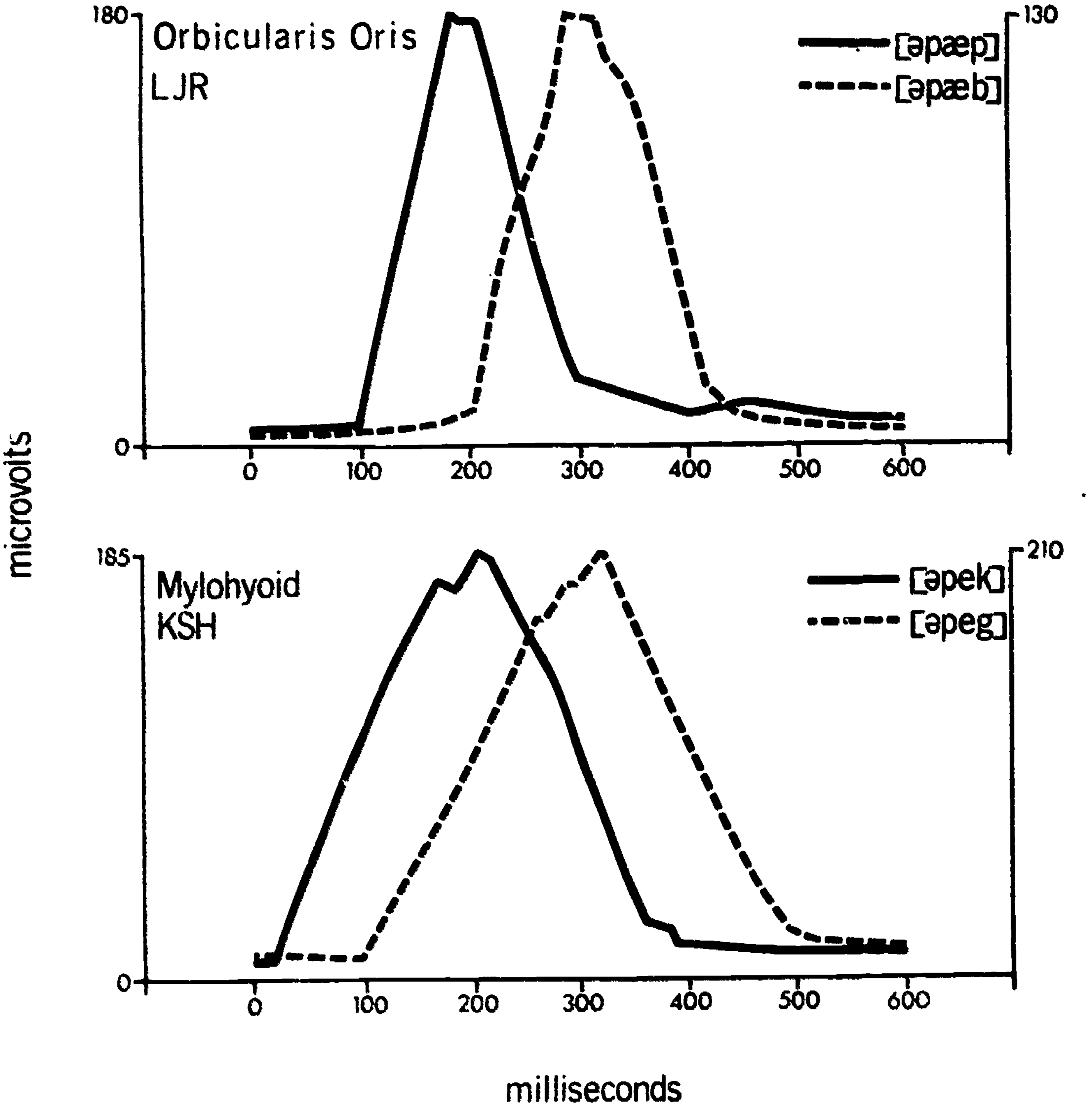


Figure 5: Paired EMG signals for voiced/voiceless syllable-final consonants.

TABLE 2: Comparison of vowel duration differences as determined spectrographically and by EMG measurements with temporal displacement of final consonant EMG peaks (for syllables ending in /p/ versus /b/).

	Duration Differences (msec)		Consonant Peak Displacement (msec)
	Vowel (Acoustic)	Vowel (EMG)	(EMG)
Subject LJR			
/i/	135	125	145
/I/	60	70	65
/e/	130	140	120
/ɛ/	55	55	65
/æ/	85	75	100
/ʌ/	30	40	45
/a/	85	80	105
/ɔ/	115	105	100
/o/	155	165	150
/ʊ/	45	55	60
/u/	110	115	110
Subject FPB			
/i/	150	140	150
/I/	85	95	75
/e/	170	175	150
/ɛ/	50	50	45
/æ/	130	---	145
/ʌ/	65	75	80
/a/	180	---	195
/ɔ/	110	---	120
/o/	205	190	185
/ʊ/	65	60	70
/u/	145	150	140

TABLE 3: Comparison of vowel duration differences as determined spectrographically with temporal displacement of final consonant EMG peaks (for syllables ending in /p/ versus /b/).

	Vowel Duration Difference (Acoustic - msec)	Consonant Peak Displacement (EMG - msec)
Subject KSH		
/i/	135	130
/I/	45	45
/e/	130	140
/ε/	60	70
/æ/	155	145
/ʌ/	45	50
/a/	185	165
/ɔ/	115	100
/o/	135	125
/ʊ/	45	60
/u/	115	105

REFERENCES

Chen, M. (1970) Vowel length variation as a function of the voicing of the consonant environment. Phonetica 22, 129-159.

Delattre, P. C. (1962) Some factors of vowel duration and their cross-linguistic validity. J. Acoust. Soc. Amer. 34, 1141-1143.

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

Elert, C-C. (1964) Phonologic Studies of Quantity in Swedish. (Uppsala: Almqvist and Wiksells).

Gimson, A. C. (1962) An Introduction to the Pronunciation of English. (London: Edward Arnold).

Hirose, H. (1971) Electromyography of the articulatory muscles: Current instrumentation and technique. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.

House, A. S. (1961) On vowel duration in English. J. Acoust. Soc. Amer. 33, 1174-1178.

Jakobson, R. and M. Halle. (1967) Tenseness and laxness. Supplement to Preliminaries to Speech Analysis, by R. Jakobson, C. G. M. Fant, and M. Halle. (Cambridge, Mass.: MIT Press).

Kenyon, J. S. (1951) American Pronunciation, 10th ed. (Ann Arbor, Mich.: George Wahr).

Kozhenikov, V. A. and L. A. Chistovich. (1965) Speech: Articulation and Perception. (Washington, D.C.: United States Department of Commerce Clearinghouse for Scientific and Technical Information).

Leanderson, R. and B. E. F. Lindblom. (1972) Muscle activation for labial speech gestures. Acta Otolaryngol. 73, 362-373.

Locke, W. N. and R-M. S. Heffner. (1940) Notes on the length of vowels II. Amer. Speech 15, 74-79.

MacNeilage, P. (1972) Speech physiology. In Speech and Cortical Functioning, ed. by J. Gilbert. (New York: Academic Press) 1-72.

Noll, J. D. (196C) The perceptual significance of certain acoustical correlates of consonant voicing contrasts. Unpublished Ph.D. dissertation, University of Iowa.

Ohala, J., S. Hiki, S. Hubler, and R. Harshman. (1968) Transducing jaw and lip movements in speech. Paper presented at the 76th meeting of the Acoustical Society of America. Abstract in J. Acoust. Soc. Amer. (1969) 45, 324.

Peterson, G. E. and I. Lehiste. (1960) Duration of syllabic nuclei in English. J. Acoust. Soc. Amer. 32, 693-703.

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.

Zimmerman, S. A. and S. M. Sapon. (1958) Note on vowel duration seen cross-linguistically. J. Acoust. Soc. Amer. 30, 152-153.

Effect of Speaking Rate on Stop Consonant-Vowel Articulation*

T. Gay[+] and T. Ushijima[++]
Haskins Laboratories, New Haven, Conn.

ABSTRACT

The purpose of this experiment was to study the effect of speaking rate on the articulation of the stop consonants /p,t/ in combination with the vowels /i,a,u/. Two speakers of American English read lists of nonsense syllables containing /p,t/ in all possible vowel-consonant-vowel (VCV) combinations with /i,a,u/ at both normal and fast speaking rates. Electromyographic (EMG) records were obtained from the orbicularis oris, superior longitudinal, and genioglossus muscles. The EMG data were analyzed using the Haskins Laboratories' data system. For both labial consonant and lingual consonant production, the effect of an increase in speaking rate was an increase in the activity level of the muscle (orbicularis oris and superior longitudinal). However, for vowel production, the effect of an increase in speaking rate was a decrease in the activity level of the genioglossus muscle. These results are discussed in relation to a general account of speaking rate control.

INTRODUCTION

In some recent experiments on the production of labial CV sequences, we showed that an increase in speaking rate involves more than a simple reordering of the timing of commands to the muscles (Gay and Hirose, 1973; Gay, Ushijima, Hirose, and Cooper, 1974). Rather, the production of both the consonant and the vowel segments of a syllable during fast speech was shown to be characterized by changes in motor organization as well as changes in motor timing. For labial consonants, the effect of an increase in speaking rate is an increase in the activity levels of the muscles that control lip closure; however, for vowels, the opposite effect occurs, i.e., an increase in speaking rate is accompanied by a decrease in the activity level of the muscle (genioglossus).

*Paper presented at the Speech Communication Seminar, Stockholm, Sweden, 1-3 August 1974, and to be published in the conference proceedings (Stockholm: Almqvist and Wiksell; and New York: Wiley).

[+] Also University of Connecticut Health Center, Farmington.

[++] Visiting from University of Tokyo, Japan.

[HASKINS LABORATORIES: Status Report on Speech Research SR-39/40 (1974)]

The purpose of the experiment reported here was to obtain additional labial consonant data from both of our previous subjects and to extend our observations of VCV articulations to lingual CV sequences.

METHOD

Subjects were two adult males, both native speakers of American English. The speech material consisted of the consonants /p,t/ and the vowels /i,a,u/ in a trisyllable nonsense word of the form /k V_1 C V_2 pə/, where V_1 and V_2 were all possible combinations of /i,a,u/, and C was either /p/ or /t/. The utterances were randomly ordered into a master list. Each utterance (preceded by the carrier phrase, "It's a....") was read at two speaking rates: normal and fast. Each rate was based on the subjects' own appraisal of comfortable slow and fast rates. On the average, the fast speech was two-thirds to three-fourths the duration of the normal speech.

For both subjects, conventional hooked-wire electrodes were implanted in the orbicularis oris, superior longitudinal, and genioglossus muscles. The orbicularis oris muscle is largely responsible for closure of the lips, the superior longitudinal is active for tongue-tip elevation (for the production of /t/), and the genioglossus is a prime mover (protruding and bunching) of the tongue. EMG data from these muscles were recorded on magnetic tape and subsequently averaged using the Haskins Laboratories' EMG data system. The basic procedure was to collect EMG data for a number of tokens of a given utterance (in this experiment, between 10 and 15 repetitions), and to average the integrated EMG signals at each electrode position.

RESULTS

The effect of speaking rate on the production of the labial CV syllables is illustrated in Figure 1. This figure shows the averaged EMG curves of the orbicularis oris and genioglossus muscles for Subject FSC during the production of the utterance /ipip/ at both normal and fast speaking rates. The orbicularis oris curves, as shown here, are associated with lip closure for the consonants, while the genioglossus curves are associated with tongue movements for the vowels. "0" on the time axis represents the time of offset of voicing of the first vowel.

This figure shows that for fast speech, orbicularis oris activity increases, while geniglossus muscle activity decreases. These changes, which coincide with our earlier findings, are consistent and occur for each subject and all utterances.

Although the existence of these effects is consistent, the magnitude of the differences varies considerably, and the data do not show any clear preconsonantal or postconsonantal vowel effects. We suspect that these inconsistencies are caused, at least in part, by trade-offs between lip closing and jaw closing.

The decrease in genioglossus muscle activity for vowel production also occurs for each subject and for all utterances. These decreases imply that the undershoot observed for vowels during faster speech is programmed into the gesture and is not the result of a too fast succession of motor commands.

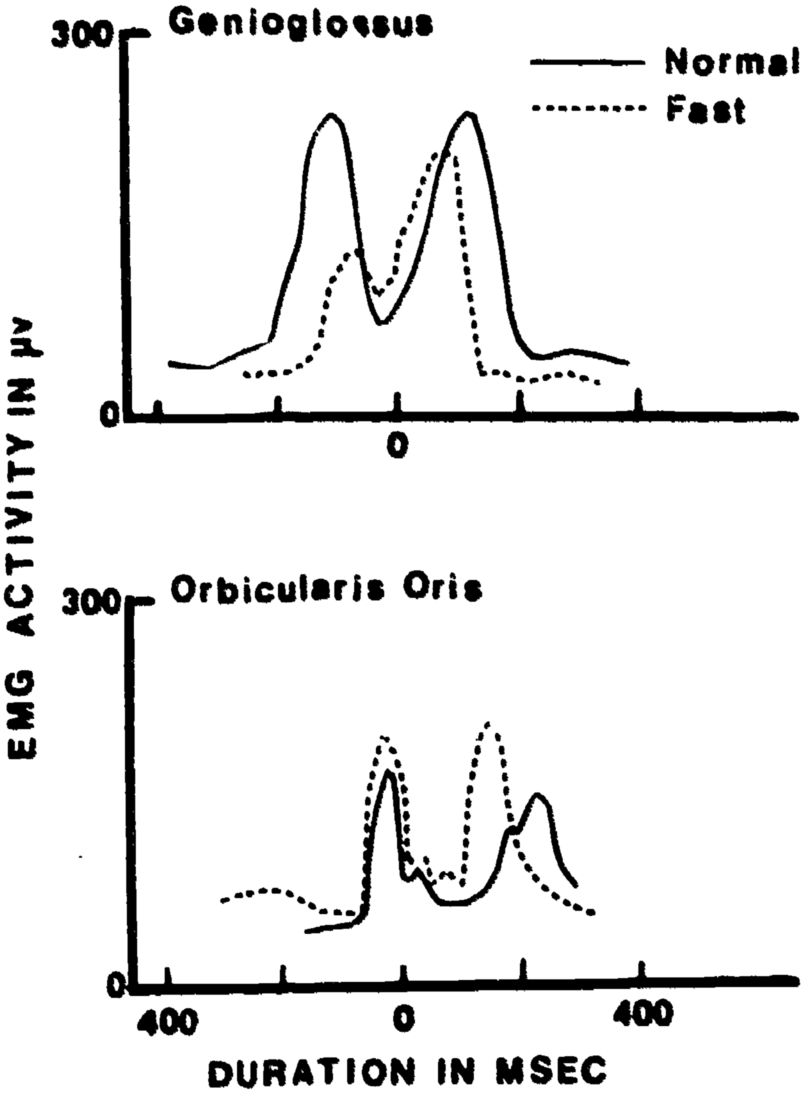


FIGURE 1

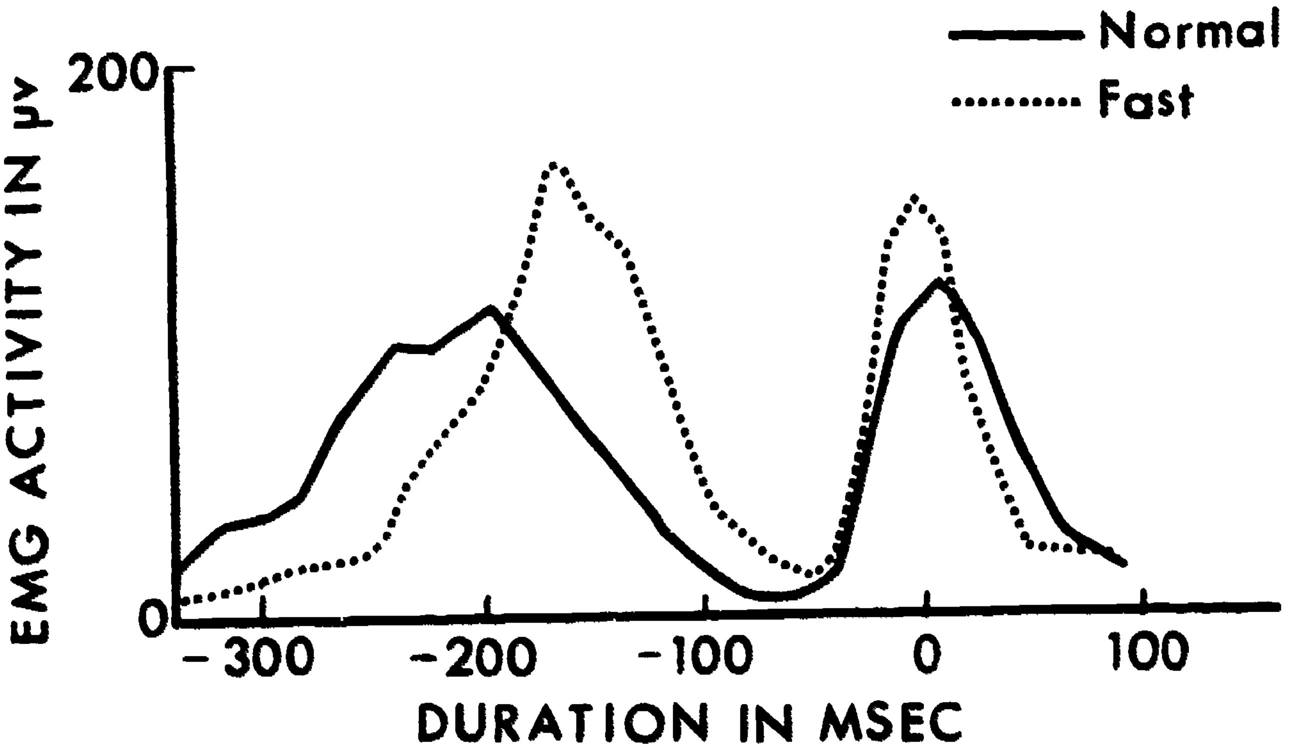


FIGURE 2

Our data also show that the lip rounding component of /u/ is likewise produced with greater levels of muscle activity during faster speech. This is illustrated in Figure 2, which shows orbicularis oris activity for the utterance /utap/ for Subject FSC.

Figure 3 illustrates the effect of speaking rate on the production of the lingual CV syllables. This figure shows the averaged EMG curves for the superior longitudinal and genioglossus muscles for the utterance /itip/, this time for Subject TG. These data show the same changes in muscle activity levels as the labial consonant data: an increase in speaking rate is accompanied by an increase in activity level for the consonant (superior longitudinal) and a decrease in activity level for the vowel (genioglossus). Again the same results occur for both subjects and all utterances, and the data do not show any consistent vowel effects. The different effects of an increase in speaking rate are especially interesting in this set of data because they demonstrate that different motor reorganization strategies can be used for different muscles of the same articulator.

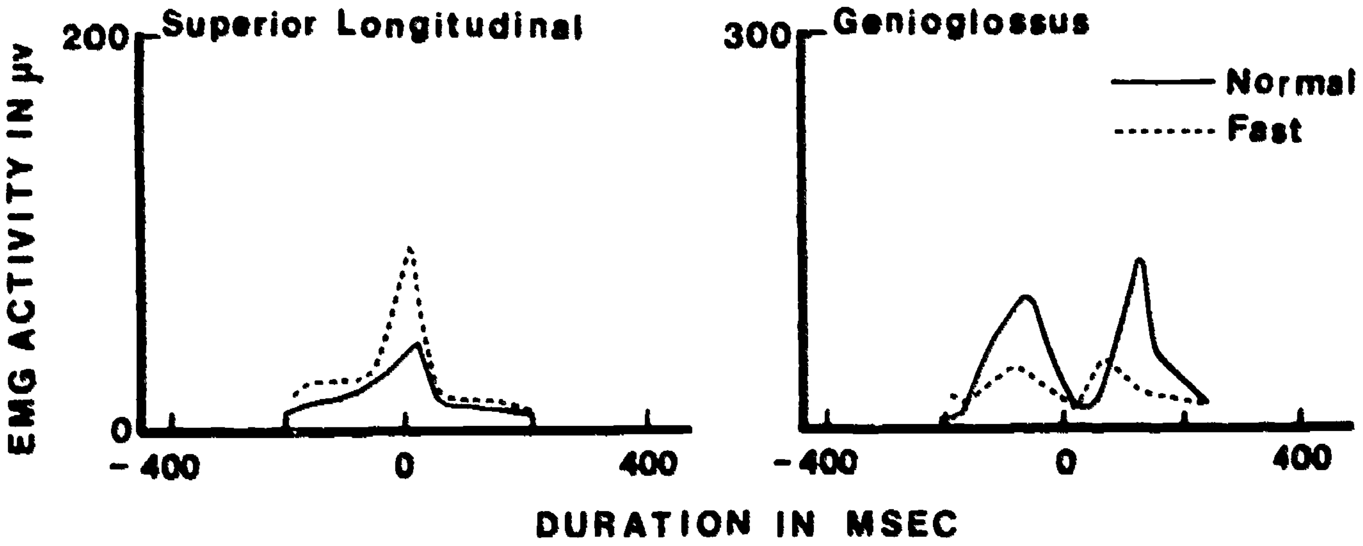


FIGURE 3

The results of this experiment can be summarized as follows. For lip movement associated with either labial consonant production or rounding for a vowel, and for tongue-tip movement associated with lingual consonant production, an increase in speaking rate is accompanied by an increase in the activity level of the muscle. For tongue movement during vowel production, in increase in speaking rate has the opposite effect: a decrease in the activity level of the muscle. The first finding implies an increase in articulatory effort and an increase in the speed of articulatory movement, while the second finding implies a decrease in articulatory effort, combined with a decrease in the speed of articulatory movement and/or a decrease in articulatory displacement.

DISCUSSION

The results for both the orbicularis oris and the superior longitudinal muscles can be explained quite readily: the production of both /p/ and /t/

requires a complete occlusion of the vocal tract; thus, under the constraints of an increase in speaking rate, the articulators move faster and with greater effort to produce that occlusion.

The data for the tongue, however, cannot be explained so straightforwardly. Obviously, the reduction in EMG activity for the vowel during faster speech is not compatible with an "extra effort" or even a "timing only" (equal-effort) control mechanism. Rather, the decrease in articulatory displacement usually associated witn fast speech is built into the planning of the gesture.

Our findings for vowel production also argue against the notion that a vowel target is internalized as a set of invariant spatial coordinates. If a vowel is organized in terms of an articulatory coordinate system, the system must be a multiple coordinate one, or one characterized by an articulatory field. Another view, however, and one that might better explain our data, would be that a vowel is internalized as a set of acoustic targets, and that the speech production mechanism uses any of a number of strategies to produce the required acoustic result. This view would also explain the differences in the tongue and lip data for /u/ during faster speech, i.e., a greater degree of lip rounding serves to compensate for the decrease in articulatory displacement of the tongue.

REFERENCES

Gay, T. and H. Hirose. (1973) Effect of speaking rate on labial consonant production. Phonetica 27, 44-56.

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

Jaw Movements During Speech: A Cinefluorographic Investigation*

T. Gay+
Haskins Laboratories, New Haven, Conn.

Jaw movement has been the subject of considerable interest in recent physiological speech research from two points of view: first, the jaw has a reputation of being largely responsible for many coarticulatory phenomena; and second, the jaw has been shown to be involved in the control of tongue height for certain vowels. The purpose of this paper is to examine in some detail the movements of the jaw during the production of speech that varies systematically in both phonetic context and speaking rate. The data reported here are part of a larger cinefluorographic study on the dynamics of both tongue and jaw movements during speech (Gay, in press).

Subjects were two adult males, FSC and TG, both native speakers of American English. The speech material consisted of the consonants /p/, /t/, and /k/ and the vowels /i/, /a/, and /u/ in a trisyllable nonsense word of the form /kipipə/, /kipapə/, /kipupə/, etc. The three consonants and vowels were thus arranged in all possible vowel-consonant-vowel (VCV) combinations. Each utterance was preceded by the carrier phrase "It's a....," and was produced at two speaking rates: normal and fast. Each rate was based on the subject's own appraisal of comfortable normal and fast rates. The subjects were instructed to speak the first two syllables of the utterance with equal stress, and the final syllable unstressed.

Lateral-view X-ray films were recorded with a 16 mm motion picture camera at a speed of 64 fps. The X-ray generator delivered 1 msec pulses to a 9 in image intensifier tube. The X-ray films were analyzed frame-by-frame, using a specialized film analyzer. The film was projected life size onto a writing surface using an overhead mirror system. Jaw movement was tracked in the vertical plane by measuring the vertical distance between the upper and lower central incisors. Measurements were made from the time of /k/ release to the time of closure for the final /p/.

The data of this experiment will be presented in the following order: jaw movement for the consonant, jaw movement for the vowel, and the effect of an increase in speaking rate on jaw movements for the entire utterance.

*Text of a paper presented at the Eighth International Congress on Acoustics, London, July 1974.

+Also University of Connecticut Health Center, Farmington.

Figure 1 shows the jaw displacement measurements, in mm, for the utterances /apa/, /ata/, and /aka/ for both subjects. "C" on the abcissa represents the time of closure for the consonant. The data in these graphs summarize the extent of differences in jaw movement for the three consonants. First, for both subjects, jaw closing is greatest for /t/ and least for /p/. The magnitude of these differences, however, varies considerably for the two subjects. Whereas the range (at time 0) from /t/ to /p/ spans approximately 8 mm for FSC, the range is only 3 mm for TG. It should be noted, moreover, that for both subjects, the range of displacement differences decreases when the vowel is either /i/ or /u/, while /a/ shows by far the greatest effects.

Jaw displacement for both /t/ and /k/ are relatively insensitive to changes in the preceding and following vowels. However, jaw displacement for /p/ shows both anticipatory and carryover coarticulation effects of the adjacent vowels; that is, jaw displacement for /p/ is greater when the preceding or following vowel is open (Sussman, MacNeilage, and Hanson, 1973). These differences are as great as 5 mm.

Figure 1 also shows individual differences in consonant effects on jaw displacement for the adjacent vowels. Subject FSC shows rather large differences (8-9 mm) in degree of displacement for the vowel, while TG shows essentially none. Again the differences here for FSC are considerably less for /i/ and virtually absent for /u/, and /a/ is the only vowel that shows consistent coarticulation effects!

Apparently, jaw displacement for the open vowel is greatest when the consonant is /p/ because the jaw is least involved in the production of this consonant. Both /t/ and /k/, on the other hand, are characterized by greater degrees of jaw closing. This probably acts to constrain the degree of opening for the following vowels.

The differences in jaw displacement for /a/ as illustrated for FSC are clearly related to differences in tongue height for the vowel. Figure 2 shows the same jaw measurements as before, this time plotted along with measurements of tongue height. The tongue height measurements represent the relative positions of a 2.5 mm lead pellet attached to the surface of the tongue at a distance of approximately 1 1/2 in from the tip. The measurements were plotted from a fixed coordinate system that used various anatomical features as landmarks.

While both sets of data for TG show identical movement patterns, the tongue measurements for FSC clearly shadow those for the jaw. Indeed, the numerical differences for the three vowel curves are similar for the graphs of both the tongue height and the jaw opening.

We now consider how jaw displacement is affected by the first and second vowels. Figure 3 illustrates the effect across the consonant of the second vowel on the displacement of the jaw for the first vowel. Both the first vowel and consonant are the same; only the second vowel is different. For both subjects, jaw displacement for the first vowel is not at all sensitive to changes in the second vowel (differences in maximum displacement never exceed 1 to 2 mm). The absence of any consistent anticipatory coarticulation effects holds up as well forall the other consonant and vowel sequences. Although jaw displacement for the first vowel in the VCV utterance is not sensitive to any right-to-left, or anticipatory, effects beyond the consonant, jaw displacement for the second

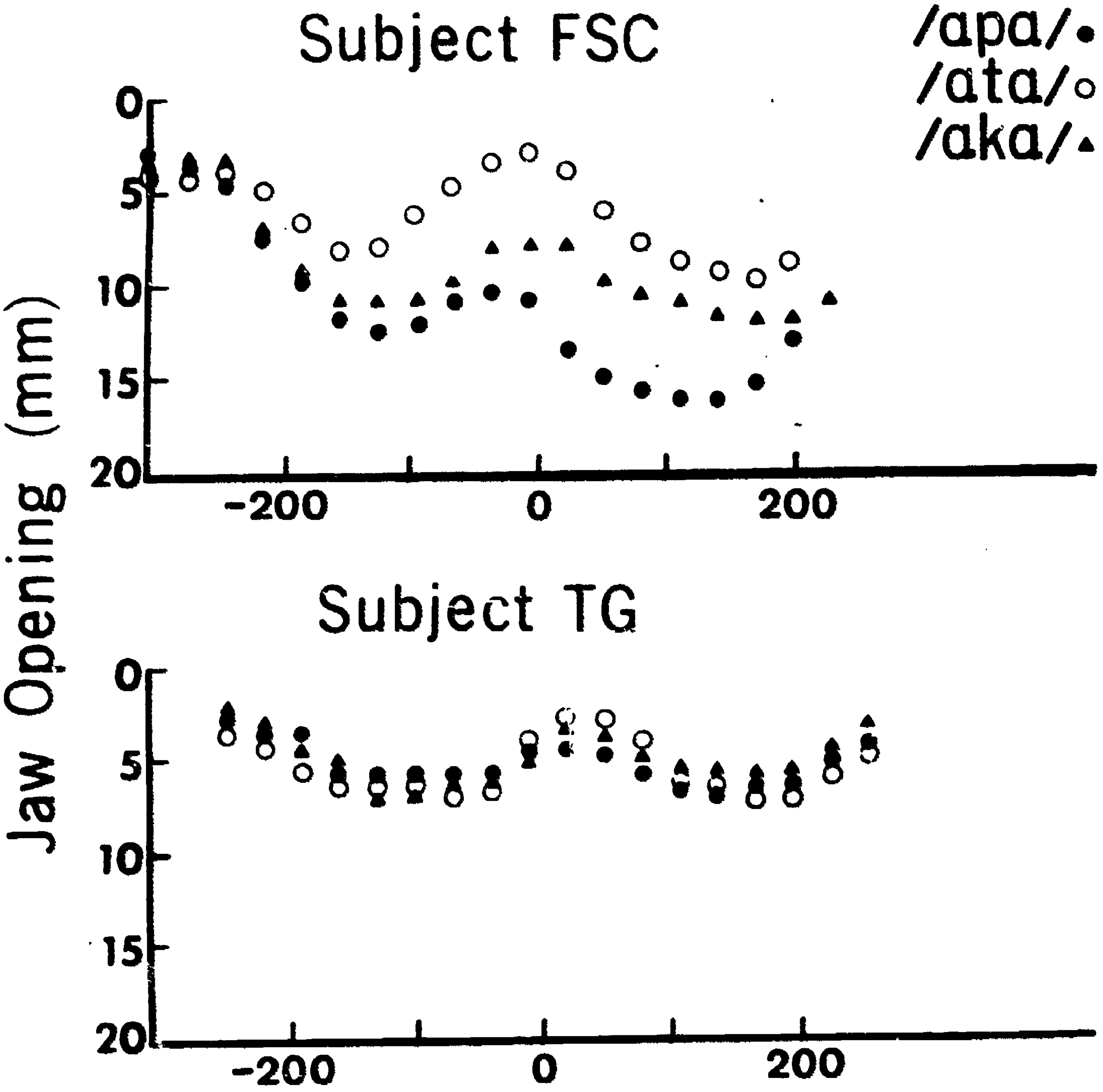

Subject FSC
/apa/●
/ata/○
/aka/▲
0
5
10
15
20
-200
0
200
Jaw Opening (mm)
Subject TG
0
5
10
15
20
-200
0
200
Duration (msec)

FIGURE 1

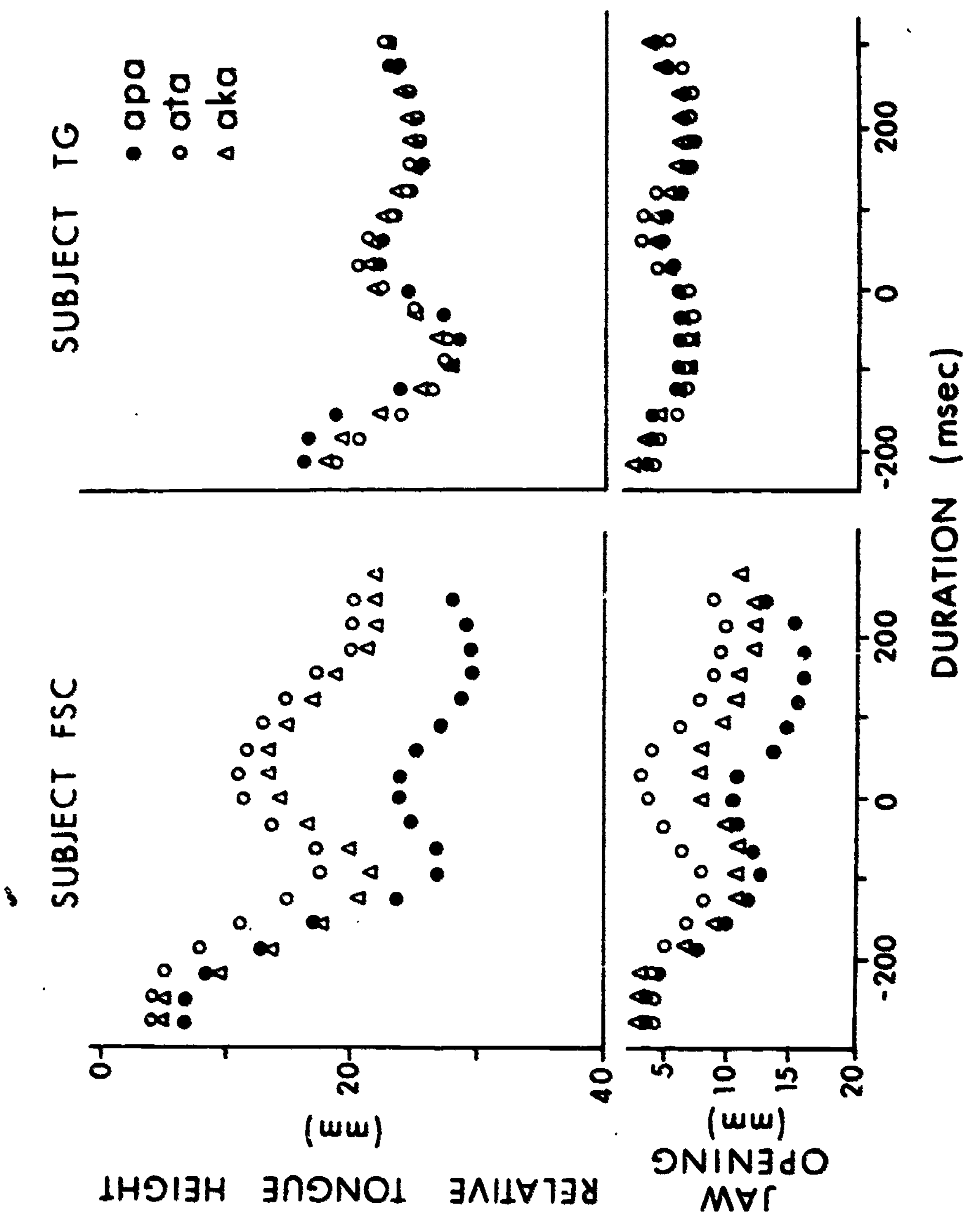

SUBJECT FSC
SUBJECT TG
● apa
○ ata
△ aka
RELATIVE TONGUE HEIGHT
(mm)
0
20
40
JAW OPENING
(mm)
5
10
15
20
-200
0
200
DURATION (msec)

FIGURE 2

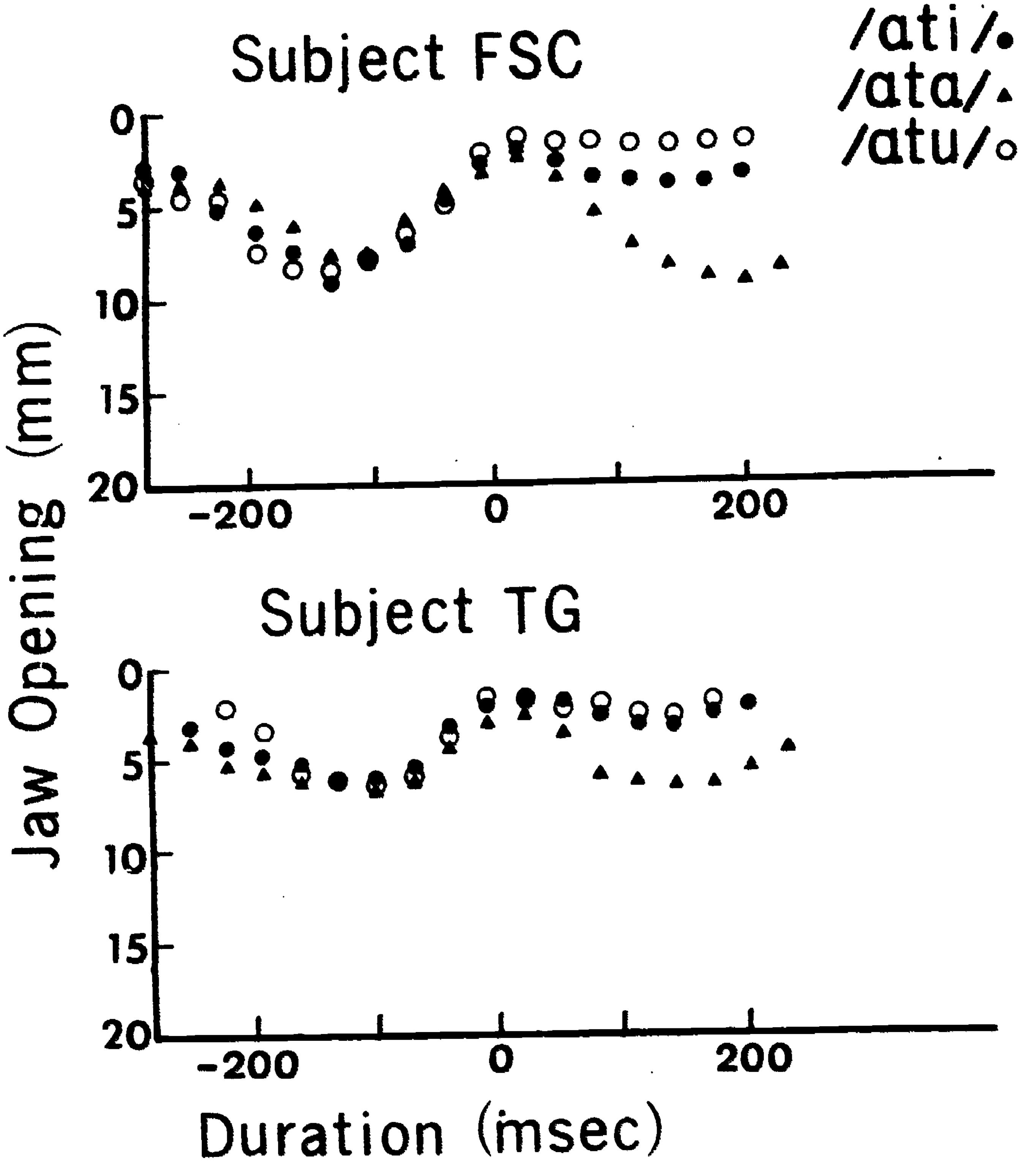

Subject FSC
/ati/ ●
/ata/ ▲
/atu/ ○
0
5
10
15
20
-200
0
200
Jaw Opening (mm)
Subject TG
0
5
10
15
20
-200
0
200
Duration (msec)

FIGURE 3

vowel is subject to some left-to-right, or carryover, vowel effects; these effects, however, are fairly complicated and are linked to the consonant.

When the consonant is /p/, the first vowel has no real effect on jaw displacement for the second vowel. Jaw displacement for all three vowels remains quite stabile. However, when the consonant is either /t/ or /k/, carryover effects appear. Although jaw displacement for /i/ and /u/ in the second vowel position remains stable, the first vowel exerts a strong influence on the displacement of the jaw for /a/, this time for both subjects. This is illustrated for the /t/ series of utterances in Figure 4.

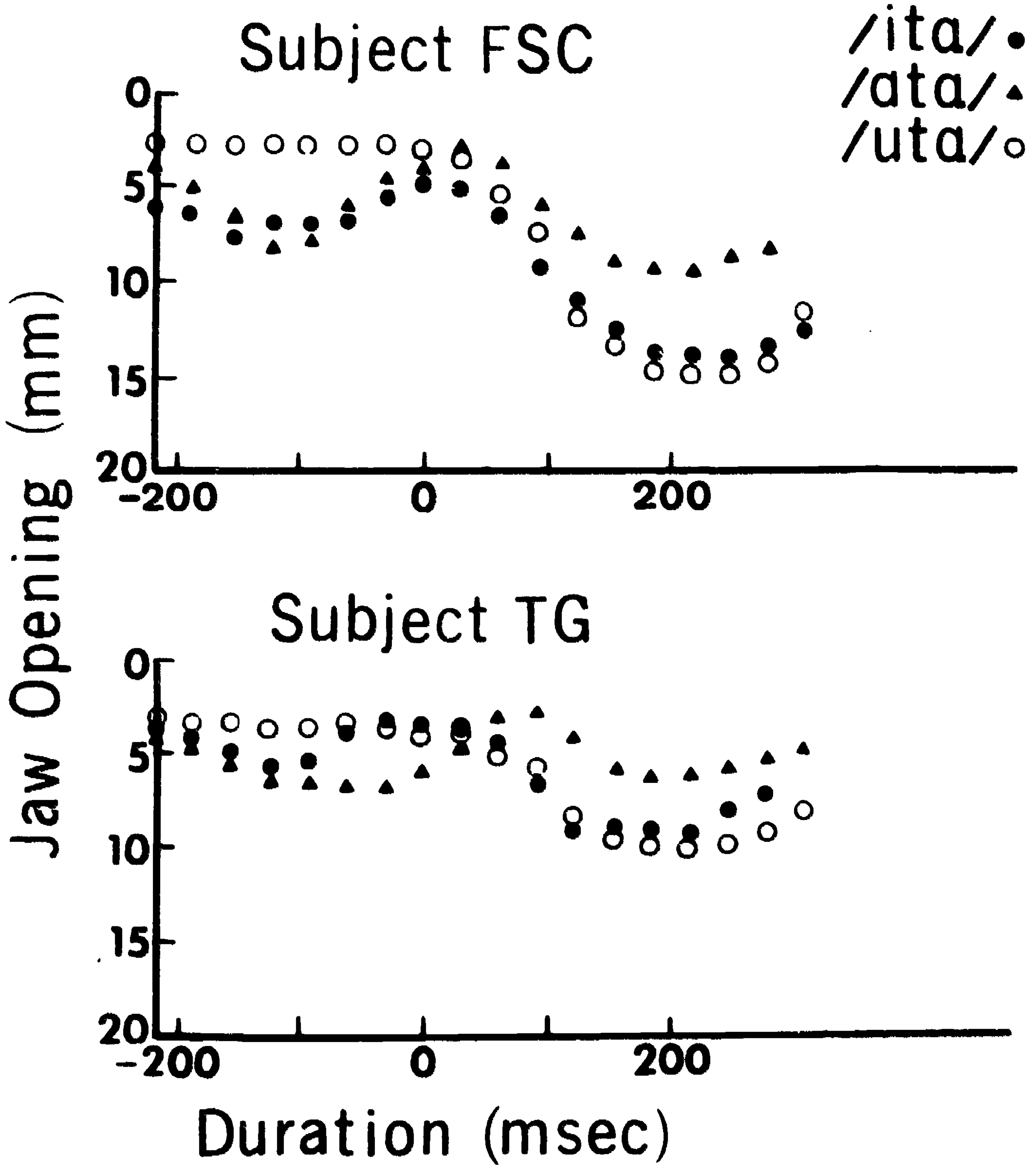
This figure shows the jaw displacement curves for the utterances /ita/, /ata/, and /uta/ for both subjects. These graphs show that there is less jaw opening for /a/ in the second position when the first vowel is /a/ than when the first vowel is either /i/ or /u/. At first glance these effects are somewhat surprising. It would seem more likely, at least intuitively, that greater degrees of opening for the second vowel would be associated with a more open, rather than a more close first vowel. However, closer inspection of these graphs can explain the effects. At about the time of closure for the consonant (0 on the abcissa), the jaw is in approximately the same position for each of the three different first vowels. At that point, however, the jaw is closing toward minimum opening from /a/ while it is already beginning to open for the second vowel from both /i/ and /u/. Thus, the jaw is moving in different directions at that point, and, in effect, has a head start towards the second vowel when the first vowel is close.

Figure 4 illustrates another aspect of jaw opening for the vowels: /u/ is the only vowel characterized by a closed jaw position. Both /a/ and /i/ are characterized by a more open jaw position--/a/ for obvious reasons, and /i/ probably to make room for the bunching of the tongue. This finding argues against an articulatory feature system that includes both the tongue and jaw in the description "close." Apparently the tongue and jaw can move independently and even in different directions at the same time.

The differences in jaw opening for the open vowel are again related to differences in tongue height for that vowel. This is illustrated in Figure 5. which shows the previous jaw measurements plotted along with the measurements for tongue height. As can be readily seen, the two sets of data closely follow each other for /a/.[1] The extent to which jaw opening controls tongue height for the open vowel can be seen even more clearly in Figure 6.

This figure shows the two sets of measurements of the previous figure plotted against each other, that is, the jaw opening measurements were subtracted from the tongue height measurements to obtain the net movement curves for the tongue. These data show two things: first, that the three tongue curves follow essentially the same paths, clearly demonstrating that the differences in tongue height were related solely to differences in jaw opening, and second, displacement for the open vowel /a/ is controlled primarily by the jaw. This is shown by the relative flatness of the curves once the tongue moves down to the floor of the mouth from the preceding phone. These data, of course, support the model proposed by Lindblom and Sundberg (1971).

[1] Note, however, the differences in tongue and jaw movement for /i/--each seems to move independently of the other.



Subject FSC
/ita/ ●
/ata/ ▲
/uta/ ○
0
5
10
15
20
-200
0
200
Subject TG
0
5
10
15
20
-200
0
200
Jaw Opening (mm)
Duration (msec)

FIGURE 4

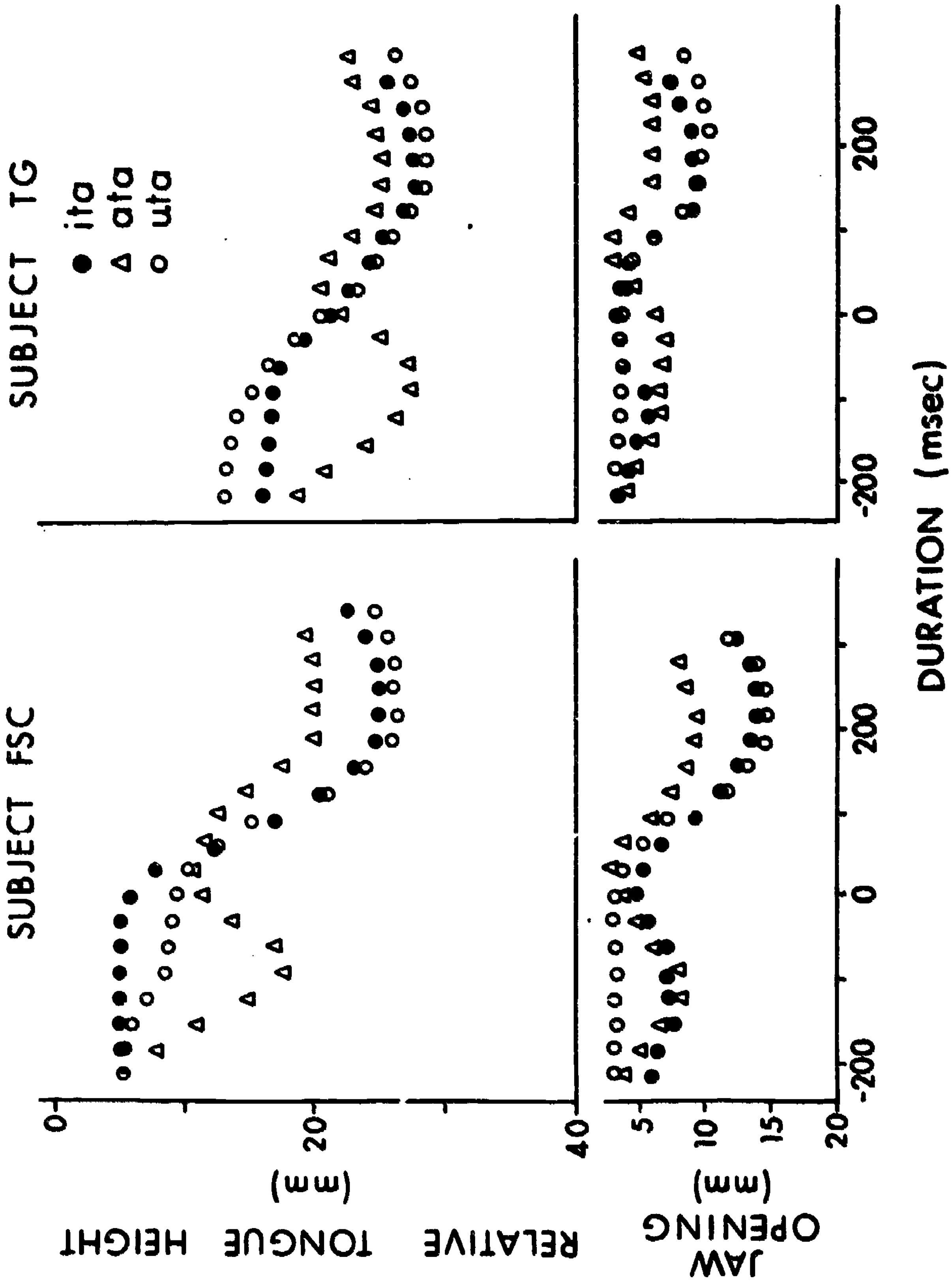

SUBJECT FSC
SUBJECT TG
● ita
△ ata
○ uta
RELATIVE TONGUE HEIGHT (mm)
0
20
40
JAW OPENING (mm)
5
10
15
20
-200
0
200
DURATION (msec)

FIGURE 5

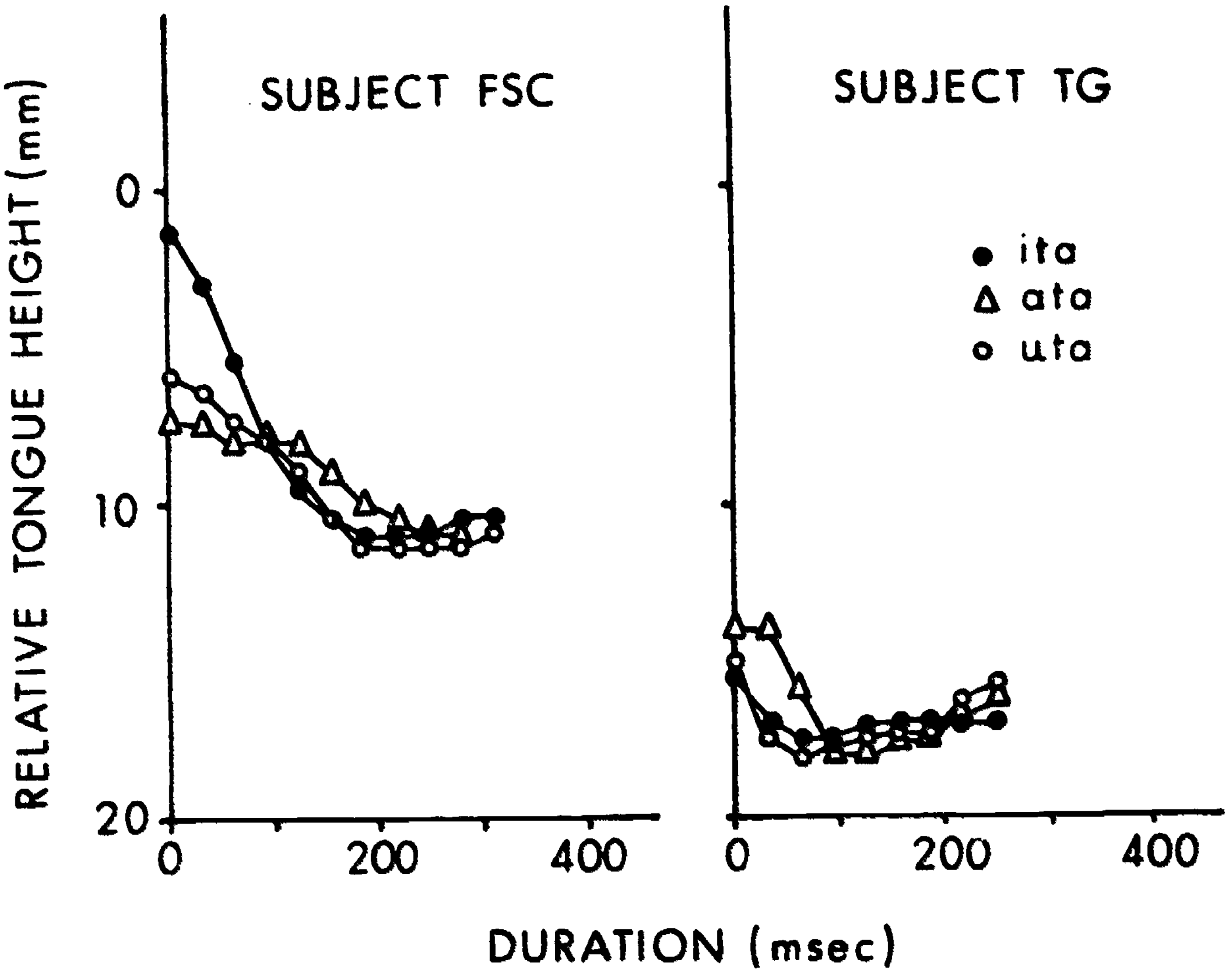

SUBJECT FSC
SUBJECT TG
RELATIVE TONGUE HEIGHT (mm)
0
10
20
0
200
400
0
200
400
DURATION (msec)
● ita
Δ ata
○ uta

FIGURE 6

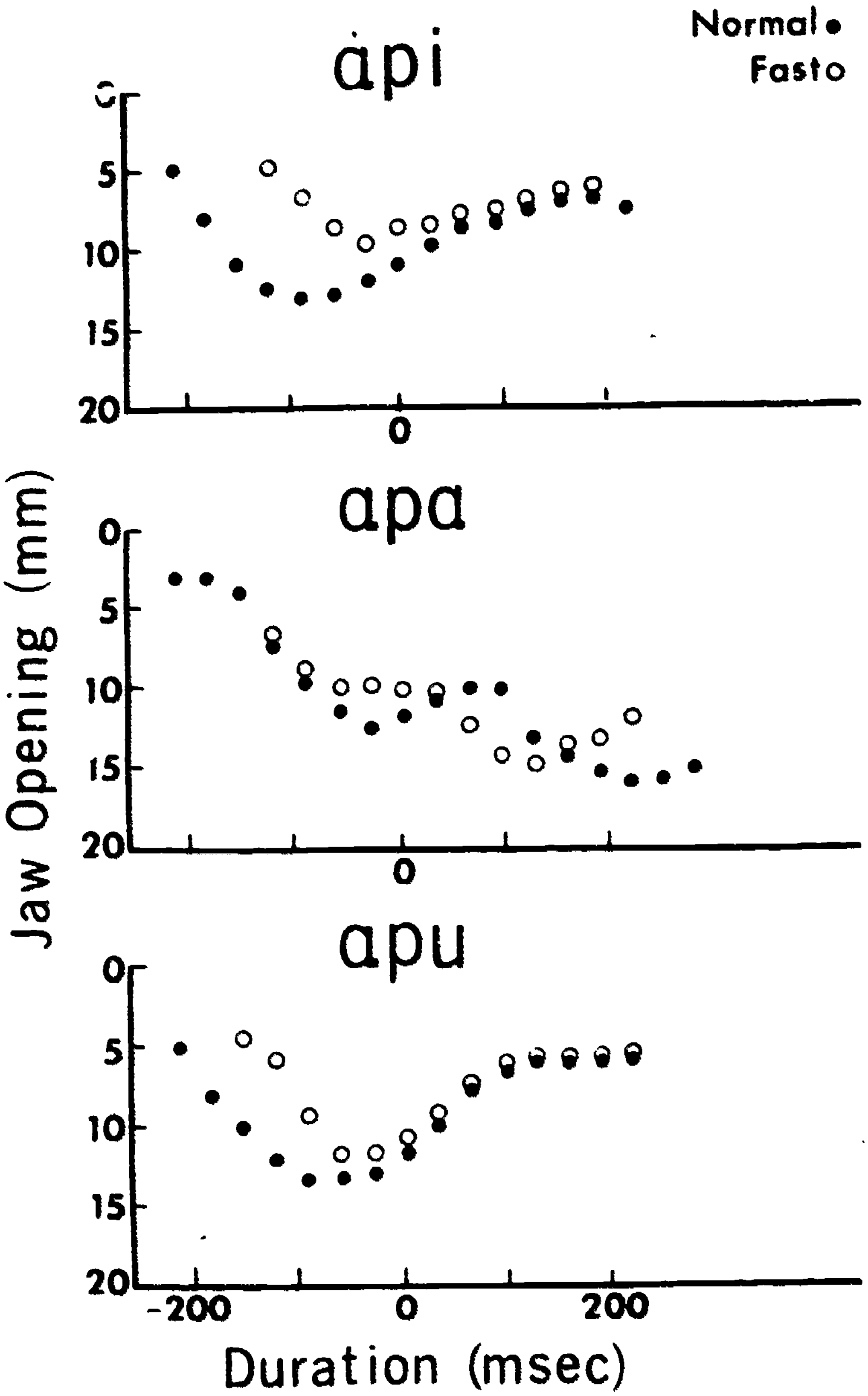

api
Normal ●
Fast ○
0
5
10
15
20
0
apa
0
5
10
15
20
0
apu
0
5
10
15
20
-200
0
200
Jaw Opening (mm)
Duration (msec)

FIGURE 7

The effect of an increase in speaking rate on the movement of the jaw is illustrated in Figure 7, which shows the jaw displacement measures of the utterances /api/, /apa/, and /apu/ for both the slow and fast speaking rates for FSC. Generally speaking, an increase in speaking rate results in a decrease in jaw displacement for the entire utterance. In other words, jaw movement during fast speech mirrors jaw movement during slow speech but from a more closed position. Also, the context effects that appeared for /a/ during slow speech were generally absent during fast speech. This is probably due to the more restricted path of movement of the jaw during fast speech.

The results of this experiment can be summarized as follows.

Jaw movement towards and jaw position for a stop consonant are different depending upon the consonant. However, the magnitude of these differences varies with the subject. Indeed, the individual differences are striking. Vowel effects on jaw displacement for the consonant occur only for /p/, apparently because the jaw is least involved in the production of /p/; /t/ and /k/ are both characterized by a more closed jaw position and are probably more resistant to coarticulatory influences.

Although a considerable amount of time was spent describing and illustrating the variability of jaw displacement for vowels, such variability is really the exception rather than the rule. Variability of jaw displacement occurred only for /a/, and then only for certain contexts. Jaw movement and displacement for /i/ and /u/ were quite stable in all environments! This would indicate that the acoustic properties of /a/ are either more insensitive to articulatory variability than those of /i/ and /u/ or that the acoustic field of /a/ is somewhat larger. The data of this study also support the model proposed by Lindblom and Sundberg (1971), that tongue height for an open vowel is controlled by the degree of jaw opening for that vowel, but the data argue against a feature system that proposes a one-to-one relationship between tongue height and jaw opening for all vowels.

The effect of an increase in speaking rate was reflected by an overall decrease in jaw displacement for the vowel, a decrease that absorbed virtually all the context-dependent coarticulation effects present during slow speech.

Finally, the major conclusions that can be drawn from the data of this experiment are that the jaw is, indeed, sensitive to certain anticipatory and carryover coarticulation effects, but only to a limited degree; individual differences in jaw movement are real and often large; and the jaw is, in a real sense, a primary articulator, controlling tongue height for an open vowel.

REFERENCES

Gay, T. (in press) A cinefluorographic study of vowel production. J. Phonetics. [Also in Haskins Laboratories Status Report on Speech Research SR-39/40 (this issue).]

Lindblom, B. and J. Sundberg. (1971) Acoustical consequences of lip, tongue, jaw, and larynx movements. J. Acoust. Soc. Amer. 50, 1166-1179.

Sussman, H., P. MacNeilage, and R. Hanson. (1973) Labial and mandibular dynamics during the production of bilabial consonants: Priliminary observations. J. Speech Hearing Res. 16, 397-420.

A Preliminary Electromyographic Study of Labial and Laryngeal Muscles in Danish Stop Consonant Production

Eli Fischer-Jørgensen[+] and Hajime Hirose[++]

INTRODUCTION

Danish has six stop consonants: *p*,*t*,*k*,*b*,*d*,*g*. Normally, *ptk* are distinguished from *bdg* only in syllable-initial position followed by a full vowel (with or without intervening *l* or *r*), e.g., *pile*/*bile* ['p^{h}i:lə, 'bi:lə], *klat*/*glat* [klad, glad], *betale*/*pedal* [be't^{h}a:ʔlə, p^{h}e'da:ʔl]. Medially before shwa only [bdg] are found (irrespectively of the orthography), and in final position [ptk] and [bdg] vary freely.[1]

In the positions where *ptk* and *bdg* are phonologically distinct, *ptk* are strongly aspirated with a voice-onset time (VOT) value of 60-80 msec in more old-fashioned speech (in modern Copenhagen speech somewhat more), and *t* is moreover affricated, whereas *bdg* are unaspirated (see Fischer-Jørgensen, 1954). Both categories are voiceless, and there is no evidence that *ptk* should be more tense than *bdg* in the narrower sense of having stronger articulatory activity. There is rather some evidence to the contrary--that *bdg* has a longer closure than *ptk* and a tendency to higher mechanical pressure at the place of articulation. The difference in duration of the closure is small (normally 20-30 msec for *b*/*p* and 40-50 msec for *d*/*t*), but it is stable and statistically significant (Fischer-Jørgensen, 1954; Frøkjær-Jensen, Ludvigsen, and Rischel, 1971). The mechanical pressure is rather variable, and the difference is significant only for some subjects. A questionnaire concerning the kinesthetic judgment of effort in stop consonants has shown that 67 percent of the speakers felt the

[+] Institute of Phonetics, University of Copenhagen.

[++] Faculty of Medicine, University of Tokyo; and visiting researcher, Haskins Laboratories, New Haven, Conn.

[1] In old-fashioned speech a distinction is sometimes made before [rə] (e.g., *kæntre*, *ændre*). The distributional relation among [ptk], [bdg], and [voγ] raises various problems for the phonological analysis which may be disregarded here. It should, however, be emphasized that we are not describing the phonetic difference between the phonemes *ptk* and *bdg*, but the difference between the "microphonemes" in syllable-initial position.

Acknowledgment: We are grateful to the staff of Haskins Laboratories for their hospitality and the use of their excellent equipment. We are also grateful to the students who undertook part of the measurements (Jentε Andresen, John Jørgensen, and Peter Molbæk Hansen). Finally, we thank our subjects who were willing to suffer for science.

[HASKINS LABORATORIES: Stutus Report on Speech Research SR-39/40 (1974)]

organic pressure to be greater in bdg and only 22 percent found it to be greater in ptk, whereas 10 percent did not feel any difference (Fischer-Jørgensen, 1972). The intraoral air pressure of ptk is about 5 percent higher than that of bdg, but the difference is found only at the end of the closure. These relations between ptk and bdg differ from what is generally thought to be characteristic of the two categories of stops.[2] It therefore seemed interesting to undertake an electromyographic (EMG) investigation of the labial muscles for p and b.

Since the main difference between Danish ptk and bdg is one of aspiration, the condition of the glottis should be of primary importance. A glottographic examination has shown that Danish ptk have a wide-open glottis with the maximum aperture close to the point of release, whereas bdg have a smaller aperture with the maximum near the beginning of the closure and practically zero aperture at the release (see Frøkjær-Jensen, 1967, 1968; and particularly, Frøkjær-Jensen et al., 1971). These findings seem to be confirmed by a preliminary fiberoptic investigation undertaken by Jørgen Rischel. Whereas the wide aperture of the glottis in ptk must be due to the neural command, the smaller aperture in bdg might (according to a hypothesis advanced by Frøkjær-Jensen et al., 1971) be due to aerodynamic forces only. In order to test this hypothesis, we made an EMG investigation of the laryngeal muscles.

EXPERIMENTAL METHOD

Electrode Insertion Technique

In the present study, hooked-wire electrodes were used. Insertion into the labial muscles was made approximately at the positions described by Leanderson (1972). During electrode insertion an oscilloscope and an amplifier system were used for monitoring the pertinent muscle activity. Insertion into the orbicularis oris superior (OOS) and inferior (OOI) was made at the vermillion border of the upper lip and the lower lip about 1 cm laterally to the midline. If EMG activity was found in OOS or OOI for protrusion of the lips or for production of labial stops, the placement was considered to be correct.

The depressor anguli oris (DAO) was reached at the point 1.5-2 cm below the angle of the mouth. Insertion into the depressor labii inferior (DLI) was made about 2-2.5 cm below the point of insertion into OOI. To verify the correct placement for these muscles, the subject was asked to pull the angle of his mouth or his lower lip downwards. One cannot always be certain in differentiating between DAO and DLI since there is a possible overlapping of the fibers of these muscles. Anatomically, however, DAO is known to be distributed more superficially than DLI. Therefore, the depth of insertion was controlled to avoid interference.

In order to reach the mentalis (MENT), the needle was inserted in the midline deeply enough to feel the bony surface of the mandible. The placement was

[2] Weaker mechanical pressure in aspirated stops was, however, found as early as 1897 by Rousselot (p. 596), and the same was found for Gujarati (Fischer-Jørgensen, 1968a:96). The measurements of organic pressure and intraoral air pressure in Danish stops have not been published in detail accept for a bilingual subject (Fischer-Jørgensen, 1968b).

considered to be correct if EMG activity occurred when the subject attempted to move the skin over the chin upwards. The technique of insertion into the laryngeal muscles was essentially the same as that described by Hirose (1971a; Hirose, Gay, and Strome, 1971). The interarytenoid (INT) and the posterior cricoarytenoid (PCA) were reached perorally, using an L-shaped probe under indirect laryngoscopy. A percutaneous approach was employed for insertion into the thyroarytenoid (VOC), the lateral cricoarytenoid (LCA), and the cricothyroid (CT).

Data Recording and Processing

EMG signals were recorded on a multichannel data recorder simultaneously with acoustic signals and automatic timing markers. The signals were then reproduced and fed into a computer after appropriate rectification and integration. The EMG signal from each electrode pair was averaged for each utterance type with reference to a line-up point on the time axis representing a predetermined speech event. In this experiment the line-up point was always at the end of the frame ([han sa:] "he said"), i.e., at the implosion of the following consonant. The data-recording and computer-processing systems used in the present experiment are described in more detail by Port (1971) and Kewley-Port (1973).

Subjects and Material

Recordings were made of six subjects: PM, EG, EFJ, TB, PH, and HA, who all speak Standard Danish, although with slight local differences. PM, PH, and TB are from Copenhagen, EG from Jutland, EFJ (one of the present authors) grew up in Funen, but has never spoken Funish dialect. HA is from Copenhagen, but has been in America since the age of 18; his Danish seems unspoiled. TB and HA have relatively long aspirations of ptk. The age of the subjects (in 1972) was 24-39 years, except for EFJ (61).

Not all subjects were able to tolerate peroral insertion into the laryngeal muscles or too many insertions in the labial area, and the material is therefore rather heterogeneous. Table 1 gives a survey of the muscles examined for the different subjects. Parentheses indicate that the interpretation of the curves is dubious. TB has much noise in his OOI, and PH's OOS was completely unusable because of noise and has been left out altogether.

The linguistic material used consisted of real Danish words spoken in the frame han sagde [han sa:] "he said." The words are listed in Table 2 in systematic groupings. The initials of the subjects who spoke the words in question are given in parentheses for each section of the list. The words in 1a and 2a were spoken by all subjects. The words in 1b were not spoken by HA and the words in 2b were not spoken by EG and EFJ. Group 3 was intended for an examination of the Danish 'stød' (see Fischer-Jørgensen and Hirose, 1974), but as it contains some words with initial p and m, it is also partly relevant for the examination of the consonants. List 3a was read by PM, TB, PH, and EFJ; 3b only by EFJ.

The words were presented in a list containing four different randomizations. The list was read four times by each subject. There are thus 16 examples of each word (but some had to be eliminated because of hesitations, artifacts, measurement error in editing the raw data, etc., at the time of computer processing). In the list read by EFJ 2a occurred only once, so that there were only four examples of each word. The consonant list was, however, spoken twice by

TABLE 1: Recorded muscles.

Subjects:		PM I	 II	EG	EFJ	TB	PH	HA
Labial muscles:	OOS	x	x	x	x	x		x
	OOI	x	x	x	x	(x)		
	DLI	x	x	x		x	x	x
	DAO	x		x				
	MENT	x	x	x	x			
	mid DLI-DAO	x		x				
Laryngeal muscles:	PCA				x			
	INT				x			x
	VOC		x			x	x	
	LCA		x					

TABLE 2: Words read by the subjects. (The words are given in phonetic transcription.)

1a. 'p^{h}anə 'banə 'p^{h}i:lə 'bi:lə 'p^{h}u:ə 'bu:ə p^{h}e'da:ʔl be't^{h}a:ʔlə
(PM, EG, EFJ, TB, PH, HA)

1b. p^{h}a'gaiʔ ba'k^{h}anʔt p^{h}u'rist bu'dist (PM, EG, EFJ, PH, TB)

2a. 't^{h}anə 'danə 'k^{h}alə 'galə (PM, EG, EFJ, PH, TB, HA)

2b. 't^{h}i:ə 'di:ə 'k^{h}ilə 'gilə t^{h}u:ə du:ə k^{h}u:lə gu:lə
(PM, TB, PH, HA)

2c. salə falə (HA)

3a. 'lɛ:sʁ 'lɛ:ʔsʁ 'p^{h}i:bʁ 'p^{h}i:ʔbʁ man manʔ (PM, TB, PH, EFJ)

3b. 'lɛsʁ 'p^{h}ibʁ ma:ʔn manʔn̩ 'k^{h}ɛ:lə 'k^{h}ɛ:ʔlʁ k^{h}ɛlʁ 'k^{h}ɛlʔʁ 'hu:ən
'hu:ʔən (EFJ)

EFJ in the same session. It was also spoken twice by PM, in two different sessions called I and II.

RESULTS AND DISCUSSION

Labial Muscles

General activity of labial muscles for labial stops. First we may ask: (a) which of the investigated muscles were found to be active for labial stops, (b) whether they are active for the closing or the opening movement, and (c) whether they are active for other sounds found in the material as well. There

TABLE 3: Survey of activity of labial muscles.

Muscle	Subjects (number) initials	closing of lips (peak at line-up point)	opening of lips (peak 100-150 msec later)	start of vowel after t,d,k,g (peak 100-200 msec after line-up point)	rounding of vowel
OOS	(5) PM, EG, EFJ, TB, HA	+: PM, EG, EFJ, TB, HA	+: EFJ (i) Ø: PM, EG, TB, HA	Ø	+: EG (PM) Ø: EFJ, TB, HA
OOI	(5) PM, EG, EFJ, PH (TB)	+: PM, EG, EFJ, TB Ø: PH	+: EG, PH (TB?) (a,i,u) +: PM (u) +: EFJ (i,u)	+: PM, EG, PH, TB (u) Ø: PM, EG, PH, EFJ (a,i); TB (t + a,i)	+: PM, EG, EFJ (PH) (TB?)
DLI	(5) PM, EG, TB, PH, HA	+: (PM) (a,i,u) +: EG (HA) (a,i) Ø: PH, TB	+: PM, EG, TB, HA (a,i) +: PH (a,i,u)	+: PM, EG, TB (PH, HA) (a,i) Ø: (u)	Ø
DAO	(2) PM, EG	+: PM, EG	Ø	Ø	Ø
MENT	(3) PM, EG, EFJ	+: PM (EG) Ø: EFJ	+: (PM) (a,i) +: EFJ (60 msec) Ø: EG	Ø	+: EFJ Ø: PM, EG

are some individual differences on this point, which may in some cases be due to the placement of the electrodes.

The results are summarized in Table 3. The symbol + indicates activity of the muscle, ∅ means "no activity." Capital letters indicate subjects and parentheses around a letter indicate that the activity is weak for this subject. The letters a, i, u indicate the vowels following the stop consonant (i includes the e of [p^{h}e'da:ʔl] and [be't^{h}a:ʔlə]). Vowels are given only in the cases where the activity is different before different vowels.

The activity for the closing movement has its peak almost at the line-up point, i.e., where the vowel [a:] of the frame ends and the implosion of the consonant takes place. It starts around 100 msec earlier and ends around 100-200 msec later than the line-up, but, particularly for OOS, the peak is generally very pronounced. OOS is active at the implosion for all subjects, OOI for all but one (PH, who, however, has a small peak in the preceding vowel). One of the subjects (EG) also has a definite activity of DLI; two others (PM and HA) have a small peak for DLI at the implosion. Both of the subjects for whom DAO was recorded show a definite peak at the implosion. The same is true for an electrode point between DAO and DLI (mid DLI-DAO in Table 1) which is not included in Table 3. MENT is clearly active for one of the three subjects and has a weak activity for one of them. This means that for the closing movement there is strong activity in OOS and OOI (with the exception of subject PH for OOI), activity in DAO for both subjects, and clear activity for one of three subjects in MENT.

The activity preceding the release (approximately 100 msec later) is quite different. Practically nobody has any activity in OOS (except for EFJ who has a small peak before i), whereas all have activity in OOI and DLI but with a characteristic distribution depending on the following vowel (see Table 4). All have activity in OOI before u and in DLI before a, i;[3] EG also has activity of OOI

TABLE 4: Activity of OOI and DLI at the release.

	OOI		DLI	
	u	a,i	u	a,i
PM	+	∅	∅	+
EG	+	(+)	∅	+
EFJ	+	∅ (+)		
TB	+?	+?	∅	+
PH	+	+	+	+
HA			∅	+

[3] Visual inspection of the lips also reveals that they are pressed somewhat forward at the release of pu, whereas the lower lip goes down in pa.

before a and i, and EFJ before i (but weaker). Only PH has no clear difference between the two muscles (except that DLI is slightly weaker before u).

This division of labor is confirmed by an examination of the activity of the two muscles at the release of consonants other than the labial ones. All subjects have activity in OOI before and during the vowel u, and all have activity in DLI when the consonant is followed by a or i. The activity of DLI is of somewhat shorter duration than the activity of OOI for u, but the peak is generally not quite as sharp as that of OOS for the closing movement. All subjects have a stronger activity of DLI after p than after t and k, and most of them (but not PM) have a tendency to stronger activity before a than before i; this means that the activity is influenced by the size of the opening movement required for the vowel. We have thus not found a definite activity of a definite muscle intended to open the lips in labial consonants. The command seems to aim at the following vowel, and the muscle that is appropriate for producing the vowel is activated. This does not, however, exclude the possibility that an investigation of other facial muscles might disclose an activity aiming specifically at the release of a labial closure.

The difference between DLI and OOI according to the following vowel also shows up in the vowel itself. It is simply the same thing. All use OOI for rounding. Even PH, who did not distinguish OOI and DLI after labial consonants, has a clear difference after tdkg. TB's OOI curve is rather noisy and difficult to interpret; he has more activity in u than in a and i after tdkg (and some activity after t in all cases), but after labial consonants there is no clear difference. Some subjects (EG and PM) also use OOS for rounding, i.e., they have a longer duration of the OOS activity for pu than for pa, but PM has hardly any activity of OOS for rounded vowels after t and k. DAO and DLI show no activity at all for rounding in the present series of experiment. The dominating muscle for rounding seems to be OOI.

The MENT is less interesting than the other muscles because the three subjects use it differently (if it is at all the same muscle that is recorded). EFJ has a peak somewhat after the implosion of the consonant, and this activity is much more pronounced before u (she has also a peak in the word huen). PM has a peak at the closing movement.

The activity of PM's labial muscles (OOS, OOI, DLI, and MENT) was recorded in two different sessions. The general pattern is very much the same, but the activity of MENT was much higher the second time (in relation to other muscles), and the relative peaks of OOS and OOI at the implosion were directly reversed. In the first session the mv value of the peak of OOS was much higher than that of OOI; in the second session it was much lower. This demonstrates that the degree of activity of two different muscles cannot be compared, since the electrodes may have been placed more or less close to the active motor units (in the illustrations no attempt has been made, therefore, to use the same mv scales for all muscles).

Figures 1-5 contain some characteristic examples illustrating what has been said on the preceding pages. Figure 1 shows averaged EMG curves of OOS, OOI, and DLI for the words [p^hanə, p^hu:ə] and Figure 2 shows those for [t^hanə, t^hu:ə] spoken by subject PM. It is evident that OOS and OOI have peaks at the implosion of p, and that OOI has a second peak at the release of p before u (continuing through the vowel), whereas DLI has a second peak at the release of

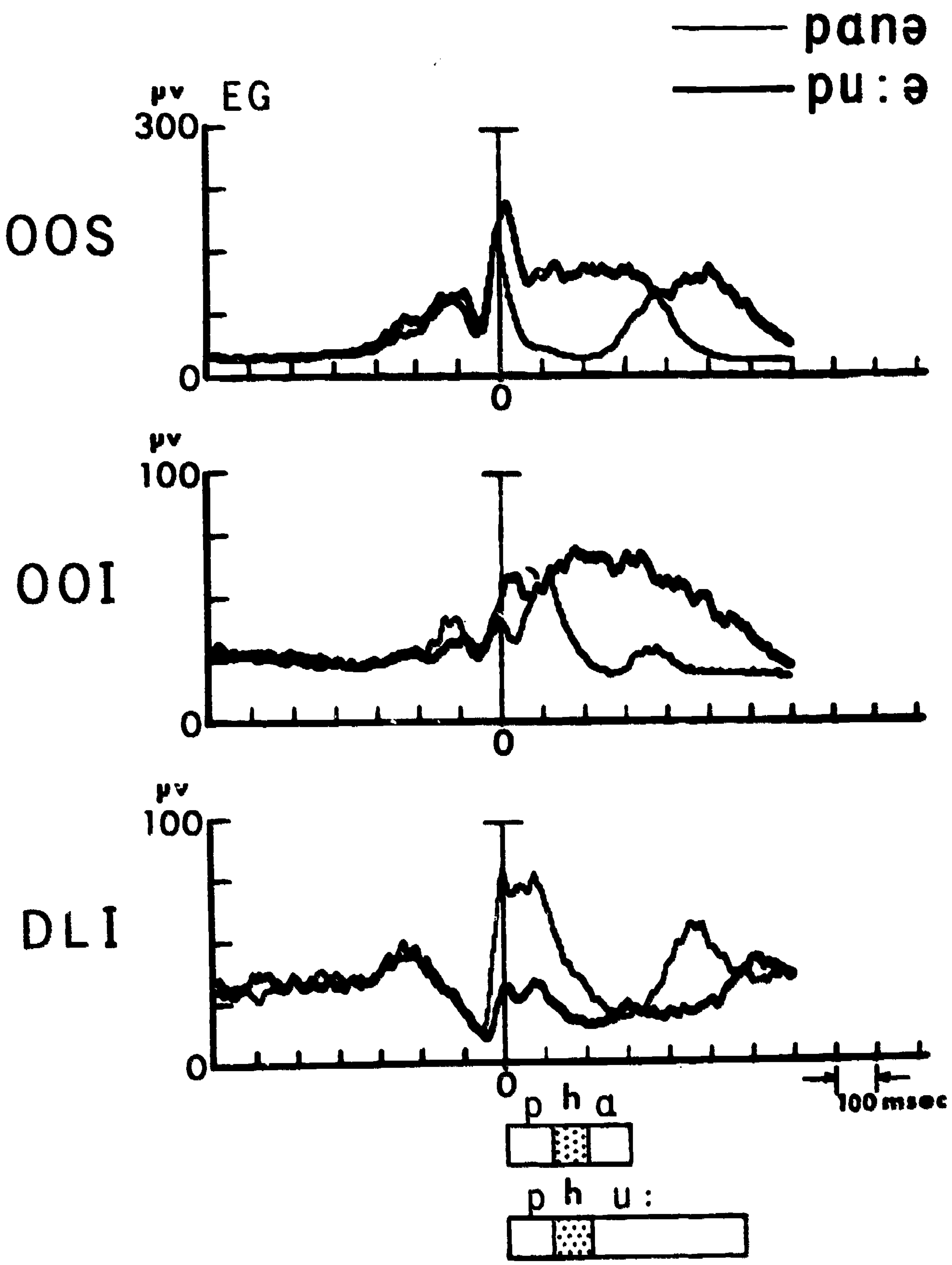

pana
pu:ə
μv
EG
300
OOS
0
0
μv
100
OOI
0
0
μv
100
DLI
0
0
100 msec
p h a
p h u:

FIGURE 1

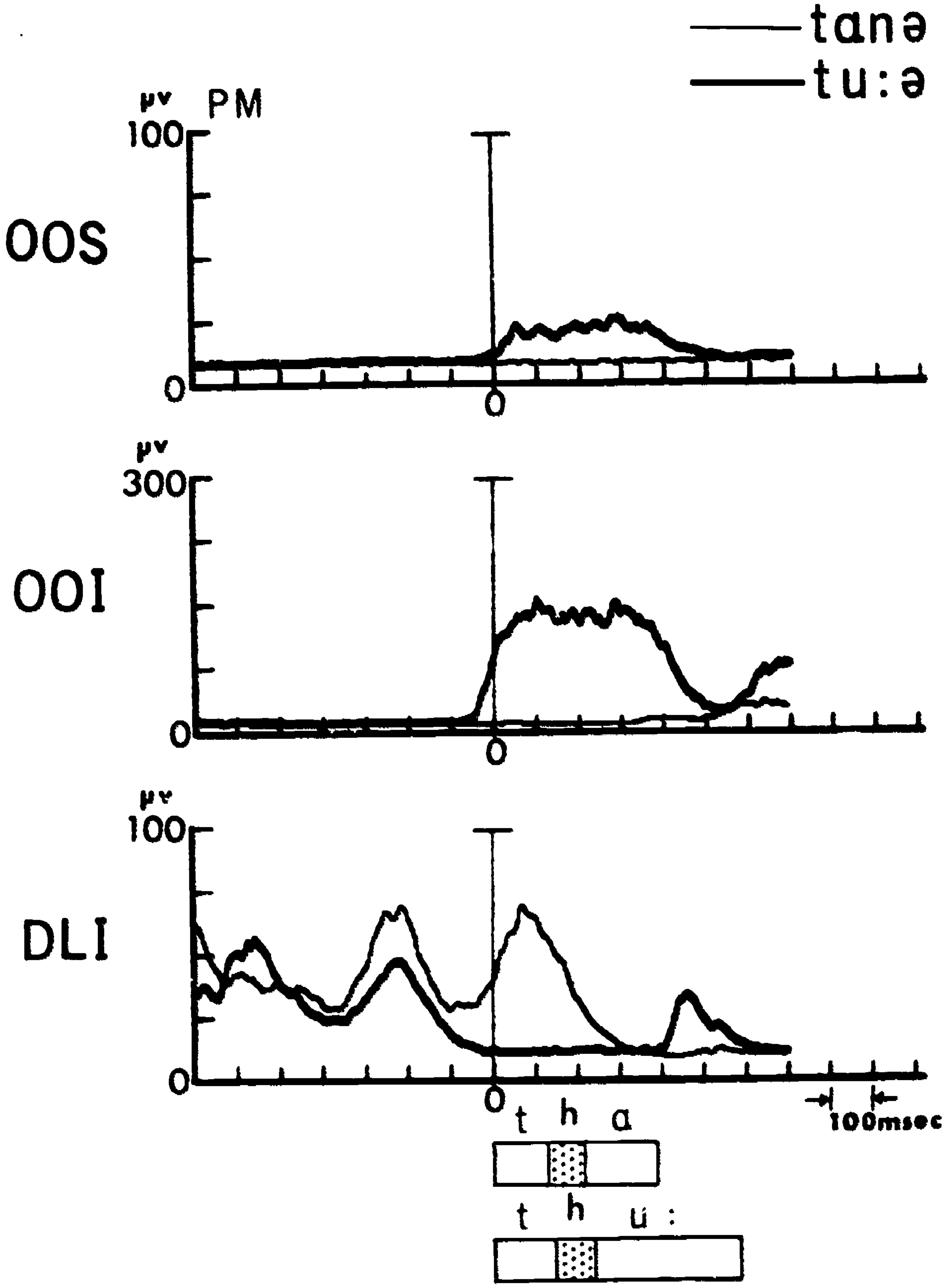

tanə
tu:ə
μv PM
100
OOS
0
0
μv
300
OOI
0
0
μv
100
DLI
0
0
t h a
t h u:
100msec

FIGURE 2

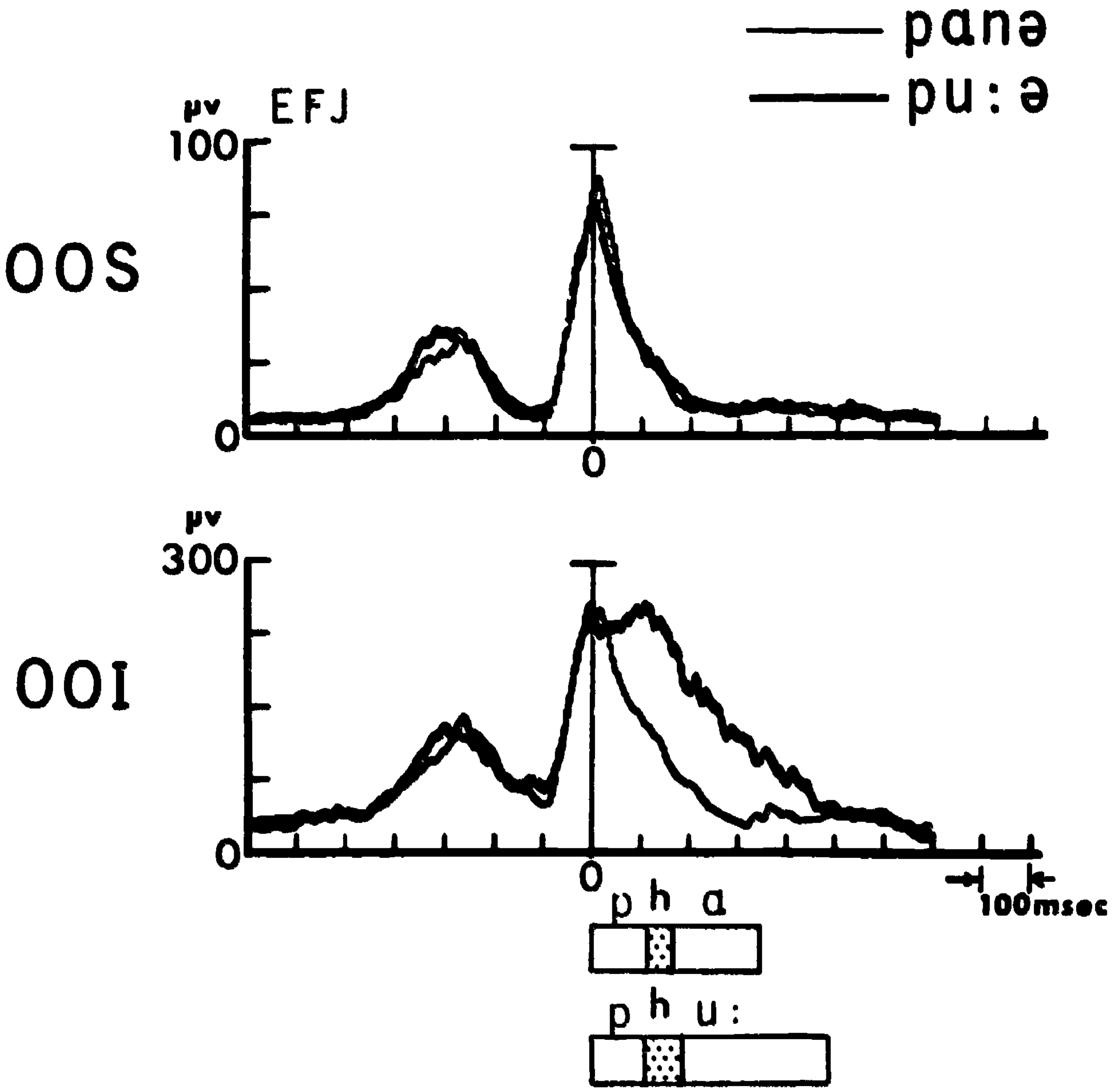


FIGURE 3

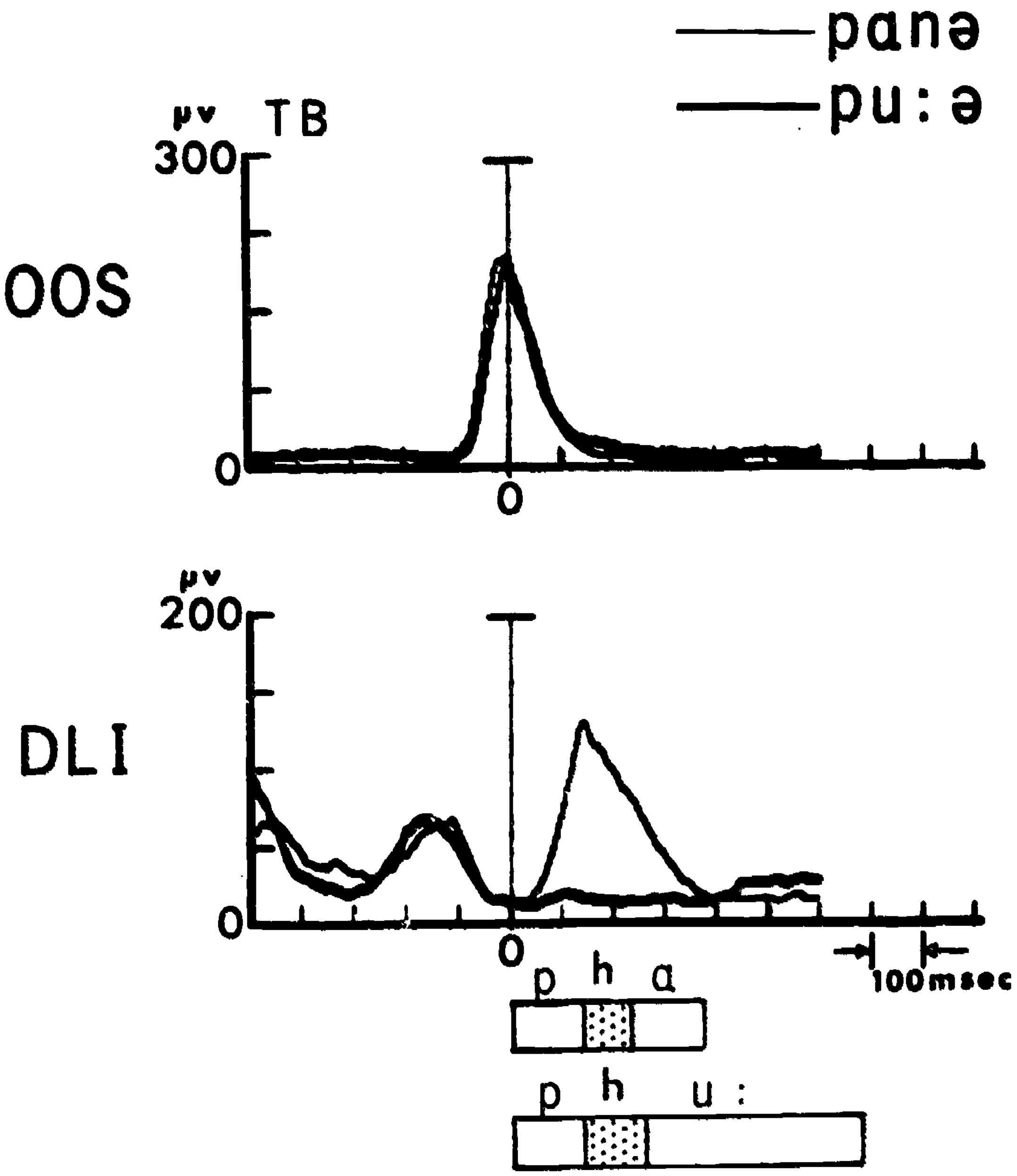


FIGURE 4

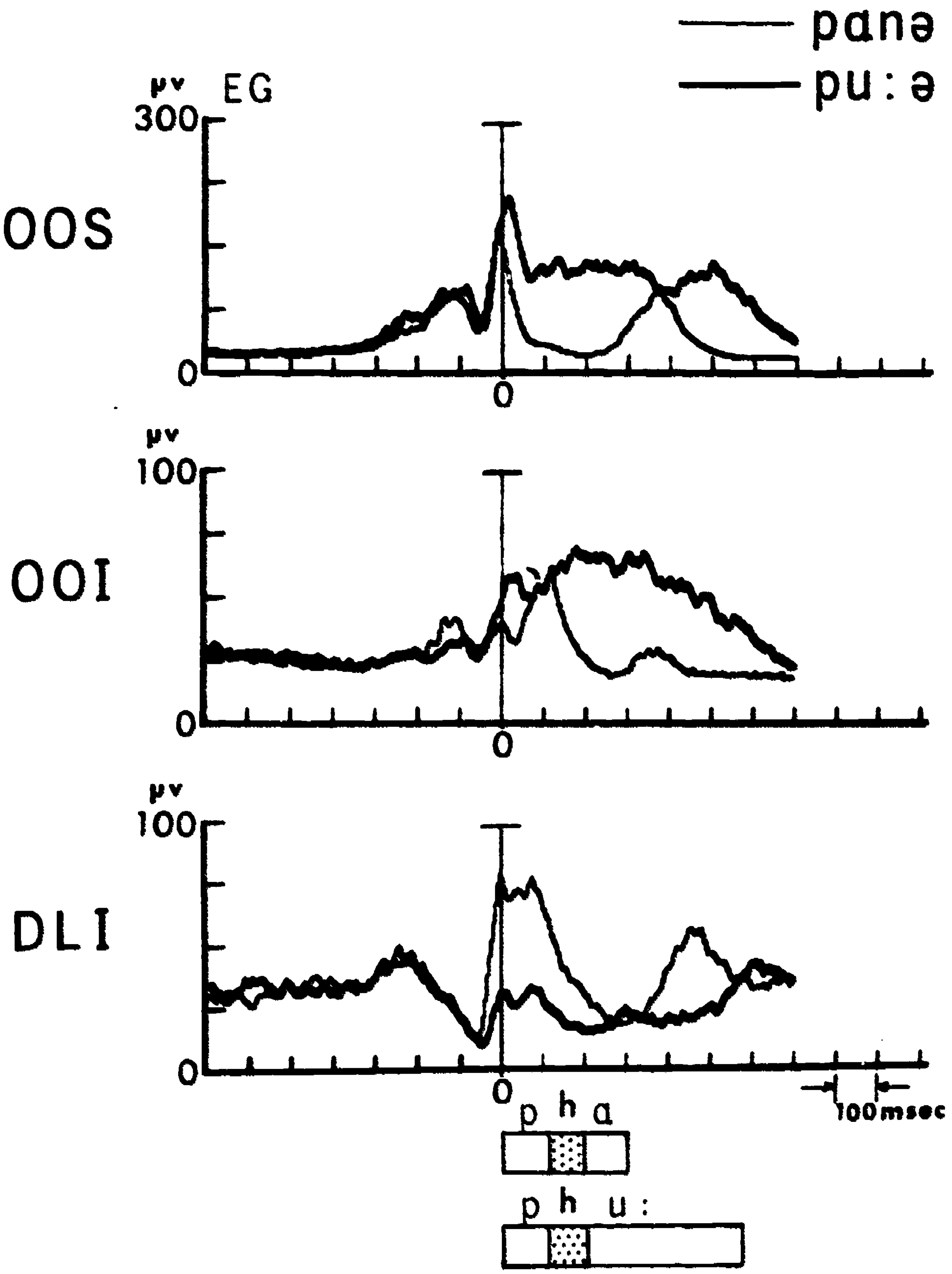

panə
pu:ə
μv EG
300
OOS
0
0
μv
100
OOI
0
0
μv
100
DLI
0
0
100msec
p h a
p h u:

FIGURE 5

p before a. Similarly OOI (and partly OOS) are active for u after t, whereas DLI is active for a after t. The DLI peak before the line-up point pertains to the start of a in [sa:] in the frame.

Figure 3 shows much the same pattern of OOS and OOI as Figure 1 for [pʰanə] and [pʰu:ə] pronounced by EFJ (the peak before the line-up point is due to a certain rounding of s), while Figure 4 shows the pattern of OOS and DLI for [pʰanə] and [pʰu:ə] pronounced by TB.

Figure 5 shows that EG differs from the other subjects in using OOS almost to the same extent as OOI for rounding, in having a high peak of DLI at the implosion besides at the release before unrounded a, and in having a second peak of OOI not only before rounded vowels, but also before unrounded vowels (although somewhat lower).

The material comprises both stressed and unstressed p and b, e.g., ['pʰanə, pʰa'gaiʔ] and ['banə, ba'kʰanʔt]. The·e is no difference in lip activity for the two stress conditions (the [b] of ['pʰibɤ] and ['pʰi:ʔbɤ] has, however, less activity than the initial [pʰ] in OOS and partly in OOI for subjects PM, EFJ, and TB).

The activity found in the labial muscles for Danish labial consonants is, in general,in agreement with what has been found for other languages. The activity of OOS at the implosion is, for example, similar to the activity found by Hirose and Gay (1971) for English stops. Öhman, Leanderson, and Persson (1965, 1966) consider DAO as a closing muscle and DLI as an opening muscle. This is confirmed by the present investigation. They consider OOS and OOI as rounding muscles. A similar argument is made by Hadding, Hirose, and Harris (in press) on Swedish rounded vowels. This is confirmed for OOI, but OOS was used by only some subjects in the present series, while both OOS and OOI function as closing muscles.

Differences between p and b. The general observations of muscle activity are the more reliable results of the investigation, whereas we did not get a clear answer to the problem that was the starting point: the difference between p and b. There are both individual differences and inconsistencies for the same speaker, and only a few of the differences between individual means of p and b are statistically significant. But some tendencies are obvious.

The words which can be directly compared are those in Table 2 (la and lb). All subjects read la, all but HA read lb, and PM and EFJ read the list twice (separate means are taken of these two readings). This gives twelve word-pairs for PM and EFJ, six for EG, TB, and PH, and four for HA.

As for the closing muscles, there are no examples for subject PH since his OOS data are bad, and he does not use his OOI as a closing muscle. EFJ has a clear tendency to more activity for b. In ten of twelve pairs b has a higher maximum of activity in OOS, and in nine of twelve pairs in OOI. All exceptions are in unstressed position. Only three pairs in stressed position have a significant difference (5 percent level), but the number of pairs showing the same difference cannot be accidental.[4] Moreover, the activity for b is normally of a

[4] There is no difference in the MENT, but its peak is later, between the normal place of closing and opening activity.

slightly longer duration, which fits well with the longer closure time. But none of the other subjects has a clear difference between *p* and *b*. The number of individual means having higher activity for *p* or for *b* is about equal, and the differences are small. This is true of HA (OOS), TB (OOS and OOI), EG (OOS and DAO), and PM (OOS, OOI, DAO, and the first peak of DLI). There are only a few deviations from this general pattern: EG and PM have more activity for *b* than for *p* in MENT (for PM ten of twelve means, for EG all six means); and EG has higher activity for *p* than for *b* in OOI (five of six means).

For the muscles that are active at the release (DLI and OOI) the picture is somewhat different. EG has no clear difference between *p* and *b*. TB's OOI is so noisy that it is difficult to see any clear peak (perhaps there is a tendency to slightly higher *p*); his DLI shows a clearly higher maximum for *b* in three of four pairs. The other subjects (PM, EFJ, PH, and HA) show a clear tendency to have a higher maximum for *b* than for *p*. The comparable pairs are PH: OOI (six pairs), and DLI (six pairs); EFJ: OOI (four pairs before *u*); HA: DLI (three pairs before *a*, *i*); and PM: OOI (four pairs before *u*), and DLI (eight pairs before *a*, *i*). Of these 31 pairs, 30 have a higher maximum for *b* than for *p*. Although only six of the individual pairs have a significant difference, the tendency is quite clear.

On the whole it must be said that the maxima are rather variable, but four of the six subjects have a clear tendency to stronger activity for *b* than for *p* at the release, and one also for the closing activity, whereas the others do not show any clear difference in this case.

A difference in the relation between *p* and *b* at the implosion and at the release has also been found by Öhman et al. (1965, 1966) for a Swedish subject: the muscles which were active at the implosion showed a higher peak for *p* than for *b*, whereas the muscles which were active at the release showed a higher peak for *b*. Similarly Harris, Lysaught, and Schvey (1965) found a tendency to a higher activity of orbicularis oris for *p* than for *b* in English at the implosion, whereas no difference was found in the activity of DLI at the release. Finally, Slis (1970) found that in Dutch the activity of orbicularis oris was significantly higher for *p* than for *b* at the implosion, whereas the activity of DLI at the release was only slightly higher for *p* than for *b*. Despite the differences among the languages the relation between implosion and release is similar in all cases, *p* being relatively stronger at the implosion than at the release.

As for the differences among the languages, we should remember that their stops are not phonetically identical: Danish has aspirated *p* and voiceless *b*, Swedish has aspirated *p* and voiced *b*, Dutch has unaspirated *p* and voiced *b*, and English has aspirated *p* and a *b* which may be voiced or voiceless. If we use a narrower phonetic transcription where [p] indicates unaspirated *p*, and voiceless *b* is indicated by a small circle, the relations at the implosion can be stated in the following way: [p > b] (Dutch, significant difference), [ph > b] (Swedish and English, tendency[5]), [b̥ = ph] (Danish). This is, as we would hope, in good agreement with findings of mechanical lip pressure: [p > ph > b] (for Armenian:

[5] This tendency has also been found for English by Lubker and Parris (1970) and by Tatham and Morton (1973), whereas Fromkin (1966) found the opposite tendency in initial position. They all measured the orbicularis oris. However, it is not clear to what extent the *b*'s of the informants were voiced.

Rousselot, 1897:599; and for Gujarati: Fischer-Jørgensen, 1968a:96), [p > b] (for French, partly significant: Fischer-Jørgensen. 1968a:71), [ph > b] (for English, tendency: Malécot, 1966; Lubker and Parris, 1970), [b̥ (>) ph] (Danish). Moreover, the only Danish subject for whom both EMG activity and mechanical pressure have been measured (EFJ) has [b̥ > ph] in both cases. Thus it seems as if unaspirated p has the strongest lip activity, voiced b the weakest activity, and aspirated [ph] is in between. Danish [b̥] is very similar to unaspirated p, but is felt as somewhat weaker; it must therefore be very close to [ph]. This description, though based on very restricted material and needing corroboration by further investigations, is in good agreement with subjective impressions.

As for the difference between implosion and release Öhman et al. (1966) suggest that the higher EMG activity at the implosion of Swedish p is conditioned by the higher air pressure, whereas at the release less muscle activity should be needed for the opening movement of p because the air pressure works in the same direction. This is an interesting hypothesis.

As far as the implosion (i.e., the activity of the orbicularis oris)is concerned, it has often been assumed that a higher lip pressure is necessary for p in order to maintain the closure against a high air pressure. In some languages p has in fact both a higher mechanical lip pressure and a higher intraoral air pressure than b, e.g., in French and Gujarati (Fischer-Jørgensen, 1968a:93 ff). This is, on the whole, the case when b is voiced, since intraoral pressure is intimately connected with voicing. The assumption of a connection between intraoral pressure and lip pressure also finds a partial support in the EMG values of m before a found in the present investigation (subjects PM, TB, and EFJ). For both OOI and OOS m has a lower peak than p and b, and in some cases the difference is significant. The difference is, however, not by far as large and clearcut as should be expected if the (very large) difference in air pressure were the conditioning factor (cf. also the very small differences between m and b found by Harris et al., 1965). Moreover, no clear difference in mechanical pressure has been found between m and bp in Danish. Various other facts speak against a close connection between air pressure and lip pressure. First, for some sounds there is an obvious lack of correlation: in Gujarati, for instance, the aspirated labial stops [ph] and [bh] were found to have a higher air pressure, but a lower mechanical lip pressure than their unaspirated cognates [p] and [b]. Moreover, Tatham and Morton (1973) did not find any correlation between air pressure and activity of the closing muscle OOS for English p and b (except for the fact that the activity of the muscle had not gone quite as far down at the release for p as for b, which can hardly be of any importance). Finally, the air pressure curve generally has its maximum close to the release, whereas the mechanical pressure has its maximum in the first half of the consonant, followed by rapid descent. This means that the mechanical pressure in the first part of the consonant is much higher than is necessary in order to maintain the closure. Tatham and Morton (1973) are probably right in assuming that this surplus of pressure permits a high degree of variability. There is thus hardly any close connection between air pressure and lip pressure at the implosion.

On the other hand, the facts about release may be interpreted in a way that supports the hypothesis of Öhman et al. (1965, 1966)--that the air pressure may contribute to the opening of the lips and thus do some of the work for the opening muscles. If, for the time being, we assume that the preliminary results found for Gujarati (lip pressure: [p > ph > b], intraoral air pressure:

[ph > p > b]) will hold,[6] then the argument may run as follows: in Swedish, where there is a large difference in air pressure between [ph] and [b] (two steps) and a relatively smaller difference in lip pressure (one step), the difference in air pressure succeeds in reversing the relation between [ph] and [b] so that we get [b > ph] for the activity of the opening muscles; in French, where the difference in lip pressure (two steps) is relatively larger than the difference in air pressure (one step), the effect of the air pressure is only to diminish the difference of activity for the release of [p] and [b] ([p] > [b]); in Danish, where there is no consistent difference in lip pressure, the (very small) difference in air pressure is sufficient to diminish the requirements on the activity of the opening muscles for [ph] so that we get the tendency [b̥ > ph] (for subjects who have stronger activity in the closing muscles for [b̥] the relation is kept for the opening muscles).[7] The influence of the air pressure also appears from the fact that the Danish subjects PM and TB show more activity for the opening of *m* than for *p* and *b*, although *m* had less activity in the closing muscles. (PH, however, has more activity of DLI for *b* than for *m*.) The argument presupposes that the lip pressure for *p* does not decrease at a faster rate than that for *b*, but in French, at any rate, it has been shown that the lip pressure for *p* is still higher than that of *b* at 70 percent of the distance from the implosion (Fischer-Jørgensen, 1968a:97). In any case, Öhman's hypothesis deserves further testing.

Laryngeal Muscles

One of the aims of the recording of the laryngeal muscles was to test the hypothesis advanced by Frøkjær-Jensen et al. (1971) that there need not be any activity in the opening muscle for *bdg*. For this purpose we especially wanted to make recordings of the abductor muscle PCA and the adductor muscle INT. Unfortunately, we were able to record from these two muscles for only one subject (EFJ), and from just INT for subject HA. Examples of averaged curves are given in Figures 6 and 7 (subject EFJ) and in Figure 8 (lower portion: subject HA).

Subjects EFJ's INT curves are very clear. In the phrases [han sa:'p^{h}anə] and [han sa:'banə] there is a large dip for *s* and *p* and a somewhat smaller dip for *b* (Figure 6). Similarly in [p^{h}a'gaiʔ] there is a larger dip for *p* and a smaller dip for *g*, and in [bak^{h}anʔt] a smaller dip for *b* and a larger dip for *k* (Figure 7). On the other hand, there is a higher peak for the vowel after *p* than after *b*. Moreover, the curves show a displacement to the right of both consonant minimum and vowel maximum for words with *ptk* compared to those with *bdg*. This is in good agreement with glottograms of the consonants in question (Frøkjær-Jensen, 1967, 1968: Frøkjær-Jensen et al., 1971), which show that the maximum aperture is found in the beginning of the lip closure for *b*, decreasing to zero at the release, whereas the maximum aperture for *p* is close to the release, decreasing during the following aspiration. The INT curves in Figures 6 and 7 are typical. There are no exceptions to the difference between the dips for *ptk* and *bdg*, and this difference is evidently significant. As for the

[6] The air pressure relation is corroborated by Nihalani (1974), but some further recordings of Indian stops by one of the present authors (EFJ) do not show a clear difference between [ph] and [p].

[7] For English the reasoning depends on the degree of voicing of *b*.

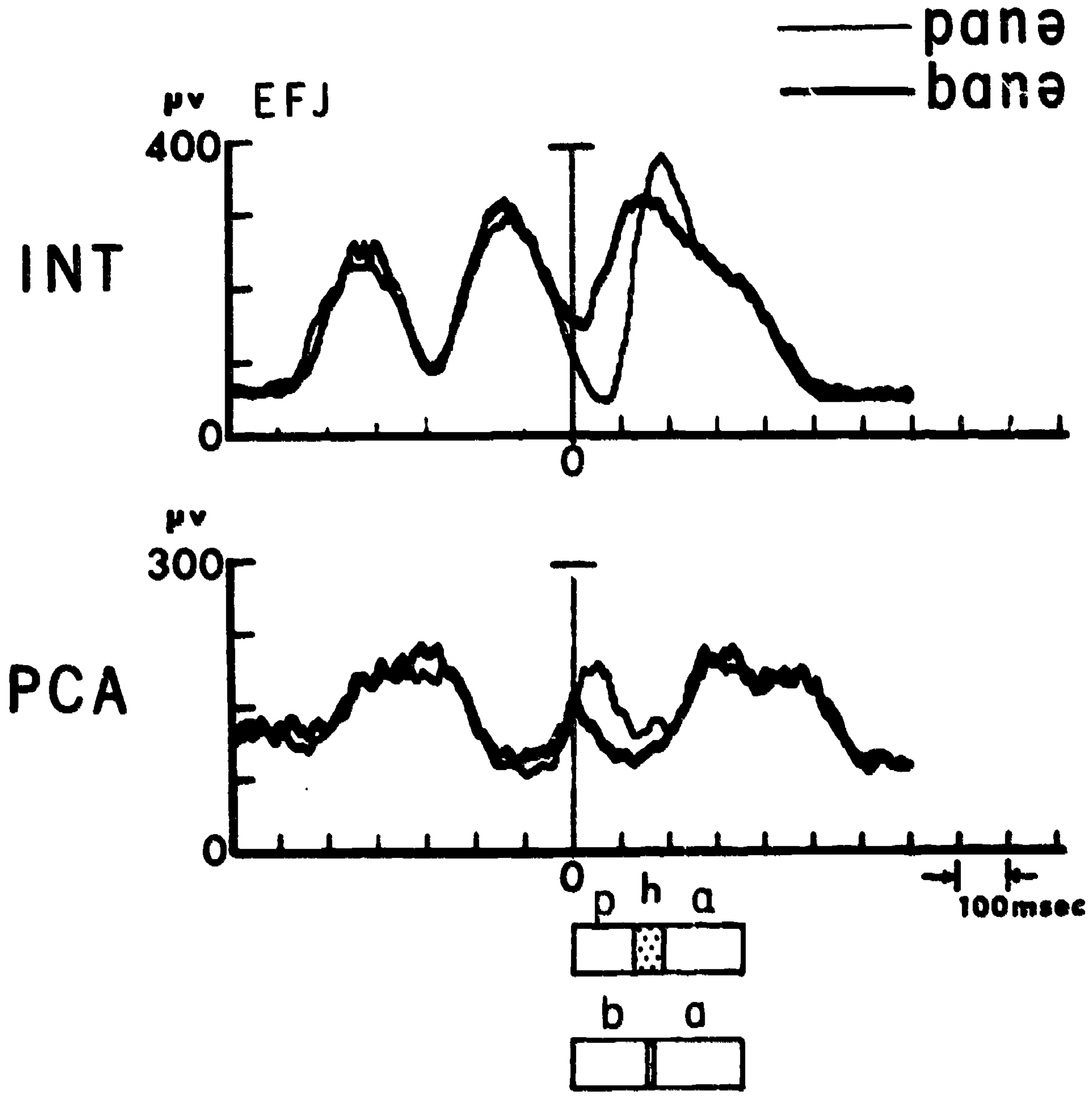


FIGURE 6

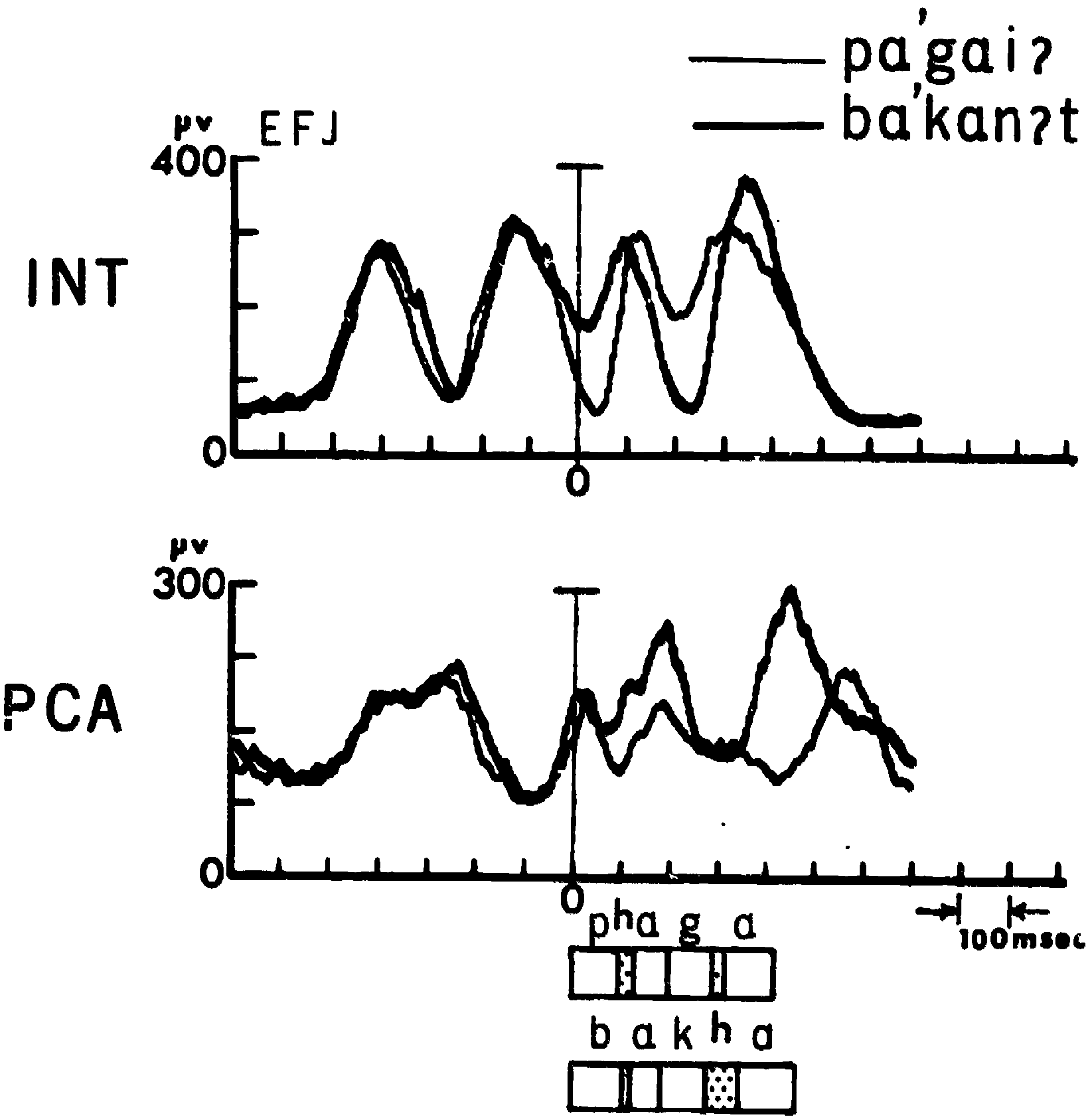

pa'gai?
ba'kan?t
μv EFJ
400
INT
0
0
μv
300
PCA
0
0
ph a g a
b a k h a
100msec

FIGURE 7

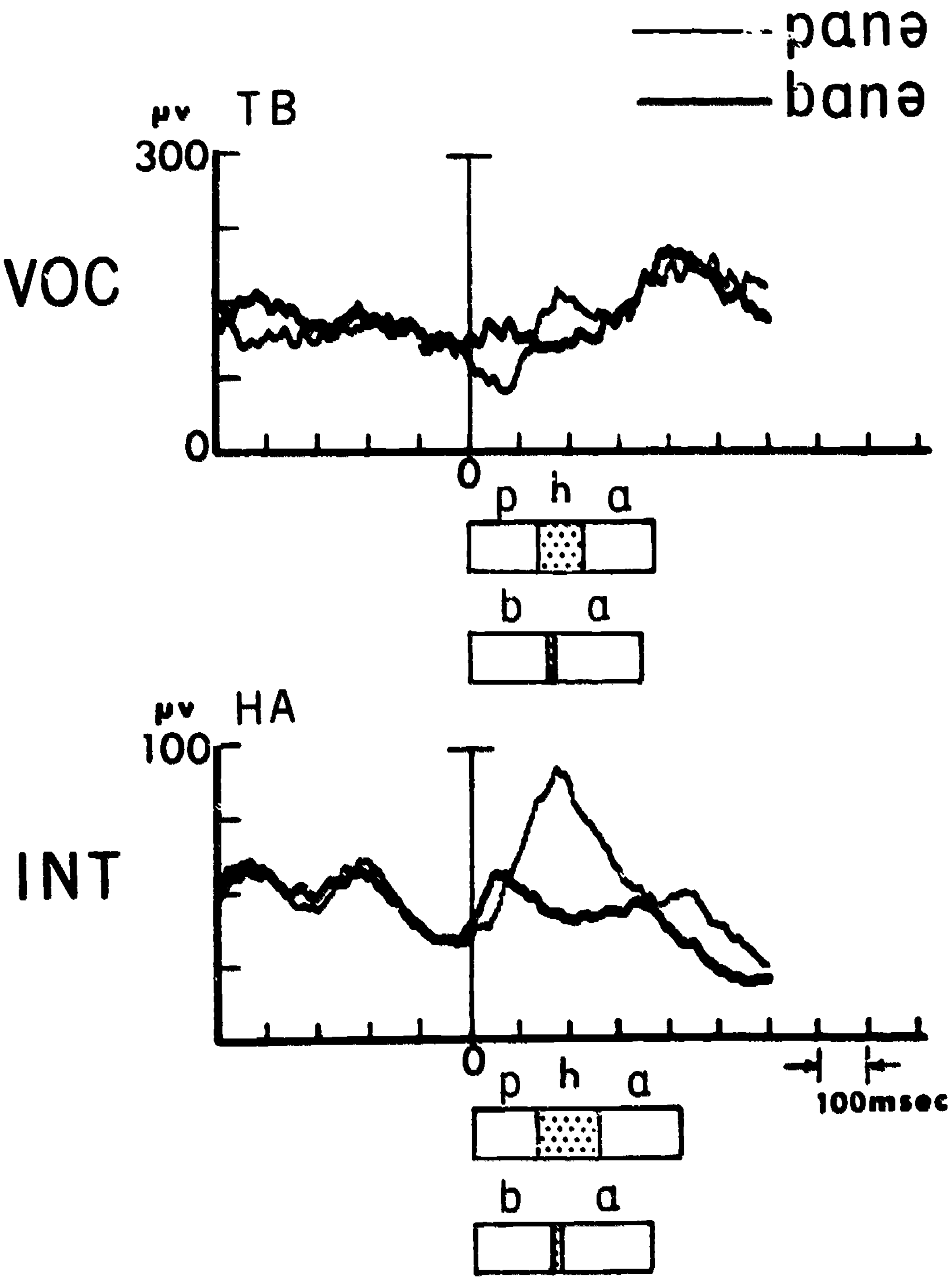


FIGURE 8

difference in vowel peaks after the stop consonants there are only two exceptions (out of 20 pairs). Although the difference between individual pairs is significant in only three cases, the general tendency is quite obvious. This means that the activity of the closing muscle is more pronounced when the preceding consonant has a larger glottis opening, i.e., the motor command and the resulting muscle activity depend on the preceding state of the vocal tract.

Subject HA's INT curves look somewhat different (Figure 8, lower portion). He shows no difference between the dips for ptk and bdg, but the difference in the following vowel is very clear and stable. The vowel has a higher peak after ptk than after bdg in all 11 pairs, and the difference is statistically significant in four individual pairs. The increase in activity for INT is thus considerable for both subjects. It is also common to the two subjects that both the minimum in the consonant and the maximum of the following vowel are delayed in ptk words in comparison to bdg words. But there are some differences in the timing in relation to the line-up point and the absolute differences between ptk and bdg, which can be seen in Table 4. Subject HA starts the relaxation of INT earlier than EFJ and reaches the minimum value earlier, but his vowel peak is, nevertheless, later after ptk.

TABLE 4: Timing in INT for stressed ptk and bdg, in msec, in relation to the line-up point.

	Start of decrease	minimum	vowel peak
EFJ	ptk -100	60	170
	bdg -100	15	120
HA	ptk -200	-30	200
	bdg -200	-50	60

The long distance from valley to peak for ptk reflects a considerably longer aspiration in the case of HA (his aspiration in [pʰanə] is, e.g., 128 msec whereas EFJ has an aspiration of 66 msec). But it is not quite clear why HA should start his relaxation of INT earlier for this purpose. For the labial muscles, his timing does not differ from that of EFJ. Air-stream curves might show whether the vowel preceding the consonant is breathy. His long aspiration and the difference in vowel peak after ptk and bdg point to a wide-open glottis in ptk, which must be produced by the abductor muscle only, since the relaxation of INT is the same for ptk and bdg, but unfortunately we did not get any curve of his PCA.

The recording of PCA for subject EFJ shows a small peak about 20 msec after the line-up point for ptk, and a slightly smaller peak at the line-up point for bdg. The peak of ptk is higher in 14 out of 16 pairs, but the difference is often small. The increase in activity starts around 50 msec before the implosion and goes rather quickly down again after the maximum (see Figures 6 and 7). The activity lasts somewhat longer for ptk than for bdg. There is no peak for m, but there is,unexpectedly, a clear peak for l (in [ˈlɛsʁ, ˈlɛ:sʁ, ˈlɛ:ʔsʁ], for which INT does not show any dip. In any case there is evidence for an active opening movement in bdg for this subject.

For subjects EFJ and HA no other laryngeal muscles have been recorded, but we obtained recordings of LCA for PM and of VOC for PM, TB, and PH (PH's VOC is, however, not reliable). As shown in Figure 9, PM has a small dip in LCA for ptk with a minimum about 65 msec after the implosion, but no dip for bdg. He has a similar dip for ptk in VOC with a minimum about 85 msec after the implosion. The VOC curve also shows a very small dip for bdg slightly later (about 95 msec after the implosion). No dips occur for m and l. TB has a somewhat deeper minimum in VOC for ptk, about 75 msec after the implosion, as shown in Figure 8 (upper portion). He has no dip for bdg. These minima are not very pronounced, but are completely regular. As the small dip in PM's VOC is later for bdg than for ptk and may presumably be later than the point of maximum aperture of the vocal cords in bdg [which, according to Frøkjær-Jensen et al. (1971), is found at a distance of 45 msec from the implosion (average of the three subjects)], it might not have anything to do with the opening of the vocal cords; probably it only has to do with a relaxation of the tension of the vocal cords. This relaxation is more pronounced for ptk than for bdg. It is thus very improbable that Danish ptk should have stiffer vocal cords than bdg (cf. Halle and Stevens, 1971).

Some of the INT curves show differences connected with stress. The s of the frame [han sa:] has a dip, but it is not quite as low as that of ptk. This might be due to the somewhat weaker stress of this word. This assumption is supported by the fact that s and f in the words [sanə] and [falə], found in HA's list only, have a dip of the same size as that of his ptk. However, clear examples of strong and weak stress, like ['p^{h}anə, p^{h}a'gaiʔ], do not show any difference in the minima of the p's. On the other hand, there is a clear difference between the peaks of the vowels in stressed and unstressed position. EFJ has 12 pairs of this type, and the difference is found in all of them. Moreover, there is a corresponding difference between, for example, the first vowel of [p^{h}a'gaiʔ] and the second vowel of [ba'k^{h}anʔt]. This double comparison could be made in eight of the twelve pairs, and there were no exceptions. The difference due to stress is, however, not as large as the difference due to the preceding consonant, and it therefore disappears if the consonant of the unstressed syllable is of the ptk-type and the consonant of the stressed syllable is of the bdg-type, whereas it is enhanced in the opposite case. For example, the INT peak for the second vowel of [ba'k^{h}anʔt] is 93 mv higher than that for the first vowel, whereas the peak for the second vowel of [p^{h}a'gaiʔ] is 3 mv lower than that for the first vowel (Figure 7). VOC and LCA do not show any stress differences, but a clear connection with pitch. They show a rise in the second syllable of ['p^{h}anə] as well as in the second syllable of [p^{h}a'gaiʔ].

The results of the present investigation of larynx muscles agree on many points with earlier findings. The aspirated Danish stops have a dip in INT and a peak in PCA like the aspirated English stops (Hirose and Gay, 1971); and the unaspirated voiceless Danish stops behave like unaspirated p as pronounced by L. Lisker (Hirose, Lisker, Abramson, 1972): they have a smaller and shorter dip in INT than aspirated ptk and a lower peak in PCA.

The activity patterns of VOC and LCA seem more complex than what has been described by one of the present authors (Hirose, 1971b) as being active in vowels only and suppressed for consonants irrespective of the type of consonants,

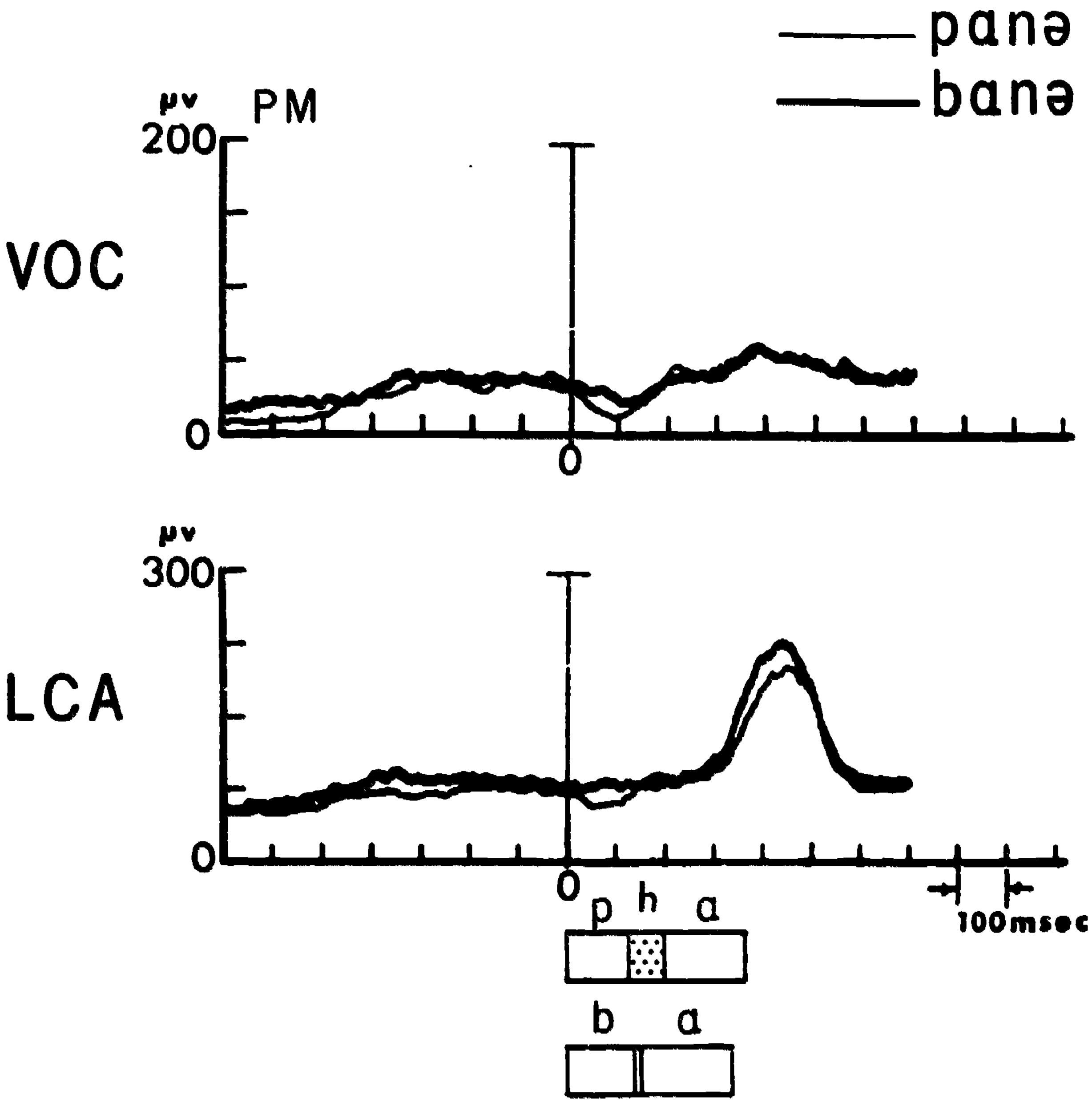


FIGURE 9

although language difference and context effects might well be taken into consideration. In the present investigation we found that LCA has a dip for ptk but not for bdg, and that VOC has a deeper dip for ptk than for bdg but none for l and m. Their activity thus seems to be more differentiated. Similar differences were found in LCA in a study by Hirose et al. (1972).

In view of the small number of subjects, particularly for the laryngeal muscles, the results of this study must be considered as preliminary. Further investigations which are in progress at the Institute of Phonetics in Copenhagen seem, however, to confirm the results described in this paper.

REFERENCES

Fischer-Jørgensen, E. (1954) Acoustic analysis of stop consonants. Miscellanea Phonetica 2, 42-59.

Fischer-Jørgensen, E. (1968a) Voicing, tenseness, and aspiration in stop consonants, with special reference to French and Danish. Annual Report of the Institute of Phonetics, University of Copenhagen (ARIPUC) 3, 63-114.

Fischer-Jørgensen, E. (1968b) Les occlusives françaises et danoises d'un sujet bilingue. Word 24, 112-153.

Fischer-Jørgensen, E. (1972) Kinesthetic judgment of effort in the production of stop consonants. Annual Report of the Institute of Phonetics, University of Copenhagen (ARIPUC) 6, 59-73.

Fischer-Jørgensen, E. and H. Hirose. (1974) A note on laryngeal activity in the Danish "stød." Haskins Laboratories Status Report on Speech Research SR-39/40 (this issue).

Frøkjær-Jensen, B. (1967) A photoelectric glottograph. Annual Report of the Institute of Phonetics, University of Copenhagen (ARIPUC) 2, 5-19.

Frøkjær-Jensen, B. (1968) Comparison between a Fabre glottograph and a photoelectric glottograph. Annual Report of the Institute of Phonetics, University of Copenhagen (ARIPUC) 3, 9-16.

Frøkjær-Jensen, B., C. Ludvigsen, and J. Rischel. (1971) A glottographic study of some Danish consonants. In Form and Substance, ed. by L. L. Hammerich, R. Jakobson, and E. Zwirner. (Copenhagen: Akademisk) 123-140.

Fromkin, V. A. (1966) Neuromuscular specification of linguistic units. Lang. Speech 9, 170-199.

Hadding, K., H. Hirose, and K. S. Harris. (in press) Facial muscle activity in the production of Swedish vowels: An electromyographic study. Haskins Laboratories Status Report on Speech Research SR-41.

Halle, M. and K. N. Stevens. (1971) A note on laryngeal features. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 101, 198-213.

Harris, K. S., G. Lysaught, and M. M. Schvey. (1965) Some aspects of the production of oral and nasal labial stops. Lang. Speech 8, 135-148.

Hirose, H. (1971a) Electromyography of the articulatory muscles: Current instrumentation and techniques. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.

Hirose, H. (1971b) An electromyographic study of laryngeal adjustments during speech articulation: A preliminary report. Haskins Laboratories Status Report on Speech Research SR-25/26, 107-116.

Hirose, H. and T. Gay. (1971) The activity of the intrinsic laryngeal muscles in voicing control: An electromyographic study. Haskins Laboratories Status Report on Speech Research SR-28, 115-142; and Phonetica 25 (1972), 140-164.

Hirose, H., T. Gay, and M. Strome. (1971) Electrode insertion technique for laryngeal electromyography. J. Acoust. Soc. Amer. 50, 1149-1150.

Hirose, H., L. Lisker, and A. S. Abramson. (1972) Physiological aspect of certain laryngeal features in stop production. Haskins Laboratories Status Report on Speech Research SR-31/32, 183-191.

Kewley-Port, D. (1973) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research SR-33, 173-183.

Leanderson, R. (1972) On the functional organization of facial muscles in speech. Thesis from the Department of Otolaryngology and Clinical Neurophysiology, Karolinska Sjukhuset, Stockholm, Sweden.

Lubker, J. F. and P. Parris. (1970) Simultaneous measurements of intraoral pressure, force of labial contact, and labial electromyographic activity during production of the stop cognates /p/ and /b/. J. Acoust. Soc. Amer. 47, 625-634.

Malécot, A. (1966) Mechanical pressure as an index of "force of articulation." Phonetica 14, 169-189.

Nihalani, P. (1974) An aerodynamic model of stops in Sindhi. Phonetics 29, 193-210.

Öhman, S., R. Leanderson, and A. Persson. (1965) Electromyographic studies of facial muscles during speech. Quarterly Progress Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) STL-QPSR 3-1965, 1-11.

Öhman, S., R. Leanderson, and A. Persson. (1966) EMG studies of facial muscle activity in speech II. Quarterly Progress Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) STL-QPSR 1-1966, 1-10.

Port, D. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research SR-25/26, 67-72.

Rousselot, J.-P. (1897) Principes de Phonétique Expérimentale I (2e édition, 1924) (Paris).

Slis, I. H. (1970) Articulatory measurements on voiced, voiceless, and nasal consonants. Phonetica 21, 193-210.

Tatham, M. A. A. and K. Morton. (1973) Electromyographic and intraoral airpressure studies of bilabial stops. Lang. Speech 16, 336-350.

A Note on Laryngeal Activity in the Danish "stød"

Eli Fischer-Jørgensen[+] and Hajime Hirose[++]

The Danish "stød" is a sort of accent connected with a definite syllable in the word (historically it is related to accent 1 in Swedish and Norwegian and in Southern Danish dialects). The stronger forms are characterized by creaky voice, which is found either at the end of a long vowel ("stød in the vowel") or at the beginning of a voiced consonant following a short vowel ("stød in the consonant"). Syllables ending in a short vowel or in a short vowel plus voiceless consonants cannot have stød. The stød is generally indicated in phonetic transcriptions by the sign for glottal stop (ʔ) after the vowel or consonant in question, but in normal standard Danish there is no closure except in very emphatic speech. The occurrence of the stød is to a large extent predictable, but there are some minimal pairs distinguished by the presence or absence of stød, e.g., [lɛ:sʁ] 'reader' vs [lɛ:ʔsʁ] 'reads,'[man] 'one, you,' (indefinite pronoun) vs [manʔ] 'man.'

The most thorough phonetic investigation of the stød was undertaken by Svend Smith (1944) on the basis of kymograms, oscillograms, pitch curves, and electromyograms of the respiratory muscles. He describes the stød as "a stress accent, a special marking movement made by a thrust-like emphasizing of sounds" (Summary, p. 6), primarily consisting in a brief and intense, rather suddenly reduced innervation of the respiratory muscles, a sort of ballistic movement, combined with a more tense articulation of the whole word, which is visible in the initial consonant. The sudden cessation of innervation of the respiratory muscles results in a reduction of subglottal pressure causing a decrease in intensity and pitch, sometimes ending in irregular oscillations. He does not find any consistent difference in the pitch movement in the beginning of the word.

Smith's acoustic description has been confirmed by later studies by Margaret Lauritsen (1968) and Pia Riber Petersen (1973). Petersen examined pitch and intensity curves based on tape recordings of six subjects. She found a very great variability in the phonetic manifestation of the stød, but a general tendency to a more extensive fall in pitch and intensity in syllables with stød sometimes ending in irregular vibrations. This latter phase of the stød (which Smith calls the second phase) is found at approximately the same distance from the beginning of the vowel, corresponding to the end of a long vowel or the beginning of a consonant after a short vowel. The stød is thus a syllabic phenomenon.

[+] Institute of Phonetics, University of Copenhagen.

[++] Faculty of Medicine, University of Tokyo, and visiting researcher, Haskins Laboratories, New Haven, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-39/40 (1974)]

However, nobody has yet tried to verify Smith's physiological description. He was not able to synchronize the electromyographic recordings with the audio signal, and it is therefore not quite certain that the activity in the respiratory muscles precedes the glottal modifications; nor is it known whether there is an active innervation of the glottis, and in the positive case, whether the activity is triggered by the respiratory activity or independent of it.

In connection with the investigation of Danish stop consonants reported in the preceding report of this volume (Fischer-Jørgensen and Hirose, 1974) some EMG recordings (using hooked-wire electrodes) were made of the activity of laryngeal muscles in Danish words with and without stød. For details of the technique applied, see the preceding report. No comparison with the respiratory muscles was made. The purpose was only to see whether there was a positive innervation of the laryngeal muscles in the stød. The subjects, PM, PH, TB, and EFJ, all have a clear stød. PM, TB, and PH are from Copenhagen; EFJ grew up in Southern Funen, where the dialect lacks stød, but she has never spoken Funish dialect. For EFJ a longer list containing words with and without stød was used, but recordings were made only of the interarytenoid muscle (INT) and the posterior cricothyroid muscle (PCA), and they did not show any difference for words with and without stød. There is a peak in PCA at the end of the word [manʔ] which may, however, be due to a more vigorous opening of the glottis at the end of the word.

The other subjects read the word pairs [lɛ:sɐ, lɛ:ʔsɐ], [p^{h}i:bɐ, p^{h}i:ʔbɐ] and [man, manʔ] in the frame [han sa:] "he said," placed in a randomized list of words used for the investigation of stop consonants. Each word appeared 16 times.

For subject PH a recording made of the vocalis muscle (VOC) did not show any differences depending on the stød. This recording, however, was not very good. In the case of subject TB a difference was found in the activity of VOC in words with and without stød (see Figure 1). The words with stød showed a higher degree of activity. It should be mentioned, however, that TB's pronunciation of the stød was somewhat exaggerated. The words with stød were pronounced with higher intensity and with higher pitch in the stressed syllable than the words without stød (this is particularly true of the pair [p^{h}i:bɐ/p^{h}i:ʔbɐ]), and the higher activity.of the VOC may be due to the rise in pitch. TB shows no difference between the words [p^{h}u'rist] and [p^{h}a'gaiʔ], belonging to the consonant list, but a somewhat lower activity in [bu'dist].

The curves of subject PM's recordings are more reliable. The recordings comprise the vocalis muscle (VOC) and the lateral cricothyroid (LCA). The subject pronounced all words with a rising pitch at the end. This explains the general rise of the curves (see Figure 2) but apart from this, the words with stød show a sudden, very clear peak in the beginning of the vowel in all three pairs. Moreover, there is a definite peak in the second syllable of the words [p^{h}a'gaiʔ, ba'k^{h}anʔt, bet^{h}a:ʔlə, p^{h}e'da:ʔl], but hardly any peak in [p^{h}u'rist] and [bu'dist]. LCA also shows a slightly higher activity, but this is not very clear. The initial consonant p̲ shows no difference, either in the inferior orbicularis oris muscle (OOI), or in the superior orbicularis oris muscle (OOS) for words with and without stød. Initial m̲ has a slightly higher average peak in OOS in the words with stød but the difference is hardly significant.

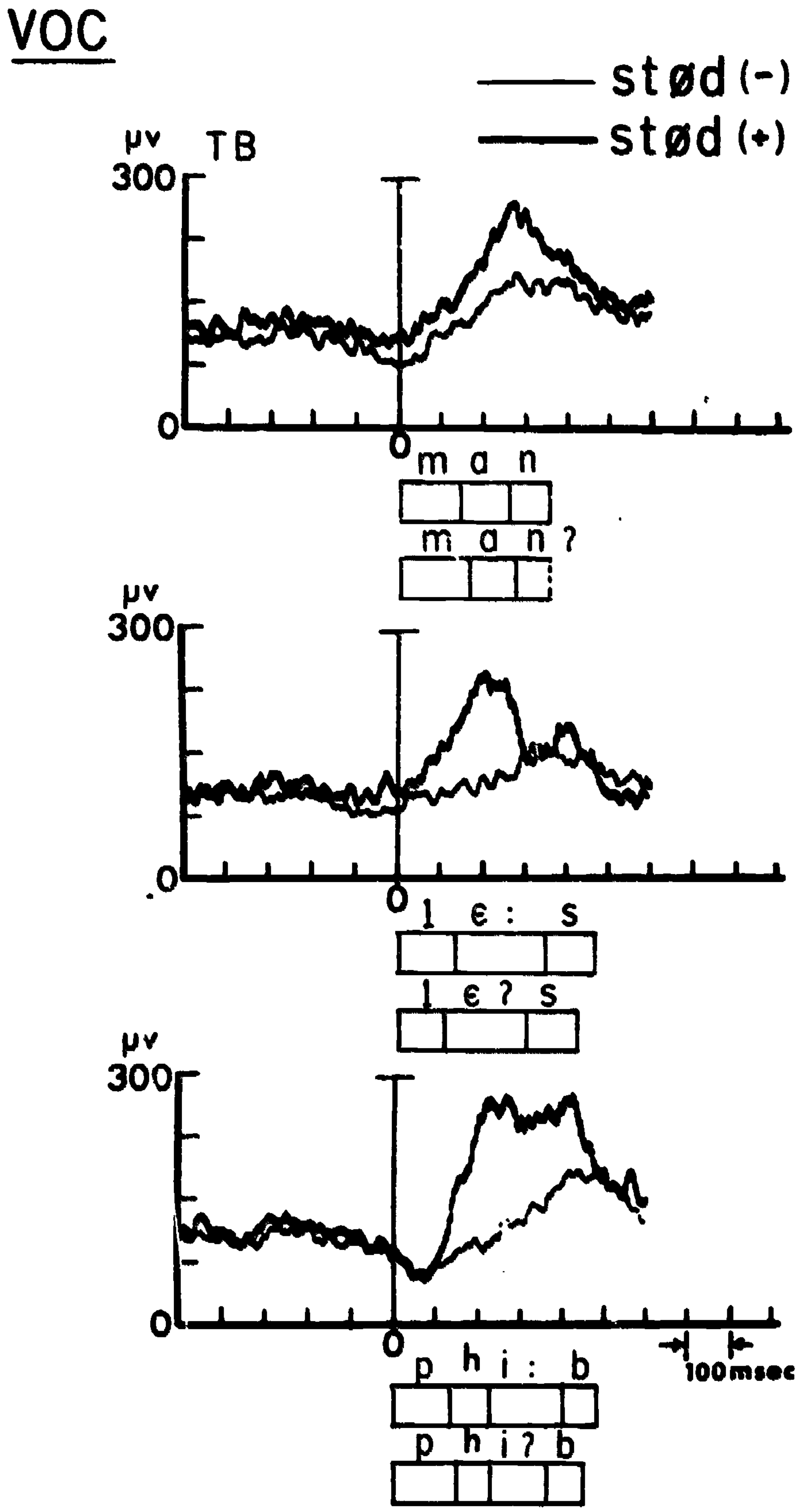

VOC
stød (-)
stød (+)
μv
TB
300
0
0
m a n
m a n ʔ
μv
300
0
0
l ɛ : s
l ɛ ʔ s
μv
300
0
0
p h i : b
p h i ʔ b
100msec

FIGURE 1 -

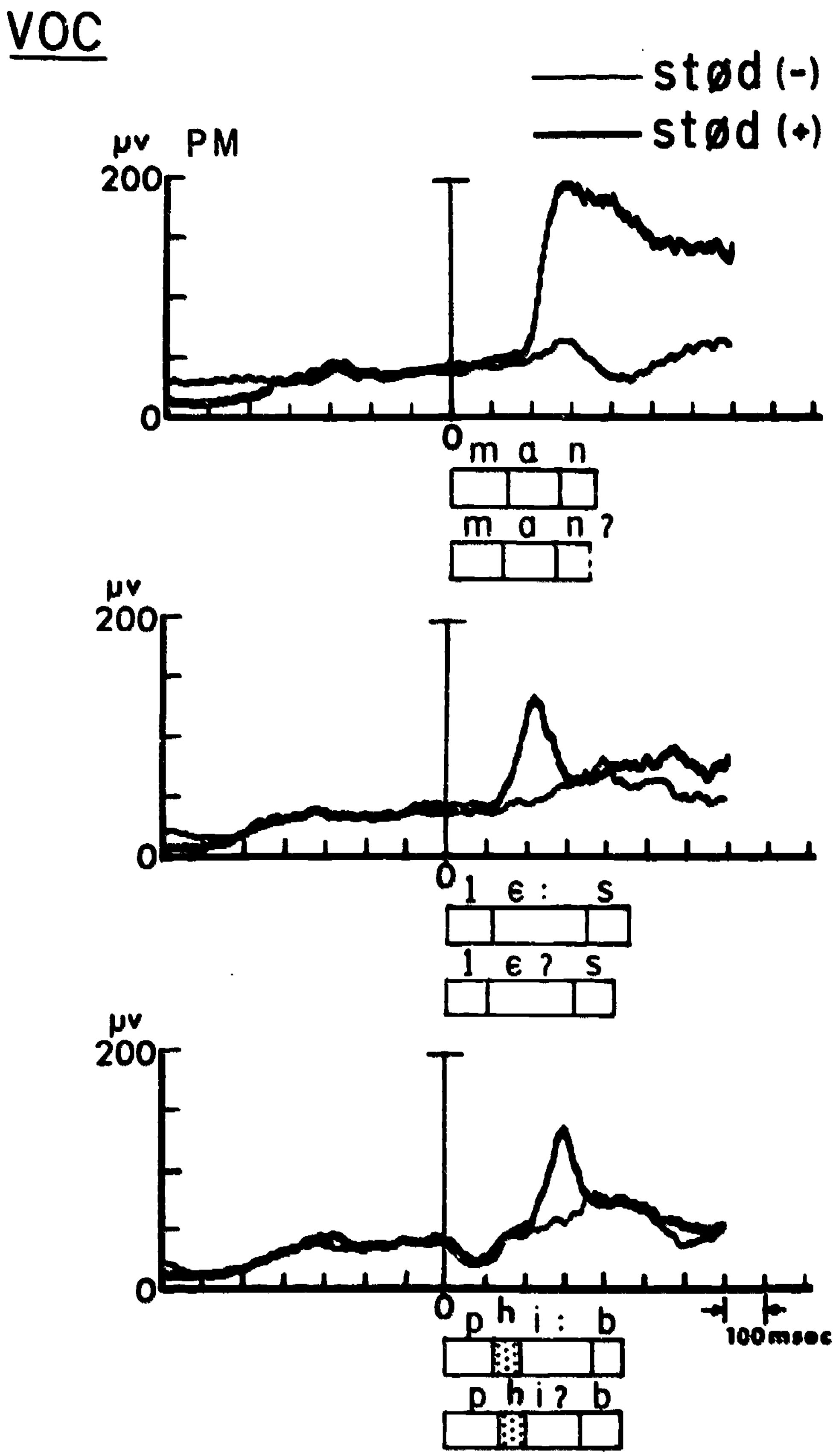

VOC
stød (-)
stød (+)
μv PM
200
0
0
m a n
m a n ?
μv
200
0
0
l ɛ : s
l ɛ ? s
μv
200
0
0
p h i : b
p h i ? b
100 msec

FIGURE 2

Thus, for one subject, whose curves are particularly reliable, a difference in innervation of the vocalis muscle has been found. The investigations are being continued at the Institute of Phonetics in Copenhagen.

REFERENCES

Fischer-Jørgensen, Eli and Hajime Hirose. (1974) A preliminary electromyographic study of labial and laryngeal muscles in Danish stop consonant production. Haskins Laboratories Status Report on Speech Research SR-39/40 (this issue).

Lauritsen, M. (1968) A phonetic study of the Danish stød. POLA Reports, Second Series 7, D1-D12.

Petersen, P. Riber. (1973) An instrumental investigation of the Danish "stød." Annual Report of the Institute of Phonetics, University of Copenhagen (ARIPUC) 7, 195-234.

Smith, S. (1944) Bidrag til løsning af problemer vedrørende stødet i dansk rigssprog [with an English summary: Contributions to the solution of problems concerning the Danish stød (Copenhagen)].

PUBLICATIONS AND REPORTS

Publications and Manuscripts

Different Speech-Processing Mechanisms Can be Reflected in the Results of Discrimination and Dichotic Listening Tasks. James E. Cutting. Brain and Language (1974) 1, 363-373.

Visual Storage or Visual Masking? An Analysis of the "Retroactive Contour Enhancement" Effect. M. T. Turvey, C. F. Michaels, and D. Kewley-Port. Quarterly Journal of Experimental Psychology (1974) 26, 72-81.

Early Apical Stop Production: A Voice Onset Time Analysis. Diane Kewley-Port and Malcolm S. Preston. Journal of Phonetics (1974) 2, 195-210.

Discrimination of Intensity Differences on Formant Transitions In and Out of Syllable Context. M. F. Dorman. Perception and Psychophysics (1974) 16, 84-86.

Hemispheric Specialization for Speech Perception in Six-Year-Old Black and White Children from Low and Middle Socioeconomic Classes. M. F. Dorman and D. S. Geffner. Cortex (1974) 10, 171-176.

On the Short-Term Retention of Serial, Tactile Stimuli. Edith V. Sullivan and M. T. Turvey. Memory and Cognition (1974) 2, 601-606.

*A Cinefluorographic Study of Vowel Production. Thomas Gay. Journal of Phonetics (1974) 2, 255-266.

Reaction Times to Comparisons Within and Across Phonetic Categories. David B. Pisoni and Jeffrey Tash. Perception and Psychophysics (1974) 15, 285-290.

Explicit Syllable and Phoneme Segmentation in the Young Child. I. Y. Liberman, Donald Shankweiler, F. William Fischer, and Bonnie Carter. Journal of Experimental Child Psychology (1974) 18, 201-212.

On 'Explaining' Vowel Duration Variation. Leigh Lisker. Glossa (1974) 8, 233-246.

Fundamental Frequency Contours of the Tones of Standard Thai. Donna Erickson. PASAA (1974) 4, 1-25.

Parallel Processing of Auditory and Phonetic Information in Speech Discrimination. Charles C. Wood. Perception and Psychophysics (1974) 15, 501-508.

On the Identification of Place and Voicing Features in Synthetic Stop Consonants. James R. Sawusch and David B. Pisoni. Journal of Phonetics (1974) 2, 181-194.

*Appears in this report, SR-39/40.

Research on Audible Outputs of Reading Machines for the Blind. F. S. Cooper, J. H. Gaitenby, I. G. Mattingly, P. W. Nye, and G. N. Sholes. Bulletin of Prosthetics Research (1973) BPR 10-20, 348-353; and (1974) BPR 10-21, 153-158.

*Speech Perception. Michael Studdert-Kennedy. To be published in Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass (Springfield, Ill.: C. C Thomas).

*The Physiological Control of Durational Differences between Vowels Preceding Voiced and Voiceless Consonants in English. Lawrence J. Raphael. Journal of Phonetics (in press).

*Speech Recognition Through Spectrogram Matching. Frances Ingemann and Paul Mermelstein. Presented at the 88th meeting of the Acoustical Society of America, St. Louis, Mo., 4-8 November 1974; to be published in the Journal of the Acoustical Society of America.

*The Tones of Central Thai: Some Perceptual Experiments. Arthur S. Abramson. To be published in Studies in Tai Linguistics, ed. by Jimmy G. Harris and James Chamberlain (Bangkok: Central Institute of English Language).

*Phonetic Segmentation and Recoding in the Beginning Reader. Isabelle Y. Liberman, Donald Shankweiler, Alvin M. Liberman, Carol Fowler, and F. William Fischer. To be published in Reading: The CUNY Conference (tentative title), ed. by A. S. Reber and D. Scarborough (New York: Erlbaum Associates).

The following two papers were presented at the Speech Communication Seminar, Stockholm, Sweden, 1-3 August 1974; and will be published in the conference proceedings (Stockholm: Almqvist and Wiksell; and New York: Wiley):

*Mechanisms of Duration Change. K. S. Harris

*Effect of Speaking Rate on Stop Consonant-Vowel Articulation. T. Gay and T. Ushijima.

Speech and the Problem of Perceptual Constancy. Donald Shankweiler, Winifred Strange, and Robert Verbrugge. To be published in Cognition, Knowledge, and Adaptation (tentative title), ed. by John Bransford and Robert Shaw (Potomac, Md.: Erlbaum Associates).

Vowel and Nasal Duration as Cues to Voicing in Word-Final Stop Consonants: Spectrographic and Perceptual Studies. Lawrence J. Raphael, M. F. Dorman, Frances Freeman, and Charles Tobin. Journal of Speech and Hearing Research (in press). [Also in SR-37/38 (1974), 255-270.]

Hemispheric Lateralization for Speech Perception in Stutterers. M. F. Dorman and R. J. Porter, Jr. Cortex (in press). [Also in SR-37/38 (1974), 117-121.]

Perception of Temporal Order in Vowel Sequences With and Without Formant Transitions. M. F. Dorman, James E. Cutting, and Lawrence J. Raphael. Journal of Experimental Psychology: Human Perception and Performance (in press) 1. [Revision of SR-37/38 (1974), 217-224.]

Categories and Boundaries in Speech and Music. James E. Cutting and Burton S. Rosner. Perception and Psychophysics (in press) 17. [Also in SR-37/38 (1974), 145-157.]

Two Left-Hemisphere Mechanisms in Speech Perception. James E. Cutting. Perception and Psychophysics (in press) 17.

Orienting Tasks Affect Recall Performance More Than Subjective Impressions of Recall Ability. James E. Cutting. Psychological Reports (in press).

Aspects of Phonological Fusion. James E. Cutting. Journal of Experimental Psychology: Human Perception and Performance (in press) 1.

The Elastic Syllable: An Acoustic View of the Stress-Intonation Link. J. H. Gaitenby. Presented at the 88th meeting of the Acoustical Society of America, St. Louis, Mo., 4-8 November 1974; Journal of the Acoustical Society of America, Supplement (Fall 1974) 56, S32(A). (To appear in SR-41.)

The Control of Pharyngeal Cavity Size for English Voiced and Voiceless Stops. Fredericka Bell-Berti. Journal of the Acoustical Society of America (in press).

Electromyographic Study of the Velum During Speech. T. Ushijima and H. Hirose. Journal of Phonetics (in press). [Also in SR-37/38 (1974), 79-97.]

Laryngeal Activity Accompanying the Moment of Stuttering: A Preliminary Report of EMG Investigations. Frances J. Freeman and Tatsujiro Ushijima. Journal of Fluency Disorders (in press). [Also in SR-37/38 (1974), 109-116.]

Velopharyngeal Function. Fredericka Bell-Berti. To be published in the Proceedings of the 2nd Annual Hayes Martin Conference on Vocal Tract Dynamics.

Effects of Mandibular Block on the Articulation of Four-Year-Olds. Gloria J. Borden. Language and Speech (in press).

Tongue Musculature and the Feature of Tension in English Vowels. L. J. Raphael and F. Bell-Berti. Phonetica (in press).

A Continuum of Lateralization for Speech Perception? Donald Shankweiler and Michael Studdert-Kennedy. Brain and Language (in press).

Auditory and Phonetic Levels of Processing in Speech Perception: Neurophysiological and Information-Processing Analyses. Charles C. Wood. Journal of Experimental Psychology: Human Perception and Performance (in press) 1. [Also in SR-35/36 (1973), thesis supplement paged separately.]

Failure of Selective Attention to Phonetic Segments in Consonant-Vowel Syllables. Charles C. Wood and Ruth S. Day. Perception and Psychophysics (in press).

*Linguistic and Nonlinguistic Stimulus Dimensions Interact in Audition but not in Vision. James E. Cutting

*Results of a VCV Spectrogram-Reading Experiment. G. M. Kuhn and R. McI. McGuire.

*Laryngeal Muscle Activity, Subglottal Air Pressure, and the Control of Pitch in Speech. Rene Collier.

*Evidence for Spectral Fusion in Dichotic Release from Upward Spread of Masking. Terrance M. Nearey and Andrea G. Levitt

*A Preliminary Elect omyographic Study of Labial and Laryngeal Muscles in Danish Stop Consonant Production. Eli Fischer-Jørgensen and Hajime Hirose.

*A Note on Laryngeal Activity in the Danish "stød." Eli Fischer-Jørgensen and Hajime Hirose.

Reports and Oral Presentations

*Word Recall in Aphasia. Diane Kewley-Port. Presented at the 87th meeting of the Acoustical Society of America, New York, April 1974.

The Tones of Central Thai: Some Perceptual Experiments. Arthur S. Abramson. Symposium on Tai Linguistics, Central Institute of English Language, Bangkok, Thailand, 16-26 April 1974.

*Jaw Movements During Speech: A Cinefluorographic Investigation. T. Gay. Presented at the Eighth International Congress on Acoustics, London, July 1974.

The Effects of Pitch on the Perception of Consonant Voicing in Thai: The Plausibility of Certain Historical Hypotheses. Arthur S. Abramson. Presented to the Siam Society, Bangkok, Thailand, 23 July 1974.

Voicing Distinctions in Thai Stop Consonants: Production and Perception. Arthur S. Abramson. Presented to the Faculty of Humanities, Ramkhamhaeng University, Bangkok, Thailand, 24 July 1974.

On Finding One's Way From Phonetic Text to Spoken Words and Back. F. S. Cooper and P. Mermelstein. Invited talk presented at the American Society for Information Science Conference, Atlanta, Ga., 14-17 October 1974.

Is it VOT or a First-Formant Transition Detector? Leigh Lisker. Presented at the Annual Meeting of the American Association of Phonetic Sciences, St. Louis, Mo., 5 November 1974. (To appear in SR-41.)

Amplitude Plus f_o in a Tone Language: Thai Perception. Arthur S. Abramson. Presented at the Annual Meeting of the American Association of Phonetic Sciences, St. Louis, Mo., 5 November 1974.

A Combined Cinefluorographic-Electromyographic Study of the Tongue During the Production of /s/: Preliminary Observation. Gloria J. Borden and Thomas Gay. Presented at the American Speech and Hearing Association meeting, Las Vegas, Nevada, 5-8 November 1974.

Pitch in the Perception of Voicing States in Thai: Diachronic Implications. Arthur S. Abramson. Presented at the Annual Meeting of the Linguistic Society of America, New York, N. Y., 27-30 December 1974.

APPENDIX

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers:

SR-21/22 to SR-37/38

Status Report		DDC	ERIC
SR-21/22	January - June 1970	AD 719382	ED-044-679
SR-23	July - September 1970	AD 723586	ED-052-654
SR-24	October - December 1970	AD 727616	ED-052-653
SR-25/26	January - June 1971	AD 730013	ED-056-560
SR-27	July - September 1971	AD 749339	ED-071-533
SR-28	October- December 1971	AD 742140	ED-061-837
SR-29/30	January - June 1972	AD 750001	ED-071-484
SR-31/32	July - December 1972	AD 757954	ED-077-285
SR-33	January - March 1973	AD 762373	ED-081-263
SR-34	April - June 1973	AD 766178	ED-081-295
SR-35/36	July - December 1973	AD 774799	ED-094-444
SR-37/38	January - June 1974	AD 783548	ED-094-445

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
Leasco Information Products, Inc.
P. O. Drawer O
Bethesda, Maryland 20014

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) Haskins Laboratories, Inc. 270 Crown Street New Haven, Connecticut 06510	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP N/A

3. REPORT TITLE

Haskins Laboratories Status Report on Speech Research, No. 39/40, July-December 1974

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; Franklin S. Cooper, P.I.

6. REPORT DATE February 1975	7a. TOTAL NO. OF PAGES 280	7b. NO. OF REFS 490

8. CONTRACT OR GRANT NO. NIDR: Grant DE-01774 NICHD: Grant HD-01994 VA/PSAS Contract V101(134)P-71 ONR Contract N00014-67-A-0129-0001 ARPA/ONR Contract N00014-67-A-0129-0002 NICHD Contract NIH-71-2420 NIH/GRS: Grant RR-5596	9a. ORIGINATOR'S REPORT NUMBER(S) SR-39/40 (1974) 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 N/A	12. SPONSORING MILITARY ACTIVITY See No. 8

13. ABSTRACT

This report (1 July - 31 December 1974) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. Manuscripts cover the following topics:

- Speech Perception
- Speech Recognition Through Spectrogram Matching
- Results of VCV Spectrogram-Reading Experiment
- Evidence for Spectral Fusion in Dichotic Release from Upward Spread of Masking
- Tones of Central Thai: Perceptual Experiments
- Phonetic Segmentation and Recoding--Beginning Reader
- Word Recall in Aphasia
- Linguistic and Nonlinguistic Stimulus Dimensions--Interaction in Vision
- Pitch Control in Speech--Laryngeal Muscle Activity, Subglottal Air Pressure
- Cinefluorographic Study of Vowel Production
- Mechanisms of Duration Change
- Physiological Control of Durational Differences between Vowels Preceding Voiced and Voiceless Consonants in English
- Effect of Speaking Rate on Stop Consonant-Vowel Articulation
- Jaw Movement During Speech: Cinefluorographic Investigation
- EMG Study of Labial and Laryngeal Muscles in Danish Stop Consonant Production
- Laryngeal Activity in the Danish "stød"

*This document contains no information not freely available to the general public. It is distributed primarily for library use.

UNCLASSIFIED
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech Perception						
Speech Recognition						
Speech Spectrograms - Matching						
Speech Spectrograms - Reading						
Dichotic Listening - Fusion						
Tone Perception - Thai						
Reading - Phonetic Activity						
Word Recall-- Aphasics						
Visual Perception						
Pitch Control in Speech						
Vowel Production - Cine Study						
Vowel Duration - EMG Study						
Speaking Rate - EMG Study						
Speech Production - Jaw Movements						
EMG Studies - Danish						