ED 071 484 FL 003 587

TITLE	Status Report on Speech Research, No. 29/30, January-June 1972.
INSTITUTION	Haskins Labs., New Haven, Conn.
SPONS AGENCY	National Inst. of Child Health and Human Development (NIH), Bethesda, Md.; National Inst. of Dental Research (NIH), Bethesda, Md.; Office of Naval Research, Washington, D.C. Information Systems Research.
REPORT NO	SR-29-30-1972
PUB DATE	Jul 72
NOTE	158p.
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	Acoustic Phonetics; Articulation (Speech); Auditory Discrimination; Auditory Perception; Blind; Child Language; *Electronic Equipment; *Experiments; Language; *Language Research; Language Skills; Listening; Phonetics; Phonology; Physiology; Reading Research; *Research Methodology; Spectrograms; *Speech

ABSTRACT

This report 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 and extended reports cover the following topics: iconic storage, voice-timing perception, oral anesthesia, laryngeal function, electromyography of speech production, encodedness and right ear effect, interference, dichotic listening, speech and reading, reading machines for the blind, speech synthesis by rule, auditory evoked potentials, infant production and perception of speech sounds, voice onset time, perception of speech and nonspeech, lag effect, and phonetic coding in Japanese. A list of publications and reports is included. (Author/MD)

SR-29/30 (1972)

SPEECH RESEARCH

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

1 January - 30 June 1972

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.)

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, Conn. 06510

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP
N/A

3. REPORT TITLE
Status Report on Speech Research, no. 29/30, January-June 1972.

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
July 1972

7a. TOTAL NO OF PAGES
149

7b. NO OF REFS
199

8a. CONTRACT OR GRANT NO.
ONR Contract N00014-67-A-0129-0001
NIDR: Grant DE-01774
NICHD: Grant HD-01994
NIH/DRFR: Grant RR-5596
NSF: Grant GS-28354
VA/PSAS Contract V101(134)P-71
NICHD Contract NIH-71-2420

9a. ORIGINATOR'S REPORT NUMBER(S)
SR-29/30 (1972)

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 (for 1 January-30 June) 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 and extended reports cover the following topics:

-Some Aspects of Selective Readout from Iconic Storage
-Voice-Timing Perception in Spanish Word-Initial Stops
-Some Effects of Oral Anesthesia upon Speech: An Electromyographic Investigation
-Laryngeal Control in Vocal Attack: An Electromyographic Study
-A Parallel Between Encodedness and the Magnitude of the Right Ear Effect
-Mutual Interference Between Two Linguistic Dimensions of the Same Stimuli
-The Phi Coefficient as an Index of Ear Differences in Dichotic Listening
-The Relationships Between Speech and Reading
-Audible Outputs of Reading Machines for the Blind
-Field Evaluation of an Automated Reading System for the Blind
-Word and Phrase Stress by Rule for a Reading Machine
-Auditory Evoked Potential Correlates of Speech Sound Discrimination
-Short-Term Habituation of the Infant Auditory Evoked Response
-Early Apical Stop Production: A Voice Onset Time Analysis
-The Discrimination of Speech and Nonspeech Stimuli in Early Infancy
-The Effect of Delayed Channel on the Perception of Dichotically Presented Speech and Nonspeech Sounds
-Phonetic Coding of Kanji

*This document contains no information not freely available to the general public. It is distributed primarily for library use.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Iconic storage						
Visual information processing						
Electromyography of speech production						
Oral anesthesia						
Laryngeal function						
Dichotic listening						
Reading and speaking						
Reading machines for the blind						
Speech synthesis by rule						
Auditory evoked potentials						
Categorical perception						
Infant production and perception of speech sounds						
Voice onset time						
Lag effect						
Perception of speech and nonspeech						
Phonetic coding in Japanese						

ACKNOWLEDGMENTS

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

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

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

National Science Foundation
Grant GS-28354

National Institutes of Health
General Research Support Grant RR-5596

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

CONTENTS

Some Aspects of Selective Readout from Iconic Storage

M. T. Turvey[*]
Haskins Laboratories, New Haven

Two experiments are reported which examined the delayed partial sampling of tachistoscopically presented displays. The first experiment compared partial report by row and partial report by color for displays of colored discs. A significant interaction was observed between selection criterion and delay of report, and partial report for both criteria was superior to whole report. These results for the arrays of colored discs were replicated in the second part of Experiment II, the first part of which also showed, however, that when the displays consisted of colored letters the decline in accuracy of partial report by row paralleled that of partial report by color. The results are discussed in terms of the distinction between preattentive and focal-attentive processes. More generally the two experiments are presented as supporting the hypothesis that performance in iconic memory tasks is jointly determined by iconic storage and short-term storage.

The notion of transient storage of visual material prior to categorization has assumed a central role in recent constructivist (Neisser, 1967) and information-processing discussions (Broadbent, 1971; Haber, 1969; Turvey, 1971) of visual perception. Neisser (1967) has suggested the term "iconic" for this kind of brief memory. The idea of an early buffer memory in perceptual systems is, of course, not new; Broadbent (1958) and Pollack (1959), for example, had pointed earlier to the theoretical need for such a concept in the analysis of auditory perception. Moreover, Woodworth (1938) essentially presaged our present conceptualizations of the iconic store:

> The primary memory image has less of a definite quality than the visual after-image and is distinguished by the fact that it does not move with the eye, but remains stationary in the place where the objects were exposed....[T]hese after-images... allow a few seconds extra for the cerebral response to the data supplied by the retina. (p. 692)

[*] Also University of Connecticut, Storrs.

Acknowledgment: The research reported here was made possible in part by support from the University of Connecticut Research Foundation Grant No. 104. The author wishes to acknowledge the assistance of David Carmichael with Experiment I, and Shari Kravetz, Edith Sullivan, and Robert Weeks who assisted with Experiment II.

Of course, the degree to which such a stimulus representation is evident varies with the perceptual task under examination. Brief, precategorical storage is more likely to be observed in tasks where an overload of items is presented for a limited duration, or where several items are presented simultaneously on a number of different channels, or where the relevant response categories are delayed (see Posner, 1963).

The procedure most favored for isolating and examining iconic storage is the delayed partial-sampling paradigm introduced by Sperling (1960), and Averbach and Coriell (1961). Essentially, the paradigm consists of displaying tachistoscopically a number of items, usually letters or digits, in excess of the memory span and following this display after a brief interval by an instruction to report a part of the display. The interesting feature of this paradigm is that this selective instruction, provided that it is given within milliseconds after the display, may give a measure of item availability superior to that obtained in a noninstructed case where S reports as many items as possible. The instructed-noninstructed difference permits the inference of a large-capacity store; the precipitous reduction of this difference with delay of instruction permits the inference of rapid decay.

In view of current theorizing on the interaction between memory systems, it is probably advisable to adopt the same attitude toward the delayed partial-sampling. or iconic memory (IM), paradigm that we have adopted toward the short-term memory (STM) distractor and probe paradigms. In short, it is argued, contrary to earlier positions, that data obtained from STM tasks are never pure indicants of the hypothesized short-term storage (STS) mechanism for categorized material. The present view is that the probability of recalling an item in a STM task is determined by the presence of the item in STS, or by its presence in long-term storage (LTS), or by both (Atkinson and Shiffrin, 1968; Waugh and Norman, 1965). Thus, by the same token, accuracy in reporting an item in an IM task is determined by the presence of a representation of that item in iconic storage (IS), or by a representation in STS, or by both.

This view of the IM task has been expressed explicitly by Averbach and Coriell (1961), who argued that performance in the delayed partial-sampling paradigm is the result of two different types of performance on the part of S. One is a nonselective readout, independent of the appearance of the instruction cue; the other is a selective readout, which occurs only subsequent to the decoding of the instruction. Nonselective readout is suggested by the fact that performance never appears to approach zero in delayed partial-sampling experiments; instead it asymptotes at the level of noninstructed, or whole, report. Therefore, we have to assume that S begins to categorize material and enter it into STS as soon as possible, at least before the instruction cue. On occurrence of the cue, some of the designated material may have been processed already; just how much depends on the size of the display and the overlap between preselected and cued items. In any event a S's cued report in an IM task can be based, in part, on STS, where STS is viewed as consisting of both an abstract visual code and a name code (see Coltheart, 1972).

The need for emphasizing that data obtained from IM tasks are not necessarily pure indicators of IS will become apparent in the two experiments reported here which took as their departure point the experiments of Clark (1969).

EXPERIMENT I

Clark (1969) investigated IM for three-by-five matrices of intermixed colored discs using the selection criteria of location and color. Instructions to report by color asked S to designate the locations in the matrix occupied by, say, red discs. Instructions to report by location, on the other hand, required that S specify the colors of the discs which occurred in the five locations of, say, the bottom row. For both means of accessing IM, partial report was significantly superior to whole report, i.e., noninstructed report. But of special interest was Clark's finding that while the accuracy of partial report by location declined with delay of the instruction cue, accuracy of partial report by color did not. Experiment I which sought to verify this observation of Clark's compared performance with the two selection criteria in a single within-Ss design; in Clark's investigation, report by location and report by color were examined in separate experiments with different Ss.

Method

Subjects. The Ss were four undergraduates at the University of Connecticut who participated in the experiment as a course requirement.

Stimulus materials and apparatus. Discs were outlined on sheets of red, green, and yellow plastic. These were then cut out and placed onto a white background for photographing. Forty-eight slides were made, each with twelve colored discs, four each of red, green, and yellow, arranged in three rows of four. Each of the forty-eight, three-by-four arrays of colored discs was constructed by assigning, at random, a colored disc to a location by the procedure of selection without replacement from the set of twelve colored discs.

A Lafayette T-2K Constant Illumination Projecting Tachistoscope was used to project the slides onto a viewing screen at a distance of 50 cm from S. The field, so viewed, subtended a visual angle of 6.0 deg vertical by 8.5 deg horizontal. At this viewing distance, the diameter of each disc subtended 1.6 deg, and the separation between discs was .8 deg within a row and 1.2 deg within a column. One channel constantly illuminated at 8 ft L a pre- and post-exposure fixation field at the center of which was a faint, but discernible, cross. The slides were exposed for 80 msec at a luminance of 20 ft L. In the partial report conditions the slide display was followed by one of the following tones--2,000 Hz, 600 Hz, 200 Hz, signalling, respectively, top, middle, or bottom row, or red, green, or yellow. Four tone delays of 0, 100, 300, and 1000 msec were used. These delays were measured from display offset. The exposure duration, tone delay, and tone duration were controlled by three Hunter timers.

Procedure. The same general procedure was followed for all trials of all conditions: S was instructed to view the cross in the fixation field until it appeared in focus, at which point S pressed a key to trigger the display of a slide. Following the display S recorded his response on a response grid using a separate response grid for each trial. In the partial-report condition a tone indicator occurred at a predetermined interval after termination of the display, cuing S to report items by row or color. In the whole-report condition no tone occurred; on termination of the display S attempted to report as many items as possible. The responses of S on each trial were scored for the number of discs reported in their correct positions. In partial report, the average proportion of discs correctly reported was taken as an estimate of the proportion of the whole display available to S, i.e., an estimate of the proportion of all the locations in the matrix for which S had correct color information.

Three days of practice on the IM task preceded the experiment proper. The purpose of the three practice days was to acquaint S with the general procedure to provide practice in discriminating between the three tones and, most important, to insure familiarity with the cuing functions of the tones. On the average, practice sessions lasted 1 to 1-1/2 hours and included a total of 100 trials divided into two blocks of 20 trials of whole report and four blocks of 20 trials of partial report (two blocks of report by row and two of report by color). At the end of each trial on Days 1 and 2, S was given feedback on the accuracy of his performance. The experiment proper was conducted on Days 4 and 5. Two Ss on Day 4 were given twelve trials of whole report followed by ninety-six trials of partial report by row, twenty-four trials at each of the four intervals randomly interspersed in the series of ninety-six, and finally twelve more trials of whole report. The other two Ss received the same number, and order, of whole- and partial-report trials but with partial report by color. Day 5 followed the same procedure with Ss receiving the partial-report condition that they had not received on Day 4.

Before each session on both days Ss were given several sequences of the tones to insure that they could reliably identify which was the high, which was the middle, and which was the low. And prior to each block of partial-report trials S was given practice in identifying the cue function of the tones relevant to that partial-report condition.

Throughout the training and experiment days, the tone-row and tone-color combinations were counterbalanced so that each tone specified each row and each color an equal number of times at each delay interval. Within blocks of partial-report trials the tones were randomized.

Results and Discussion

Essentially, the logic behind this type of experiment is that if a selection criterion is efficient the whole-partial difference should be significant. We may suppose that efficient selection criteria, so defined, reflect the character of the iconic store. Either they identify those properties of stimulation that are already available or they point to those properties which can be most rapidly ascertained at this stage in the flow of visual information in the nervous system.

The results of the experiment are shown in Figure 1. A repeated-measures analysis of variance was conducted on the proportion of items reported in the whole-report and the 0-sec-delay partial-report conditions. This difference was significant: $F(2,6) = 18.21$, $p < .001$. A second analysis was performed on the total number of discs correctly repcrted by each S at each indicator delay for both kinds of selection. This analysis showed that the main effect of delay, $F(3,9) = 7.74$, $p < .01$, and the interaction between selection criterion and indicator delay, $F(3,9) = 5.33$, $p < .025$, were significant but selection criterion as a main effect was not, $F(1,3) = 1.06$, $p > .05$.

In the main these results confirm the major findings of Clark: color and location were both efficient selection criteria and the temporal course of report

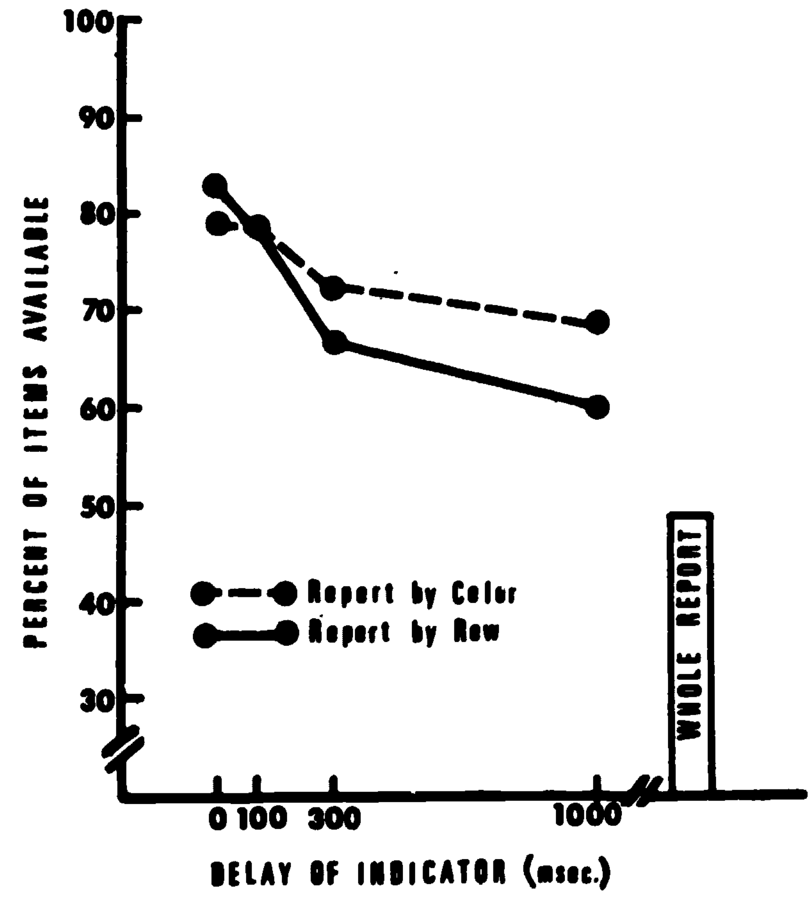


Figure 1: Partial report by color and by row of disc arrays as a function of indicator delay in Experiment I.

by color differed from that of report by location. The present experiment, however, did not show that report by location was superior to report by color, suggesting that the superiority of location evident in Clark's data was probably due to between-Ss differences. In addition, Clark reported that report by color was invariant with indicator delay but a separate analysis of the color selection data of the present experiment revealed a significant decline in performance across delay intervals, $F(3,9) = 4.45$, $p < .05$.

There are at least three hypotheses which speak to the question of why different IM functions were obtained with report by location and by color. One hypothesis proposes that the decay rate of the iconic material differed for the two modes of selection. This possibility seems unlikely given that the same set of stimulus features, color and location, was required for report in both selection modes. Moreover this hypothesis contradicts the view that the decay parameter of IS is a structural property of the system and, in keeping with the distinction drawn by Atkinson and Shiffrin (1968), is not modifiable by control processes available to S (cf., Doost and Turvey, 1971).

Another hypothesis views the different IM functions as the result of the different types of uncertainties operating in the two conditions. In report by color, item uncertainty is zero but spatial uncertainty is high; in report by row, on the other hand, item uncertainty is higher than spatial uncertainty. In view of the slight but nonsignificant tendency for color selection to be better than row selection it might have to be argued, on this hypothesis, that item uncertainty is more detrimental than spatial uncertainty in the present IM task. However, it is far from clear how these differences in item and spatial uncertainty would produce the obtained interaction between selection criterion and indicator delay.

A third hypothesis, the one favored here, derives from two distinctions: the distinction drawn by Neisser (1967) between preattentive and focal-attentive processes and the distinction made earlier between IM and IS.

Preattentive processes are viewed as relatively crude preliminary operations that segregate the optical array into units which are then acted upon by focal attention, a process which makes extensive contact with LTS and is essential for pattern recognition. In the present experiment, report by location requested S to name the colors of the discs in each of a specified set of locations. Report by color, on the other hand, requested S to specify the location occupied in the matrix by discs of a designated color. The requirement to name the disc colors in report by location suggests the involvement of focal attention; by contrast, performance in the report-by-color condition could have been mediated primarily, if not solely, by the products of preattentive mechanisms. All that is needed in the report-by-color condition is that the elements in the matrix be segregated into patterns by color, which is not an unreasonable demand since preattentive processes are analogous to the modes of perceptual organization described by Gestalt psychology (Neisser, 1967). The organizing principle invoked here is that of grouping according to similarity. Given the resolution of the array into color patterns, S may now simply enter these patterns into STS without awaiting the partial-report instruction. Indeed, only two patterns need be entered--which, of course, is well within the limits of STS capacity--since the disc locations of the remaining color class could be remembered by elimination. [An interpretation of this kind was proposed by Keele and Chase (1967) for the failure of Eriksen and Steffy (1964) to obtain a decline in partial-report accuracy with indicator delay when the stimuli were six-item binary arrays.] On this interpretation the advantage of partial report by color over whole report lies in the reduction of information requiring rehearsal, i.e., central processing capacity (Posner, 1966), or in the reduction of output interference.

The idea, therefore, is that performance in the present IM task with color as the selection criterion was determined primarily by STS. Of course entering patterns into STS was an option open to S in report by location; in that case, however, encoding the stimulus in this way would not be especially useful since reporting the colors of a row would require a relatively complicated decoding operation. In short, the difference between the two selection modes, on this view, is that report by color involves only preattentive processes and was relatively more dependent on STS than on IS, while report by location required focal-attentive operations and had to depend more on the less persistent IS representation.

At first blush it might seem that focal attention or figural synthesis (Neisser, 1967) is the sine qua non for determining the establishment of more persistent modes of representation, i.e., for effecting the translation from IS into forms suitable for storage in STS and LTS. There are, however, several reasons for doubting this.

Most notable is the fact that one can remember certain gross characteristics of a visual event some considerable time after the event has occurred and without having known more detailed or categorical properties of the event. For example, one can remember that something occurred, without ever having known what that something actually was, or one can remember that a particular location was occupied, without having known the identity of the occupant. In delayed partial-sampling experiments Ss may have a rough idea of how many items were presented in the array without knowing what they were (Eriksen and Rohrbaugh, 1970). In dichotic listening experiments Ss can report, after a relatively lengthy delay, the voice quality of an unattended message but not know the semantic content of the message (e.g., Cherry, 1953). In brief, the products of preattentive processes like those of focal-attentive processes can enjoy the privileges of post-iconic, categorical stores.

EXPERIMENT II

If the interpretation of the selection criteria by delay interaction of Experiment I is basically correct, then it should be possible to eliminate this interaction by requiring focal-attentive processes in both report by location and report by color. In the second experiment, report by location and by color were examined, with letters as the to-be-reported items. Report by location asked: what were the letters in these locations? While report by color asked: where were the letters of this color and what were they? In the selection-by-color condition of this second experiment, each location occupied by an object of a particular color would have to be examined in order to synthesize/identify the object's name. Thus, the selection-by-color condition of the second experiment differed from that of the first in that focal attention was needed to produce the required response. Entering the color patterns into STS would give little advantage since the process of naming would have to make use of the content of IS. In brief, IM performance with letter arrays for both selection criteria in Experiment II should be determined primarily by IS.

Experiment II was conducted in the same manner as Experiment I and included a replication of Experiment I for purposes of comparison. The experiment was conducted in two parts.

Method

Subjects. The Ss were three University of Connecticut graduate students, who volunteered their services, and the author. The same four Ss participated in both parts of the experiment.

Stimulus materials and apparatus. Using the same method and materials used for making the disc slides, forty-eight new slides were made each with twelve colored letters, four each of red, green, and yellow arranged in three rows of four. The twelve letters were: C, F, H, J, L, N, P, S, T, U, X, Z, and no letter was repeated within a slide. Because of the large number of slides required for complete counterbalancing of letter, color, and location, the following procedure was used. For any given slide twelve letters were assigned at random to the twelve locations by the method of selection without replacement from the set of twelve letters. A color was then randomly assigned to a letter in a location by a similar procedure. The letters in the display subtended 1.6 deg vertical and on average 1.2 deg horizontal. The average separation between columns was .8 deg and between rows it was 1.2 deg. The apparatus, tones, delay intervals, and all other viewing measurements were the same as those of Experiment I. The three-by-four disc slides used in the second part of the experiment were the same as those described previously in Experiment I.

Procedure. The two parts of the experiment were conducted over seven days with the first three days as training days. On Days 4 and 5 the experiment proper was conducted with the letter slides (Part I). Days 6 and 7 were used to run the Experiment I replication with the disc slides (Part II). The procedure used over these seven days followed the pattern outlined in Experiment I, except that all training was done on the letter slides.

Results and Discussion

Each S on each trial of both Parts I and II was scored for the number of items reported in their correct location. In the partial-report conditions, the average percentage of items correctly reported in a cued row, or of a cued color, was taken as the estimate of the total number of locations in the array for which S had correct letter (Part I) or color (Part II) information. Figure 2 shows the relation between these percentages and indicator delay for both selection criteria; also included are the whole-report means for both Parts I and II. Inspection of Figure 2 lends support to the hypothesis under test: the relation between the two selection criteria in Part I differed fundamentally from that in Part II. A repeated-measures analysis of variance performed on the Part I data showed that report by location was superior to report by color, $F(1,3) = 54.74$, $p < .01$, and that the main effect of indicator delay was significant, $F(3,9) = 64.10$, $p < .001$; however, there was no significant interaction between the selection-criteria functions ($F < 1$). Quite to the contrary was the outcome of the same kind of analysis of the Part II data which showed that the selection criterion by delay interaction was significant $F(3,9) = 15.14$, $p < .001$. Also, although the difference between report by color and report by location was not significant, $F(1,3) = 7.15$, $.05 < p < .10$, inspection of Figure 2 suggests

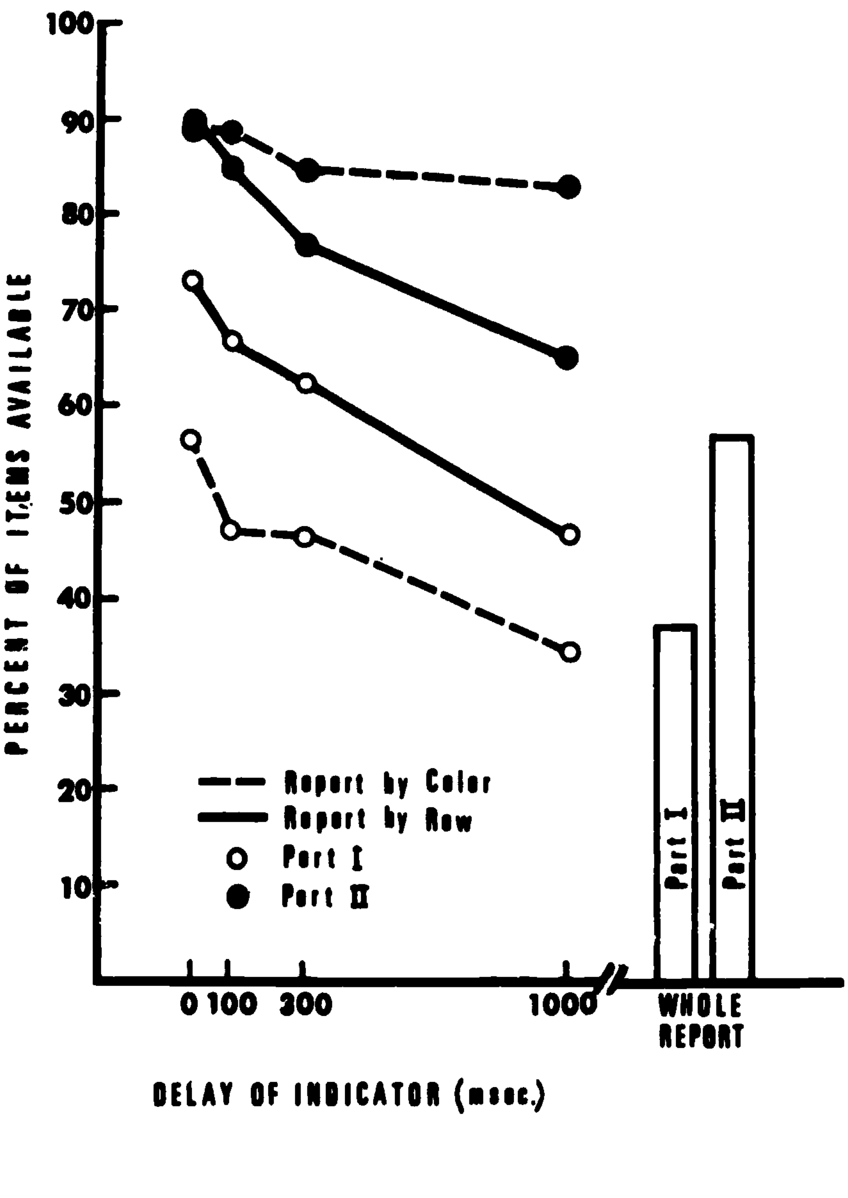


Figure 2: Partial report by color and by row of letter (Part I) and disc (Part II) arrays as a function of indicator delay in Experiment II.

that, if anything, color was superior to location as a selection criterion, contrary to Part I. And of course, as inspection of Figure 2 further suggests, the main effect of indicator delay was significant, $F(3,9) = 16.52$, $p < .001$. All of these results for the disc arrays replicate those of Experiment I.

Two further separate analyses were conducted. One compared partial report of letters by color at 0-msec delay with the whole report for the letter arrays and found a significant difference, $F(1,3) = 15.93$, $p < .05$. The other showed that accuracy of report by color in Part II declined significantly with delay of indicator, $F(3,9) = 4.21$, $p < .05$.

In sum, these analyses lend support to the interpretation that disc selection by color was conducted differently from disc selection by location or letter selection by either criterion. The hypothesis advanced above argues that performance in the latter three conditions required focal attention and, therefore, was more dependent on IS, while performance in the former condition did not require focal attention and relied for the most part on a representation in STS.

In Part I of the present experiment selection by location was superior to selection by color, and the magnitude of the difference between the two was relatively constant across all delays of indicator. Both conditions involved high item uncertainty but spatial uncertainty was pronounced only in the selection-by-color condition, and this might have accounted for the difference in performance between the two conditions (see Bennett, 1971, for a discussion of item and spatial uncertainty in IM tasks). We may compare the situation in Part I to that existing in Part II. There, item uncertainty was limited to selection by location, and selection by color involved only spatial uncertainty. Although the difference was not significant, selection by color tended to be superior to selection by location, an observation corroborated by Experiment I. What this implies is that spatial uncertainty _per se_ could not have accounted for the inferior selection-by-color performance in Part II of the present experiment. More probably the inferior performance was due to spatial uncertainty coupled with item uncertainty and the resulting extra or at least different demands on processing that this coupling brought about. We might suppose, as we did above, that the presence of item uncertainty in the selection-by-color condition of Part I (as opposed to the selection-by-color condition with the disc arrays in Part II) prohibited _S_ from taking advantage of the patterns entered into STS following preattentive segregation. Indeed, entering the segregated color patterns into STS may well be prohibitive under the task demands of identifying letters.

GENERAL DISCUSSION

The present paper has suggested that both IS and STS have to be considered in the analysis of IM performance. Recently in a series of papers by Holding (1970, 1971) and Dick (1971) the existence, or at least utility, of IS has been questioned and the implication has been made that IM performance is based solely on STS. Holding (1970) argued that in IM tasks where the selection criterion is row, if _S_ could predict which row would be cued and if he then fixated on the expected row, the estimate of available items derived from partial report would be inflated. In support of his argument Holding demonstrated that performance in an IM task varied systematically with the predictability of the cue. His conclusion therefore was that the concept of selection from IS was not needed co explain the difference between partial report and whole report. However, while Holding's experiments suggest caution in the construction of cue sequences when the selection criterion is row, his explanation of the partial-whole difference in terms of fixation strategies cannot apply to IM situations where nonspatial selection criteria such as color, size, brightness, or shape are used (e.g., Turvey and Kravetz, 1970; von Wright, 1968, 1970).

Evidence that Ss are using different strategies, perhaps relying differently on IS and STS under conditions of spatial and nonspatial selection criteria, is strongly implied by Table I. Table I shows the number of times Ss in Part I of Experiment II responded with zero, one, two, three, or four letters in their correct locations as a function of cue delay for both selection by row and by color.[1] (The data on one S are not included in Table I because her overall

TABLE I

Frequencies of Response Categories for Report by Row and by Color[a]

Delay (msec)		Category				
		0	1	2	3	4
0	Row	3	6	23	9	30
	Color	6	20	24	15	5
100	Row	6	16	15	14	21
	Color	11	22	21	14	3
300	Row	6	15	20	14	17
	Color	12	20	25	11	4
1000	Row	17	22	18	4	11
	Color	18	32	15	6	1

[a] The sums of the frequencies are not identical across the rows of the table because Ss sometimes misinterpreted the tone indicator. This resulted in the occasional loss of a trial.

[1] This analysis was suggested by Joel Kleinberg.

performance was so high that most of her responses were in the four category for both selection criteria.) At all delays with row as the selection criterion, there were more perfect responses in the four category than there were responses with three items correct. Moreover, the frequency of three items correct was consistently less at all delays than the frequency of two items correct. Thus the distribut:ons at each delay were relatively normal with the exception of the four category. By comparison the distributions for selection by color show no irregularity in the four category, and at all delays the distributions are normal. Obviously Ss are not behaving in the same way in the two selection criterion conditions; selection by row does seem to take advantage of fixational or attentional biases of the kind suggested by Holding. Moreover we may conclude from inspection of Table I that partial report by row made greater use of STS than partial report by color.

In a similar vein, Holding (1970) and Dick (1971) have pointed out that because more items have to be output in whole report than in partial report, there is more output interference in the former condition than in the latter, resulting in an artifically inflated difference between the two. As we have noted, Holding's view is that IS does not exist or at least cannot be accessed, i.e., it is not useful to S. Thus partial-report performance characterizes STS just as whole-report performance does, and any difference between the two performances represents nothing more than some artifact of measurement. Against this argument, however, are two kinds of evidence which show that partial report and whole report do not reflect completely identical memorial representations. In the first place there are the data of Averbach and Sperling (1961) and Keele and Chase (1967) which show that luminance conditions affect partial report but do not affect whole report. In the second place Sharf and Lefton (1970) have shown that an after-coming pattern mask which does not impair whole-report performance at delays greater than 50 msec (cf., Sperling, 1963) impairs partial-report performance even when delayed 250 msec. In sum, there is good reason to believe that the delayed partial sampling of a visually presented array of items depends on information available in a storage medium other than STS.

REFERENCES

Atkinson, R. C. and R. M. Shiffrin. (1968) Human memory: A proposed system and its control processes. In K. W. Spence and J. T. Spence, eds., The Psychology of Learning and Motivation. Vol. 2. (New York: Academic Press).

Averbach, E. and A. S. Coriell. (1961) Short-term memory in vision. Bell Syst. tech. J. 40, 309-328.

Averbach, E. and G. Sperling. (1961) Short-term storage of information in vision. In C. Cherry, ed., Symposium on Information Theory. (London: Butterworth).

Bennett, I. F. (1971) Spatial effects in visual selective attention. Human Performance Center Technical Report No. 32, University of Michigan, Ann Arbor.

Broadbent, D. E. (1958) Perception and Communication. (New York: Pergamon).

Broadbent, D. E. (1971) Decision and Stress. (London: Academic Press).

Cherry, E. C. (1953) Some experiments on the recognition of speech with one and with two ears. J. acoust. Soc. Amer. 25, 975-979.

Clark, S. E. (1969) Retrieval of color information from the preperceptual storage system. J. exp. Psychol. 82, 263-266.

Coltheart, M. (1972) Visual information processing. In P. C. Dodwell, ed., Psychology 1972. (Baltimore: Penguin).

Dick, A. O. (1971) On the problem of selection in short-term visual (iconic) memory. Canad. J. Psychol. *25*, 250-253.

Doost, R. and M. T. Turvey. (1971) Iconic memory and central processing capacity. Percept. Psychophys. *9*, 269-274.

Eriksen, C. W. and J. W. Rohrbaugh. (1970) Some factors determining efficiency of selective attention. Amer. J. Psychol. *83*, 330-342.

Eriksen, C. W. and R. A. Steffy. (1964) Short-term memory and retroactive interference in visual perception. J. exp. Psychol. *68*, 423-434.

Haber, R. N. (1969) Information processing analyses of visual perception: An introduction. In R. N. Haber, ed., *Information Processing Approaches to Visual Perception*. (New York: Holt, Rinehart and Winston).

Holding, D. (1970) Guessing behavior and the Sperling Store. Quart. J. exp. Psychol. *22*, 248-256.

Holding, D. (1971) The amount seen in brief exposures. Quart. J. exp. Psychol. *23*, 72-81.

Keele, S. W. and W. G. Chase. (1967) Short-term visual storage. Percept. Psychophys. *2*, 383-386.

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

Pollack, I. (1959) Message uncertainty and message reception. J. acoust. Soc. Amer. *31*, 1500-1508.

Posner, M. I. (1963) Immediate memory in sequential tasks. Psychol. Bull. *60*, 333-349.

Posner, M. I. (1966) Components of skilled performance. Science *152*, 1712-1718.

Scharf, B. and L. A. Lefton. (1970) Backward and forward masking as a function of stimulus and task parameters. J. exp. Psychol. *84*, 331-338.

Sperling, G. (1960) The information available in brief visual presentations. Psychol. Monogr. *74*, (Whole No. 498).

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

Turvey, M. T. (1971) On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. Haskins Laboratories Status Report *SR-28*, 1-91.

Turvey, M. T. and S. Kravetz. (1970) Retrieval from iconic memory with shape as the selection criterion. Percept. Psychophys. *8*, 171-172.

von Wright, J. M. (1968) Selection in visual immediate memory. Quart. J. exp. Psychol. *20*, 62-68.

von Wright, J. M. (1970) On selection in visual immediate memory. Acta Psychol. *33*, 280-292.

Waugh, N. and D. A. Norman. (1965) Primary memory. Psychol. Rev. *72*, 89-104.

Woodworth, R. S. (1938) *Experimental Psychology*. (New York: Holt).

Voice-Timing Perception in Spanish Word-Initial Stops[*]

Arthur S. Abramson[+] and Leigh Lisker[++]
Haskins Laboratories, New Haven

In the general phonetic literature it is commonly stated that languages use such phonetic features as voicing, aspiration, glottalization, implosion, "tensity," etc., to distinguish consonants produced at the same supraglottal place of articulation. In previous work we have argued (Lisker and Abramson, 1971) and to some extent demonstrated (Lisker et al., 1969; Sawashima et al., 1970) that some of these features are entirely or largely explainable in terms of laryngeal control. Our view has been that the timing of events at the glottis relative to supraglottal articulation provides a simple description of how this laryngeal control is manifested (Abramson and Lisker, 1970a).[1] In our earlier work on this subject (Lisker and Abramson, 1964), we measured voice onset time (VOT) in word-initial stop consonants across a number of languages. VOT, the interval between the release of the stop and the onset of phonation as shown in spectrograms, was the simplest single measure we could find in the acoustic signal of the timing of laryngeal adjustments. The dimension proved efficacious in acoustically differentiating stop consonants in most of the languages with two, and even three, phonological categories at each place of articulation.[2]

In the present study we wanted to determine the nature of the relations between VOT as varied in synthetic speech and the labeling and discrimination behavior of Spanish speakers whose two stop categories differ phonetically from the two of English. This is a continuation of studies reported earlier (Abramson and Lisker, 1965, 1970b; Lisker and Abramson, 1970).

[*] This is a revised version of a paper given at the 83rd Meeting of the Acoustical Society of America, Buffalo, N. Y., 18-21 April 1972.

[+] Also University of Connecticut, Storrs.

[++] Also University of Pennsylvania, Philadelphia.

[1] Recent electromyographic work on laryngeal muscles lends support to this view for English consonants (Hirose and Gay, in press). A helpful schematic picture of the temporal relations is given by P. Ladefoged (1971:10).

[2] A fourth category examined, voiced aspiration, clearly involves glottal adjustments but not of the kind that is discernible on the VOT dimension. Our current electromyographic work with Hajime Hirose, however, does show that this category is distinguished from the others, at least in part, by temporal factors in the contraction of intrinsic muscles of the larynx.

To control VOT in measured increments we used the Haskins Laboratories formant synthesizer. Our basic pattern was three steady-state formants for a vowel of the type [a]. Labial, apical, and velar stop releases were simulated by means of appropriate formant transitions. We synthesized thirty-seven VOT variants ranging from 150 msec before the release to 150 msec after it. For voicing before the release (voicing lead), we used only low-frequency harmonics of the buzz source. For voice onset after release (voicing lag), the interval between release and onset of the periodic source was excited by hiss alone, with suppression of the first formant to simulate the well-known first-formant "cutback" (cf., Liberman et al., 1958). Three conditions of VOT for synthetic labial stops are shown in Figure 1. The thirty-seven VOT variants thus generated were recorded on tape in eight random orders for each place of articulation and played to a total of twelve native speakers of Latin American Spanish who, using Spanish orthography, were to identify the stimuli with their stop phonemes. Instructions were prepared in Spanish and given to the subjects to help insure that they would apply Spanish categories to the stimuli.

The twelve subjects used in the identification experiments were not dialectally homogeneous, coming from Puerto Rico and some six nations of Central and South America. To the best of our knowledge, there is not enough information about phonetic variation in the Spanish dialects of Latin America with regard to the voicing feature to help explain individual differences in our data.[3] For our part, we had too small a sampling of subjects from each of the areas represented to make any dialectological statements based on the results of our experiments. The subjects were all more or less bilingual in Spanish and English, having studied English for some years. At the time of starting the experiment, most of them had been in the United States[4] no more than one year, two of them less than six months, and one for five years. Although they varied considerably in English

[3] A search of the literature, with the much-appreciated bibliographical help of Gardiner H. London of the University of Connecticut, yields no statement describing instability of voicing in word-initial /b d g/ or, for that matter, unexpected aspiration in /p t k/. This is true of general works (e.g., Lope Blanch, 1968) and descriptions of varieties of Spanish represented in our sampling of test subjects: Argentinean (Malmberg, 1950; Vidal de Battini, 1964), Colombian (Flórez, 1964), Cuban (Lopez, 1971), Mexican (Lope Blanch, 1964; Harris, 1969), and Puerto Rican (Navarro, 1948). These authors call attention only to dialectal differences in the positional and lexical distribution of stop and fricative allophones. Harris (1969:41) affirms, at least for the cultivated speech of Mexico City, that voicing lead is "clearly audible under good acoustical conditions." Lope Blanch (1964:88) comments that in the Yucatan Spanish of Mexico, stops are glottalized because of the Mayan substratum. It is possible, of course, that certain recent trends in pronunciation have not been documented. Malcah Yaeger of the University of Pennsylvania has observed (personal communication) devoicing of /b d g/ in some regional and social dialects.

[4] Our single Puerto Rican subject had been on the mainland close to fifteen months. He may well have had much more contact with English than the others, but no effects on his Spanish were discerned.

Three Conditions of Voice Onset Time
Synthetic Labial Stops

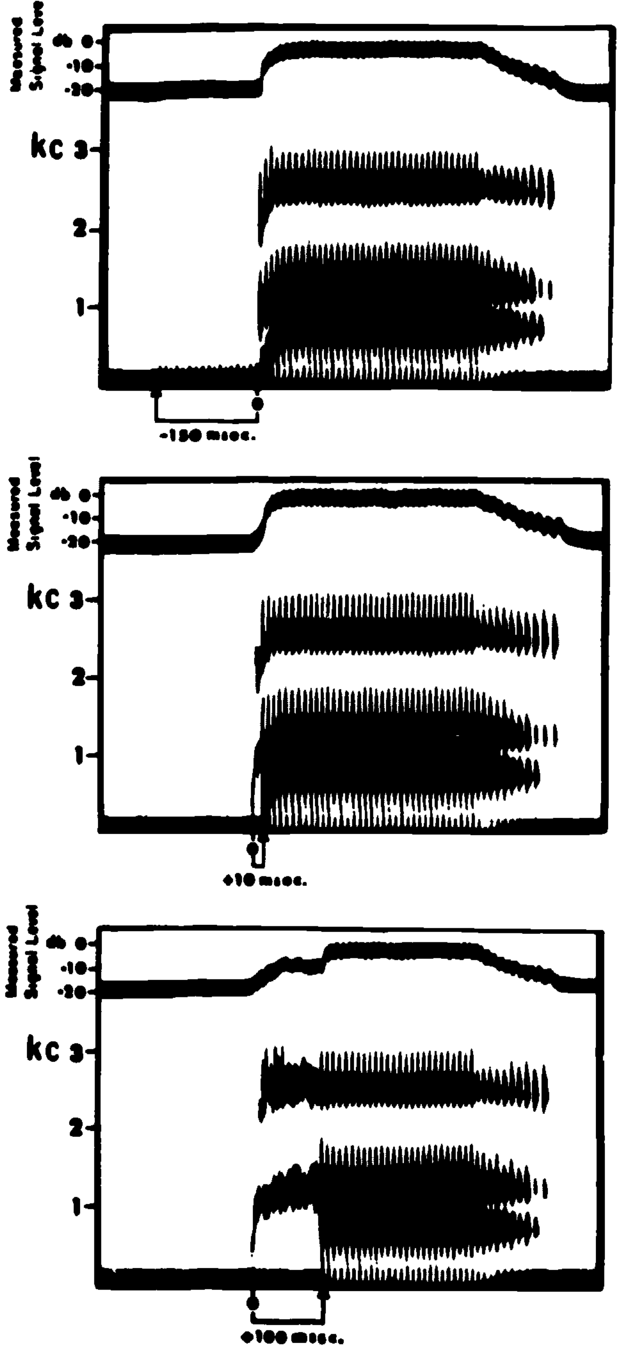


Figure 1: From top to bottom, spectrograms of voicing lead, s.ight lag, and long lag.

proficiency, all but one of them showed marked Spanish phonic and syntactic interference in their English. The one exception was an excellent bilingual with a barely detectable Spanish accent and seemingly native English grammar. To help insure against the probability of English interference in the Spanish of our subjects, we chose them with the aid of Spanish language consultants at Queens College of the City University of New York and the University of Connecticut, where the tests were run. Our screening of the subjects, done in hiring interviews by our consultants, was perhaps too superficial to rule out entirely the possibility of any phonic interference from their exposure to English, but for the very recently arrived individuals, at least, the likelihood was small.

Figure 2 gives the results of these tests. On the abscissa, negative numbers are assigned to voicing lead and positive numbers to lag, while the moment of stop release is labeled zero. The stimuli varied in 10-msec steps, except for the range of -10 to +50, where we made them in 5-msec steps. For each place of articulation, the identification curves are functions of VOT values. The synthetic patterns clearly provided enough cues for two good perceptual categories at each place of articulation. The 50 percent crossover points are given in the table below with the comparable English points, reproduced from our earlier work (Lisker and Abramson, 1970; Fig. 2) for comparison.[5]

Spanish and English Category Boundaries
in Perception of Voice Timing

(msec)

	Spanish	English
Labial	+14	+25
Apical	+22	+35
Velar	+24	+42

The Spanish perceptual crossovers have lower VOT values than the English for all places of articulation. This is consistent with the fact that English initial /p t k/ show considerable voicing lag, i.e., aspiration, in stressed syllables, while Spanish /p t k/ show little or no voicing lag and are unaspirated; furthermore, Spanish /b d g/ are characterized by voicing lead, i.e., voicing during the occlusion, whereas their English counterparts seem normally to show VOT values of about zero (Lisker and Abramson, 1964:392, 394).

[5] The 1970 study also includes VOT identification functions for the three-way voicing distinction of Thai. Perceptual data derived from tests with somewhat similar stimuli have been presented for Dutch (Slis and Cohen, 1969).

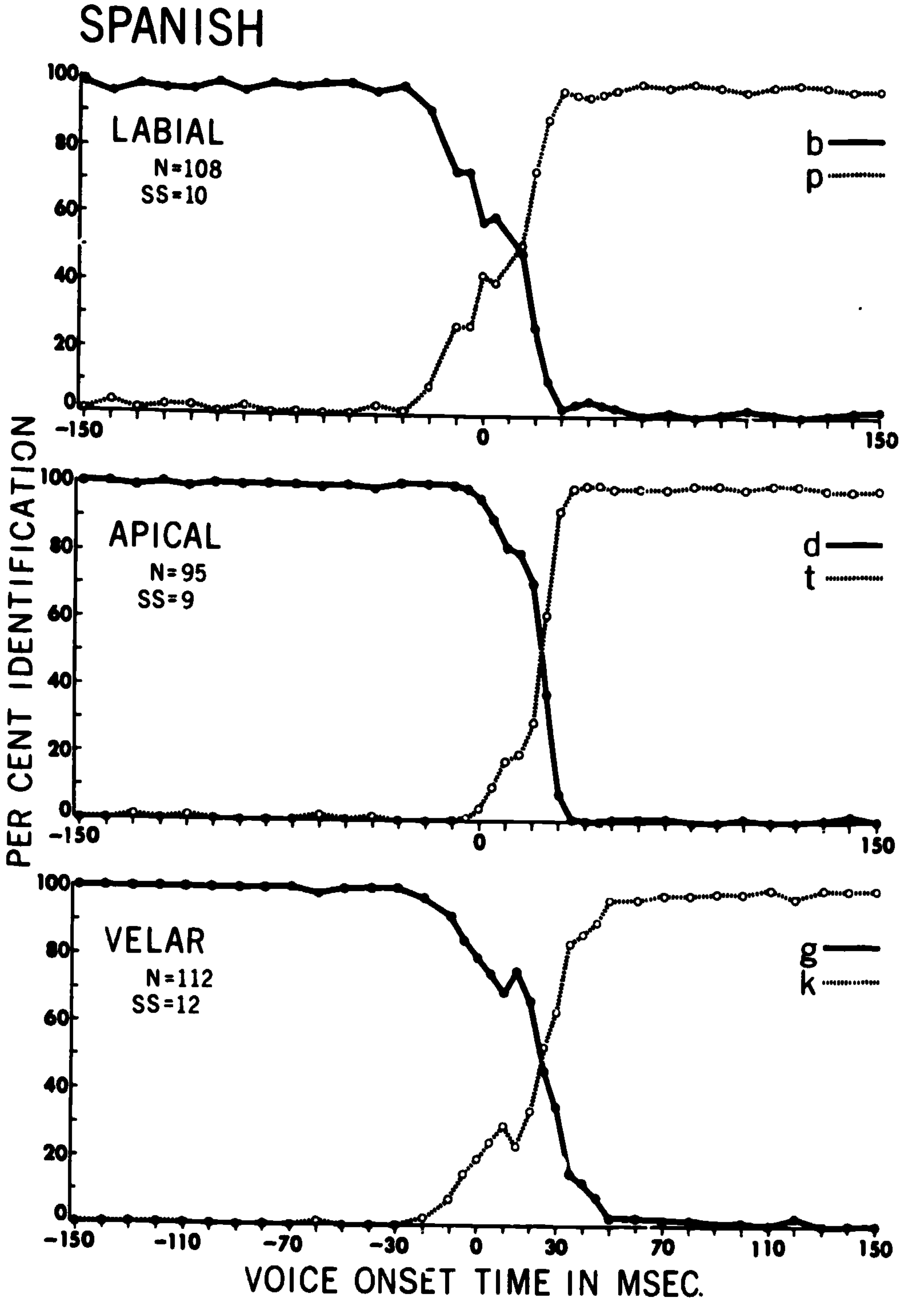


Figure 2: Perceptual identification of VOT variants by native speakers of Spanish. Pooled data.

A long-standing interest in the effects of linguistic experience upon the discriminability of variants along a phonologically relevant continuum (Liberman et al., 1957) has in recent years been investigated across languages (Stevens et al., 1969). Our own work along these lines, the testing of discriminability of VOT variants in English and Thai (Abramson and Lisker, 1970b), has been extended to Spanish in the present stuay. For this we used thirty-one of the syllables described earlier, covering the span from -150 to +150 msec in steps of 10 msec. We presented these variants in triads as an oddity task. In each triad two stimuli were identical and one was different. The task was to decide whether the odd one was in first, second, or third position. The triads were made by pairing stimuli at 2-, 3-, and 4-step intervals along the continuum, thus comparing differences of 20, 30, and 40 msec. Several permutations of the triads and randomizations of the test series were presented to some of the native speakers of Spanish who had taken the identification tests. Very few of the subjects were able to stay with the experiment over a long enough period of time to accumulate a large number of data points for each comparison;[6] therefore, we have not pooled our data but rather presented them for individual subjects and only for two places of articulation.

At the top of Figure 3 we see labial discrimination curves for all three levels of difficulty for Subject MP. Each point on a curve is placed equidistant between the two VOT values being discriminated. The line placed perpendicular to the time axis at +22 msec shows MP's 50 percent perceptual crossover point in the labial identification task. This point is almost precisely at the discrimination peak for all three levels, indicating considerable correlation with the phonological boundary. Note, however, the additional one or two small peaks for higher values of voicing lag.

At the bottom of Figure 3 we see the discrimination data for LQ. At his 50 percent crossover point, shown by the vertical line at -15 msec, there is a 4-step discrimination peak of 77 percent. There are, however, two other large discrimination peaks at +20 msec and +70 msec, and the one at +20 msec is 95 percent, condiderably higher than the peak at the phoneme boundary. Both subjects, then, especially LQ, seem to show effects other than the linguistic.

The discrimination of VOT in velar stops is shown for Subject EL at the top of Figure 4. His identification crossover at +27 msec is under a discrimination peak that reaches 86 percent. In addition, his voicing lag discrimination is generally quite high with another peak of 70 percent at about +70 msec. EL, by the way, is the one subject described earlier as an excellent bilingual with hardly any Spanish interference in his English. VOT measurements of his initial /g/ and /k/ in recordings of Spanish words[7] yield a boundary that corresponds

[6] Strange and Halwes (1971) have shown, using our VOT stimuli, that the use of confidence ratings in the oddity task can save much testing time. By the time they had shown this, we were too far along in our Spanish experiments to modify our discrimination procedures. Had we used confidence ratings, we might have been able to salvage another two or three subjects. An important discussion of discrimination procedures in experiments on the perception of speech sounds is found in Pisoni (1971).

[7] We failed to obtain voice recordings of the other subjects. We made a special

SPANISH LABIAL DISCRIMINATION

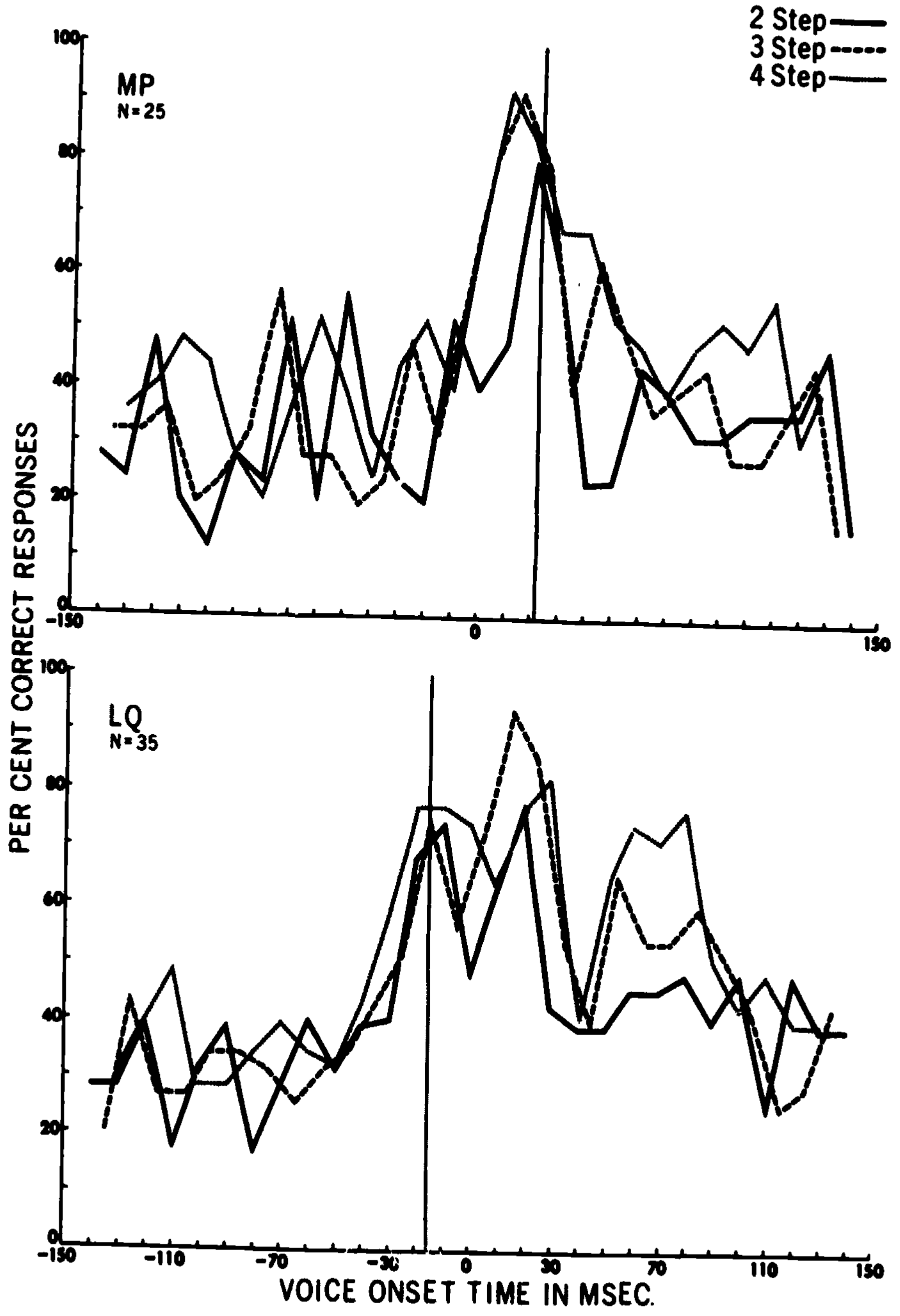


Figure 3: Labial discrimination functions for two individual subjects.

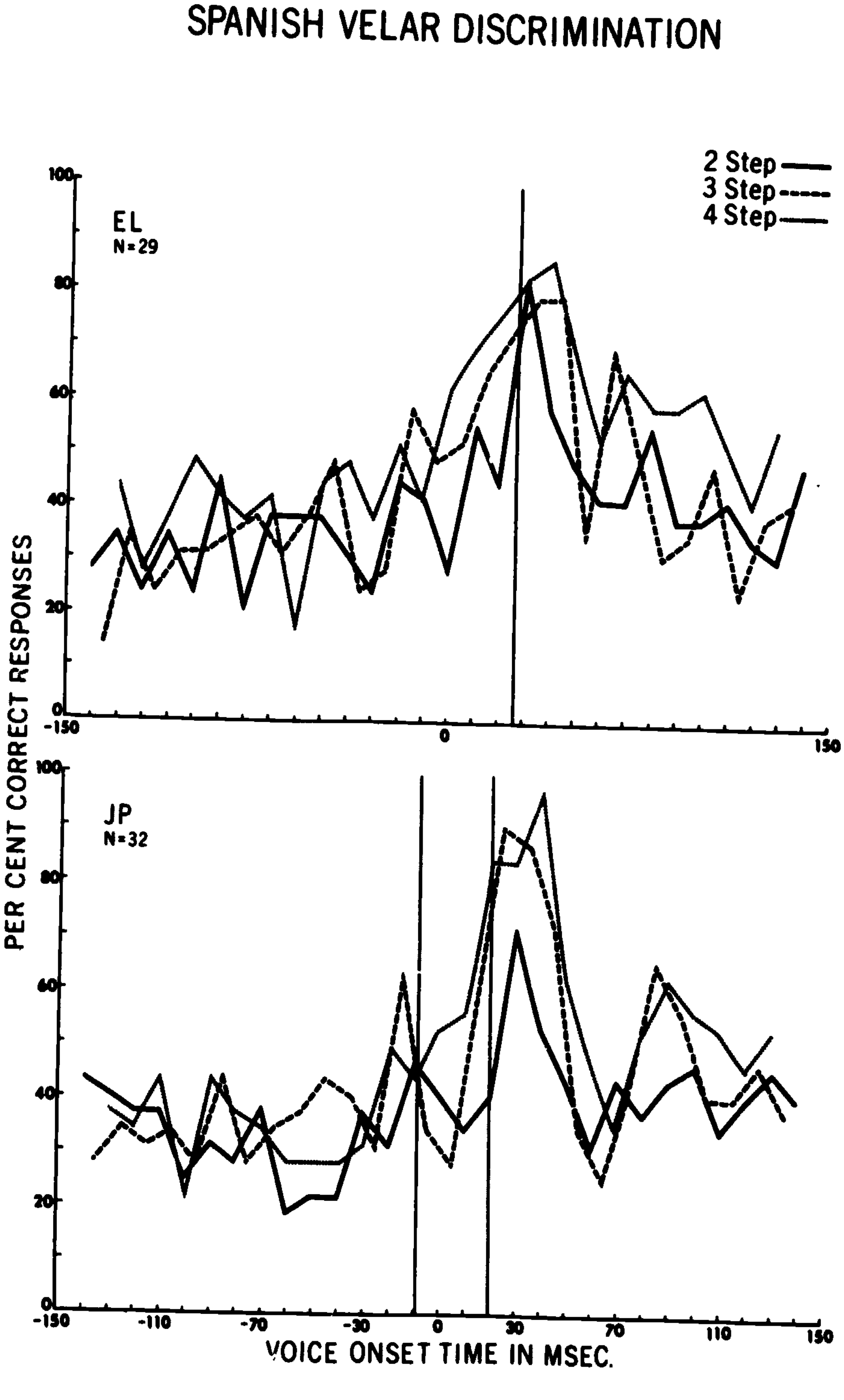


Figure 4: Velar discrimination functions for two individual subjects.

with his identification crossover point. That is, his /g/ and /k/ ranges about at +30 msec. He deviates from previously examined Spanish speakers (Lisker and Abramson, 1964:402) in producing instances of /g/--indeed, 56 percent of the time--with no voicing lead. His background makes it hard to rule out English interference. He had his elementary and secondary schooling at an American school in Lima, Peru, where he studied English for thirteen years. In addition, he had spent four and a half years in the United States when the experiment began.

The velar data for JP, shown at the bottom of Figure 4, are somewhat more complicated. His velar identification data do not reveal a single 50 percent crossover point; rather they show a zone of ambiguity between /g/ and /k/ from -8 msec to +20 msec. We have placed two vertical lines on the time axis to show this span. A discrimination peak reaching 97 percent straddles the right end of the crossover zone, while a smaller peak straddles the left end; there is a third peak around +90 msec.

The perceptual efficacy of VOT as a sufficient cue for distinguishing the voiced and voiceless stops of Spanish seems established. The possible information-bearing value of other particular acoustic features sometimes associated with voicing distinctions, e.g., pitch (Haggard et al., 1970; Fujimura, 1971) and F1 transitions (Cooper et al., 1952:600; Stevens and Klatt, 1971), we believe, is also ascribable to the relative timing of events at the larynx and the supraglottal place of articulation. The question of the influence of linguistic categories on the performance of discrimination tasks, at least as far as the present study is concerned, is more complicated. The presence of a phonological boundary certainly has an effect, more with some subjects than others, but there are also discrimination peaks remote from the phonological boundary and indeed always in the lag end of the continuum where spectral variation is somewhat more complex. That is, even though in Spanish and in many other languages the presence or absence of voicing lead is an important cue to a phonological category, stop variants with voicing lag are just easier to discriminate on some psychoacoustic basis. Our earlier work with English and Thai showed similar effects, but they are much more striking here. Of course, there is also the possibility that the psychoacoustic effect is combined with a linguistic one in the sense that large values of lag may sound so aspirated to the Spanish ear that they are considered foreign by the listener and therefore well discriminated from others judged to be more Spanish-like. We have no theoretical rationale for predicting how naive listeners might process a range of speech sounds which lies well outside the norms of their native language, whether they treat these sounds in effect as belonging to a single category of "foreign" speech sound or as some sort of nonspeech continuum.[8]

effort to record EL's speech because of his unusual background. His English stops, recorded in a separate session, look native by and large but seem to show a slight Spanish interference.

[8] For either way of processing these sounds, we do not know what kind of discrimination function the subjects would show. See the discussion in Mattingly et al. (1971:152-154).

REFERENCES

Abramson, A. S. and L. Lisker. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. Proc. 5th Intl. Cong. Acoustics, ed. by D. E. Commins, A51 (Liège: Imp. G. Thone).

Abramson, A. S. and L. Lisker. (1970a) Laryngeal behavior, the speech signal and phonological simplicity. Actes du X^e Congrès International des Linguistes. Vol. VI, 123-129 (Bucarest: Editions de l'Académie de la République Socialiste de Roumanie).

Abramson, A. S. and L. Lisker. (1970b) Discriminability along the voicing continuum: Cross-language tests. Proc. 6th Intl. Cong. Phon. Sci., Prague, 1967, 569-573 (Prague: Academia).

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

Flórez, L. (1964) El español hablado en Colombia y su atlas lingüístico. Presente y futuro de la lengua española: Actas de la Asamblea de Filologia del I Congreso de Instituciones Hispanicas. Vol. I, 6-77 (Madrid: Ediciones Cultura Hispanica).

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

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

Harris, J. W. (1969) Spanish Phonology. Research Monograph No. 54. (Cambridge, Mass.: M.I.T. Press).

Hirose, H. and T. Gay. (in press) The activity of the intrinsic laryngeal muscles in voicing control: An electromyographic study. Phonetica. (Also in Haskins Laboratories Status Report on Speech Research SR-28, 115-142.)

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

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within the across phoneme boundaries. J. exp. Psychol. 53, 358-368.

Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1958) Some cues for the distinction between voiced and voiceless stops in initial position. Lang. Speech 1, 153-167.

Lisker, L. 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. Proc. 6th Intl. Cong. Phon. Sci., Prague, 1967, 563-567 (Prague: Academia).

Lisker, L. and A. S. Abramson. (1971) Distinctive features and laryngeal control. Language 47, 776-785.

Lisker, L., A. S. Abramson, F. S. Cooper, and M. H. Schvey. (1969) Transillumination of the larynx in running speech. J. acoust. Soc. Amer. 45, 1544-1546.

Lope Blanch, J. M. (1964) Estado actual del español en Mexico. Presente y futuro de la lengua española: Actas de la Asamblea de Filologia del I Congreso de Instituciones Hispanicas. Vol. I, 82-91 (Madrid: Ediciones Cultura Hispanica).

Lope Blanch, J. M. (1968) El español de America. (Madrid: Ediciones Alcala).

Lopez Morales, H. (1971) Estudios sobre el español de Cuba. (New York: Las Américas).

Malmberg, B. (1950) *Etudes sur la phonétique de l'espanol parlé en Argentine*. (Lunds Univ. Årsskrift. N.F.Avd.1. Bd 45. Nr 7).

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

Navarro, T. (1948) El español en Puerto Rico. Contribución a la geografia linguística hispanoamericana. Ed. de la Univ. de Puerto Rico, Río Piedras, P. R.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Ph.D. dissertation, University of Michigan. (Also Supplement to Haskins Laboratories Status Report on Speech Research, November 1971).

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

Slis, I. H. and A. Cohen. (1969) On the complex regulating the voiced-voiceless distinction, II. Lang. Speech 12, 137-155.

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

Stevens, K. N. and D. H. Klatt. (1971) The role of formant transitions in the voice-voiceless distinction for stops. J. acoust. Soc. Amer. 50, 146-147 (A).

Strange, W. and T. Halwes. (1971) Confidence ratings in speech perception research: Evaluation of an efficient technique for discrimination testing. Percept. Psychophys. 9, 182-186.

Vidal de Battini, B. E. (1964) El español de la Argentina. (Buenos Aires: Consejo Nacional de Educación).

Some Effects of Oral Anesthesia upon Speech: An Electromyographic Investigation[*]

Gloria Jones Borden[+]
Haskins Laboratories, New Haven

INTRODUCTION

It has been a long-observed fact that when one comes from the dentist's office there is often a disturbance of clearly articulated speech until the effect of the anesthesia has disappeared. It is understandable, therefore, that investigators interested in afferent control of speech should block the sensory nerves of normal speakers with anesthesia in order to study the relationship between feedback from the oral area and articulation of speech. Presumably all feedback channels are used to develop language, audition, taction, and proprioception. The question is whether skilled speakers need depend upon these feedback possibilities during ongoing speech and to what degree or under what circumstances each channel may play a role. Is learned speech centrally patterned, with little or no need under normal circumstances for peripheral control? A series of studies during the 1950s and '60s dealt with this subject. It was found that bilateral mandibular and intraorbital injections of anesthesia increased the number of judged errors in articulation of adult speakers (McCroskey, 1958; Ringel and Steer, 1963). The speech distortions were found to be subtle and were most evident in the production of fricatives and affricates (Scott, 1970; Borden, 1971; Gammon, Smith, Danilof, and Kim, 1971). It was assumed by the investigators that the speech effect was the result of decreased oral sensation as a result of blocking sensory feedback from the tongue via the lingual nerve. A phonetic analysis of the speech effect under anesthesia revealed two factors which prompted further investigation; first was the variability of effect among speakers, with some subjects unaffected by the nerve block, although oral sensation was reported to be lost, and the second factor was the predominance of articulatory distortions among the sibilants and affricates, especially /s/ in consonant clusters, in those subjects who were affected (Borden, 1971). It was decided to study electromyographically the contraction of some of the muscles thought to be implicated in lingual movement under conditions of nerve block and under normal conditions.

FIRST ELECTROMYOGRAPHIC STUDY

Two separate electromyographic (EMG) experiments were conducted in an attempt to find out what happens to certain suprahyoid muscles as subjects speak under conditions of trigeminal nerve block. Since the nerve block seemed to

[*] This paper reports a portion of the research carried out for a Ph.D. dissertation accepted by the City University of New York in 1971.

[+] Also City College of the City University of New York.

produce an /s/ effect, muscles which are thought to contribute to tongue elevation were reviewed (Van Riper and Irwin, 1958; Hirano and Smith, 1967; Zemlin, 1968). The muscles which were accessible, clearly identifiable, and of interest for this study were the genioglossus, geniohyoid, mylohyoid, and the anterior belly of the digastric muscles. The orbicularis oris was included as a reference (Figure 1).

Method

The monopolar electrodes used were DISA concentric needle electrodes with a diameter of .45 mm. Needle placement was made through the cutaneous tissue under the chin to the depth required. Correct placement was checked by observing the oscilloscope while protruding the tongue for genioglossal activity, saying "ta" for geniohyoid activity, lowering the mandible for digastric activity, and saying "ka" for mylohyoid activity. Correct placement was checked periodically throughout each run.

The subject for the first experiment was a normal adult speaker. Two runs were produced, the first without nerve block, and the second with bilateral mandibular blocks. A total of 7.5 cc of 2% xylocaine was injected by a dentist, 3 cc in each side and an additional 1.5 cc on one side. The technique was similar to that used by McCroskey (1958), the model for all previous studies. A partial run was recorded with a medial nasopalatine block of 1 cc and an anterior palatine block of 2 cc added, but this part of the study was not analyzed, as the speech effects were not noticeably different from the run with the bilateral mandibular blocks alone. It seems that loss of sensation from the anterior portion of the hard palate and the alveolar ridge adds very little to the speech effect evidenced with the mandibular blocks.

For the EMG studies, material was selected from the utterances used in our previous work. Eleven utterances in sentence form, using the format "It could be the _________," were used to permit the necessary rapid connected speech. Each utterance was represented twice in a randomized list of twenty-two utterances. There were ten such lists, each individually randomized. Each utterance was spoken twenty times during the course of one run. The utterances were as follows:

It could be the snowballs splashing.
It could be the cat's whiskers.
It could be the fixed sweater.
It could be the school blocks.
It could be the thirsty wasp.
It could be the sleeping taxi.
It could be the spider string.
It could be the squirrel nest.
It could be the rooster scratch.
It could be the spring grapes.
It could be the stove smell.

The 220 utterances for each run were printed and mounted on large cards which were flipped as the subject read them, with equal stress attempted on each of the final two words.

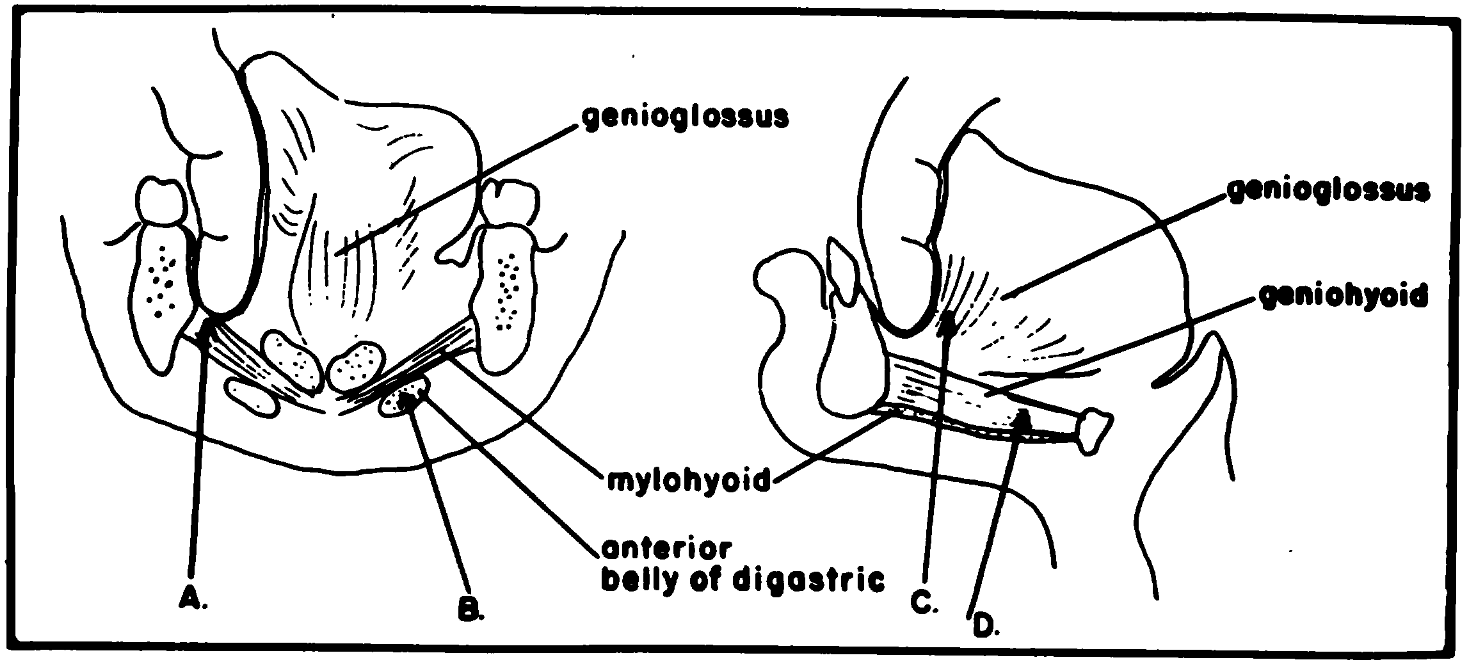


Fig. 1

Muscles examined in EMG study. Arrows indicate direction of needle insertions. Frontal and sagittal views.

A 16-channel magnetic tape was produced, recording the electrical output of the muscles, which were monopolar recordings; that is, the difference was recorded between the active tissue of the muscles and the inactive tissue of the earlobe. Some of the channels were used for audio signals, such as the utterances produced by the subject and the comments for record-keeping produced by the experimenters. Each utterance was numbered by a pulse code which was laid down on the tape and eventually on the computer output.

The output of the channels was put onto paper tape both at the time of the run and later for locating and inspecting the individual tokens. Each utterance was represented twenty times during each run, and a single point in time, a line-up point, was selected so that all of the tokens of a single type could be averaged by computer for each electrode. The line-up point was chosen at a point of particular interest and marked on the simultaneous recording of the subject's audio recording.

Each tape was subjected to five computer programs to check that the code pulses were in order, to set the gains of the playback amplifiers at levels appropriate for the analog-to-digital converter, to make control tapes of the line-up points and distances from point zero for each utterance, to set each EMG channel at the optimum level, and finally to average the data on the control tapes.

The paper output of this process is a list of numbers for each channel, indicating the averaged value of each electrode in microvolts every 5 msec. The three runs were hand plotted.

Results and Discussion

Inspection of the data reveals that the muscular activity recorded during speech under normal conditions remained high during the nerve-block condition with the exception of two muscles. After the nerve-block injections, it was observed by the experimenters that the activity on the oscilloscope of the mylohyoid muscle and the anterior belly of the digastric muscle dropped dramatically to a state of relative inactivity. The electrodes were checked and found to be in place, but as long as the anesthesia was effective those muscles were in effect "paralyzed." The speech of the subject under nerve block revealed the typical mandibular block effect of distorted sibilants, the /s/ clusters being most prominently affected. Compare the graph of the two affected muscles during the production of the utterance "sleeping taxi" under normal conditions (Figure 2) with the graph of the same electrode placements during nerve block. All eleven utterances showed the same drop in activity for the mylohyoid muscle and the anterior belly of the digastric during anesthesia.

A closer look at the anatomy at the injection area showed us that we should not have been surprised. The mandibular injection which has traditionally been used for these studies deposits half of the solution in the area of the lingual nerve, then moves on to deposit the rest of the solution in the area of the inferior alveolar nerve. It happens that just before the inferior alveolar nerve enters the mandibular foramen into the mandibular canal, it gives off the nerve fibers of what is known as the mylohyoid nerve, the only purely motor component of the otherwise sensory inferior alveolar branch of the trigeminal nerve (Figure 3). The mylohyoid nerve is motor to the mylohyoid muscle and to the anterior belly of the digastric muscle, the two muscles which dropped in activity during the nerve-block condition.

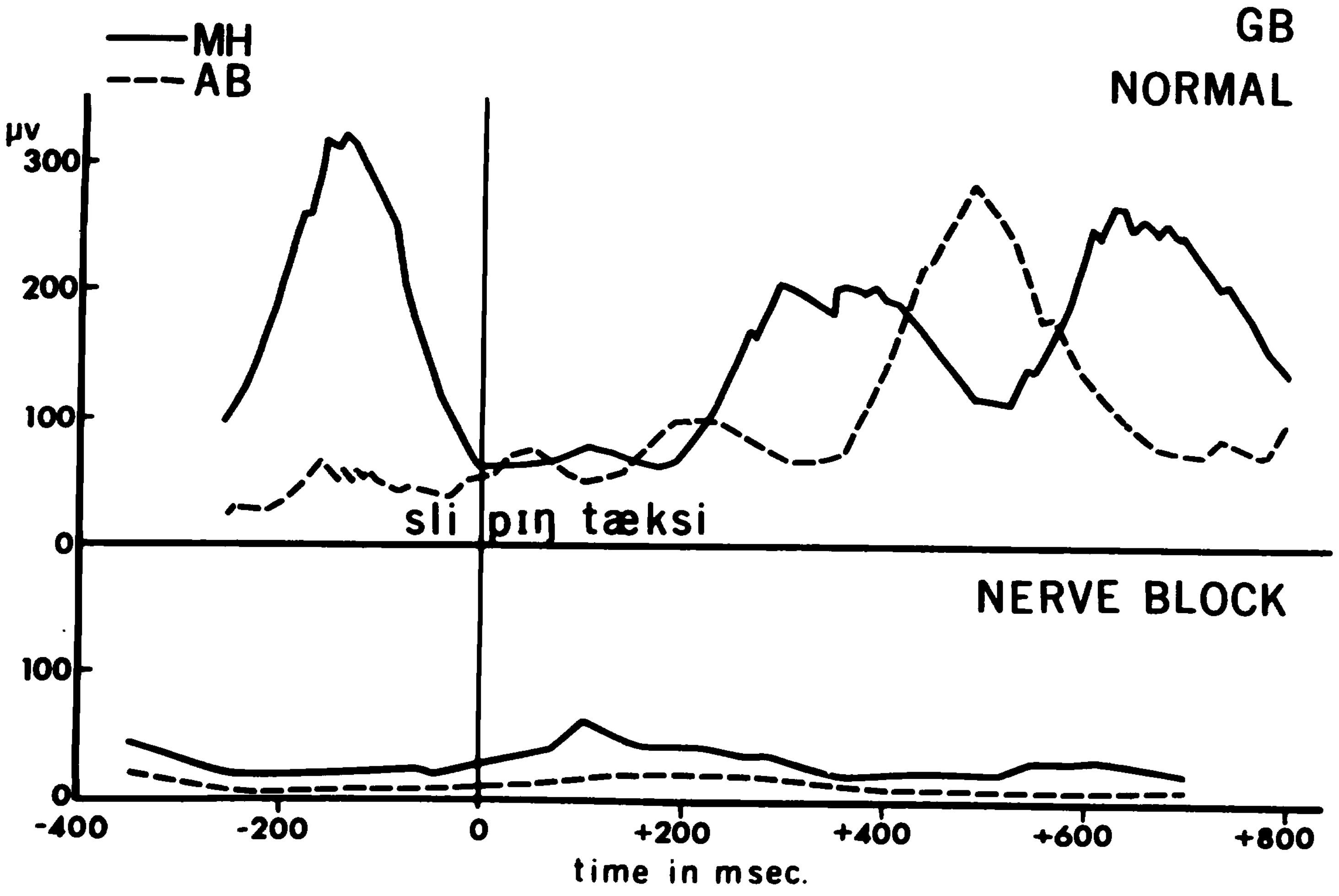
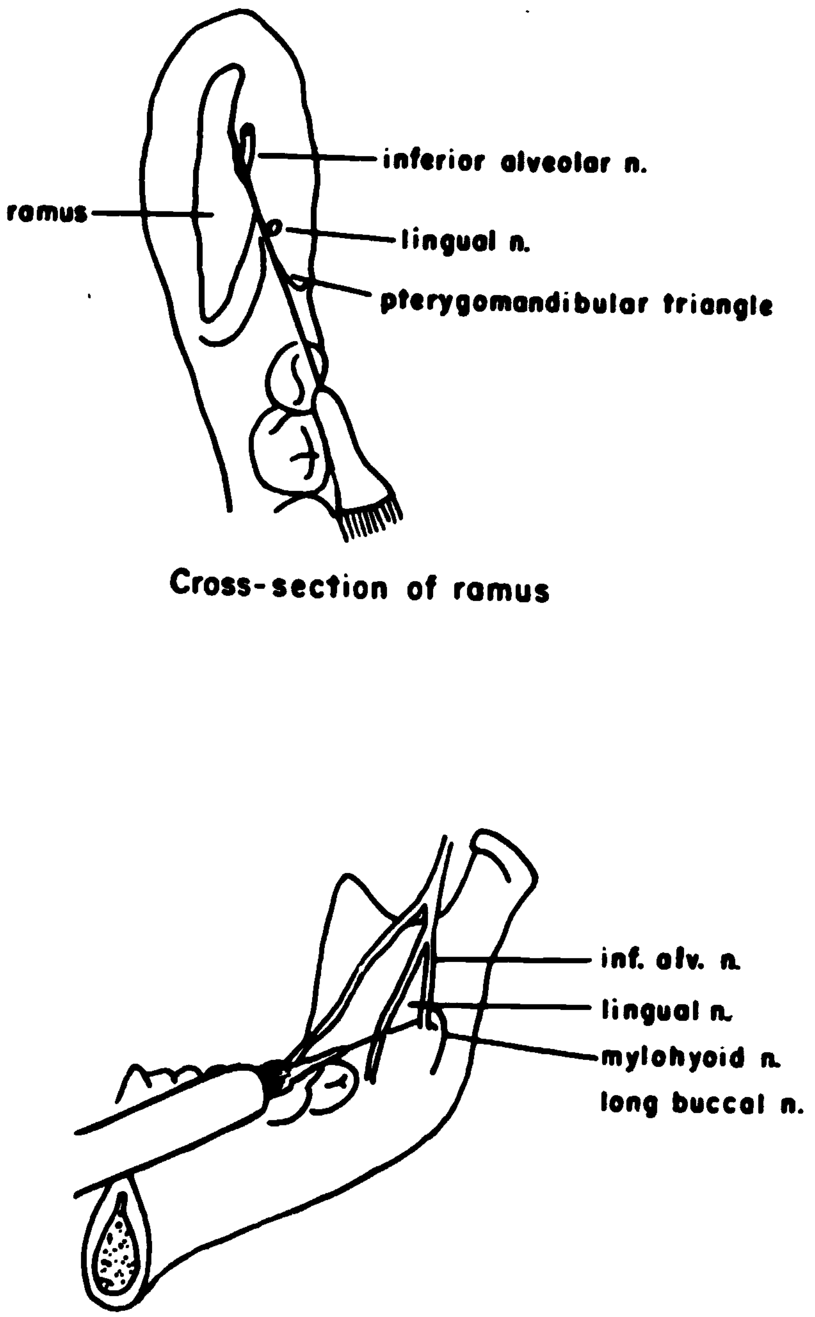


Fig. 2

EMG recording of the mylohyoid muscle and the anterior belly of the digastric muscle during normal and nerve-block conditions.



Cross-section of ramus

Inner surface of ramus with needle in the right mandibular sulcus.

Fig. 3

The next consideration was whether the inactivity of either of these muscles could have contributed to the noted speech deterioration. If the speech effect is primarily due to sensory loss, then loss of feedback from the tongue tip region would probably be responsible. If it is due to motor loss, however, then the anterior belly of the digastric muscle and the mylohyoid muscle are probably responsible.

The normal function of the anterior belly of the digastric muscle is to open the jaw. EMG data on this muscle, obtained by recording muscle activity during simple "CVp" utterances, showed no action for /i/ and /u/ and a large peak for /ɑ/ (Harris, 1971). Since there was no perceptible speech effect of the nerve block upon vowels, and since the action of the anterior belly would not reasonably be expected to affect the spical gestures which deteriorated under nerve block, it seems unlikely that its motor loss could have caused the speech effects observed. It may be that other mouth-openers compensate.

The normal function of the mylohyoid muscle was found by both Harris (1971) and Smith (1970) to be highest for the production of /k/. Its contraction seems to lift the body of the tongue. In the more complex utterances of the present study, it can be seen that the mylohyoid muscle peaked normally in preparation for the /s/ consonant clusters and for the velars (Figure 4). Notice the activity at the beginning of "spring," "spider," and "string," and at the end of "grapes" and "string." Observe the drop in activity of the mylohyoid muscle during the nerve-block condition. The peaks of activity under normal speaking conditions, then, coincided with the speech distortions produced under the nerve-block condition, with the exception of the velars.

The nerve block did not distort the velars sufficiently to be perceived as a distortion. The production of /k/ remained intact, as had been reported in all previous nerve-block experiments. The explanation may lie in the comparatively gross production of /k/ and the fact that we, as listeners, accept as /k/ a less precise gesture than we do as /s/.

It seems, therefore, that the effected "paralysis" of the mylohyoid muscle might reasonably be related to the speech effect, since, for this subject, the mylohyoid muscle appears to be important in lifting and steadying the body of the tongue for consonant clusters, especially those with /s/ (Table 1). This subject produces /s/ with the tongue tip down, making it imperative that the body of the tongue be raised to produce the friction. Deprived of motor ability in the mylohyoid and deprived of lingual sensation, the /s/ clusters were distorted. It is impossible to conclude which of these factors, if not both, is responsible for the distorted speech, but it cannot be assumed, as it has in previous studies, that the effect is due to loss of sensory feedback.

In summary, the clear conclusion of this first EMG experiment was that a motor component existed in what was previously assumed to be a sensory deprivation. The motor loss was evident in two of the suprahyoid muscles, the mylohyoid muscle and the anterior belly of the digastric muscle. One of these muscles, the mylohyoid, is normally active for this subject for /s/ clusters and velars. Since this subject produced /s/ with a high dorsum, it is reasonable to assume that the motor loss in the mylohyoid muscle may have contributed to the speech deterioration during anesthesia.

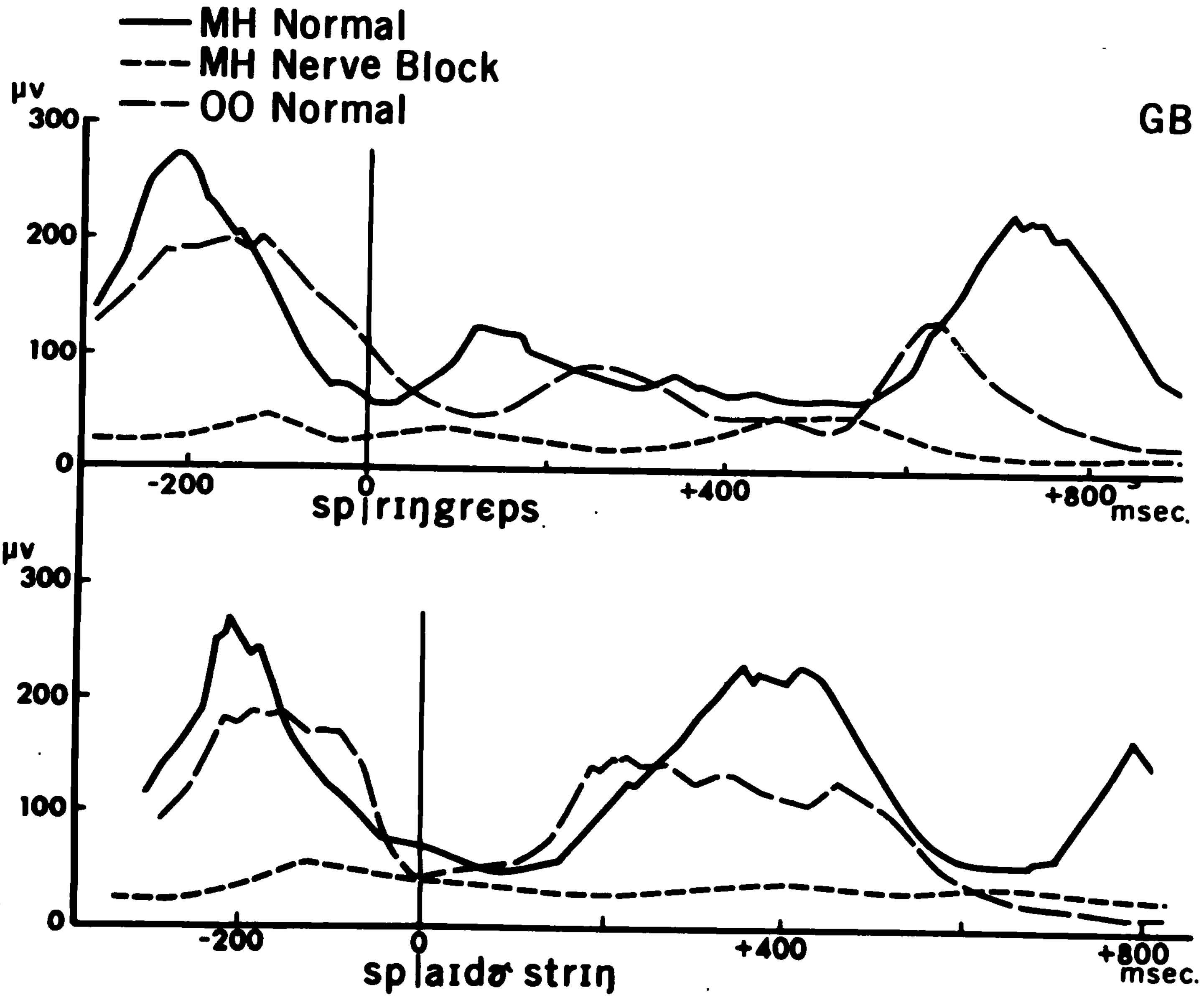


Fig. 4

Mylohyoid muscle peaked in this subject under normal conditions for /s/ consonant clusters and for velars.

TABLE 1: Peak values in microvolts for mylohyoid muscle in first EMG experiment during nerve-block and normal conditions.

	springrapes		
Normal	345	155	285
NB	30	35	20
msec	(-225)	(125)	(715)

	catswhiskers			
Normal	315	355	380	370
NB	35	40	40	20
msec	(-800)	(-505)	(-140)	(200)

	thirstywaspj	
Normal	185	310
NB	30	35
msec	(-855)	(-255)

	stovesmell	
Normal	335	355
NB	30	50
msec	(-215)	(325)

	snowballssplashing		
Normal	415	340	430
NB	30	55	25
msec	(-140)	(500)	(900)

(/ng/ not plotted)

	roosterscratch			
Normal	175	200	310	370
NB	30	40	40	20
msec	(-775)	(-440)	(-125)	(325)

	fixedsweater		
Normal	485	210	140
NB	45	45	15
msec	(45)	(325)	(585)

	schoolblocks	
Normal	380	400
NB	50	30
msec	(-145)	(640)

	squirrelnest	
Normal	215	150
NB	50	25
msec	(-175)	(635)

	spiderstring		
Normal	355	300	210
NB	35	40	25
msec	(-210)	(365)	(790)

	sleepingtaxi		
Normal	425	265	355
NB	30	40	40
msec	(-155)	(300)	(635)

SECOND ELECTROMYOGRAPHIC STUDY

The purpose of the second EMG study was to verify the result of the first study, which was that mylohyoid motor loss accompanied the distorted speech during the nerve-block condition, and also to study further the changes in muscle activity by comparing the electrical potential in normal speech with the electrical potential during nerve block.

Method

It was necessary to use a second subject for this experiment. The material consisted of thirty utterances in the frame "the ______." They were randomized into four lists repeated alternatively four times, making sixteen lists of thirty utterances each. Fifteen of the utterances were chosen from the Scott (1970) list in an attempt to observe the muscle changes in the distorted speech which might explain the phonetic changes which she had transcribed. The other fifteen utterances were selected from the sentences in the first study and from the perceptual study. Two runs were produced. Done on the same morning, the first one was conducted under normal conditions, the second under blocked condition.

The electrodes were .0002-inch wires hooked to remain in place. Correct placement was checked by observing the oscilloscope while lifting the tongue for genioglossal activity, tensing the floor of the mouth while relaxing the tongue for geniohyoid activity, saying "ka" for mylohyoid activity, opening the mouth with jaw effort for anterior belly of digastric activity, saying "pa" for orbicularis oris activity, and lifting the head or opening the mouth under pressure for sternohyoid activity. The genioglossus and geniohyoid were also checked during swallowing, as their activity differs in timing (Hirose, 1971). Electrodes were placed in both sides of the mylohyoid muscle and in both anterior bellies of the digastric muscle.

After the normal run, a total of 7.5 cc of 2% xylocaine was injected into the oral region of the subject. There are two general types of dental injections, supraperiosteal injections and block injections. A supraperiosteal injection, sometimes called an infiltration, is a procedure in which the anesthetic solution is deposited in the periosteum opposite the roots of certain teeth. The solution is carried by diffusion through the periosteum and bony plate to the nerves. The only infiltration injection used in this experiment was the anesthetization of the posterior superior alveolar nerve. A block injection is one in which the anesthetic solution is deposited between the brain and the field of operation. The solution penetrates the nerve trunk or nerve fibers and blocks either the sensations coming from the distal field or the motor impulses coming from the brain. All of the injections used in this study were nerve blocks, except the one to the posterior superior alveolar nerve (Cook-Waite Labs, 1971). A summary of the injections is given in Table 2.*

A rough check of two-point discrimination was made, and when the experimenters and subject were satisfied that sensation was lost in the tongue and the palate, Ringel's (1969) fifty-five-item oral discrimination test of ten plastic forms was administered. When the subject had returned to normal, the Ringel test was again administered. The subject made nine errors in normal condition and fifteen errors in the nerve-block condition, the difference being errors of shape, not size. Confusion of shape occurred three times in normal condition and nine times in nerve-block condition. Nevertheless, the experimenters were surprised that there was so little difference in performance on this test. It was noted that the subject used the usual tongue manipulations during normal condition but relied on deep pressure against the palate when sensation was decreased. This technique was reported as the method used by successful subjects in the previously mentioned study on the effect of anesthesia on oral stereognosis (Mason, 1967).

The multichannel magnetic tapes which were produced for each of these runs were analyzed in much the same way as the first experiment. There were some

*The reason that such extensive injections were administered was to enable the experimenters to compare results with the Scott data. In the present study, the dentist attempted to hit the lingual nerve and to avoid the mylohyoid nerve. The intent was to produce a purely sensory block without any motor effects.

TABLE 2: Injections of anesthesia administered in the second EMG study.

Cranial Nerve	Branch	Amount of Solution	Location of Injection	Area of Sensation
V (mand.)	Inf. Alveol. n.	1.5 cc ea. side	pterygomand. triangle	mand. alv. ridge, lip, gums
	Lingual n.			ant. 2/3 tongue
V (mand.)	Long Buccal n.	.5 cc ea. side	1st molar	buccal
V (max.)	Infraorbital Ant. Sup. Alv. Middle [illegible] Alv.	.5 cc ea. side	infraorbital foramen	upper lip alv. ridge ant. teeth
V (max.)	Nasopalatine n.	.5 cc midline	post. to central incisors	ant. 1/3 palate
V (max.)	Post. Sup. Alv. n.	.5 cc ea. side	2nd molar	molars
V (max.)	Greater Palatine n.	.5 cc ea. side	palate 3rd mol.	post. 2/3 palate

refinements in the computer programs. A concise description of the analysis procedure is reported by Port (1971).

Results and Discussion

As a result of the first EMG experiment, the investigators were particularly interested in this second study in the activity of the mylohyoid muscle. Since there were bilateral placements of electrodes in both the mylohyoid muscle and the anterior belly of the digastric muscles, the investigators had an opportunity to study the activity on both sides of these muscles. During the normal run, before the injections of anesthesia, the mylohyoid and the anterior bellies showed activity similar to the first subject. The anterior belly peaked for mouth opening and the mylohyoid for velar gestures and somewhat for the /s/ clusters.

During the condition of nerve block, however, there was a decrease in activity in both muscles on the right side. The right anterior belly of the digastric was in all cases significantly less active than normal after anesthesia. The right mylohyoid was consistently less active than normal for velar gestures, but for the /s/ clusters, it was sometimes less active and sometimes more active than normal. The decreased activity on the right side in this experiment was not as pronounced as it had been in the first EMG study, indicating that the attempt on the part of the dentist to avoid the motor mylohyoid nerve was partially successful. The limited effect on the right side was presumed by the investigators to be the result of some infiltration of the anesthetic in the area of the mylohyoid nerve.

In contrast with the decreased activity observed on the right side of the mylohyoid and anterior belly of the digastric muscles, the left side of these muscles were usually more active than normal while the anesthesia was in effect. Figure 5 demonstrates the asymmetry of effect. The right peak in each of the four graphs represents the labial closing for /p/ in "duckpond." It can be seen that the right side of both muscles was quite active during normal speech but dropped in activity during speech with nerve block. The left electrode placement in the mylohyoid was in a slightly less active field than the right side. That is, there were fewer motor units firing near the electrodes on the left side. The left-side placement of the electrode into the anterior belly of the digastric was in a particularly inactive field. The problem of electrode placement into a more or less active field of the muscle is less important in this study than in many, because our interest is in comparing the activity recorded at a single site under two different conditions, normal and nerve block. Relative values, therefore, are more important than absolute values. A final look at Figure 5 shows both muscles on the left side to be more active during nerve block than they were normally.

We have no explanation of these results except to assume that the anesthetic solution had a motor effect on one side of the subject and that there was some reorganization of motor function on the opposite side to compensate for the motor loss. Typically, bilateral injections of anesthesia result in some asymmetry of effect. In the perceptual study we sometimes had to reinject a subject on one side, due to insufficient loss of sensation. The subject for the first EMG study required an additional 1.5 cc of xylocaine on one side to equalize the desensitivity. It is reasonable to assume, therefore, that there would be the same possibilities for asymmetry of motor effect, depending upon the amount of infiltration of the anesthetic solution into the fibers of the motor mylohyoid nerve.

The most prominent result of this study, therefore, was that despite considerably less anesthesia and an attempt to avoid the mylohyoid nerve, there was a unilateral drop in mylohyoid and anterior belly of digastric activity during anesthesia, although the other apparently unaffected side demonstrated efforts at compensation, by showing more than normal activity.

A second interesting result of this EMG experiment was that the subject's articulation appeared to be clear under nerve block. There were no discernible phonetic distortions. The speech sounded as acceptable under the nerve-block condition as under the normal condition. The utterances were louder under nerve block and produced with what might be described as overarticulation.

This variability of nerve-block effect among subjects was observed during the perceptual part of this series of studies. It is unclear why there was no speech effect. It might be a difference in muscle use, as this subject produces /s/ with tip of the tongue raised and might not rely on mylohyoid muscle activity as much as the first subject, who produces /s/ with dorsum of the tongue raised, keeping the tip down. Another explanation for no speech effect might be a difference in anesthesia, either in amount or in technique of injection. It is customary in these studies to inject anesthesia until the subject reports loss of sensation. In the mandibular block, loss of sensation is reported immediately when the lingual nerve has been hit directly, as it was in the case of this second subject. Only 1.5 cc of xylocaine solution was injected into each side, whereas 4.5 cc in each side was necessary before the subject of the first experiment lost sensation. The solution presumably anesthetized the

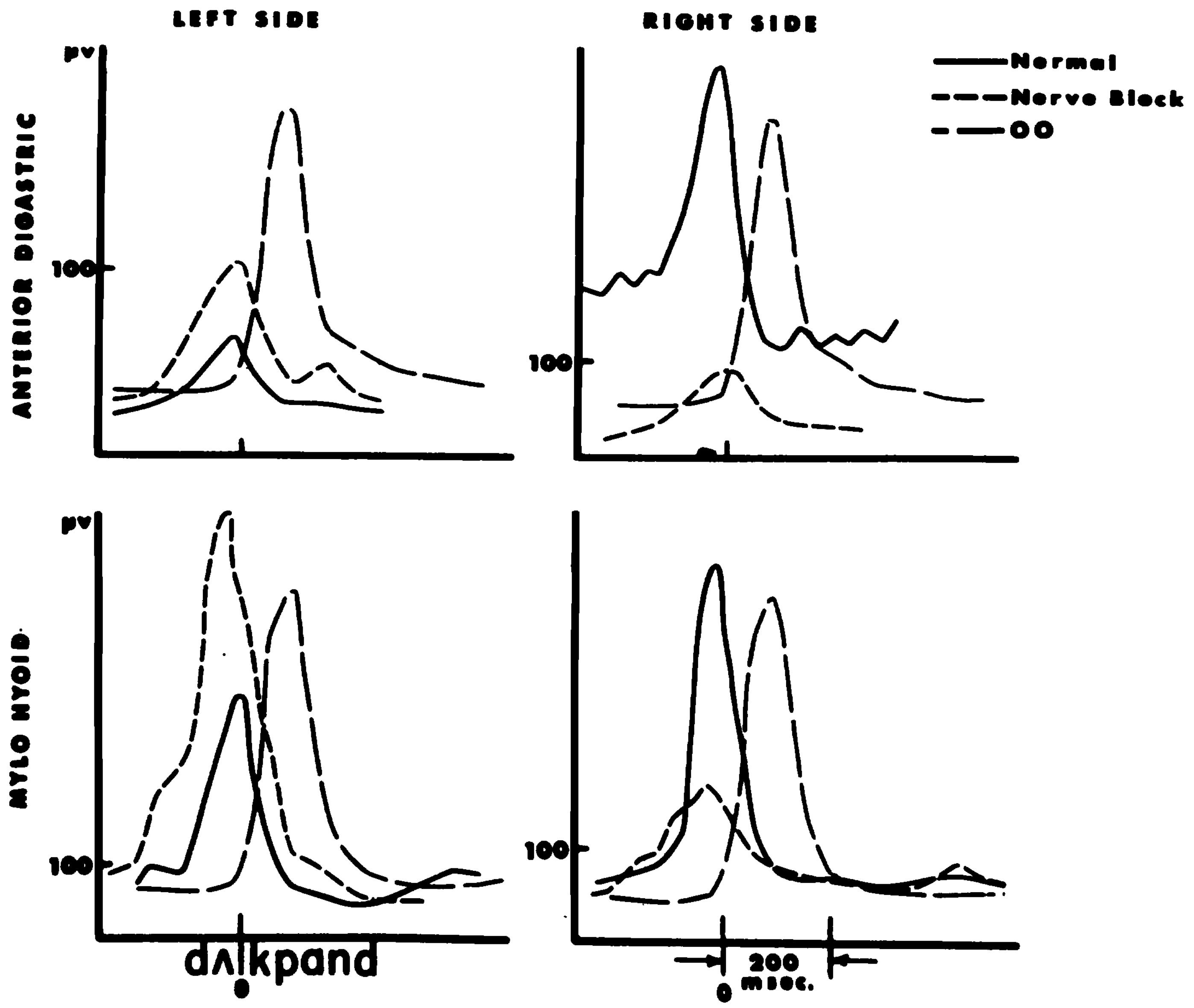


Fig. 5

Decreased right side activity and increased left side activity during nerve block of the mylohyoid and anterior belly of the digastric muscles. Obicularis oris is included as a reference.

mylohyoid nerve of the first subject, as we have indicated mylohyoid muscle and anterior belly of the digastric muscle inactivity. In this subject there was less anesthesia needed to effect loss of sensation and the solution apparently did not penetrate the mylohyoid motor nerve fibers on one side.

The third result of the second EMG study was a fairly consistent pattern of muscle reorganization under nerve block. Table 3 summarizes the muscle activity in general for each utterance during nerve block as it compares to its own activity normally. The orbicularis oris was usually either the same as normal or more active than normal. The genioglossus tended to be the same. Inexplicably, the geniohyoid was less active during nerve block. The rest of the muscles follow a reasonable pattern of adjustment. The right side of the mylohyoid muscle and the anterior digastric lost activity during nerve block, as previously discussed. The scatter plot of the differences in peak values of the right anterior digastric during the two conditions is clear. It was always higher normally than during anesthesia (Figure 6). The right mylohyoid showed the same decreased activity for the normally active velar gestures, but for the high front gestures such as /t/ or /s/ clusters, there was increased activity during nerve block (Figure 7).

TABLE 3: Relative muscle activity during the nerve-block condition for each utterance.

	More Active Than Normal	Less Active Than Normal	Same As Normal	Different Than Normal
OO	11	5	13	1
GG	1	8	21	
GH		29	1	
SH	22		7	1
MHR	4	14	7	5
MHL	24		6	
ADR		30		
ADL	15		15	

Shifting our attention to the left mylohyoid and anterior digastric, we see that again the anterior belly clearly increases activity during nerve block, perhaps as compensation for the less active left side (Figure 8). The left mylohyoid, however, is somewhat more complex. It, too, was more active during nerve block. Notice that for the less active front consonant gestures, there was less increase in activity during nerve block than for the already normally active velars (Figure 9). When the right side of the mylohyoid dropped for the velars, the left side soared. Finally, the sternohyoid was interesting as it was consistently more active during nerve block than under normal conditions and might reasonably be expected to compensate for the inactivity of the anterior digastric. The anterior belly of the digastric opens the jaw, as does the sternohyoid (Figure 10). In summary, the muscles do seem to be behaving

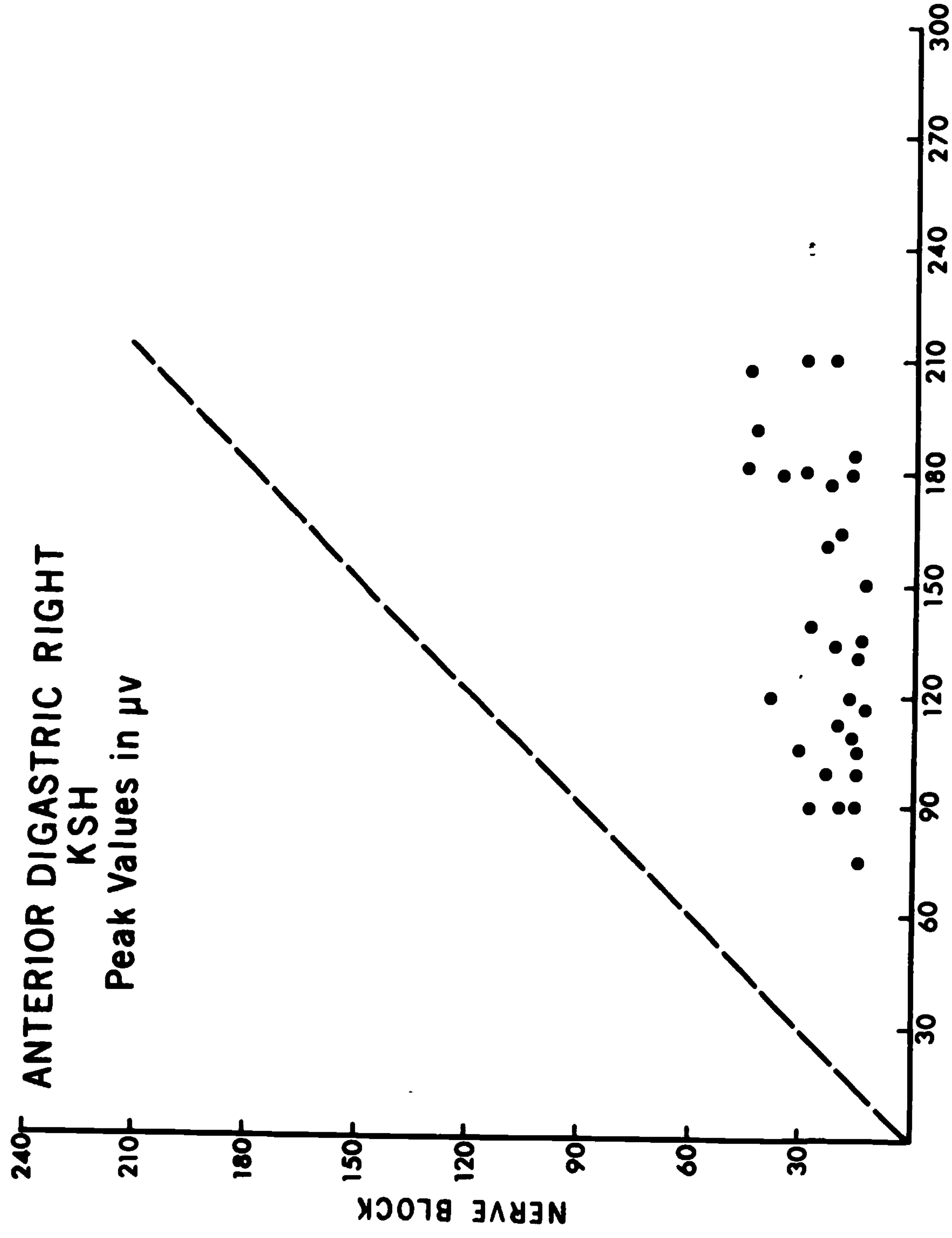

ANTERIOR DIGASTRIC RIGHT
KSH
Peak Values in µv
NERVE BLOCK
240
210
180
150
120
90
60
30
NORMAL
30
60
90
120
150
180
210
240
270
300

Fig. 6

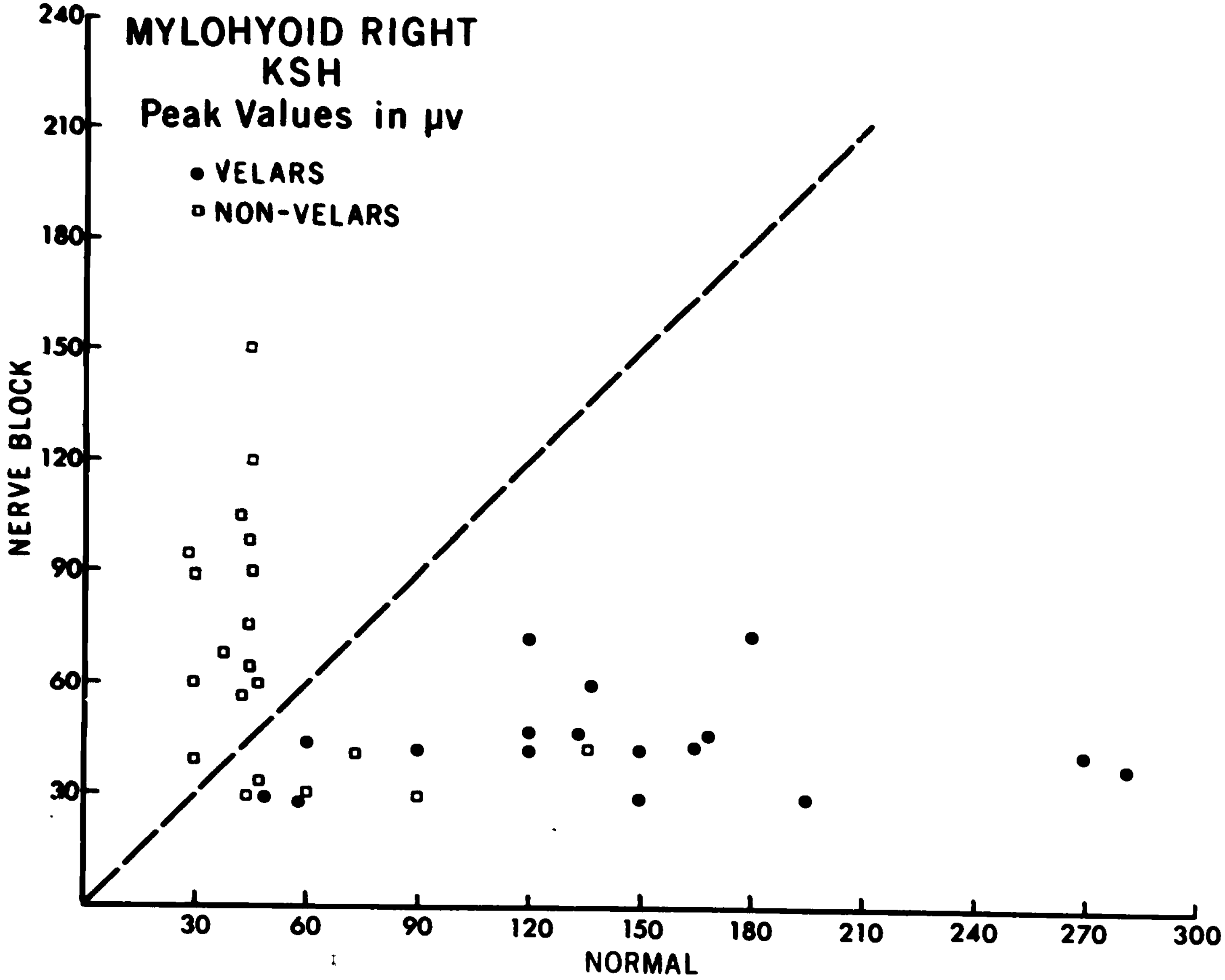

MYLOHYOID RIGHT
KSH
Peak Values in µv
● VELARS
□ NON-VELARS
NERVE BLOCK
240
210
180
150
120
90
60
30
30
60
90
120
150
180
210
240
270
300
NORMAL

Fig. 7

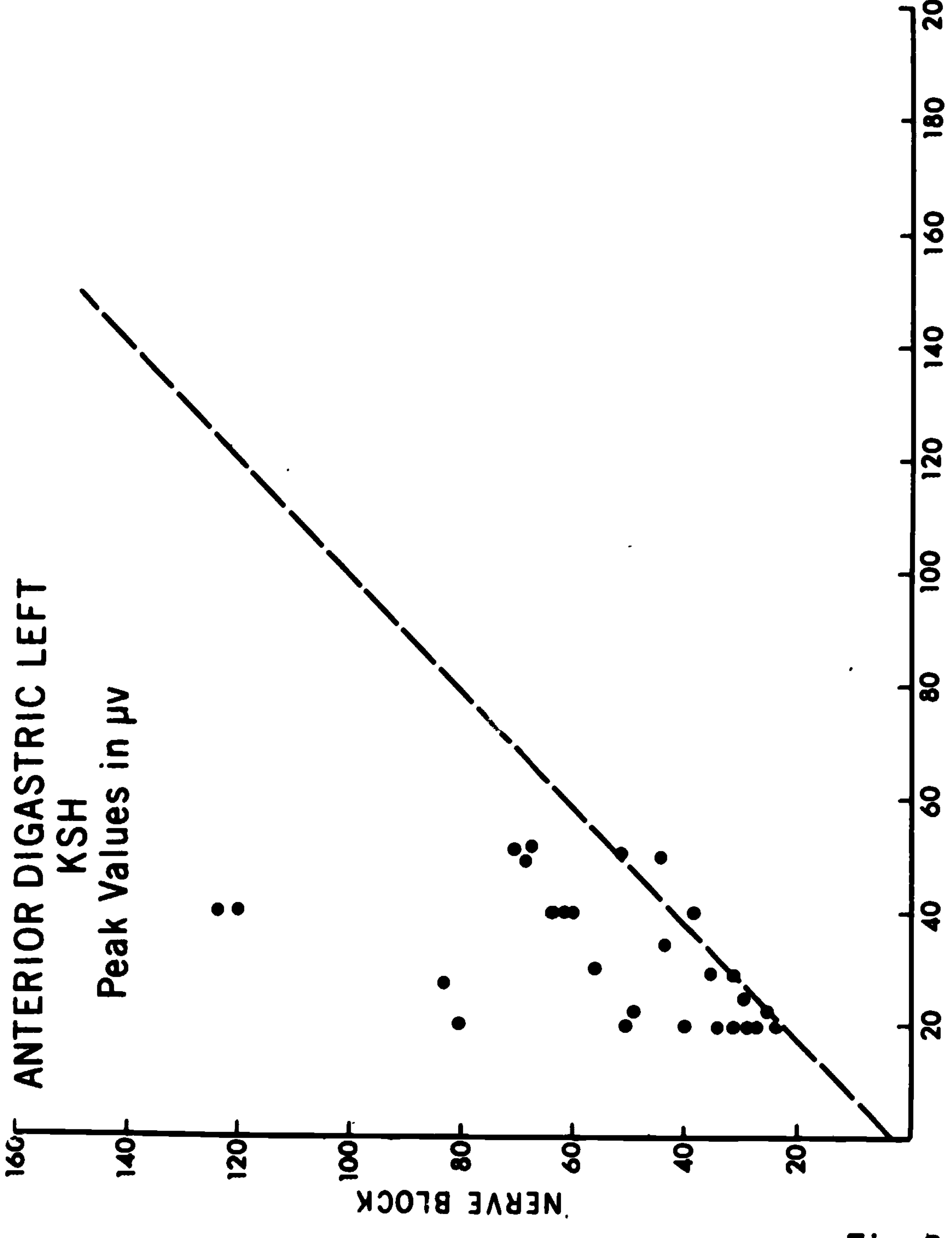

ANTERIOR DIGASTRIC LEFT
KSH
Peak Values in μv
NERVE BLOCK
160
140
120
100
80
60
40
20
NORMAL
20
40
60
80
100
120
140
160
180
200

Fig. 8

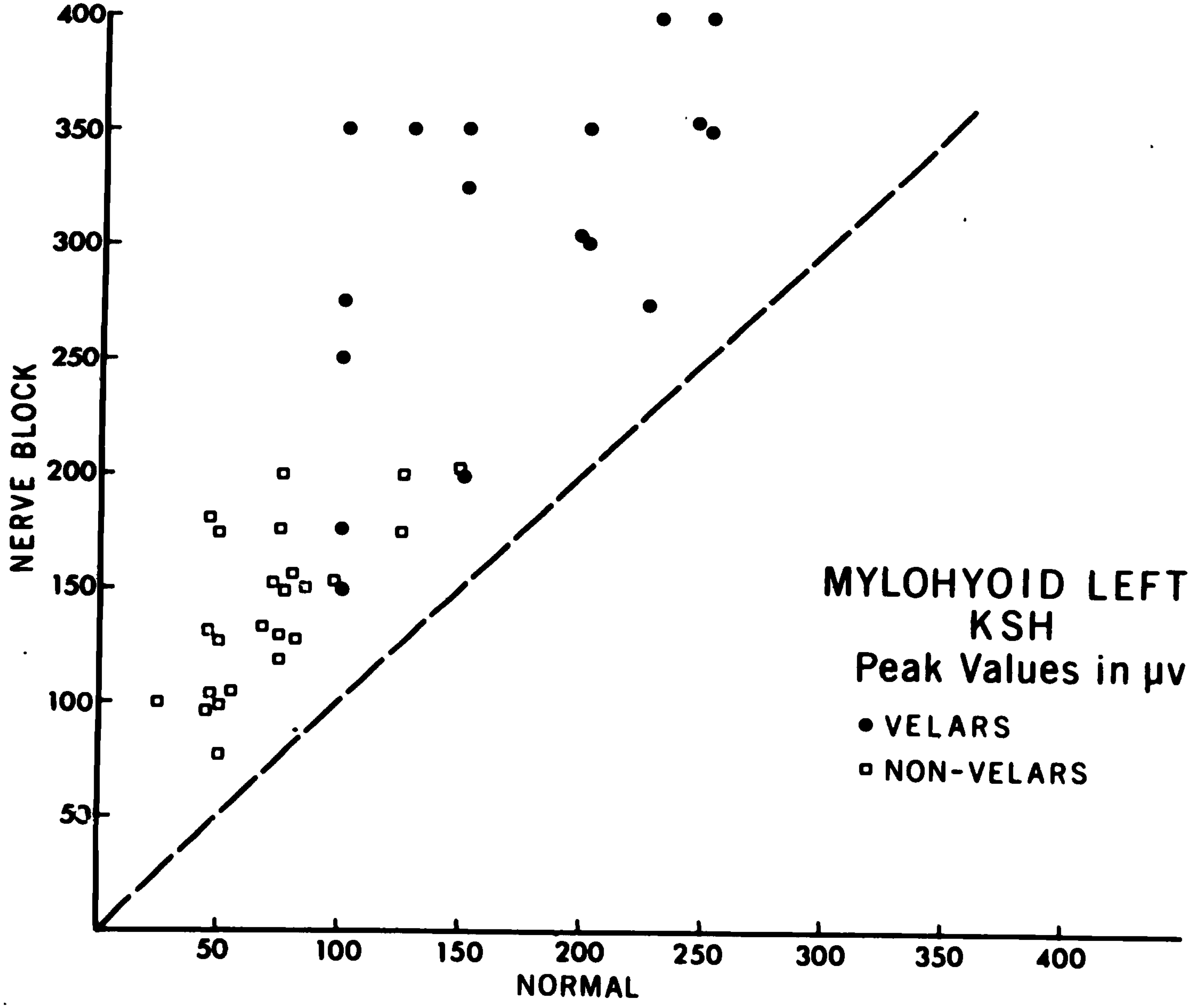

MYLOHYOID LEFT
KSH
Peak Values in µv
● VELARS
□ NON-VELARS
NERVE BLOCK
400
350
300
250
200
150
100
50
NORMAL
50
100
150
200
250
300
350
400

Fig. 9

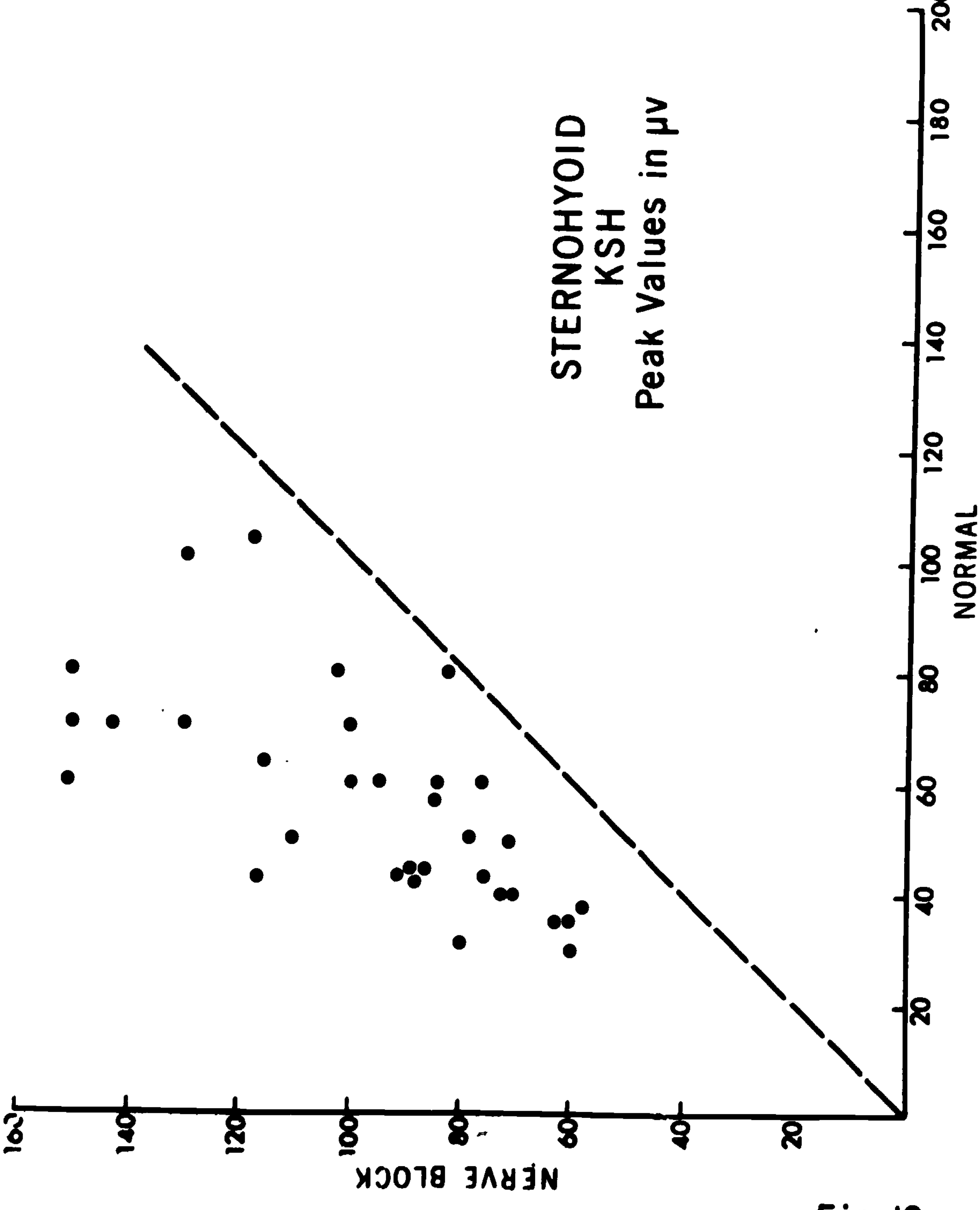

STERNOHYOID
KSH
Peak Values in µv
NORMAL
20
40
60
80
100
120
140
160
180
200
NERVE BLOCK
20
40
60
80
100
120
140
160

Fig. 10

differently during the nerve-block condition. They do not seem to change their pattern of function so much as their amplitude. Finally, there are some instances which look like compensatory action as a result of decreased activity on the opposite side or in another muscle.

Whether the speech of this subject might have remained sharp and clear even had the mylohyoid nerve been bilaterally affected by the anesthesia resulting in mylohyoid muscle "paralysis," as seemed to be the case with the first subject, remains unclear. Is the speech deterioration, when it exists, related to a loss of tactile and kinesthetic sensation, as has traditionally been suggested? Or might it be related to a loss of motor function, which these studies force us to consider? Or might it possibly be related to some reorganization of the unaffected muscles in an attempt to compensate for the motor loss, an attempt which perhaps succeeds except on phonemes demanding rapidity and precision of gesture such as /s/ and /r/ in consonant clusters?

The conclusion which we must draw from these experiments is that there is an error of method in the experimental technique which has been traditionally used to study tactile and kinesthetic loss of sensation on speech. The most critical injection of anesthesia, the mandicular block, produces not only a sensory loss but a motor loss. These studies have demonstrated, furthermore, that the use of EMG is important in experiments on sensory deprivation as a direct check on motor function.

REFERENCES

Borden, G. J. (1971) Some effects of oral anesthesia upon speech: A perceptual and electromyographic analysis. Ph.D. dissertation, City University of New York.

Cook-Waite Labs, Inc. (1971) Manual of Local Anesthesia in General Dentistry. New York, Rev. 2nd ed.

Gammon, S. A., P. J. Smith, R. G. Daniloff, and C. W. Kim. (1971) Articulation and stress/juncture production under oral anesthetization and masking. J. Speech Hearing Res. 14, 271-282.

Harris, K. S. (1971) Action of the extrinsic musculature in the control of tongue position: Preliminary report. Haskins Laboratories Status Report on Speech Research SR-25/26, 87-96.

Hirano, M. and T. Smith. (1967) Electromyographic study of tongue function in speech: A preliminary report. U.C.L.A. Working Papers in Phonetics 7, 46-56.

Hirose, H. (1971) Electromyography of the articulatory muscles: Current instrumentation and techniques. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.

Mason, R. M. (1967) Studies of oral perception involving subjects with alterations in anatomy and physiology. In Symposium on Oral Sensation and Perception, ed. by J. F. Bosma. (Springfield, Ill.: Charles C. Thomas)

McCroskey, R. L. (1958) The relative contribution of auditory and tactile cues to certain aspects of speech. Southern Speech J. 24, 84-90.

Port, D. K. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research SR-25/26, 67-72.

Ringel, R. L., A. S. House, K. W. Burk, J. P. Dolinsky, and C. M. Scott. (1970) Some relations between orosensory discrimination and articulatory aspects of speech production. J. Speech Hearing Dis. 35, 3-11.

Ringel, R. L. and M. D. Steer. (1963) Some effects of tactile and auditory alterations on speech output. J. Speech Hearing Res. 6, 369-378.

Scott, C. M. (1970) A phonetic analysis of the effects of oral sensory deprivation. Doctoral dissertation, Purdue University.

Smith, T. J. (1970) A phonetic analysis of the function of the extrinsic tongue muscles. Doctoral dissertation, University of California.

Van Riper, C. and J. V. Irwin. (1958) Voice and Articulation. (Englewood Cliffs, N.J.: Prentice-Hall Inc.)

Zemlin, W. R. (1968) Speech and Hearing Sciences: Anatomy and Physiology. (Englewood Cliffs, N. J.: Prentice-Hall Inc.)

Laryngeal Control in Vocal Attack: An Electromyographic Study

Hajime Hirose* and Thomas Gay**
Haskins Laboratories, New Haven

SUMMARY

Multichannel EMG recordings were obtained from the intrinsic laryngeal muscles of four American English speakers for three different types of vocal attack: breathy, soft, and hard. The data were processed by a digital computer to obtain an average indication of overall muscle activity.

The results indicate that the three different types of vocal attack are characterized by coordinated actions of the abductor and adductor muscles of the larynx, and further, that these muscles work in reciprocal fashion for each type of attack.

INTRODUCTION

The mechanism of laryngeal control for different types of vocal attack[1] has long been a subject of interest in the fields of laryngology and experimental phonetics. Three types of vocal attacks are generally recognized: (1) breathy or aspirate, (2) soft or simultaneous, and (3) hard or glottal.

Various experimental techniques have been used to investigate these types of vocal attack--high-speed cinematography (Moore, 1938; Werner-Kukuk and von Leden, 1970), aerodynamic study (Isshiki and von Leden, 1964; Koike, 1967; Koike et al., 1967), and electromyography (EMG) of the intrinsic laryngeal muscles (Gay et al., in press; Hirano, 1971; Koike, 1967; Sawashima et al., 1958). Among these, electromyography is particularly useful, as it provides the most direct information on the actions of the individual muscles responsible for vocal attack.

Most of the previous studies in laryngeal physiology generally support the classical division of the intrinsic laryngeal muscles into three functional groups: abductor, adductor, and tensor. However, there still are many unanswered questions concerning the function of individual laryngeal muscles in different modes of laryngeal adjustment.

*On leave from Faculty of Medicine, University of Toyko.

**Also University of Connecticut Health Center, Farmington.

[1] The term "attack" usually refers to vocal initiation in singing; if we use this classification to refer to speech utterances consisting of /C + V/ sequences, breathy attack should be equivalent to the utterance initiated with /h/, soft attack to that with voiced consonant, and hard attack to that with glottal stop.

The first EMG study of vocal attack was attempted by Faaborg-Andersen (1957), who compared the activity of the vocalis and cricothyroid in the production of /hop/, /bop/, and /op/, representing breathy, soft, and hard attack, respectively.[2] He stated that the time interval between the start of the increase in activity in the two muscles and the onset of the tone (▲t) was greater for hard attack than for either the breathy or soft attack in both muscles.

Koike (1967) later examined the EMG activity of the vocalis and the cricothyroid in his extensive study of vocal attack and claimed that ▲t was largest for hard attack but that values were variable for soft and breathy attacks. He also claimed that the amplitude of the pre-phonatory activity of these two muscles seemed to serve as a more reliable index for differentiating the type of vocal attack than ▲t values.

Hirano (1971) repeated the first two studies using trained singers who were asked to begin phonation on a signal. He was unable to distinguish ▲t values for the three vocal attack conditions. He suggested, rather, that the mode of activity of the adductors (the lateral cricoarytenoid and vocalis in this case) of the larynx, particularly during the pre-phonatory period, is the most essential factor for differentiating the type of vocal attack.

These previous EMG reports dealt solely with the adductor and the tensor groups of the larynx, and no attempt was made to clarify the participation of the abductor muscle, the posterior cricoarytenoid, in vocal attack. Furthermore, most previous studies were based on the observation of limited numbers of raw EMG traces. It would seem reasonable, then, that a detailed, systematic EMG study of all the intrinsic muscles of the larynx is needed to provide a complete description of the muscle control mechanism of vocal attack.

The purpose of the present study was to investigate systematically the actions of all the intrinsic laryngeal muscles in different types of vocal attack. Particular attention was directed to comparing the temporal characteristics of the EMG activity patterns for the abductor and adductor muscles.

PROCEDURES

Subjects

The subjects were four adults, three male and one female, all native speakers of American English. The female subject (AP) was a trained singer.

For each subject, an attempt was made to record from the five intrinsic muscles simultaneously. However, this goal was reached for only two of the four (LJR and LL). Unsatisfactory recordings were obtained for the posterior cricoarytenoid and the cricothyroid of one subject (TG), and the posterior cricoarytenoid, the interarytenoid, and the vocalis of another (AP).

[2] These test utterances can be transcribed phonetically as (hɔp), (bɔp), and (ʔɔp), respectively.

Recording and Processing of Data

Conventional hooked-wire electrodes consisting of a pair of insulated platinum-iridium alloy wires with a short hook at the tip were used in the present experiment (Hirano and Ohala, 1969). The wires were threaded in a hypodermic needle and inserted into the muscle with the needle. After the tips of wires were located in the muscle, the needle was withdrawn, leaving the wires in place.

The electrodes were inserted percutaneously through the skin of the anterior neck into the lateral cricoarytenoid (LCA), the vocalis (VOC), and the cricothyroid (CT), while the insertions into the posterior cricoarytenoid (PCA) and the interarytenoid (INT) were made perorally by indirect laryngoscopy under topical anesthesia. A specially designed curved probe was used for the peroral insertions.

The basic data-processing procedures followed in the present experiment were to collect EMG data for a number of tokens of each type of vocal attack and, using a digital computer, average the integrated EMG signals at each electrode position. EMG data were recorded on a multichannel tape recorder together with the acoustic signal and digital code pulse (octal format). This pulse was used for identifying each utterance for the computer during processing. In the present experiment, the line-up point for averaging was the onset of voicing of each utterance. A more detailed description of both the data-recording and data-processing techniques can be found elsewhere (Gay et al., in press; Hirose, 1971; Hirose et al., 1971; Port, 1971).

Experimental Conditions

Isolated monosyllabic words /ha/, /ba/, and /ʔa/ were used to represent breathy, soft, and hard attacks, respectively. The subjects were required to repeat each test utterance sixteen times. Vocal intensity and frequency were kept at normal levels.

RESULTS

Figure 1 shows the averaged EMG curves of the five intrinsic laryngeal muscles for subject LJR. It is clearly demonstrated in this figure that the pattern of activity of the individual laryngeal muscles differs depending upon the type of vocal attack.

In breathy attack, PCA stays active throughout the pre-phonatory period up to the point immediately before the onset of voicing (in this example, the activity starts to decrease approximately 150 msec before the onset of voicing). Its activity then decreases steeply and remains suppressed during the period of voicing. Conversely, the activity of the other four muscles appears to be suppressed during the pre-phonatory period and then increases steeply when PCA activity begins to decrease, peaking at about the time of voice onset.

In hard attack, PCA activity decreases well before the onset of voicing. PCA then shows a transient increase in activity just before the onset of voicing after which it is suppressed again for the period of voicing. The adductors, LCA in particular, show a very characteristic pattern of activity

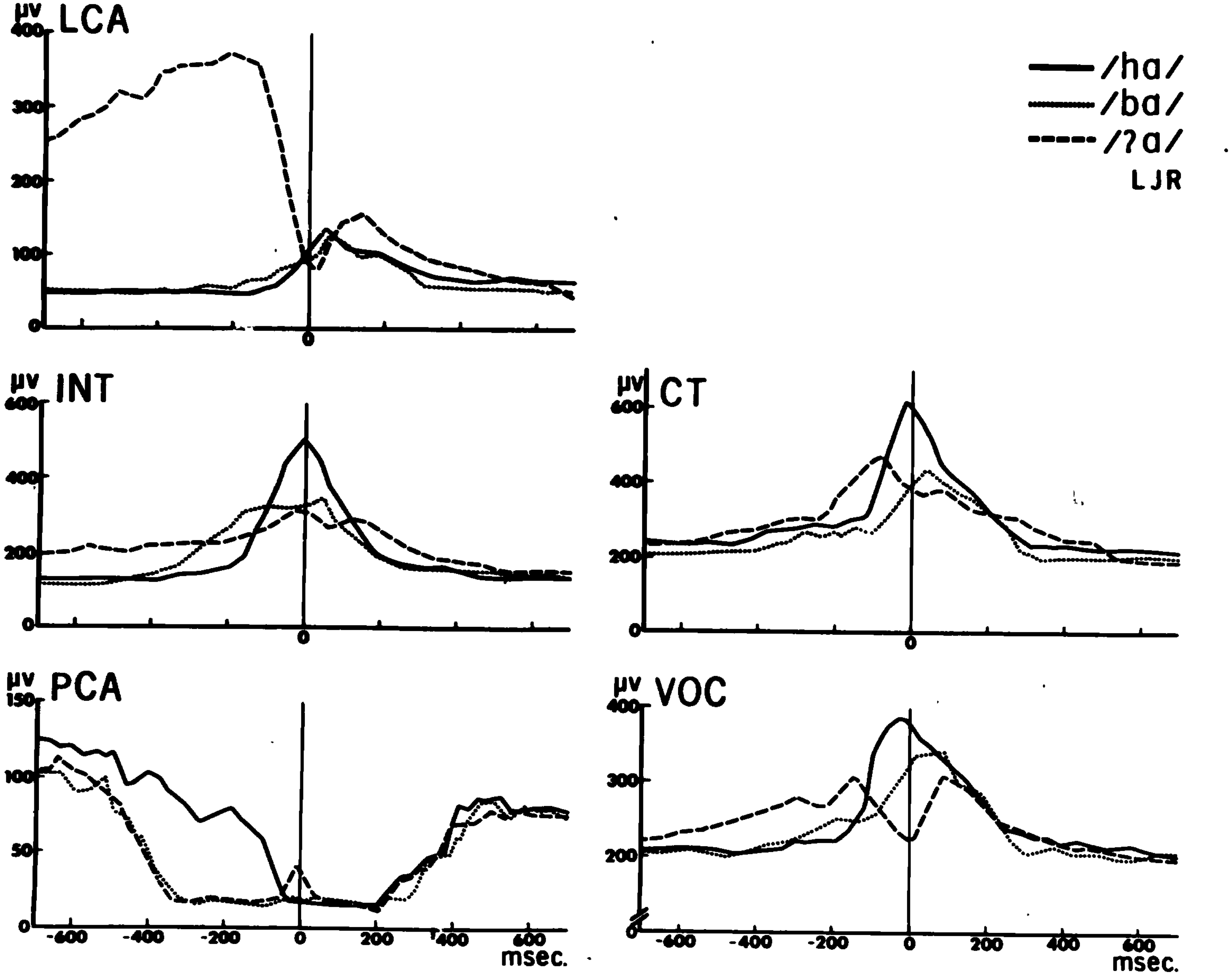


Figure 1: Averaged EMG curves (in microvolts) of the intrinsic laryngeal muscles of subject LJR for three different types of vocal attack. Zero on the time axis, for this and all subsequent curves, is the onset of voicing. Muscle identifications are LCA: lateral cricoarytenoid; INT: interarytenoid; PCA: posterior cricoarytenoid; CT: cricothyroid; VOC: vocalis.

for hard attack. LCA activity increases markedly long before (in this example more than 700 msec prior to) the onset of voicing and stays high during the pre-phonatory period. It then shows a steep fall immediately before the onset of voicing, followed by a less pronounced rise for the voicing period. INT, VOC, and CT also show activity during the pre-phonatory period followed by a fall at approximately the onset of voicing.

In soft attack, PCA activity is suppressed throughout the pre-phonatory period. The general pattern of activity of PCA in soft attack is similar to that in hard attack, except that there is no temporary increase before the onset of voicing. The activity of the adductors and CT increases gradually, reaching a peak after the onset of voicing.

The pattern of activity of the individual laryngeal muscles examined in the other three subjects was essentially similar to that observed in the first subject.

Figure 2 illustrates the averaged EMG curves of a second subject (LL) for three types of vocal attack. The temporal characteristics of PCA activity of the second subject are quite similar to those of the first subject with respect to the following points: (1) in breathy attack, PCA stays active throughout the pre-phonatory period up to the moment immediately preceding the onset of voicing after which it shows a steep fall; (2) in soft and hard attack, PCA activity starts to decrease well before the onset of voicing; for hard attack there is a transient increase before the voice onset, while for soft attack, it stays suppressed throughout; (3) PCA activity is higher for the pre-phonatory period than for the period of voicing regardless of the difference in the type of vocal attack.

When we compare the temporal characteristics of adductor activity of the second subject to those of the first subject, it is observed in both cases that adductor activity in breathy attack remains suppressed during the pre-phonatory period and then increases steeply for initiation of voicing. In hard attack, LCA shows a marked increase in activity during the pre-phonatory period in the second subject too, although the timing of the onset of the increase is somewhat later than that in the first subject. In soft attack, the adductors show a gradual increase in activity toward initiation of voicing. In the second subject, however, the increase starts earlier for soft attack than in the case of breathy attack.

Figures 3 and 4 show the averaged EMG curves for subjects TG and AP, respectively. In subject TG, all the adductors show more or less similar temporal characteristics for each type of vocal attack. In hard attack, in particular, both INT and VOC also showed considerable pre-phonatory activity followed by a fall immediately after. There is a general tendency of gradual increase in activity in soft attack. It is noted in breathy attack that there is temporary suppression of activity preceding the steep rise for initiation of voicing for all the three muscles.

The temporary dip in activity just before steep rise in breathy attack is also found in subject AP, both in LCA and CT. The general pattern of LCA activity of subject AP for each type of vocal attack is essentially similar to that of subject TG, though the activity increases more steeply near the voice onset, both for breathy and soft attacks.

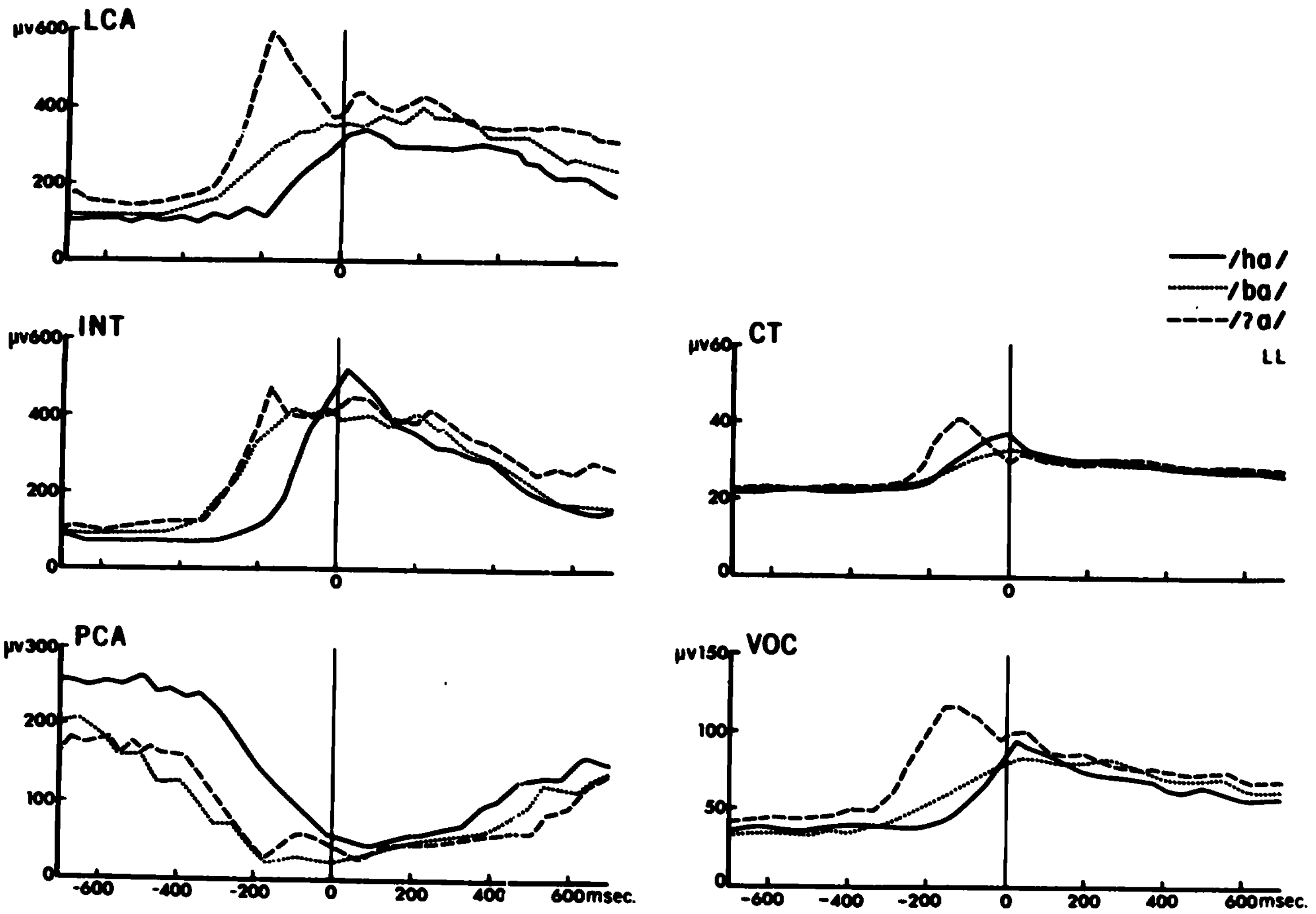


Figure 2: Averaged EMG curves of the intrinsic laryngeal muscles of subject LL for three different types of vocal attack.

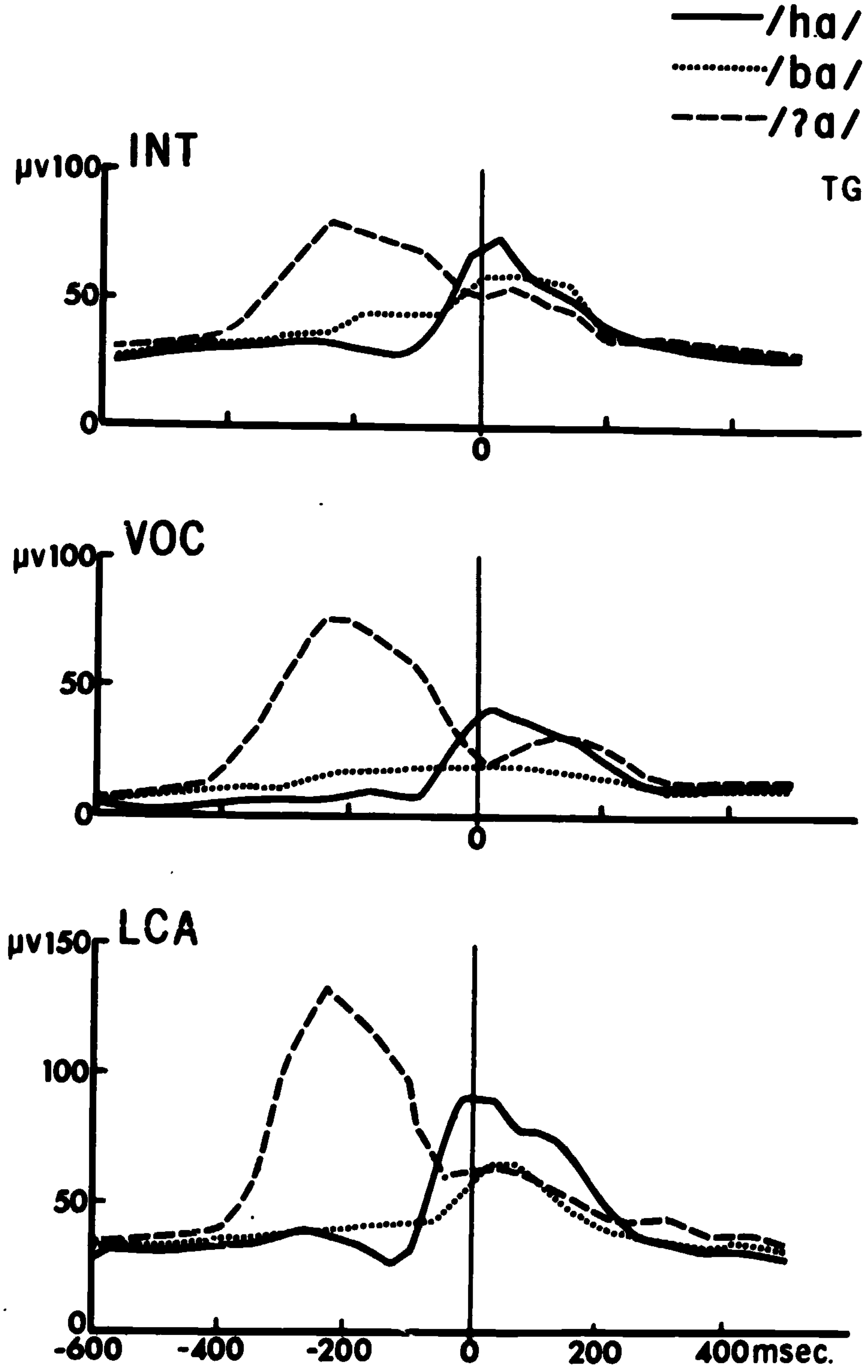


Figure 3: Averaged EMG curves of the intrinsic laryngeal muscles of subject TG.

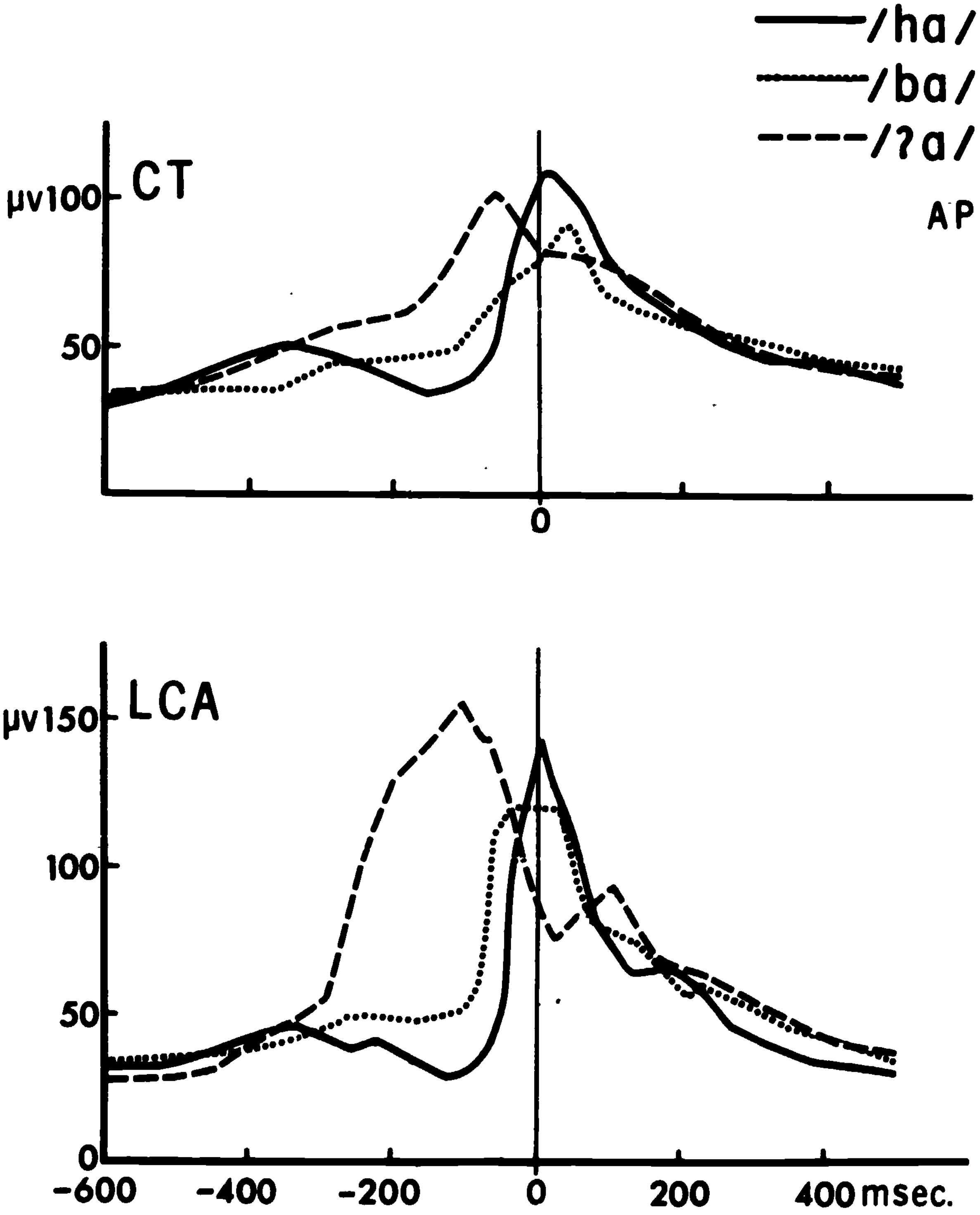


Figure 4: Averaged EMG curves of the intrinsic laryngeal muscles of subject AP.

DISCUSSION

The present study revealed that coordinated actions of the abductor and adductor muscles of the larynx characterize each type of vocal attack.

In breathy attack, PCA shows a characteristic pattern during the pre-phonatory period, where it stays active until just before the onset of voicing. It has been observed by both high-speed cinematography (Werner-Kukuk and von Leden, 1970) and by fiberoptic observation (Sawashima, 1968) that the glottis remains open during the production of initial /h/. Presumably, the relatively high pre-phonatory activity of PCA in conjunction with low adductor activity is the physiological correlate of the open glottis for initial /h/ production. The activity of the adductors in breathy attack appears to increase toward the onset of voicing as in soft attack. However, it shows a steeper increase near voice onset in breathy attack. Hirano (1971) stated that the adductors often show a temporary dip in activity preceding the steep rise in breathy attack in singing. In the present study, it appears that there is a temporary dip in LCA activity in breathy attack in subjects TG and AP (see Figures 3 and 4), but the finding is not consistent for the others. The dip may well be interpreted as a temporary suppression of adductor activity for the production of /h/, which has begun increasing slightly in order to maintain preparatory muscle tonus. Also, differences may exist between singing and speech articulation in the degree of preparatory muscle tonus as well as of temporary suppression for /h/ production. It is worth noting further that the maximum value of INT activity is higher in breathy attack than in the other two attacks.[3] It is generally agreed that EMG activity grossly represents the muscle action necessary for obtaining effective force or displacement, although either isometric or isotonic conditions in a strict sense are hardly expected in reality. As we reported elsewhere (Hirose and Gay, in press), the INT is considered to play a principal role in glottal adduction during speech. Since glottal width immediately prior to voice onset is obviously larger in breathy attack, it should be reasonable to expect that in order to accomplish the larger displacement for the glottal closure after /h/, INT activity must necessarily be higher. On the other hand, the activity of the LCA or the VOC is not always higher for breathy attack than for soft attack. This would suggest that these two muscles are not simply adductors of the vocal cords but might have additional functions, such as supplying medial compression or tension to the vocal folds.

What appears to characterize hard attack is the temporal pattern of LCA activity. The marked increase in LCA activity during the pre-phonatory period appears to be related to the strong medial compression or constriction of the glottis prior to release. A steep fall in LCA activity accompanied by a brief pulse cf PCA activity appears to be the physiological mechanism controlling abrupt glottal release after the period of constriction. In subject TG, VOC and INT also appear to contribute to strong constriction of the glottis during the pre-phonatory period.

It is characteristic in soft attack that PCA activity decreases gradually toward the onset of voicing, while the adductors appear to show gradual increase

[3] Subject TG showed highest INT activity for hard attack during pre-phonatory period. However, INT activity for voicing appears to be highest in breathy attack.

in activity for initiation of voicing. In the present study, the test utterance which was used for soft attack was initiated by /b/. It is conceivable, therefore, that the vocal folds hardly start vibrating before articulatory release even if they are adducted near to the midline. There may, however, be a difference in the action pattern for "soft attack" depending cn whether the utterance is initiated by a voiced consonant or a vowel.

In their recent study on the activity of the intrinsic laryngeal muscles in voicing control in speech, the present authors reported that PCA and INT show a reciprocal pattern of activity in voicing control of speech (Hirose and Gay, in press). It was revealed that PCA shows marked activity for the production of a voiceless consonant, while INT is suppressed. Conversely, INT generally shows higher activity for the production of vowels and voiced consonants, while PCA is reciprocally suppressed. It was further revealed that the other adductors, LCA and VOC, show a different pattern of activity in voicing control when compared with INT. Namely, LCA and VOC showed increasing activity for vowel production, while appearing inactive for consonant production regardless of the voiced vs. voiceless distinction, at least in that particular context.

In the present study, it is shown that INT shows a different pattern of activity in vocal attack from that of LCA and VOC in subjects LJR and LL. In subject TG, on the other hand, the three adductors show more less similar temporal patterns of activity, in which participation of INT in tight glottal closure in hard attack appears more dominant than in the other subjects.

There seems to be very little difference between the activity patterns of LCA and VOC in either of our two studies. The present data suggest there is no qualitative but perhaps some quantitative difference in their activity patterns. In previous studies, Hirano et al. (1970) have suggested that the two muscles function differently in register control for trained singers; differences between their results and ours may be due to the different tasks of the two groups of subjects. In any event, further study on various vocal maneuvers is needed to determine any possible functional differentiation within the adductor muscle group of the larynx.

CT is generally considered as a prime pitch raiser acting by tensing the vocal folds (Gårding et al., 1970; Gay et al., in press; Hiroto et al., 1967; Simada and Hirose, 1971). Gårding et al. (1970) reported that there is apparent antagonism between VOC and CT in the production of a glottal stop in one of their two Swedish subjects, in which CT activity appeared to be suppressed at the moment of maximum activity of VOC for the period of glottal closure. However, their data might not be comparable to the present data because the test utterance used in that particular experiment included variations of word accent and intonation in addition to glottal stop productions. The apparent suppression of CT activity in their data can be correlated to the falling in pitch toward the period of glottal closure. The present study revealed that CT shows more or less similar patterns of activity with VOC in subject LJR and LL and with LCA in subject AP in respect to the difference in the type of vocal attack when pitch is not changed.

In the present study, the measurement of so-called Δt was not attempted because of the ambiguity in defining the onset of EMG activity relative to the onset of the acoustic representation of voicing. For example, it is not unusual to observe in raw EMG records of the intrinsic laryngeal muscles a good amount of continuous resting discharge even during the period of silence between utterances. Thus, it appears difficult to define the onset of EMG activity from the raw EMG traces and, as a result, it is difficult to specify general rules for these measurements.

On the other hand, the temporal patterns of averaged muscle activity present in this paper certainly give no less information than simple comparisons of Δt and can be considered as more appropriate for comparisons of such activity patterns.

As shown in the previous figures, it is generally observed that laryngeal muscle activity, except for that of PCA, starts to increase earlier for hard attack than for the other two. This confirms findings reported in previous reports. However, what seems to differentiate the various types of attacks is not simply the pre-phonatory activity of the adductors but rather the overall activity patterns of all the intrinsic laryngeal muscles. In other words, different coordinated actions of the intrinsic laryngeal muscle systems, working in reciprocal fashion, determine each type of vocal attack.

REFERENCES

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 tone: A preliminary report on an EMG study. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, Univ. of Tokyo) No. 4, 45-54.

Gay, T., H. Hirose, M. Strome, and M. Sawashima. (In press) Electromyography of the intrinsic laryngeal muscles during phonation. Ann. Otol. Rhinol. Laryng.

Hirano, M. (1971) Laryngeal adjustment for different vocal onsets: An electromyographic investigation. J. Otolaryng. Japan. 74, 1572-1579.

Hirano, M. and J. Ohala. (1969) Use of hooked-wire electrodes for electromyography of the intrinsic laryngeal muscles. J. Speech Hearing Res. 12, 362-373.

Hirano, M., W. Vennard, and J. Ohala. (1970) Regulation of register, pitch and intensity of voice. Folia phoniat. 22, 1-20.

Hirose, H. (1971) Electromyography of the articulatory muscles: Current instrumentation and technique. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.

Hirose, H. and T. Gay. (In press) The activity of the intrinsic laryngeal muscles in voicing control: An electromyographic study. Phonetica.

Hirose, H., T. Gay, and M. Strome. (1971) Electrode insertion techniques for laryngeal electromyography. J. acoust. Soc. Amer. 50, 1449-1450.

Hiroto, I., M. Hirano, Y. Toyozumi, and T. Shin. (1967) Electromyographic investigation of the intrinsic laryngeal muscles related to speech sounds. Ann. Otol. Rhinol. Laryng. 76, 861-872.

Isshiki, N. and H. von Leden. (1964) Hoarseness: Aerodynamic studies. Arch. Otolaryng. 80, 206-213.

Koike, Y. (1967) Experimental studies on vocal attack. Oto-Rhino-Laryng. Clin. Kyoto 60, 663-688.

Koike, Y., M. Hirano, and H. von Leden. (1967) Vocal initiation: Acoustic and aerodynamic investigation of normal subjects. Folia phoniat. 19, 173-182.

Moore, P. (1938) Motion picture studies of the vocal folds and vocal attack. J. Speech Dis. 3, 235-238.

Port, D. K. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research SR-25/26, 67-72.

Sawashima, M. (1968) Movements of the larynx in articulation of Japanese consonant. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, Univ. of Tokyo) No. 2, 11-20.

Sawashima, M., M. Sato, S. Funasaka, and G. Totsuka. (1958) Electromyographic study of the human larynx and its clinical application. Jap. J. Otol., Tokyo 61, 1357-1364.

Simada, Z. and H. Hirose. (1971) Physiological correlates of Japanese accent patterns. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, Univ. of Tokyo) No. 5, 41-49.

Werner-Kukuk, E. and H. von Leden. (1970) Vocal initiation. Folia phoniat. 22, 107-116.

A Parallel Between Encodedness and the Magnitude of the Right Ear Effect

James E. Cutting*
Haskins Laboratories, New Haven

Early studies in dichotic listening (Broadbent, 1956; Kimura, 1961) presented different digits simultaneously to each ear. The results showed that this task overloaded the perceptual system, and numerous errors occurred. The errors, however, were differentially distributed; more errors occurred in recalling digits presented to the left ear than to the right. The superior performance of the right ear over the left is known as the right ear effect and has been explained, in part, as a reflection of linguistic capabilities of the cerebral hemispheres. In the dichotic situation it appears that linguistic information can best travel the path from the right ear to the left hemisphere (see Studdert-Kennedy and Shankweiler, 1970). We have known since the mid-nineteenth century that the left hemisphere of the brain is primarily responsible for language functions. Nevertheless, it was not known what aspects of dichotic stimuli were responsible for the right ear effect. Paired digits differ in duration, phonemic encodedness, syllabic form, and many other aspects. Any one of these differences might have been responsible for the ear effect.

Shankweiler and Studdert-Kennedy (1967) showed that the right ear effect was closely related to certain parts of the sound pattern of speech, but not to others. The identification of stop consonants in dichotic consonant-vowel (CV) syllables showed a large right ear effect. The identification of steady-state vowels, on the other hand, showed no significant ear effect.

Other classes of phonemes have been tested dichotically and appear to show results which are intermediary between stops consonants and vowels. Liquids and semivowels (Haggard, 1971) have been shown to give a right ear effect, but the magnitude appears to be smaller than that usually found for stops. Fricatives (Darwin, 1971) have been shown to give a small right ear effect when they have formant transitions, but no ear effect when the transitions are removed.

A possible synthesis of the results of these studies is to propose an ear-effect continuum which parallels an encodedness continuum (see Day, in press). Liberman et al. (1967) have used the term "encodedness" to describe the amount of acoustic restructuring a phoneme undergoes in various speech contents. Highly encoded phonemes (e.g., stops) undergo considerable change in their acoustic form as a function of their environments, whereas less encoded phonemes (e.g., fricatives, vowels), on the other hand, undergo little change. Thus the phonemes might be arrayed in parallel along an encodedness

*Also Yale University, New Haven.

continuum and an ear-effect continuum in the following manner: stop consonants are the most highly encoded phonemes and generally give the largest right ear effects in dichotic listening; liquids and semivowels are less encoded than stops and generally show smaller right ear effects; fricatives are less encoded than liquids and generally show a small right ear effect; and vowels are the least encoded of the phoneme classes and usually show no ear effect.

Thus far the existence of an ear-effect continuum and any parallel it might have with an encodedness continuum have been only speculative. No study has tested the various phoneme classes with the same group of subjects and made the appropriate comparisons. The present study attempts to make these comparisons using stops, liquids, and vowels combining them within CCV nonsense syllables.

GENERAL METHOD

Stimuli. Eight consonant-consonant-vowel (CCV) syllables were prepared on the Haskins parallel resonance synthesizer. There were three phoneme classes within each syllable: stops, liquids, and vowels. Each phoneme class was represented by two phonemes: /g/ and /k/ were the stops; /l/ and /r/, the liquids; and /ɛ/ and /æ/, the vowels. All possible combinations were used: /glɛ, klɛ, grɛ, krɛ, glæ, klæ, græ, kræ/. The stimuli were 455 msec in duration and had the same falling pitch contour. The duration of the formant transitions in the stop + liquid clusters was 210 msec followed by 245 msec of the steady-state vowel.

Subjects. Sixteen Yale undergraduates served as subjects in both experiments. They were all right-handed native American English speakers with no history of hearing trouble. Subjects were tested in groups of four, with stimuli played on an Ampex AG500 tape recorder and sent through a listening station to Grason-Stadler earphones.

EXPERIMENT I: IDENTIFICATION

A brief identification test was run to assess the quality of the stimuli.

Procedure. The subjects listened to two tokens of each stimulus to familiarize them with the synthetic speech. They then listened to a binaural identification tape of sixty-four items. Each of the eight stimuli was presented eight times in random sequence with a three-second interstimulus interval. Subjects were asked to identify each stimulus, writing their responses using the following orthography: GLEH, KLEH, GREH, KREH, GLAA, KLAA, GRAA, KRAA.

Results. The stimuli were highly identifiable. Subjects correctly identified the stimuli on more than 97% of the trials.

EXPERIMENT II: EAR MONITORING

Tapes and Procedure. The same eight stimuli were used; however, this time, instead of presenting one stimulus at a time, two stimuli were presented simultaneously, one to each ear. Dichotic tapes were prepared using the pulse code modulation system (Cooper and Mattingly, 1969). Each stimulus was paired with all other stimuli, but not with itself. There were 112 dichotic items per tape: (28 possible pairs) X (2 channel arrangements per pair) X (2 replications). Two such tapes were prepared with different random orders. Both tapes had a four-second interval between pairs. Subjects listened to two passes through each 112-item tape for a total of 448 trials. They were told to listen to one ear at a time and to write down which of the eight stimuli they heard presented to that ear. The order of ear monitoring was done in the following manner: half the subjects attended first to the left ear for a quarter of the trials, then to the right ear for half the trials, and finally back to the left ear for that last quarter (LRRL). The other half of the subjects attended in the reverse order (RLLR). There was a brief rest between blocks of 112 trials. The order of the ear monitoring, the order of the channel-to-ear assignments, and the order of the tapes were counterbalanced across subjects.

Results and Discussion

There are two major levels at which to analyze the data, the syllable level and the phoneme level.

Syllable level. Although the subjects were familiar with the eight stimuli many errors occurred in reporting the correct syllable in the monitored ear. A syllable was scored correct when all three phonemes were correctly reported. Overall performance was 58% correct. Subjects performed significantly better when they monitored the right ear than when they monitored the left [$F(1,15) = 20.96$, $p < .001$]; they were 62% correct in reporting the syllable when they attended the right ear and only 53% correct when they attended the left, a net 9% ear difference.

Phoneme level. Since there was a stop, a liquid, and a vowel in each stimulus, The syllable can be parsed to look at the overall performance and ear effects for each phoneme class.

If we consider each phoneme as a stimulus, there are two types of trials, *contrast* trials and *identity* trials. Considering the stops, there are those trials in which the two stimuli share the same stop, for example GREH/GLAA and KRAA/KLAA, and those which have different stops, for example GREH/KLAA and KRAA/GLAA. The first type of trial is a *stop-identity* trial, the second type is a *stop-contrast* trial. (Note that when considering a given phoneme class, we temporarily disregard the other phoneme classes.) There are also two types of liquid trials. GLEH/KLAA is a *liquid-identity* trial and GLEH/KRAA is a *liquid-contrast* trial. Vowels may be treated in the same manner; there are *vowel-identity* trials (e.g., KLAA/GLAA), and *vowel-contrast* trials (e.g., KLAA/GLEH). It is on the *contrast* trials that most (92%) of the errors occurred, and it is those which we will discuss first.

Contrast trials. First consider the stops. Subjects were 66% correct in reporting the stop in the monitored ear. There was a large, significant right ear effect [$F(1,15) = 22.55$, $p < .001$]: subjects were 72% correct in reporting the stops when monitoring the right ear and only 60% correct when monitoring the left, a net 12% difference. Eight of the sixteen subjects had significant right ear effects, and none had significant left ear effects as shown in Figure 1. These results were calculated using a chi square index discussed below.

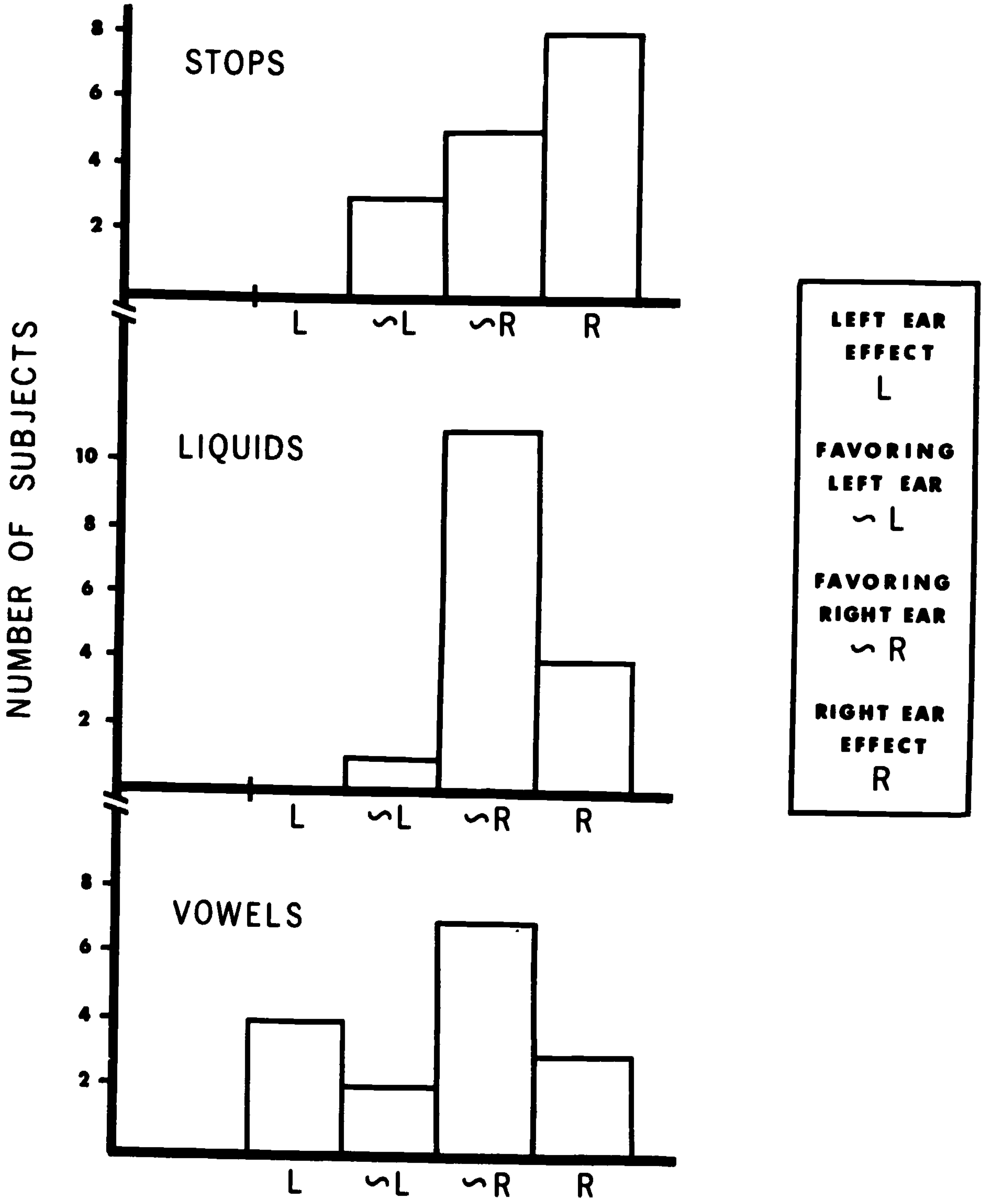


Figure 1: Distribution of subjects' ear effects for the three phoneme classes, calculated by the chi square index.

The liquids showed a pattern similar to the stops. Subjects correctly identified the liquid in the monitored ear on 64% of the trials. Again there was a significant right ear effect [$F(1,15) = 13.33$, $p < .005$], but somewhat smaller than that for the stops: subjects were 68% correct in reporting the liquid when monitoring the right ear and only 59% correct when monitoring the left, a net 9% difference. Figure 1 shows that, unlike the stops, only four subjects had right ear effects, but again, none had a significant left ear effect.

Vowels showed a very different pattern of results. Overall performance was considerably higher: subjects were 81% correct in identifying the vowel in the monitored ear. Furthermore, there was no ear effect for the group data. But the group average is misleading. Seven subjects did have significant ear effects: three had a right ear effect, and four had a left ear effect, as shown in Figure 1.

Chi square index and phoneme class comparisons. To pursue the idea of an ear-effect continuum for stops, liquids, and vowels, we must be able to compare ear effects for the three phoneme classes. To do this we must compensate for the different performance levels. A chi square analysis takes this consideration into account.[1] The analysis is performed on a 2 X 2 contingency table. The cell entries are the number of trials for a) right ear correct, b) left ear correct, c) right ear incorrect, and d) left ear incorrect. A chi square is computationally always positive. However, if we arbitrarily assign positive values to right ear effects and negative values to left ear effects, we have an index which distinguishes between the two results. A two-way chi square index was computed for each subject for each phoneme class with $p < .025$ as the criterion for rejecting the null hypothesis. Since the chi square index is a monotonic transformation of the original data, the chi square indices are suitable for further analysis.

Figure 2 shows the ear effects and ranges for the stops, liquids, and vowels arrayed in the order of their encodedness from high to low. The data is plotted in terms of percent right ear correct minus percent left ear correct. Thus left ear advantages yield negative scores. Note that the array appears to show an ear-effect continuum: right ear effect for the stops is greater than that for liquids, which in turn is greater than that for vowels. This linear trend from a large right ear effect for stops to no ear effect for vowels is also reflected in the ranges of the phoneme classes. A trend test (Winer, 1962) showed that this linear relationship was significant [$F(1,45) = 9.56$, $p < .005$] by analysis of variance on the subjects' chi square indices.[2] Furthermore, nine subjects showed this relationship: stops greater than liquids, greater than vowels. By chance alone this is a very unlikely outcome ($z = 3.91$, $p < .0005$). Only one subject had ear effects in the reverse order.

[1] I would like to thank Gary Kuhn for many suggestions which led to the use of of this statistic.

[2] In this type of analysis it is also necessary to consider other possible trends. Since there are only three phoneme classes, the only possible trends are linear and quadratic. The quadratic trend did not approach significance [$F(1,45) = .77$, ns].

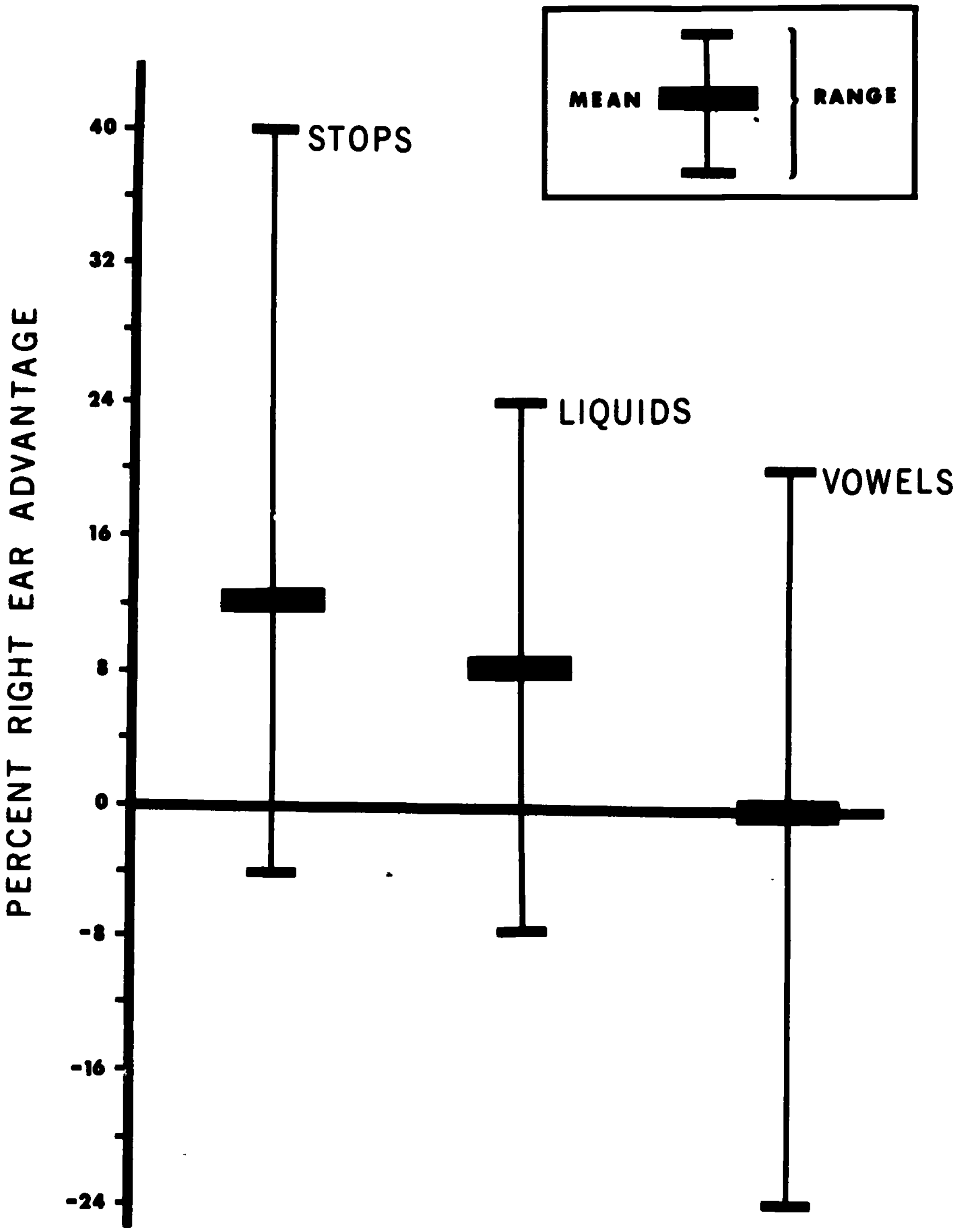


Figure 2: Means and ranges of subjects' ear effects for the three phoneme classes arrayed in the order of their encodedness.

A possible difficulty with the present study is the correspondence between phoneme class and temporal order. Stops are always first, liquids second, and vowels third. Nevertheless, the average ear effects shown in Figure 2 compare favorably with the results from other studies. Shankweiler and Studdert-Kennedy (1967) found a 14% right ear advantage for stops in CV syllables. They also found a net 9% right ear advantage for both initial and final stops in CVC syllables (Studdert-Kennedy and Shankweiler, 1970). Haggard (1971) found a 5% right ear advantage for initial liquids and semivowels in CVC syllables. For steady-state vowels Shankweiler and Studdert-Kennedy (1967) found a non-significant 4% right ear advantage and Chaney and Webster (1966) found a 1% right ear advantage.

Identity trials. Most of the errors (92%) occurred on *contrast* trials where phonemes of a given class were not shared. The remaining 8% occurred on *identity* trials. Few errors occurred on these trials because the same phoneme was presented to both ears. If /k/ was part of the stimulus in both ears, the subjects had little trouble in identifying the /k/ as part of the stimulus in the monitored ear. That errors occurred at all was probably a result of acoustic differences between the two instances of the same phoneme For example, a /k/ before /læ/ is slightly different than a /k/ before /rε/. Although *identity*-trial errors were relatively few, significantly more errors were made when subjects monitored the left ear than when they monitored the right ($z = 3.52$, $p < .0005$). There were no differences among the stop, liquid, and vowel classes for these errors. All had significant right ear effects.

Individual phonemes. Using a chi square analysis, we can also assess ear effects for individual phonemes within each phoneme class. There was no difference between the two phonemes in either the stop or the vowel classes. Both /g/ and /k/ had similar right ear effects. Both /ε/ and /æ/ had no ear effects.

There was, however, a difference between the liquids. Subjects had a 12% right ear advantage for /l/ and a 5% right ear advantage for /r/. This difference was significant ($p < .05$) by a Wilcoxon test on the chi square indices (Siegal, 1956). The occurrence of this differential ear effect for /l/ vs. /r/ is puzzling. The liquids often present puzzling problems in speech perception and speech productions; for a description of other interesting phonomena see Cutting and Day (in press).

Summary. Sixteen subjects were tested in a dichotic ear-monitoring task using stop-liquid-vowel nonsense syllables. The results showed that

1) There was an overall right ear effect for reporting the monitored syllables.

2) The ear effects for stops, liquids, and vowels were arrayed along a continuum. There was a larger right ear effect for stops than liquids, and a larger right ear effect for liquids than vowels. This relationship parallels an encodedness continuum for the same phoneme classes. Stops undergo more context condition variation than liquids, and liquids undergo more variation than vowels. Thus, the present study lends evidence for a parallel between the two continua.

REFERENCES

Broadbent, D.E. (1956) Successive responses to simultaneous stimuli. Quart. J. exp. Psychol. 8, 145-162.

Chaney, R.B. and J.C. Webster. (1966) Information in certain multidimensional sounds. J. acoust. Soc. Amer. 40, 2, 447-455.

Cooper, F.S. and I.G. Mattingly. (1969) Computer controlled PCM system for the investigation of dichotic speech perception. J. acoust. Soc. Amer. 46, 115(A). (Also in Haskins Laboratories Status Report on Speech Research, SR-17/18, 17-21.)

Cutting, J.E. and R.S. Day. (in press) Dichotic fusion along an acoustic continuum. J. acoust. Soc. Amer. (Also in Haskins Laboratories Status Report on Speech Research, 1972, SR-28, 103-113.)

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

Day, R.S. (in press) Engaging and disengaging the speech processor. In Hemispheric Asymmetry of Function, Marcel Kinsbourne, ed. (London: Tavistock).

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

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

Liberman, A.M., F.S. Cooper, M. Studdert-Kennedy, and D. Shankweiler. (1967) Perception of the speech code. Psychol. Rev. 74, 6, 431-461.

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

Siegal, S. (1956) Non-parametric Statistics. (New York: McGraw-Hill), p. 75-83.

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

Winer, B.J. (1962) Statistical Principles in Experimental Design. (New York: McGraw-Hill), pp. 132ff.

Mutual Interference Between Two Linguistic Dimensions of the Same Stimuli[*]

Ruth S. Day[+] and Charles C. Wood[++]

A single speech stimulus can be considered as a composite of values along many different dimensions. For example, a token of the syllable /ba/ will have a particular fundamental frequency, overall intensity, initial second-formant transition, formant values for the vowel, and so on. We are interested in the extent to which a given dimension can be processed independently of the others. An interesting and efficient way to study this problem is to select two dimensions and pit them against each other in a choice reaction-time paradigm. Subjects must attend to one dimension and ignore the other.

The dimensions studied in the present experiment were stop consonants (differing in place of articulation) and vowels. On each trial a single syllable was presented binaurally. In one task subjects had to target for the stop consonants, while in the other they had to target for the vowels.

Stop Consonant Task. Subjects pressed button #1 as soon as they heard /b/ and button #2 as soon as they heard /d/. This task was performed under two conditions of stimulus variation as shown in Figure 1. In the One-Dimension Condition, the target dimension (place of articulation for stop consonants) was the only one that varied: the stimuli were /ba/ and /da/.[1] A mean reaction time of 400 msec was obtained. In the Two-Dimension Condition, both stop consonants and vowels varied: the stimuli were /ba, bæ, da, dæ/. Again, subjects had to identify stop consonants, but they also had to ignore irrelevant variation in vowels. They had difficulty in doing so, as reflected by increased reaction time: the mean rose to 450 msec.

Vowel Task. Subjects pressed button #1 as soon as they heard /a/ and button #2 as soon as they heard /æ/. The same two conditions of stimulus variation were used (Figure 1). In the One-Dimension Condition, the target dimension (formant values for vowels) was the only one that varied: the stimuli were /ba/ and /bæ/.[2] The mean reaction time here was 348 msec. In the Two-Dimension Condition, the same four stimuli as in the Stop Consonant Task were used. This time, subjects had to identify vowels and ignore irrelevant variation in stop

[*] Paper presented at the 83rd meeting of the Acoustical Society of America, Buffalo, N. Y., April 1972.

[+] Haskins Laboratories and Yale University, New Haven.

[++] Neuropsychology Laboratory, Veterans Administration Hospital, West Haven, and Yale University, New Haven.

[1] Or, in another block of trials, /bæ/ and /dæ/.

[2] Or, in another block of trials, /da/ and /dæ/.

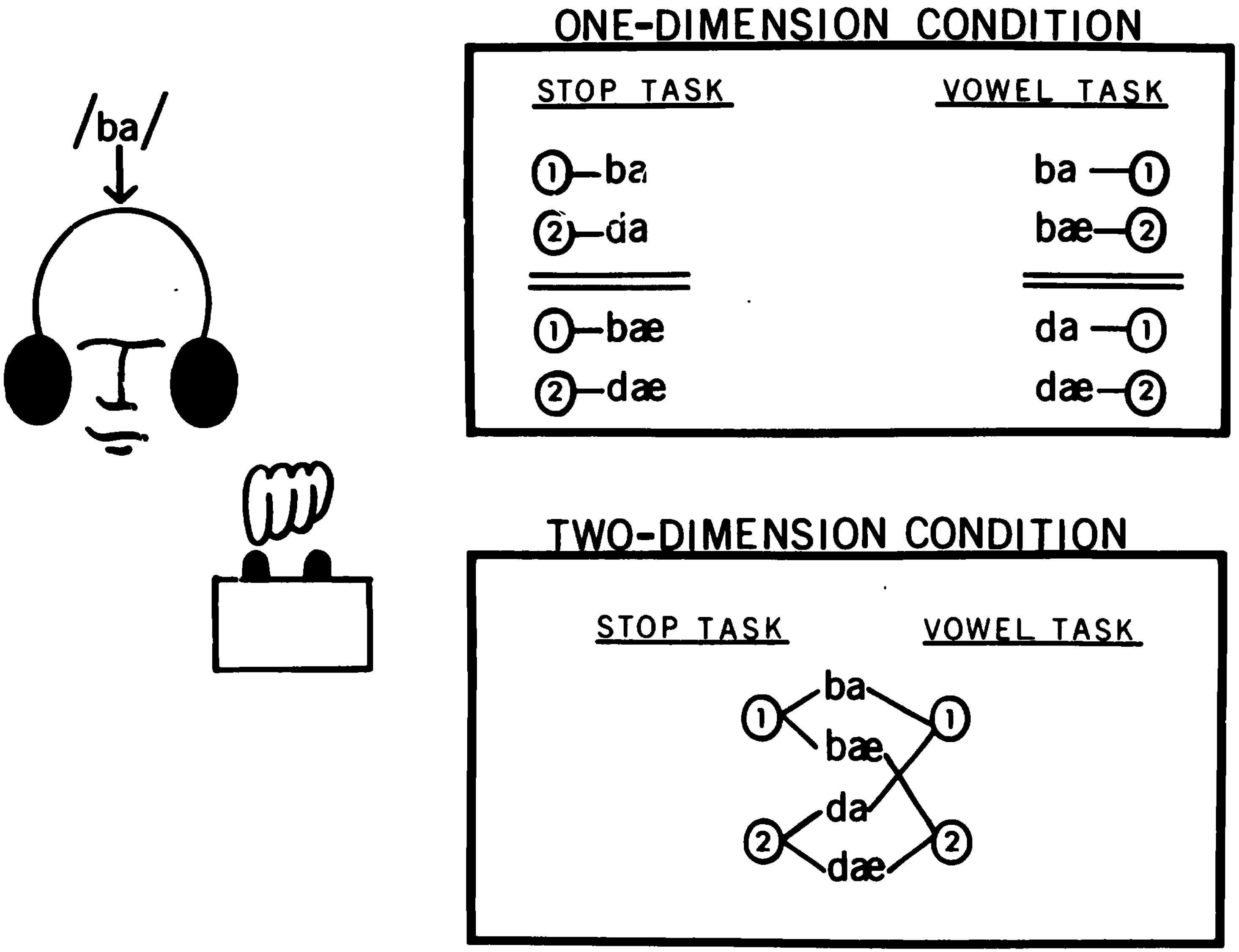


Fig. I

Figure 1: Stimulus arrays and button-press responses for each task and condition.

consonants. Again, this was a difficult task: mean reaction time rose to 414 msec.

The results of the experiment are summarized in Figure 2. We are interested in the extent to which reaction time for each target dimension increased in the Two-Dimension Condition. Both tasks yielded a sizeable increase. Thus there was a mutual interference effect: irrelevant variation in either dimension interfered with perception of the other.

How might these results be explained? One possibility is that the perceptual processors for place of articulation and for vowels are strongly interdependent. Such perceptual interdependence may well reflect known interdependence at articulatory and acoustic levels.

If this analysis is correct, then dimensions whose processors are not so strongly interdependent ought to yield a different pattern of results in this paradigm. Recently, we reported such a study (Day and Wood, 1972) in which the same stop consonants served as the linguistic dimension while fundamental frequency served as the nonlinguistic dimension. (Fundamental fraquency is nonlinguistic in the sense that it does not carry linguistic information at the phoneme level in English.) The results are shown in Figure 3. Again, the Stop Consonant Task showed a large increase in reaction time in the Two-Dimension Condition. However, the corresponding increase for the Fundamental Frequency Task was much less: it barely reached statistical significance.[3] Thus there was a unidirectional interference effect, in that it was much more difficult to ignore irrelevant variation in fundamental frequency while identifying stop consonants than vice versa.

The pattern of results for the stop consonant vs. fundamental frequency experiment suggest that these two dimensions behave in very different ways in the two-choice identification paradigm. When processing stop consonants, the listener cannot disengage his processing operations for fundamental frequency; however, when processing fundamental frequency, he can, to a considerable extent, disengage his linguistic processing operations. In fact, some subjects report that they are "unaware" of what consonants are occurring during the Fundamental Frequency Task; no one reports being unaware of pitch differences in the Stop Consonant Task.

It is also important to consider cases where both dimensions are nonlinguistic. Wood (1972) used stimuli that varied in both fundamental frequency and overall intensity and obtained a mutual interference effect: both dimensions interfered with each other to the same extent. These results are comparable to those of the present experiment. Note that in the present experiment there were two linguistic dimensions, while in that of Wood there were two nonlinguistic dimensions. A mutual interference effect may be a direct consequence of an interdependence of perceptual processors for the two dimensions. Thus far, this effect has occurred only for cases where both dimensions were from the same general class, that is, both linguistic or both nonlinguistic. The only cases where a unidirectional effect has occurred employed a dimension from each

[3] In a recent replication of this experiment, Wood (1972) obtained no increase for the Fundamental Frequency Task.

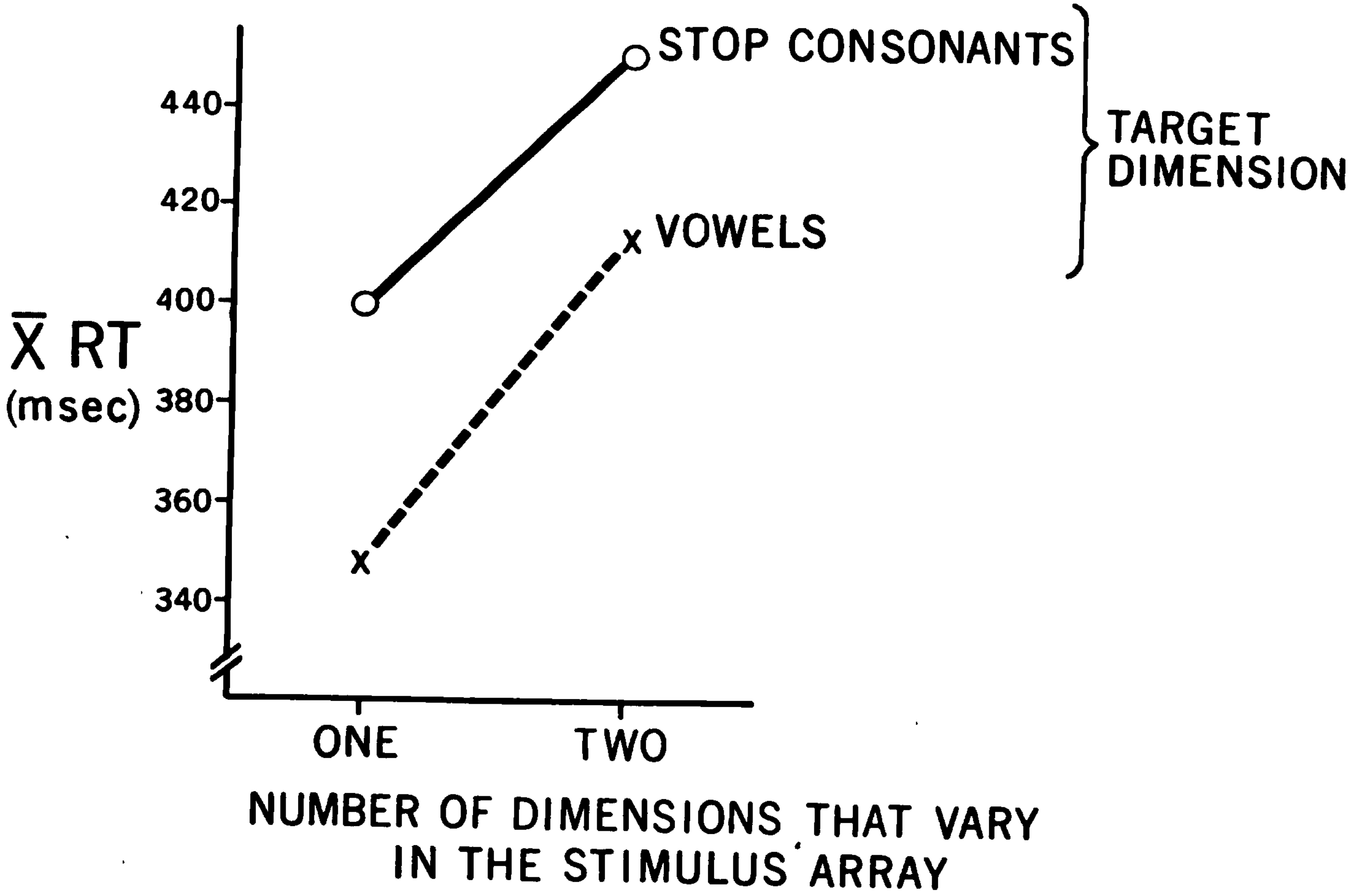


Fig. 2

Figure 2: Mean reaction time for identifying stop consonants and vowels under two conditions of stimulus variation.

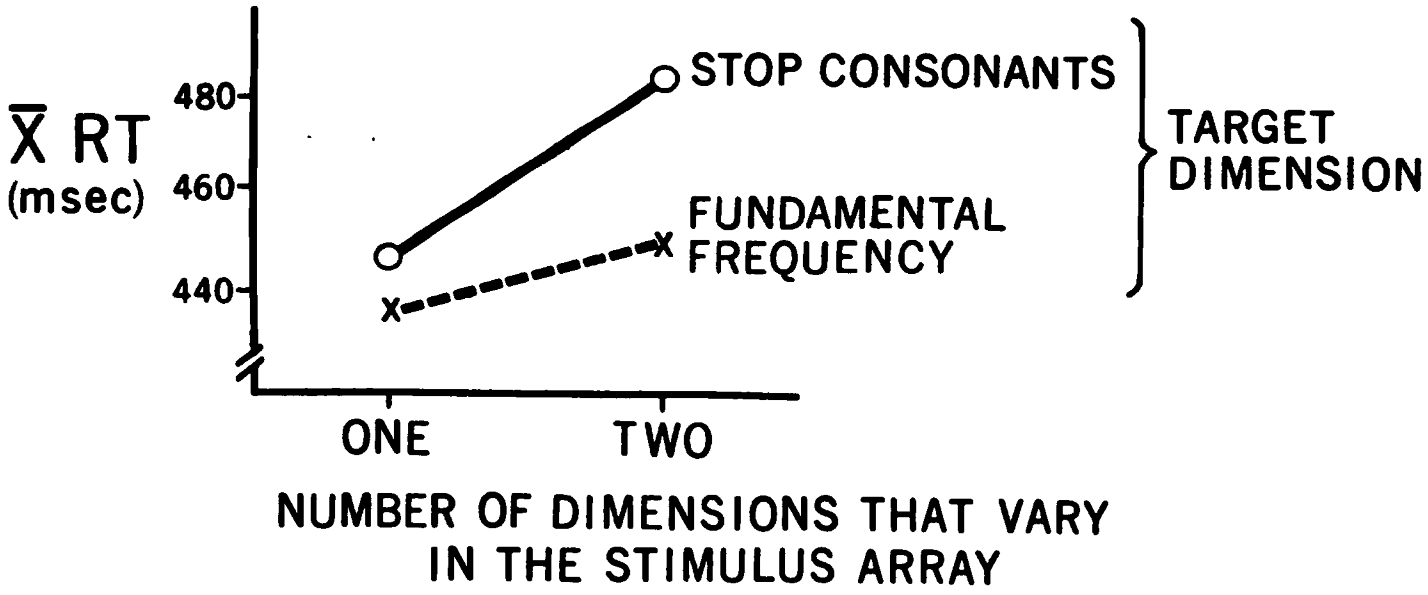


Fig. 3

Figure 3: Mean reaction time for identifying stop consonants and fundamental frequency under two conditions of stimulus variation. (After Day and Wood, 1972)

of the two general classes.

The status of each dimension as linguistic or nonlinguistic, then, appears to be important in predicting outcomes of these two-choice reaction-time experiments. There are, however, other factors that may be involved. In the experiments reported thus far, information about both dimensions is available from the onset of the stimulus. Situations where this is not the case may behave very differently. For example, variation in voice onset time vs. fundamental frequency would delay the onset of fundamental frequency information relative to stop consonant information. By studying such a situation we will be able to determine the extent to which temporal processes are important in perceiving various dimensions of the speech signal.

Another factor of possible interest here is the relative discriminability of each dimension. Thus far we have used pairs of dimensions that are of roughly comparable discriminability. It will be of interest to see whether decreased discriminability of certain dimensions will alter the basic pattern of results more than others.

Summary. Subjects listened to simple consonant-vowel syllables that varied along two dimensions: place of articulation for stop consonants and formant values for vowels. When they had to identify values along one dimension, it was difficult to ignore irrelevant variation in the other dimension. This was true for both dimensions to the same extent. These results are compatible with an explanation that emphasizes the degree of interdependence between processors for linguistic and nonlinguistic dimensions.

REFERENCES

Day, R. S. and C. C. Wood. (1972) Interactions between linguistic and nonlinguistic processing. J. acoust. Soc. Amer. 51, 79(A). (Also in Haskins Laboratories Status Report on Speech Research, 1971, SR-27, 185-192.)

Wood, C. C. (1972) Levels of processing in speech perception: Neurophysiological and cognitive analyses. Unpublished Ph.D. thesis, Yale University (Psychology).

The Phi Coefficient as an Index of Ear Differences in Dichotic Listening

Gary M. Kuhn*
Haskins Laboratories, New Haven

INTRODUCTION

Studdert-Kennedy and Shankweiler (1970) have applied the index $\frac{R-L}{R+L}$, where

R = the number of correct right-ear responses
L = the number of correct left-ear responses

to measure ear differences in dichotic listening tests. If the task of the subject in such tests is to identify one stimulus on each dichotic presentation, as was the case above, then this index may yield its maximum value of $\pm$ 1 regardless of the subject's level of performance. But if the task is to identify both stimuli on each dichotic presentation, then the maximum value of the index decreases rapidly, as cverall performance rises abovc 50%.

In order to avoid this ceiling effect in the two-response paradigm, Studdert-Kennedy has suggested that only those trials on which one stimulus is correctly reported should be included in the computation of the ear advantage (see the review in Halwes, 1969:24). There are, however, a few points about the effect of applying the index in this way that may be worth while to keep in mind.

First, the number of one-correct trials may vary considerably across subjects. As a result of such differences in sample size, we cannot necessarily have equal confidence, in the statistical sense, in ear advantages of equal reported magnitude.

Second, the proportion of one-correct trials may vary considerably across subjects. Thus the ear advantage reported over one-correct trials could look very similar for two subjects whose advantages over all trials (measured for statistical significance) were of very different magnitude.

Third, it may be that the proportion of one-correct trials varies systematically across levels of performance. But a measure of ear difference that does not take performance into account is one that assumes that overall performance

*Also University of Connecticut, Storrs.

Acknowledgment: B. McK. Johnson, L. Katz, C. Darwin, and D. Goldstein. Special thanks are extended to M. Studdert-Kennedy and D. Shankweiler for criticizing and discussing an earlier draft of this paper.

gives no information about ear advantage. Clearly this would be an interesting assumption to test.

APPLICATION TO TWO RESPONSES PER TRIAL

For these reasons it may be desirable to apply over all trials in a two-response paradigm an index of ear difference whose computed values can be compared directly for statistical significance, independent of the level of performance. Such an index may be derived from the X^2.

If we let

L = the number of correct right-ear responses
R = the number of correct left-ear responses
T = the number of dichotic presentations or "trials"

we may establish as the null hypothesis that the ears contribute equally to any observed R+L. Then for a two-response paradigm, we can express the expected outcome of a subject's performance in the following contingency table:

Ear of Presentation

		Left	Right	
Response Category	Correct	$\frac{R+L}{2}$	$\frac{R+L}{2}$	R+L
	Incorrect	$T - \frac{R+L}{2}$	$T - \frac{R+L}{2}$	2T - (R+L)
		T	T	2T

However, a subject's observed performance will be distributed according to the following table:

Ear of Presentation

		Left	Right	
Response Category	Correct	L	R	R+L
	Incorrect	T - L	T - R	2T - (R+L)
		T	T	2T

(1)

To compute the X^2 of the difference between the observed and expected frequencies of these two tables we sum the values of the following table:

		Ear of Presentation	
		Left	Right
Response Category	Correct	$\frac{(L - \frac{(R+L)}{2})^2}{\frac{R+L}{2}}$	$\frac{(R - \frac{(R+L)}{2})^2}{\frac{R+L}{2}}$
	Incorrect	$\frac{((T-L) - (T - \frac{(R+L)}{2}))^2}{T - \frac{R+L}{2}}$	$\frac{((T-R) - (T - \frac{(R+L)}{2}))^2}{T - \frac{R+L}{2}}$

where each cell is of the form

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

and

O = the observed number of responses for cell_{ij}
E = the expected number of responses for cell_{ij}

The sum of the evaluated expressions of these four cells is a X^2 with one degree of freedom, which can be used as a measure of the observed ear advantage and as a test of the H_o.

Note that the sum of the correct row equals

$$\frac{2(R - \frac{(R+L)}{2})^2}{\frac{R+L}{2}} = \frac{\frac{(R-L)^2}{2}}{\frac{R+L}{2}} = \frac{(R-L)^2}{R+L}$$

and that the sum of the incorrect row equals

$$\frac{2((T-L) - (T - \frac{(R+L)}{2}))^2}{T - \frac{R+L}{2}} = \frac{\frac{(R-L)^2}{2}}{T - \frac{R+L}{2}} = \frac{(R-L)^2}{2T - (R+L)}$$

A simplified form for stating the four-cell sum is then

$$X^2 = \frac{(R-L)^2}{R+L} + \frac{(R-L)^2}{2T - (R+L)}$$

This sum is a linear transform of the index $\frac{R-L}{R+L}$ with the y intercept changing as a function of performance.

If the observed number correct actually is the same for the two ears, then the four-cell sum will take on a value of Ø. At the other extreme, if one ear reports every stimulus correctly and the other ear none, then the generated value will reach its maximum of

$$\frac{T^2}{T} + \frac{T^2}{T} = 2T$$

We may normalize the scale of possible values of the four-cell sum. A normalizing procedure which reduces the maximum value of the sum to 1.Ø and its distribution under the H_o to that of a normal variate consists of dividing the computed X^2 by N, (N = 2T here), and then taking the square root of the resulting value.

$$X^2_{\text{norm.}} = \sqrt{\frac{X^2_{\text{comp.}}}{N}}$$

It turns out that the values of this normalized four-cell sum are equal in magnitude to those of the "phi coefficient," since

$$X^2 = N\ \text{phi}^2$$

(Walker and Lev, 1953:272). The phi coefficient is a correlation coefficient for two independent, dichotomously measured dimensions. If dimension 1 is either R or L and dimension 2 is either C or I then the 2x2 contingency table from which the strength of their phi could be evaluated would look like the following:

		1		
		L	R	
2	C	b	a	a+b
	I	d	c	c+d
		b+d	a+c	a+b+c+d

where a, b, c, and d are cell frequencies. For the special instance where the column totals are equal,

$$a+c = b+d$$

the computational formula for the phi coefficient reduces to

$$\text{phi} = \frac{a-b}{\sqrt{(a+b)(c+d)}}$$

(Walker and Lev, 1953:272).

Given this relationship between phi and the X^2 and the fact that the column sums in the observed frequencies contingency table (1) are indeed identical, we may compute the value of the normalized X^2 index from

$$\text{phi} = \frac{R-L}{\sqrt{(R+L)(2T - (R+L))}}$$

One reason for favoring computation of the value of this index through the phi formula is that the direction of the ear advantage, i.e., its sign, is retained. Computed in this way, the index can be thought of as yielding a value of correlation between correct performance and "right earedness": a negative value indicates a left-ear advantage.

By way of example, suppose that in a dichotic listening test of 100 trials, where the task was to report both stimuli on each trial, a subject gave the following performance:

$$L = 72$$
$$R = 88$$

His ear advantage would be

$$\text{phi} = \frac{R-L}{\sqrt{(R+L)(2T - (R+L))}} = \frac{88-72}{\sqrt{(160)(40)}} = .20$$

Critical Values

A table of the smallest significant or "critical" values of the index may be constructed by letting

X^2 = the value of $X^2{}_{1\ df}$ with the desired level of significance
N = 2T, i.e., the total number of responses

and solving the equation

$$\text{phi} = \sqrt{\frac{X^2}{N}}$$

The following table has been calculated in this fashion and is included here for illustrative purposes. From the table we see that the probability of obtaining an ear advantage as large as the one observed for the hypothetical subject above is <.01.

	Probability under the H_0 that phi $\geq$ PHI		
2T	.05	.02	.01
48	.282	.335	.371
80	.219	.260	.288
96	.200	.237	.262
100	.195	.232	.257
120	.178	.212	.235
160	.154	.183	.203
192	.141	.167	.185
200	.138	.164	.182
240	.126	.150	.166
320	.109	.130	.144
384	.100	.118	.131
400	.097	.116	.128
480	.089	.106	.117
640	.077	.091	.101
960	.063	.075	.083
1000	.061	.073	.081

A Comparison of Indices

For the sake of comparison, the ear advantages of five hypothetical subjects have been computed using three indices:

1. $\frac{R-L}{R+L}$ over all trials
2. $\frac{R-L}{R+L}$ over one-correct trials
3. phi over all trials

If the data for the 100 trials of a two-response paradigm are

Subject	Over all trials		Over one-correct trials	
	L	R	L	R
1	87	93	2	8
2	77	83	2	8
3	67	73	2	8
4	57	63	2	8
5	47	53	2	8

then the values of the ear advantages would be

Subject	$\frac{R-L}{R+L}$ all	$\frac{R-L}{R+L}$ one-correct	phi	P(phi ≥ PHI)
1	.033	.600	.100	not significant
2	.037	.600	.070	"
3	.042	.600	.065	"
4	.050	.600	.061	"
5	.060	.600	.060	"

50% Performance

If a subject's performance level over a given set of trials is 50%, then his errors equal his number correct

$$2T - (R+L) = R+L$$

and his total number of responses equals twice his performance

$$N = 2T = 2(R+L)$$

His four-cell sum is then

$$X^2 = \frac{(R-L)^2}{R+L} + \frac{(R-L)^2}{R+L} = \frac{2(R-L)^2}{R+L}$$

And since

$$phi = \sqrt{\frac{X^2}{N}}$$

we have

$$\text{phi} = \sqrt{\frac{\frac{2(R-L)^2}{R+L}}{2(R+L)}} = \sqrt{\frac{2(R-L)^2}{2(R+L)^2}} = \frac{R-L}{R+L}$$

Since the values of the two indices are identical for the case of 50% performance, it might appear that values of $\frac{R-L}{R+L}$ as computed over one-correct trials could be compared directly for statistical significance. This is not so, of course, if the size of the one-correct sample varies from subject to subject.

ONE RESPONSE PER TRIAL

Using the same computational formula,

$$\text{phi} = \frac{R-L}{\sqrt{((R+L)(2T - (R+L))}}$$

the phi index could also be applied to the data of a one-response directed recall listening test.

In this application, only a correct response from the requested ear is counted as correct under either condition of recall. Also, the quantity T is set equal to the number of trials under either condition of recall.

The null hypothesis here is that the ears contribute equally to the requested R+L. Phi is computed once over both conditions.

It does not seem to be appropriate to apply the phi to the data of a one-response, free-recall paradigm, since the performance of either ear may conceivably exceed half the total number of responses, or looked at another way, since an incorrect response cannot be assigned to either ear.

SUMMARY AND CONCLUSION

The phi correlation coefficient is proposed as an index of ear differences in dichotic listening tests. It is proposed specifically for the two-response paradigm, where, as an index of ear difference over all trials, it would be statistically appropriate for correlation with overall performance.

Using the same computational formula, the phi index may also be applied to the results of a one-response, directed-recall listening test.

The interest of the phi index lies in the fact that for a constant size of response set and number of dichotic trials, its values may be directly compared for statistical significance, independent of the level of performance.

REFERENCES

Halwes, T. G. (1969) Effects of dichotic fusion on the perception of speech. Supplement to Haskins Laboratories Status Report on Speech Research.

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

Walker, H. M. and J. Lev. (1953) Statistical Inference. (New York: Holt, Rinehart, and Winston).

The Relationships Between Speech and Reading[*]

Ignatius G. Mattingly[+] and James F. Kavanagh[++]

For scientists who have a special concern with language--researchers in linguistics, phonetics, speech science, experimental psychology, and communications engineering--no subject in the school curriculum arouses as much interest as reading. It is impossible to speculate very deeply about reading without touching on the nature of thought and language, and on the fundamental role that reading plays in this society. At first, of course, because his own experience of learning to read is so far in the past, the speculator takes his literacy for granted, just as he does his ability to speak and to listen to language. It is regrettable that some have speculated no further and rashly issued *ex cathedra* directives about the proper methods of reading instruction. Those who do consider a little further realize that reading is really a rather remarkable activity which could hardly have been predicted from what is presently known about the production and perception of speech and language.

Recent research by linguists in generative grammar and by experimental phoneticians in speech perception has, if anything, made reading seem even more remarkable. The form of natural language, as well as its acquisition and function, Chomsky (1965) tells us, are biologically determined. There is good reason to believe, according to Liberman et al. (1967), that linguistic communication depends on some very special neural machinery, intricately linked in all normal human beings to the vocal tract and the ear. It is therefore rather surprising to find that a substantial number of people can also, somehow, perform linguistic functions with their hands and their eyes. Reading seems more remarkable still when one considers that only in modern Western culture is it a basic social skill. Some civilizations have attained a high level of culture without being literate at all; in many others, reading and writing were the prerogatives of the hierarchy or the skills of the specialist. But this society insists that *everyone* learn to read and, if he wishes to obtain or retain middle-class credentials, to read in silence, rapidly and efficiently. In Augustine's (397 A.D.) *Confessions* (Book VI), he records his amazement on finding that when his teacher, Ambrose, was reading, "his eye glided over the pages, and his heart searched out the sense, but his voice and tongue were at rest...the preserving of his voice (which a very little speaking would weaken) might be the...reason for his reading to himself." How surprised Augustine would be if he could see millions of children learning to do Ambrose's little trick.

[*] Paper presented at the International Reading Association Convention, Detroit, Mich., May 1972.

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

[++] National Institute of Child Health and Human Development, Bethesda, Md.

Just about a year ago, a group, including researchers in all the disciplines mentioned earlier, met under NICHD sponsorship at Belmont, the Smithsonian Institution conference center in Maryland, for three days of papers and discussion on the relationships between speech and reading.[1] For the most part, they were people who had specialized not in the study of reading but in areas related to it in interesting ways: speech production and perception, phonology, information processing, language acquisition, memory. But the group also included a few people who had carried on research in reading for many years.

The original purpose of the conference was to consider speech and reading from the psychological and linguistic points of view, but the cultural role of reading came in for some heated discussion as well. In retrospect, it seems that there was one question which recurred throughout the conference. The question arose in various guises which may seem quite dissimilar at first. Its most familiar guise is the question of reading readiness: just what, besides competence in his native language, is necessary before the child can learn to read? Another version is, can reading and listening, as Bloomfield (1942) and Fries (1962) thought, be regarded simply as parallel processes in different modalities, converging at some point on a common linguistic path? Or, finally, one can put the question very abstractly: is it really possible to represent the relationships between speech and reading in the form of a nontrivial block diagram?

To answer these questions, or at least to understand them better, it seemed worthwhile to consider a number of differences between speech perception and reading that are interesting because they cannot be attributed merely to differences in modality.[2] To begin with, listening is easy and reading is hard. All living languages are spoken languages, and every normal child acquires through maturation a tacit knowledge of the grammatical rules of his native tongue and can speak and understand it. In fact, we are forced to conclude that the child has in some sense an innate ability to perceive speech, for without some such ability he could not collect the linguistic data that Chomsky (1965) asserts are required to infer these grammatical rules. Indeed, some recent work by Eimas et al. (1971) suggests that a four-week-old infant is capable of

[1] The conference was entitled "Communicating by Language--The Relationships Between Speech and Learning to Read." Those who attended or contributed to the conference included, in addition to the present authors, William F. Brewer, John B. Carroll, Carol Conrad, R. Conrad, Franklin S. Cooper, Robert Crowder, Eleanor J. Gibson, Philip B. Gough, Morris Halle, James J. Jenkins (co-chairman), Edward S. Klima, Paul A. Kolers, David LaBerge, Joe L. Lewis, Alvin M. Liberman (co-chairman), Isabelle Y. Liberman, Lyle L. Lloyd, John Lotz, Samuel E. Martin, George A. Miller, Donald A. Norman, Wayne O'Neil, Monte Penney, Michael I. Posner, Merrill S. Read, Harris B. Savin, Donald Shankweiler, and Kenneth N. Stevens. The conference proceedings will be published in September 1972 as Language by Ear and by Eye (J. F. Kavanagh and I. G. Mattingly, in press). (The papers given by Cooper, by I. Y. Liberman and Shankweiler, and by Mattingly appeared in SR-27.)

[2] These differences were pointed out by Liberman at an earlier NICHD conference (Kavanagh, 1968).

phonetic discrimination. On the other hand, relatively few languages in the history of the world have been written languages, and the alphabet seems to have been invented only once. In general, children must be deliberately taught to read, and despite this teaching, many of them fail to learn. Someone who has been unable to acquire language by listening--for example, a congenitally and profoundly deaf child--will hardly be able to acquire it by reading; on the contrary, a child with a language deficit owing to deafness will have great difficulty learning to read properly.

Secondly, the form in which information is presented is basically different for the listener and the reader. The listener is processing a complex acoustic signal in which the speech cues lie buried. (A "speech cue" is a specific acoustic event that carries linguistic information, for example, the aspiration that distinguishes voiceless /p, t, k/ from voiced /b, d, g/.) The cues are not discrete events, well separated in time and frequency; they blend into one another in complex ways. The segmental sounds the listener perceives quite often have no obvious segmental counterparts in the signal. To recover the phonetic segments, the listener has first to separate the speech cues from a mass of irrelevant detail. The process is largely unconscious; and in many cases a listener is quite unable to hear a speech cue as a purely acoustic event; he hears only phonetically (Mattingly et al., 1971). The complexity of the listener's task is indicated by the fact that no scheme for speech recognition by machine has yet been devised that can perform it properly. The reader, on the other hand, is processing a series of symbols which are quite simply related to the physical medium which conveys them. The marks in black ink are information; the white paper is background. The reader has no difficulty in seeing the letters as visual shapes if he chooses to, and optical character recognition by machine, though it is a very challenging problem for the engineer, is one that can be solved.

If reading and listening differed only in modality, one would expect that a visual presentation of speech that preserved the essential linguistic information could be easily read and, conversely, that an acoustic representation of written text which clearly differentiates the sounds representing the letters would be easy to listen to. But neither prediction is correct. It is possible to display speech visually in the form of a sound spectrogram, which shows the distribution of energy in the acoustic frequency range over time. We know that a spectrogram contains most of the essential linguistic information, for it can be converted back to acoustic form without much loss of intelligibility (Cooper, 1950). Yet reading a spectrogram is very slow work at best, and at worst, impossible. The converse task, "reading" written characters represented in acoustic form, is somewhat easier but not very fast. For example, Morse Code, or the various acoustic al habets for the blind reader, can be understood only at rates much slower than a typical listening rate for speech.

Finally, the number of different sounds used in speech in all the languages of the world is relatively small. These sounds can be classified in terms of their component phonetic features--voiced or voiceless, stop or fricative, labial or dental or velar--and the number of these features is very small--fifteen or twenty at most (Stevens and Halle, 1967). But the situation with the writing systems of the world, as one can verify by spending an hour or two looking at the plates in David Diringer's book, The Alphabet (1968), is very different. Formally speaking, the symbols used in writing systems have an endless variety, and so do conventions for arrangement of symbols on the page. Swift (1727) does not exaggerate in his description of the writing system of the Lilliputians in

Gulliver's Travels: "Their manner of writing is very peculiar, being neither from the left to the right, like the Europeans; nor from the right to the left, like the Arabians; nor from up to down, like the Chinese; nor from down to up like the Cascagians, but aslant from one corner of the paper to the other, like ladies in England." (Book I, Chap. 6)

However, if one looks at a writing system not just as an ensemble of visible marks but as a representation of some linguistic level, one finds a more orderly variation. The possible levels seem to range from the morphemic to the phonetic. Chinese characters are essentially morphemic; no information about pronunciation is given. If one wishes to read aloud in some dialect of Chinese one must have memorized the phonetic values of the characters in that dialect. The English writing system, as Chomsky (1970) has remarked, is essentially morphophonemic. Thus we use the letter *s* for the regular plural morpheme even though it is phonetically realized not only as [s] in *cats* but also as [z] in *cans* and as [əz] in *cases*. The orthography preserves the morphological relationship between *sign* and *signature* even though the phonetic vowel written as *i* is different in the two words and the *g* is pronounced in *signature* but silent in *sign*. But, as Martin points out in his conference paper, English, unlike Chinese, does not always define the morpheme boundaries clearly. Are *misled*, *molester*, and *bedraggled* to be read as *mis+led*, *molest+er*, and *be+draggled* or as *misl+ed*, *mole+ster*, and *bed+raggled*? Still other writing systems are fairly close to the phonetic level, for instance those used for Finnish or Spanish. Either their morphology is less complex than that of English, or some of the morphological complexity is masked by the written language for the sake of phonetic regularity. In his conference paper, Klima explores this range of orthographic variation from a theoretical standpoint, proposing several conceivable orthographic conventions for representing morphological and phonological content of sentences.

Twenty years ago, it could have been said that the range of writing systems spread over most of the known linguistic domain and that in principle there was no interesting restriction on the linguistic levels they represented, but the findings of the generative grammarians and the experimental phoneticians compel a drastic revision of this view. It is now clear that there are extensive areas in semantics, syntax, and speech perception which are part of the speaker's competence in his native language. Yet, except for the purpose of examples in the literature of linguistics and phonetics, one does not encounter writing consisting of deep structure tree diagrams and transformations, or, on the other hand, writing consisting of articulatory patterns, narrow phonetic transcriptions, distinctive features, or spectrographic patterns.[3] Thus, it now appears possible to make a significant generalization about writing systems. They actually represent, as Cooper pointed out at the conference, a relatively narrow linguistic stratum. Moreover, this stratum does not include the level at which the listener perceives speech. In short, writing tends to represent language at the morphemic,

[3] There have been a few interesting exceptions to this generalization. The Hankul alphabet of the Koreans (described by Martin in his paper for the conference) and the experimental writing systems of Wilkins (1668) and A. G. Bell (1867) described by Dudley and Tarnoczy (1950) represent each speech sound by a symbol depicting articulation, and Potter, Kopp, and Green (1947) used a moving spectrographic display in a project to teach the deaf to read speech sounds.

morphophonemic, or broad phonetic level, while speech represents language at the acoustic level.

The differences which have been listed indicate that even though reading and listening are both clearly linguistic and have an obvious similarity of function, they are not really parallel processes. Instead, a rather different account of the relationship of reading to language is proposed. This account depends on a distinction between primary linguistic activity itself and the speaker-hearer's awareness of this activity. Primary linguistic activity consists of the processes of producing, perceiving, understanding, rehearsing, or recalling speech. Many investigators have come to think that these processes are essentially similar, since they all require the construction or reconstruction of utterances in both phonetic and semantic form (Neisser, 1967). As a cover term for all these processes, the term synthesis may be used.

Having synthesized some utterance, the speaker-hearer is conscious not only of a semantic experience (understanding the utterance) and perhaps an acoustic experience (hearing the speaker's voice) but also of experience with certain intermediate linguistic processes. Not only has he synthesized a particular utterance, but he is also aware of having done so and can reflect upon this experience as he can upon his experiences with the external world.

If language were deliberately and consciously learned, this linguistic awareness would hardly be surprising. One would suppose that development of such awareness is needed to learn language, but language seems to be acquired through maturation. Linguistic awareness seems quite remarkable when one considers how little introspective awareness we have of the intermediate stages of other forms of complex behavior, for example, walking or seeing. The speaker-hearer's linguistic awareness is what gives linguistics its special advantage over other forms of psychological investigation. Taking his informant's awareness of particular utterances, not at face value but as a point of departure, the linguist constructs a description of the informant's intuitive competence in his language which would be unattainable by purely behavioristic methods.

However, linguistic awareness is far from being evenly distributed over all phases of linguistic activity. As Klima points out in his conference paper, some stages of linguistic activity are more "accessible" than others. Much of the process of synthesis takes place well beyond the range of immediate awareness (Chomsky, 1965) and must be determined inferentially. The speaker-hearer is unaware of the deep structure of utterances or of the processes of speech perception. He is aware of phonetic events and easily detects deviations, and this awareness can be increased with proper phonetic training. At the morphophonemic level, reference to various structural units is possible. Words are perhaps most obvious to the speaker-hearer, and morphemes hardly less so, at least in highly inflected languages. Syllables, depending on their structural role in the language, may be more obvious than morphophonemic segments. In the absence of appropriate psycholinguistic data, any ordering of this sort must be very tentative, and in any case it would be a mistake to overstate the clarity of the speaker-hearer's awareness and the consistency with which it corresponds to a particular linguistic level. But it seems safe to say that, by virtue of this awareness, he has an internal image of the utterance, and this image probably owes more to the morphophonemic representation than to any other level.

Linguistic awareness can become the basis of various language-based skills. Secret languages, such as Pig Latin (Halle, 1964) form one class of examplcs. In such languages a further constraint, in the form of a rule relating to the morphophonemic representation, is artifically imposed upon production and perception. If one has synthesized a sentence, an additional mental operation is required to perform the encipherment; and to carry out the process at a normal speaking rate, one has not only to know the encipherment rule but to have developed a certain facility in applying it. A second class of examples are the various systems of versification. The versifier is skilled in synthesizing sentences which conform not only to the rules of the language but also to an additional set of rules relating to certain phonetic features (Halle, 1970). To listen to verse, one needs at least a passive form of this skill to distinguish correct from incorrect lines without scanning them syllable by syllable. Like Pig Latin, versification requires awareness of the phonetics and phonology of the language.

It would appear that there are clear differences between language-based skills, such as Pig Latin and versification, and primary linguistic activity. For one thing, there seems to be considerable individual variation in linguistic awareness: some speakers are very conscious of linguistic patterns and exploit their awareness with obvious pleasure in verbal play (punning and charades) and verbal work (linguistic and phonetic research). Others seem never to be aware of much more than words and are surprised when quite obvious linguistic patterns are pointed out to them. This variation contrasts markedly with the relative uniformity among different individuals in the primary linguistic activity. Moreover, if one were unfamiliar with Pig Latin or with a system of versification, one might fail to understand what the Pig Latinist or the versifier was up to, but one would not suppose either of them to be speaking an unfamiliar language. And even after one catches on to the trick, the sensation of engaging in something beyond primary linguistic activity does not disappear; one continues to feel a special demand upon one's linguistic awareness. In short, synthesis of an utterance in primary linguistic activity is one thing; the awareness of this process of synthesis is qui another.

The conclusion suggested here is that reading is not a primary linguistic activity but a secondary language-based skill, and so requires a degree of linguistic awareness. The form in which a written sentence presents itself to the reader is determined not by the actual linguistic information to be conveyed by the sentence but by the writer's linguistic awareness of the process of synthesizing the sentence, an awareness which he wishes to impart to the reader. Since the reader has much the same linguistic awareness as the writer, and is familiar with the conventions of the writing system, he can synthesize something approximating what the writer intended, and so understand the sentence.

Since the writing system of English is, as has been said, essentially morphophonemic, the reader probably forms something like a morphophonemic representation as he reads. Does he also form a phonetic representation? Though it might seem needless to do so in silent reading, there is reason to think he does. In view of the complex interaction that must take place in primary linguistic processing, it seems unlikely that the reader could omit this step at will. Many information-processing experiments suggest that words and sentences are stored in phonetic form in short-term memory during the mysterious process by which the understanding of utterances takes place. Moreover, even though the writing system may be essentially morphophonemic, linguistic awareness is in part phonetic. Thus a sentence which is phonetically bizarre--"The rain in Spain falls mainly in the plain," for example--will be spotted by the reader. Again, many of those who manage to read and write ordinary text without "inner speech" or any signs of vocalization have to mumble their way through numerical computations, though the numerals, unlike alphabetic words, have no overt phonetic structure. Finally, Erickson et al. (in press) have shown that in a test of recall from short-term memory, Japanese subjects confuse kanji characters that are homophones, even though the kanji, like numerals, have no overt phonetic structure.

In conclusion, the questions raised earlier in this paper can be reconsidered. What is required for reading readiness? Apparently some degree of linguistic awareness, in particular (for English, at least) awareness of morphophonemic segments. Two of the conference papers directly support this view. Shankweiler and I. Y. Liberman found that a group of poor readers could often identify the first segment of a word like /bæg/ but usually failed to segment the entire word correctly. Savin reported that his subjects, poor readers in Philadelphia schools, could not master Pig Latin and shied away from any word game involving segmentation, but they were happy enough in games where syllable recognition was a sufficient skill. One begins to understand why the alphabet was invented only once.

Are reading and listening parallel processes? Evidently not. Reading appears rather to be parasitical on spoken language, exploiting the reader's awareness of the contents of short-term memory. And finally, can the processes of reading and speech be represented on a single block diagram? Not very easily, because one of the boxes in a block diagram of reading must itself include the kind of partial knowledge of the block diagram of listening and speaking that has here been called linguistic awareness.

REFERENCES

Augustine. (397 A.D.) The Confessions. Tr. Edward B. Pusey. Harvard Classics Edition. (New York: P. F. Collier, 1933).

Bell, A. M. (1867) Visible Speech: The Science of Universal Alphabetics. (New York: Van Nostrand).

Bloomfield, L. (1942) Linguistics and reading. Elementary English Rev. pp. 125-130 and 183-186.

Chomsky, N. (1965) Aspects of the Theory of Syntax. (Cambridge, Mass.: M.I.T. Press).

Chomsky, N. (1970) Phonology and reading. In Basic Studies on Reading, Harry Levin and Joanna Williams, eds. (New York: Basic Books).

Cooper, F. S. (1950) Spectrum analysis. J. acoust. Soc. Amer. 22, 761-762.

Diringer, David. (1968) The Alphabet. Third edition. (London: Hutchinson).

Dudley, Homer and Thomas H. Tarnoczy. (1950) The speaking machine of Wolfgang von Kempelen. J. acoust. Soc. Amer. 22, 151-166.

Eimas, Peter D., Einar R. Siqueland, Peter Jusczyk, and James Vigorito. (1971) Speech perception in infants. Science 171, 303-306.

Erickson, Donna M., I. G. Mattingly, and Michael Turvey. (In press) Phonetic coding in kanji. J. acoust. Soc. Amer. (A).

Fries, C. C. (1962) Linguistics and Reading. (New York: Holt, Rinehart and Winston).

Halle, M. (1964) On the bases of phonology. In The Structure of Language, J. A. Fodor and J. J. Katz, eds. (Englewood Cliffs, N.J.: Prentice-Hall).

Halle, M. (1970) On metre and prosody. In Progress in Linguistics, M. Bierwisch and K. Heidolph, eds. (The Hague: Mouton).

Kavanagh, J. F., ed. (1968) Communicating by Language: The Reading Process. (Bethesda, Md.: National Institute of Child Health and Human Development).

Kavanagh, J. F. and I. G. Mattingly, eds. (In press) Language by Ear and by Eye. (Cambridge, Mass.: M.I.T. Press).

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

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

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

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

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, W. Wathen-Dunn, ed. (Cambridge, Mass.: M.I.T. Press).

Swift, Jonathan. (1727) Gulliver's travels. In Gulliver's Travels A Tale of Tub, Battle of the Books, etc., W. A. Eddy, ed. (New York: Oxford University Press, 1933).

Wilkins, John. (1668) An Essay Towards a Real Character and a Philosophical Language. (London: Printed by J M for S. Gellibrand, etc.).

Audible Outputs of Reading Machines for the Blind*

Franklin S. Cooper, Jane H. Gaitenby, Ignatius G. Mattingly,+ Patrick W. Nye, and George N. Sholes
Haskins Laboratories, New Haven

The goal of research on reading machines for the blind at Haskins Laboratories is to produce by machine methods an output of clear, audible English from an input of ordinary printed text. The core problem--generating acceptable speech from phonetic spellings--seems very near a successful solution through synthesis-by-rule methods. There is still much to be done by way of evaluating and improving the synthetic speech, but the research can now turn to some of the other problems involved in setting up a complete Reading Service Center for the blind. During the six months covered by this report, attention has been focused on two main endeavors: evaluation studies of the reading machine output have continued with blind students, and further progress has been made toward automation of the entire print-to-speech generating process.

Evaluation by Blind Students

Continuing the work reported in the previous issue of the Bulletin, two studies have been made of student reactions to hearing some of their regular textbook assignments in the medium of synthetic speech. For the first study, with the help of faculty at the University of Connecticut, ten recorded passages totaling 2-1/2 hours of listening time were administered to six blind students. These passages covered chapters in psychology and psychiatry as well as ancient and modern literature. The content fell broadly into two classes: either basically simple prose style or more elaborate composition demanding close analytical attention.

Following these trial readings, the comments of the blind students showed general agreement on five points. First, all the students found the speech intelligible, and although an occasional word was missed, they had no trouble in following the meaning of the simple prose; however, some students found difficulty in concentrating on the subject matter of the more complex material. Second, all students were favorably impressed by the stress and intonation aspects of the speech. Third, all students complained about the "cold-in-the-head" quality of the speech, but the samples used were too short to determine whether the students would acclimate to this aspect of voice quality. Fourth, all students thought that the speed of presentation of the samples was too slow. [The rates ranged from 109 to 156 words per minute. The

*Summary prepared for the Bulletin of Prosthetics Research BPR 10-17, Spring 1972.

+Also University of Connecticut, Storrs.

latter is within the normal of human speaking rates but the long silences (2 to 8 sec.) between some sentences in these early recordings made the overall rate seem slow. These undue silences were eliminated in subsequent recordings.] Finally, long and often unfamiliar polysyllabic words were recognized easily. The words missed were usually monosyllables embedded in sequences of other short words.

In a second study during the late fall of 1971, another series of tapes was prepared at substantially higher word rates (164-221 wpm). These tapes received the benefit of more recent refinements in the rules for producing and recording synthetic speech. The most detailed comments on the second tests were obtained from two female students who voiced opposing views that were, however, typical of the group as a whole. Both students noted the improvements in naturalness compared with the earlier tapes. The first student indicated that, for passages that were complex (in topic, grammar, or vocabulary), she might well have preferred a slower rate. This student noted that her difficulty in focusing attention on the content (rather than on the voice quality) might disappear with longer experience in listening, but she was uncertain about how well she could use synthetic speech as a primary study tool. The second student, who listened to a text having a simple narrative style, was enthusiastic. She claimed to have missed only two words in a 15-minute recording that "spoke" at 221 wpm. She felt that she could make use of synthetic speech as a primary study tool.

Plans for Further Testing and for a Reading Service Center

More sophisticated tests are scheduled for the spring of 1972. In addition, a faculty committee at the University of Connecticut is now actively planning further steps toward the development of a Reading Service Center which will be located on the campus and will utilize the Haskins Laboratories speech synthesis facilities. These plans call for a two-part program commencing with a 12- to 18-month study of the human, economic, and technical factors involved in the operation of such a Reading Service Center. Haskins Laboratories will be involved in this study as a supplier of synthetic speech material to the University, using the automated facility currently being developed. The University researchers will be responsible for distributing the tape recordings and conducting sequential listening tests, both with blind students (some of them veterans) enrolled at the University and with blind students in schools and colleges throughout Connecticut. The second part of the program will incorporate the data from these studies to make decisions on the size and type of computer and optical character recognition equipment required for an on-campus Reading Service Center and to seek funding for its implementation during the 1973-74 academic year.

Automating Text Preparation

At the Laboratories, the task of automating the production of synthetic speech from an input of printed text continues. Enquiries are in progress toward the acquisition (on lease) of a limited-font optical character recognizer. Needed is a suitable machine for converting text that has been typed in OCR-A or -B (upper and lower case) type face into an alphanumeric code on magnetic tape. (The choice of a "simple" OCR machine and human typists for the initial production phase is based primarily on cost considerations.

Multifont machines to read book pages directly are available and their higher cost will be justified when a higher level of text production is wanted.) Optical character recognition represents the first of six stages involved in the production of a synthetic speech recording. These stages are shown in Figure 1.

Following the recognition stage, the words of the text are converted into phonemic form by means of a dictionary look-up (Stage 2). This dictionary now contains about 150,000 entries which are distributed in three compartments, with room for several-fold expansion. The first compartment contains a few hundred of the most frequently used words such as "the," "of," etc. In the second compartment is stored the overwhelming bulk of all entries. To facilitate access, this main store has been divided into functional "pages," which are referenced from a page-size table of contents. Locating an entry in the main store entails a two-part search, first through the table of contents, then through a page. The third compartment contains all oversize words (length greater than sixteen letters).

Each word entry, in both the high-frequency and main stores, contains the orthographic spelling, the phonetic respelling, and an indication of the word's usual grammatical functions. The initial version of the main store has now been completed, and programs for searching it are being written. These programs allow for editorial intervention to introduce new words that are not now available in the dictionary, as well as to correct errors.

Stress and Intonation

In the third stage, the phonemic string generated by the dictionary search is processed to introduce the stress and intonation features required to guide the synthesis program. Each dictionary word is (by the rules we are using) a member of one of five main stress classes: Low Stable, Low Unstable, Mid Unstable, High Stable, High Unstable. (Words with unstable stress shift their stress grade in specified contexts.) In general, low-stressed words are the so-called function words of speech (articles, prepositions, auxiliary verbs, many pronouns, connectives); words with mid stress are modifiers and verbs in the past tense (and past participles); high-stressed words are nouns (or multi-use words that can be nouns), words of four or more syllables, numerals, certain emphatic words, comparative and superlative forms of adjectives, and a small number of semantically special words that tend to receive full stress in normal speech.

In the fourth stage, the phonetic strings from Stage 2 and the stress and intonation assignments from Stage 3 are combined into a series of syllable-generating digital instructions by the computer program. These instructions are realized as a synthetic speech wave form by the synthesizer (Stage 5) which is recorded as a series of audible sentences in the final stage.

Recent work has centered on adjusting the specifications for the basic American English sounds (the phonemes) for better compatibility at fast word rates (above 150 wpm), modifying the speech program to provide pauses of various lengths, and refining the stress assignment rules for complex texts.

THE TEXT-TO-SPEECH PROCESSOR

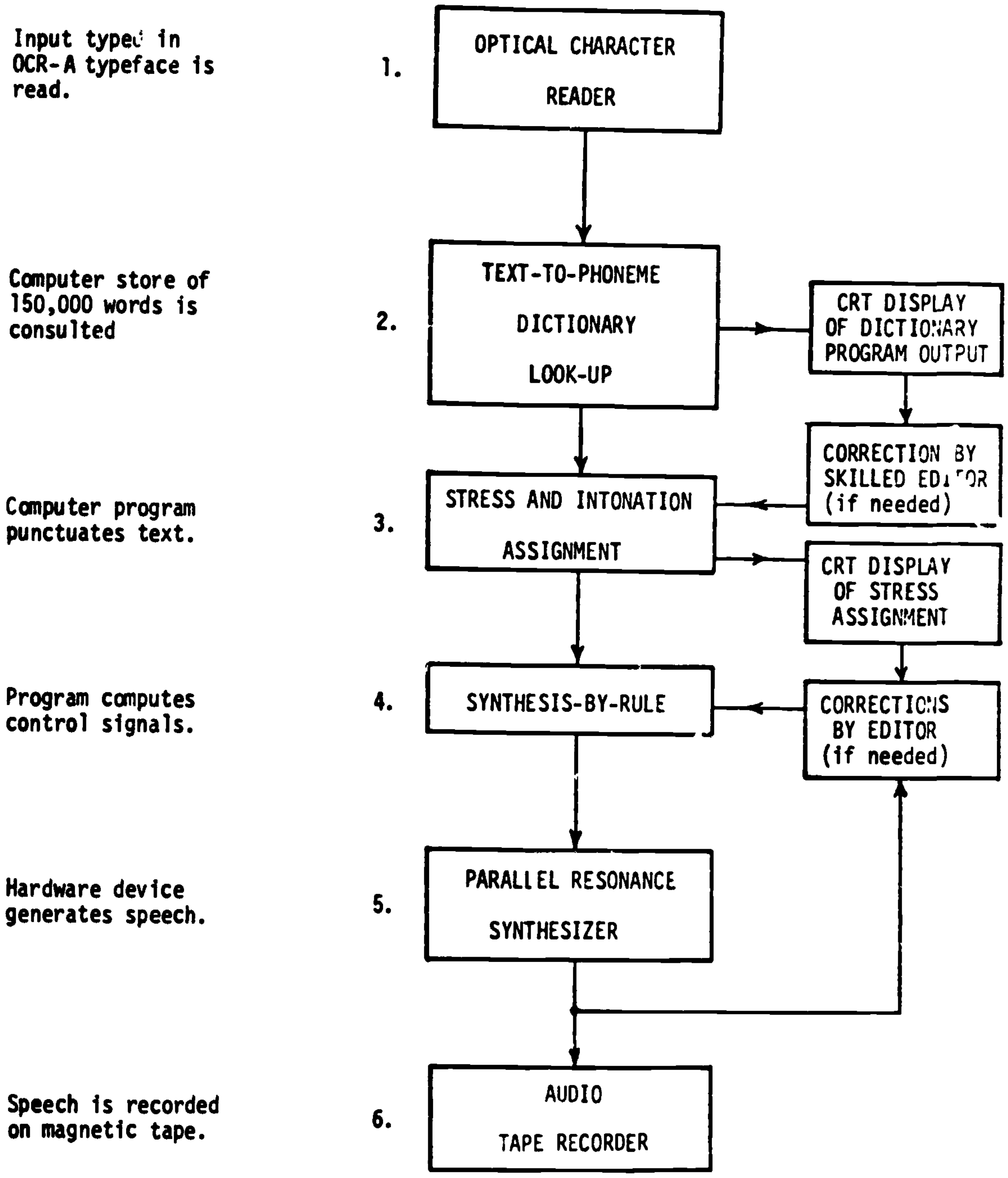

Input typed in OCR-A typeface is read.
1.
OPTICAL CHARACTER READER
Computer store of 150,000 words is consulted
2.
TEXT-TO-PHONEME DICTIONARY LOOK-UP
CRT DISPLAY OF DICTIONARY PROGRAM OUTPUT
CORRECTION BY SKILLED EDITOR (if needed)
Computer program punctuates text.
3.
STRESS AND INTONATION ASSIGNMENT
CRT DISPLAY OF STRESS ASSIGNMENT
Program computes control signals.
4.
SYNTHESIS-BY-RULE
CORRECTIONS BY EDITOR (if needed)
Hardware device generates speech.
5.
PARALLEL RESONANCE SYNTHESIZER
Speech is recorded on magnetic tape.
6.
AUDIO TAPE RECORDER

Fig. 1

Field Evaluation of an Automated Reading System for the Blind[*]

P. W. Nye,[+] J. D. Hankins,[++] T. Rand,[+,++] I. G. Mattingly,[+,++] and F. S. Cooper[+]

SUMMARY

After more than two decades of research it is now possible to construct a high-performance reading system for the blind which will produce synthetic speech from printed text. The entire process can be carried out automatically by computer and associated special-purpose devices. As a first step toward the eventual deployment of a reading system, we have begun an evaluation study in collaboration with faculty and students at the University of Connecticut and with trainees at the Veterans Administration Eastern Blindness Rehabilitation Center. Questions to be answered concern the comprehensibility and educational uses of the output and the technical and economic resources required to make automated reading services accessible to progressively larger groups of blind people.

INTRODUCTION

Our objective is to obtain an improvement in the variety, the quantity, and the speed of delivery of spoken material to blind people. We believe that this objective can be achieved through the use of an automated reader. With so many literate human readers presumably available, it may appear somewhat incongruous that effort should be spent on developing a machine to read books to blind people. A few words of explanation, therefore, are in order.

The justification for the work we are undertaking stems from the fact that the direct supply of reading matter to the blind is currently extremely slow and inadequate. This situation imposes severe educational, vocational, and recreational handicaps on blind people of all ages, but to blind college students, in particular, the limitation is critical. At the present time, blind students rely chiefly on Braille for note taking and on Braille, Talking Books, taped voice recordings, or face-to-face readers for more extensive material. However, text books are often unavailable in Talking Book or Braille form or are

[*] This is a revised version of a paper presented at the 1972 Conference on Speech Communication and Processing, Newton, Mass., 24-26 April 1972.

[+] Haskins Laboratories, New Haven.

[++] University of Connecticut, Storrs.

Acknowledgment: The evaluation study described in this paper is supported, in part, by the University of Connecticut Research Foundation.

Field Evaluation of an Automated Reading System for the Blind[*]

P. W. Nye,[+] J. D. Hankins,[++] T. Rand,[+,++] I. G. Mattingly,[+,++] and F. S. Cooper[+]

SUMMARY

After more than two decades of research it is now possible to construct a high-performance reading system for the blind which will produce synthetic speech from printed text. The entire process can be carried out automatically by computer and associated special-purpose devices. As a first step toward the eventual deployment of a reading system, we have begun an evaluation study in collaboration with faculty and students at the University of Connecticut and with trainees at the Veterans Administration Eastern Blindness Rehabilitation Center. Questions to be answered concern the comprehensibility and educational uses of the output and the technical and economic resources required to make automated reading services accessible to progressively larger groups of blind people.

INTRODUCTION

Our objective is to obtain an improvement in the variety, the quantity, and the speed of delivery of spoken material to blind people. We believe that this objective can be achieved through the use of an automated reader. With so many literate human readers presumably available, it may appear somewhat incongruous that effort should be spent on developing a machine to read books to blind people. A few words of explanation, therefore, are in order.

The justification for the work we are undertaking stems from the fact that the direct supply of reading matter to the blind is currently extremely slow and inadequate. This situation imposes severe educational, vocational, and recreational handicaps on blind people of all ages, but to blind college students, in particular, the limitation is critical. At the present time, blind students rely chiefly on Braille for note taking and on Braille, Talking Books, taped voice recordings, or face-to-face readers for more extensive material. However, text books are often unavailable in Talking Book or Braille form or are

[*] This is a revised version of a paper presented at the 1972 Conference on Speech Communication and Processing, Newton, Mass., 24-26 April 1972.

[+] Haskins Laboratories, New Haven.

[++] University of Connecticut, Storrs.

Acknowledgment: The evaluation study described in this paper is supported, in part, by the University of Connecticut Research Foundation.

delivered after long delays, sometimes up to several months. Ultimately, we believe, the solution lies in the establishment of a national network of Reading Service Centers utilizing reading machines capable of generating synthetic speech at a rate many times faster than natural speech (a rate of twenty to thirty times faster would be practical) and of recording the output on tapes moving at a proportionally fast rate. The speech could then be replayed by the listener at a normal speed. Braille could be provided as well when desired. These centers could be based in large regional libraries and would rapidly provide taped synthetic speech in response to requests made by mail, telephone, or in person. The service they would provide could not fail to have significant economic and social value by permitting a far larger segment of the blind population to contribute their skills to society.

The goal of a national network is, of course, far reaching and, in comparison with current expenditure on reading services for the blind, is likely to be considered expensive. But the problems to be faced in establishing such a network are not only economic. As stated elsewhere (Nye and Bliss, 1970), there are often many other difficulties to be met in our society in establishing an effective interface between a technical capacity and its potential field of application, and these are exemplified in the field of sensory aids for the blind. In fact, most of the difficulties are acutely visible in the whole bioengineering field (Task Group, 1971).

The research work on the development and improvement of synthetic speech has been in progress for a number of years. Further progress can be expected. Nevertheless, we believe that the point has now been reached when it is necessary to evaluate our progress and to determine whether the speech is good enough to apply in its present form. Our reasons are the following. First, synthetic speech, although not yet perfectly natural, has been developed to the point where it is intelligible to people who have received no prior exposure to synthetic speech or training in its use. Moreover, this is true of synthetic speech delivered at rates in excess of 150 wpm. No other reading machine output intended for use by the blind can make such a claim. It can be argued, therefore, that the value of synthetic speech has already been established and the question of how it may be deployed to provide a useful service can now be given serious consideration. Second, there is an immediate need in the blind community (particularly among students) for an increase in the supply and speed of delivery of spoken text. A reading machine is ideally suited to the task of producing large volumes of material quickly and can already start to fill the gap in present services by supplementing the material produced by human readers. Third, although synthetic speech appears at present to be at an economic disadvantage when compared with naturally produced speech, the costs of operating reading machines can be expected to fall in the future, whereas human labor costs will certainly increase. The eventual widespread use of reading machines is therefore inevitable. This conclusion leads to our fourth point which is that the initial entry of automated techniques into any new arena can always be expected to be met by new and often unforeseen problems. Such problems are usually amenable to solution. However, they first need to be identified and then time must be allocated to find ways of circumventing each difficulty. We believe this to be true for reading machines and that, in the interest of comprehensiveness, we cannot afford to delay any longer the task of evaluating synthetic speech with blind people under field trial conditions.

Thus the University of Connecticut and Haskins Laboratories are collaborating in the development of an evaluation program leading to the construction of a pilot Reading Service Center on the University campus--initially to serve the blind students enrolled there, but with the eventual goal of extending the service to other blind people statewide. The pilot center will be a first step toward setting up similar centers elsewhere. The evaluation and development work that we are undertaking builds upon the research carried out at Haskins Laboratories under Veterans Administration support. Initial reading tests with synthetic speech texts have already been conducted on blinded veterans, as well as with blind students, with encouraging results.

THE TEXT-TO-SYNTHETIC-SPEECH PROCESS

The speech synthesis system which we will use was constructed at Haskins Laboratories, both as a research tool for studies on the perception of speech and as a step toward the development of a reading machine for the blind. Figure 1 shows the sequence of steps involved in text-to-speech conversion. The characters, which are recognized by the optical reader, are grouped into words and recoded into phonemic form by means of an automatic dictionary. The phonemic text is "punctuated" with stress and intonation assignments and then transformed by another program into instructions for the control of a terminal analog speech synthesizer. Synthetic speech output from the synthesizer is then recorded on tape for use by the blind reader. A substantial part of this system--the speech synthesis procedure which embraces the last three steps of Figure 1--is already fully operational. Input to this completed portion of the system (by way of a phonetic keyboard) at present requires considerable hand labor. This work will be avoided when the first three stages, which are currently under development, are made operational.

Synthetic speech is currently being produced at Haskins Laboratories by a Honeywell DDP-224 computer which controls a hardware synthesizer designed by Cooper. To make the machine speak, a phonetically trained typist must transliterate the printed text into a phonemic text and type it on a keyboard attached to the computer. Stress and intonation markers are assigned by programmed rules to "punctuate" the phonemic text, as described by Gaitenby, Sholes, and Kuhn (in press). The typed phonemic symbols and punctuation are then displayed on a storage oscilloscope which allows the operator to examine the input to the computer and to correct typographical errors if necessary. Using this phonemic input, the computer calculates values for the dynamically controlled parameters of the synthesizer on the basis of programmed rules devised by Mattingly (1968). These values are then fed to the synthesizer at a rate set by the operator. In practice, speech can be generated at rates from 60 words per minute (wpm) to over 300 wpm. However, a passage of speech lasting for ten minutes at a normal presentation rate may well take the phonetic typist at least on hour to prepare. The way in which we propose to avoid the excessive labor and delay involves the addition of three major component steps which will automate to a large degree the tasks now performed by the phonetic typist and should greatly speed the process of transliteration. These steps will enable us to generate the relatively large volumes of reading material required to provide a reading service.

The first step employs an Optical Character Recognition (OCR) machine. Primarily because the size of our evaluation study does not merit the use of a multi-font OCR capable of reading proportionally spaced ink print, we plan to have text material retyped in an OCR-A upper- and lower-case type face and read

THE TEXT-TO-SPEECH PROCESSOR

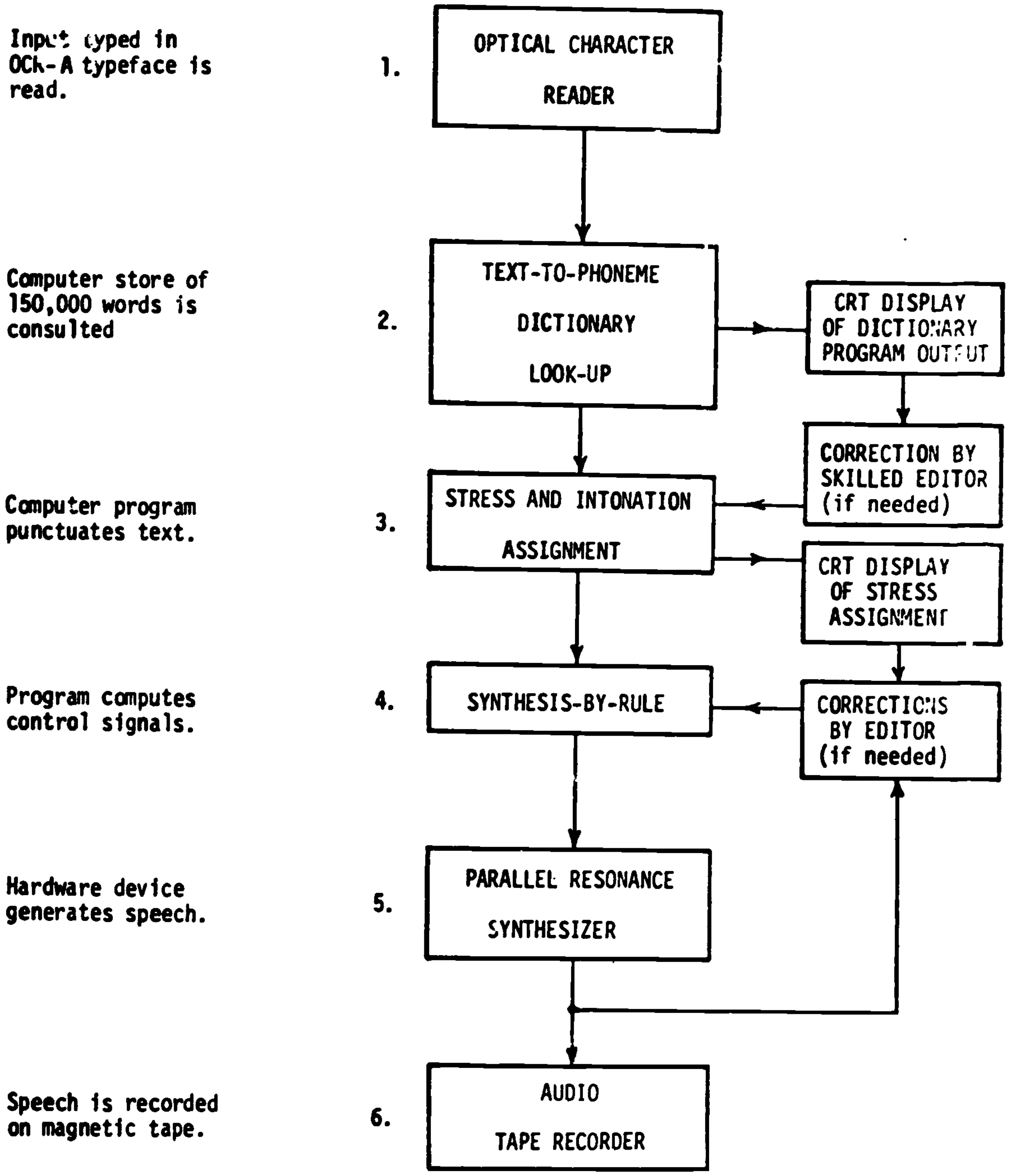

Input typed in OCR-A typeface is read.
1.
OPTICAL CHARACTER READER
Computer store of 150,000 words is consulted
2.
TEXT-TO-PHONEME DICTIONARY LOOK-UP
CRT DISPLAY OF DICTIONARY PROGRAM OUTPUT
CORRECTION BY SKILLED EDITOR (if needed)
Computer program punctuates text.
3.
STRESS AND INTONATION ASSIGNMENT
CRT DISPLAY OF STRESS ASSIGNMENT
Program computes control signals.
4.
SYNTHESIS-BY-RULE
CORRECTIONS BY EDITOR (if needed)
Hardware device generates speech.
5.
PARALLEL RESONANCE SYNTHESIZER
Speech is recorded on magnetic tape.
6.
AUDIO TAPE RECORDER

Fig. 1

by one of the smaller limited-font machines. The output will be recorded on digital magnetic tape for subsequent use by the computer. There are several other reasons why we prefer to use even a limited-capacity optical reader for input rather than an on-line typewriter, punched paper tape, or punched cards. The first is that much larger volumes of reading matter than we have used before must be fed into the computer as rapidly as possible. A good typist can typically work on straightforward text more rapidly and accurately than can a keypunch operator. Moreover, if the work is performed off-line, it need not occupy the computer during the text production process. Once a large volume of typescript has been prepared, an OCR reader can convert it into an alpha-numeric code expeditiously and cheaply. A second reason for our interest in OCR input lies in the opportunity it provides to obtain some introductory experience of current OCR technology so that we can better judge which machines and techniques best meet the needs of blind people and what problems still require solution. It is already apparent that the specifications of OCR devices designed for commercial applications do not fully satisfy the requirements of reading machines for the blind. For example, almost all of the commercial multi-font OCR development is geared toward high-speed and high-accuracy operation on an input medium for which some or all of the following features are closely specified: size and shape of the page, color, type styles, print quality, and the position of the printed text within the page. In contrast, a reading machine for the blind must be flexible with respect to each of these input specifications. It is possible that by deliberate design this flexibility could be gained at the expense of recognition accuracy which, for a reading machine application, may be a little less stringent than that required for business purposes. However, because of the limited market potential of the blind community it is unlikely that the commercial sector will show an interest in solving their special problems in the near future. The solutions must be sought by those who have a direct interest in the eventual development of automated reading services.

The second step required to speed our production of reading materials is the addition of an automated dictionary for converting the alphabetic representation of each word into its corresponding phonemic representation. In the compilation of our now-completed dictionary we are greatly indebted to the work of Dr. June Shoup of the Speech Communications Research Laboratory. This dictionary will initially include approximately 150,000 words but may have to be expanded. Optimum dictionary size can best be determined through actual use in a practical production system. Throughout this process the system's performance will have to be carefully watched to insure a proper balance between the size of the dictionary and search time or production rate.

The third step entails automated stress and intonation marking. Following its assembly by the dictionary search routine, the phonemic string will be punctuated with stress and intonation symbols by a program based on the rules of Gaitenby, Sholes, and Kuhn (in press). In the final system the output from the dictionary and the stress assignment routines will be displayed on a storage oscilloscope and monitored by an editor-phonetician. Corrections of errors will be made by the editor, who will also note the circumstances in which errors occur. By this procedure we expect to produce useful text and at the same time to detect defects and omissions in the system and readily correct them.

When the combination of phoneme string and stress markings has been formed, it will be processed by the speech synthesis program that is already in operation, converted to synthetic speech, and recorded.

EVALUATION STUDIES

The completed production system will generate synthetic speech from text in sufficient quantity to meet the needs of an evaluation study. The purpose of the trial is to determine whether the operation of a Reading Service Center is economically feasible. The question of economic feasibility is not easy to answer because of the intangible human values involved, but it is obvious that we need to identify costs and benefits as accurately as possible and to assess them in relation to available resources. To find the answers we need, we propose to operate the production system we have just described, to perform some analytical tests, and to transcribe into synthetic speech some actual reading assignments required by blind students at the University. These assignments will be supplied in a manner similar to that in which existing services operate at the University to provide recordings of natural speech. By examining the way in which these materials are used, we expect to find the answers to two broad classes of questions. The first relates to human factors, the second to technical and economic factors. Figure 2 illustrates the areas to be explored.

In the area of human factors, we are concerned with the relative comprehensibility of synthetic speech and natural speech over a range of speaking rates. The key question here is whether any differences in comprehensibility that may emerge are significant enough to affect the educational utility of synthetic speech. Analytical testing procedures will be used. The basic strategy for assessing comprehensibility involves the presentation of a lively passage of general interest followed by a series of questions which seek measures of the number of facts retained (i.e., names, places, distances, colors, etc.) and also the ability of the reader to derive logical inferences from the information. We propose to apply such tests using synthetic speech and natural speech controls (with appropriate counterbalancing) and then to compare the performance of the students. In a series of interviews designed to assess acceptability, we plan to gather data on such subjective factors as the relative preference for synthetic speech versus natural speech, the comparative comfort in use of the media, judgments regarding the aptness of different media for various fields of study, and the influence of delivery rate on all of these factors.

In the area of technical factors, we are concerned with establishing an accurate assessment of the overall demand that a Service Center will be required to meet, the technical quality of the synthetic speech medium required to produce acceptable performance at reasonable cost, the turn-around time which is both acceptable and economic, and the range of speaking rate required of the output. From these data an optimum equipment configuration can be determined and labor and operating costs can be estimated.

CONCLUSION

In this paper, we have argued that in order to provide better educational, vocational, and recreational opportunities for the blind population of this country, faster and more flexible reading services are required. Moreover, the technical resources are now available to supplement existing services through the use of reading machines located in Reading Service Centers. We believe that the time is now ripe to make a determined effort to move this technical capability out of the laboratory and into the community it could serve. The preliminary evaluative work we have described here is extensive and time consuming. Nevertheless, it is essential that an exploration of the extent to which the results of our research meet the needs of blind people be carried on in parallel with

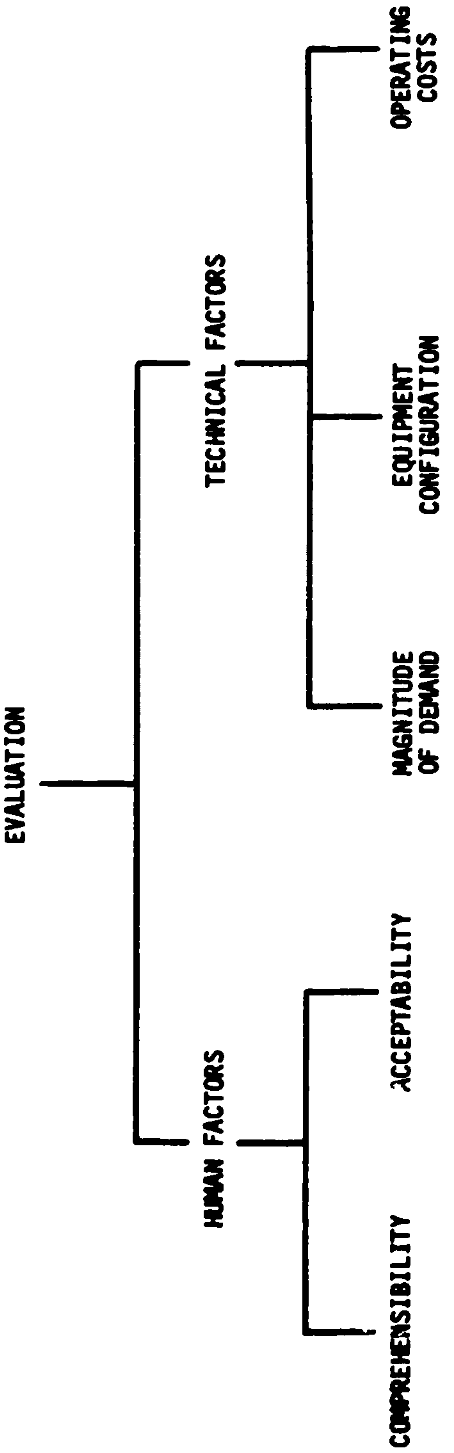

EVALUATION
HUMAN FACTORS
TECHNICAL FACTORS
COMPREHENSIBILITY
ACCEPTABILITY
MAGNITUDE OF DEMAND
EQUIPMENT CONFIGURATION
OPERATING COSTS

Fig. 2

continued research. This strategy must be followed if we are to apply effectively a laboratory-developed technology to a socio-economic problem as complex as blindness.

REFERENCES

Gaitenby, J. H., G. N. Sholes, and G. M. Kuhn. (in press) Word and phrase stress by rule for a reading machine. IEEE Trans. Audio and Electroacoustics. (See also this issue of the Status Report.)

Mattingly, I. G. (1968) Experimental methods for speech synthesis by rule. IEEE Trans. Audio and Electroacoustics AU-16, 198-202.

Nye, P. W. and J. C. Bliss. (1970) Sensory aids for the blind: A challenging problem with lessons for the future. Proc. IEEE 58, 1378-1898.

Task Group on Industrial Activity of CIEBM. (1971) An Assessment of Industrial Activity in the Field of Biomedical Engineering (National Academy of Engineering: Washington, D.C.).

Word and Phrase Stress by Rule for a Reading Machine[*]

Jane H. Gaitenby, George N. Sholes, and Gary M. Kuhn[+]
Haskins Laboratories, New Haven

ABSTRACT

A blind listener, using a reading machine that produces synthetic English speech, must receive an auditory version of a printed text that is intelligible, reasonably natural, and as fast as he likes. Good, fast, synthetic speech by rule (SSBR) has existed for some time; but stress and intonation tag assignment, being dependent partly upon the specific synthesis program used, has evolved slowly. The present stress assignment rules (written for Mattingly's SSBR) require a large stored dictionary, but the context rules are fairly simple and work quite passably in the majority of English prose constructions. These rules are explained briefly (and, at the conference, were demonstrated in synthetic speech).

The preceding paper by Dr. Nye and colleagues dealt with a plan for evaluating the synthetic speech output of a reading machine for the blind--the machine designed and developed at Haskins Laboratories. Our purpose here is dual: first, to explain some of the technicalities of providing the machine with the capability of stressing its words and phrases in a manner appropriate to General American English and, second, to demonstrate the particular approach currently in use.

By way of background, a short comparison of the reading process as done by humans and as done by machine seems necessary. It is obvious that the goal in reading machine research on synthetic speech outputs is to produce clear English that is as acceptable to a blind listener as a reading made by a human would be. How does a human read aloud? No one knows exactly how this is done, but an attempt at a quick description follows.

The good human reader comes fore-armed to the reading task with years of experience in conversational English (listening and speaking). A graphic word is equivalent to a familiar spoken word to him, and the word has one or another familiar meaning depending on its context. The reader is accustomed to the normal stress patterns of English, and to phrase and sentence grammatical and intonational structure. Since the written word is merely a symbolic substitute

[*] This is a revised version of a paper presented at the 1972 Conference on Speech Communication and Processing, Newton, Mass., 24-26 April 1972. It appeared as Paper A4 in the Conference Record.

[+] Also University of Connecticut, Storrs.

for the spoken word, using the same vocabulary and syntactic rules, the human, when reading aloud, utters the spoken words indicated in the text in the same serial order as they appear in print. As he reads aloud, he continuously scans the print ahead by eye, processes the visual information into chunks of meaning, and uses stress and intonation that are appropriate to the word order, to the punctuation, and to the vocabulary content.

A requirement of an electronic apparatus that reads text aloud is that it produce verbal results similar to the speech of a good human reader. Where possible, the reading machine has been equipped with parallels to human faculties. In place of eyes, an optical character recognizer will scan the words of the printed page in serial order. Substituting for the human's speaking and listening experience, the large dictionary stored in the computer memory will match the scanned incoming words with their usual phonetic equivalents (in machine code), and the stress patterns/levels of words in probable stress contexts will be assigned by rule (to be described below). From a table of American English phonemes and a program that combines the phonemes into syllables, the machine will digitally manufacture acoustic specifications for the words of the text. Finally, the machine will provide intonation, by rule, to the sentences it generates in synthetic speech (taking assigned stress into account).

The machine is capable of converting print to speech rapidly, for hours or days at a time, without voice fatigue. And the machine's output words can be produced at any one of a range of rates--depending on the blind user's preferred listening rate.

But the machine cannot think. In contrast to the skilled human reader who consults the deep structure of the language (that is, the meaning) as well as the surface structure (the phonetics suggested by the spelling, word order, punctuation, etc.) as he reads, the machine operates only on the surface structure level. The machine is not equipped to make stress or intonational adjustments that are signalled by overall meaning. But the machine can be programmed to treat categories of words in special ways, and to modify stress in certain positional contexts. This being so, stress assignment by machine depends on classification of the English vocabulary by predictable or probable stress patterns. (That which is unpredictable is, _ipso facto_, of no utility to the machine.)

Basic to English phrasal stress are the inherent (lexical) stress _pattern_[1] of each _polysyllabic_ word and the inherent stress _level_ of each _monosyllabic_ word (as a phrasal constituent).[2] The stress relationships of syllables within a

[1] The relative levels of sequential syllables within a polysyllabic word constitute its stress pattern.

[2] A monosyllabic word has, of course, no internal syllabic stress contrasts (stress pattern). The stress _level_ of a monosyllabic word has been inferred (and assigned) on the basis of its normal (probable) level within a normal phrase. The normal phrase can be construed as resembling a polysyllabic word: the phrase tends to have a stress pattern. The phrase, however, is made up of free morpheme units and is consequently rather less stable in its stress pattern than the polysyllabic word.

given polysyllabic word are ordinarily maintained in any sentence location and syntactic circumstance. English shows favorite word stress patterns: HIGH-LOW for two-syllable words, as in "cattle"; HIGH-LOW-MID for three syllables, as in "catalogue". Stress patterns for longer words are usually predictable if stress-st[illegible]fixes such as "-ation" are present. However, there are numerous exceptions to the common English stress patterns (due largely to abnormal stress in some compound words and words borrowed from Romance languages). This fact has made it worthwhile to store an entire English dictionary, with each lexical stress indicated, in the memory of the machine.

The stored dictionary contains a phonetic word in digital form (along with the lexical stress) to match each text word recognized by the optical scanner. (Rare proper names, recent coinages, and other words not contained in the dictionary will be generated by letter-to-sound rules.)

Before discussing the actual stress rules, Ignatius Mattingly's Speech Synthesis by Rule Program, which the word and phrase stress rules operate upon and within, will be briefly described. In order to illustrate the program for synthesis in the simplest way, it will be described in synthetic speech itself, with its stress, intonation, and phonetics generated entirely by rule. The textual input to the machine used in generating the demonstration tape was typed on a phonetic typewriter (simulating the print-to-phonetics conversion in the dictionary, as well as the automatic implementation of the stress rules). The typed phonetic data went directly to the computer where the synthesis program determined the computation of the acoustic features for the sentences. The computed material was then synthesized by a parallel formant generator.

> [SYNTHETIC SPEECH DEMONSTRATION 1] "This is the voice of the synthesizer at Haskins Labs. There are two main parts of Mattingly's Speech Synthesis Program. The first part consists of a table of standard American English phonemes, and the second part consists of digital instructions for combining the phonemes into syllables with reasonable intonation. There are four grades of stress possible in the program, of which three are being used in this demonstration. Mattingly's Rules compute the intonation for each breath group on the basis of the punctuation given in the printed text input, and on the basis of the stresses assigned."

(At the meeting it was then explained that the Mattingly program includes a choice of three intonational contours: the Fall, Fall-Rise, and the Rise.)

> [SYNTHETIC SPEECH DEMONSTRATION 2] (This consisted of words "yes" and "no" played in each of the three intonational contours, at three word rates.)

To get to the stress rules, as mentioned above the stress assignment procedure takes as its point of departure a dictionary that includes inherent stress along with the phonetic word, in machine language. (That is no small endowment!) The dictionary words are also tagged as members of stress categories that are compatible with the synthetic speech program.

Three grades of stress are being used. Stress III (LOW), the unstressed grade, is realized in synthesis as low in pitch (F_0) and short in duration. Stress II (MID, or secondary stress) has longer duration than Stress III.

Stress I (HIGH, or primary stress) has the same relative duration as MID stress, but the fundamental frequency of a HIGH syllable is raised above the basic intonational contour. (Intensity is not manipulated in the SSBR program to indicate stress except in stop consonants. Intensity does change as a function of the spectral properties intrinsic to individual phones, howeve:. In the case of vowels, in particular, this spectral property is in itself a strong cue to stress.)

To group words into what has turned out to be five major word categories (and several subcategories) two features have been ascribed to each word: stress stability and probable stress level (stress level of the lexically stressed syllable, if there is one). Words that are classed as stable maintain their assigned stress in all contexts; unstable words alter their dictionary stress level in specified contexts, in specified ways.

Stress Group A consists of many of the so-called function words--mostly monosyllables--(for the most part, these are personal pronouns, auxiliary verbs, and conjunctions). These words are usually unstressed in speech. They are among the 100 most frequent English words in texts and are of major importance to the rhythm and intonation of speech, even when they are barely audible. Words belonging to Group A are classed as unstable in stress. They are unstressed (LOW)--except in pre-pause position where they become MID. There are exceptions to this rule that apply to some sequences of function words such as two successive prepositions (in which case the first preposition receives more stress than the second). Prepositions and similar words are accordingly placed in a subcategory. Other exceptions are words like "the," "of," "and," "as," and "him." The phonetic shape of these words is dictated by immediate context, positional (e.g., initial in breath group) in some cases, and phonetic (e.g., followed by a consonant) in others.

[SYNTHETIC SPEECH DEMONSTRATION 3] (Two sentences exemplifying stress Group A were played.)

Stress Group B contains only a few words--some pronouns in the objective or possessive case, such as "me," "my," "their," "us," and several contractions: "I'll," "it's," etc.--words that are rarely final, or else very seldom stressed unless italicized. This group is classed as stable in stress and LOW.

Group D contains a mixed bag of words from the point of view of grammatical role. (Group C will come later.) D is characterized as a stable stress group and HIGH. This means that the lexically stressed syllable in a D word receives HIGH stress in all locations. The words comprising this group are: all words of four or more syllables (a primary stress seems to be a requirement in long words); all spelled-out numerals (but the word "one" is a special case); all words of two or more syllables ending in "-ing" and "-ings" (except for the auxiliary verbs of Group A); polysyllables ending in "-ion" or "-ions," ending in "-al" or "-als," or ending in "-ic" or "-ics"; and a list of specific words (seventy-five such words at last count) most of which deal semantically with limit or extent, e.g., "both," "else," "fully," "rare," "similar," "single," "entire." Also in Group D are all comparative and superlative adjectives. (A special feature of this group is that it causes an adjacent single noun to the right to lower its stress. This feature is actually stated as a noun stress shift rule.)

There are two more stress groups. Group E consists of all nouns not contained in Group D. Root (or present-tense) forms of verbs that also may function as nouns are classed as nouns alone, since nouns as a class occur more frequently than do verbs. This group is unstable in stress. It receives MID stress when preceded by a Group D word as in "four books," "better idea," "every day." It also receives MID stress when preceded by its own class, E-- as in "book store," "market basket," "cotton mill." Otherwise the words in Group E receive HIGH stress. (Proper names, capitalized initially, are an exception to this rule. The rightmost word of a string of initially capitalized words receives HIGH stress by rule; those at the left are assigned MID stress.)

[SYNTHETIC SPEECH DEMONSTRATION 4] (Five sentences demonstrating Groups D and E were played.)

The last general stress category, Group C, contains all the words not already classified. Group C thus contains a host of adjectives, adverbs, past-tense verb forms and past participles (many of which are also adjectives). This group can be viewed as intermediate on a scale of information content, generally less significant in a message than nouns and other semantically powerful words. MID stress has proven appropriate for this group in all positions except pre-pausal, where a C word receives HIGH stress. Group C is thus unstable in stress.

[SYNTHETIC SPEECH DEMONSTRATION 5] (Group C words were demonstrated in three sentences. All of the synthetic speech in Demonstrations 1 through 5 was played at rates within a 120-140 words per minute range.)

There are few words that do not fit fairly well, in practice, within one of these stress groups, but there are some words that require additional rules. Special rules are being developed as grouped exceptions are noted in the course of producing texts for blind students. (Capitalized words have already been mentioned, and hyphenated words are another special case. A single hyphen, not line final, calls for a HIGH-MID stress sequence for the words it joins. Two hyphens, joining three words, signal a HIGH-LOW-MID stress sequence.)

[SYNTHETIC SPEECH DEMONSTRATION 6] (Examples of special cases were played.)

The criteria for establishing the membership of a word in one of the stress groups described are not elegant, it must be admitted. To recapitulate: function words fall within Group A; certain pronouns and contractions, for the most part, make up Group B; words of four or more syllables, comparative and superlative adjectives, numerals, and a list of special words comprise Group D; nouns not belonging to Group D fall into Group E; and all the remaining words are gathered in Group C (with exceptions as noted earlier).

But if the categories are regrouped according to the stress grade initally assigned them, some order appears. Groups D and E, assigned HIGH stress (nouns and the like) are high in information and low in predictability. Most of these words occur as subjects or objects in sentences and are customarily stressed in the spoken chain. In contrast, the highly frequent and low in information words of Groups A and B that have been assigned LOW stress are the barely audible connective tissue of spoken sentences. And Group C, with MID stress, containing all the other words, consists chiefly of modifiers and verbs denoting completed action (and these words may be considered intermediate in information content).

The five stress groups are rules for stress assignment in running speech, and they work well in view of their relative simplicity. Parsing is not necessary. (The inherent stress grade assigned to each word in the stored dictionary represents a form of parsing.) Tagging the thousands of words in the dictionary by stress type, however, has presented interesting problems in programming. At present the stored dictionary contains excess information. (The dictionary was "inherited," so to speak, and came to Haskins with an embarrassment of grammatical riches attached.)

In conclusion, here is one more synthetic speech sample using the stress rules (word categories and stress shifts due to context) that have been described. The text to be heard is one that was requested by a blind student for a reading assignment in psychology at the University of Connecticut. The tape will be played at 120 words per minute, then at 150, and finally at 200 words per minute.

[SYNTHETIC SPEECH DEMONSTRATION 7] (A 132-word text sample was played.)

Auditory Evoked Potential Correlates of Speech Sound Discrimination[*]

Michael F. Dorman[+]
Haskins Laboratories, New Haven

Numerous studies have indicated that the sounds of speech enjoy a special mode of perception, distinct from that of nonspeech signals (Liberman, 1970; Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). One set of investigations supporting this view has examined the relationship between identification and discrimination of speech and nonspeech signals. Listeners can discriminate many more nonspeech stimuli than they can identify absolutely (Miller, 1956; Pollack, 1952). However, certain speech sounds, the stop consonants [b,d,g,p,t,k], tend to be discriminated no better than they can be identified (Pisoni, 1971; Studdert-Kennedy, Liberman, Harris, and Cooper, 1970). This unique relationship between identification and discrimination is termed "categorical perception."

In a typical experiment, Lisker and Abramson (1970) presented to Ss for identification and discrimination a series of computer-synthesized stop consonants which differed solely along the physical continuum of voice onset time (VOT).[1] Listeners identified these stimuli exclusively as members of the phonetic category [ba] or [pa]. Ss discriminated almost perfectly between stimuli which were assigned to different phonetic categories. However, when physically different stimuli were drawn from the same phonetic category, discrimination was only slightly better than chance. Thus, equal acoustic differences (for example, 20 msec) along the VOT were not equally discriminable. Only when

[*] This paper reports a portion of the research carried out for a Ph.D. dissertation accepted by the University of Connecticut in 1971.

[+] Currently Herbert Lehman College of the City University of New York.

Acknowledgment: The author wishes to thank Drs. Cooper and Liberman for making available the facilities of the Haskins Laboratories for the generation of the stimulus material. Mr. R. Hoffmann and Dr. P. Morse are due acknowledgment for their part in the execution of this study. Dr. B. Karmel graciously allowed the use of his laboratory for the study.

[1] VOT refers to the relative timing of the release of supraglottal closure and the onset of laryngeal pulsation or "voicing." Abramson and Lisker (1970) have argued that the acoustic features of explosion energy, amount of aspiration, and first-formant intensity may all be derived from the single articulatory variable of VOT. In sound spectrograms VOT is reflected by the onset of the first formant relative to the second and third formants and, for stop consonants with a delayed onset of the first formant, the presence of aspiration in the upper formants in the period preceding the onset of voicing.

stimuli were drawn from different phonetic categories could listeners discriminate accurately between physically different stimuli.

In contrast to the categorical perception of the stop consonants, nonspeech signals and steady-state vowels are perceived "continuously." Signals drawn from the same nonspeech or vowel category are discriminated equally well or poorly as signals drawn from different categories (Mattingly, Liberman, Syrdal, and Halwes, 1971).

The purpose of the present study was to determine whether components of the human cortical averaged auditory evoked response (AER) would reflect the categorical perception of different stop consonant signals or the equal physical differences between the different signals.

Very few studies have explored AERs to speech stimuli (Cohen, 1971; Greenberg and Graham, 1970; Wood, Goff, and Day, 1971). However, previous studies have indicated that when an auditory stimulus, a click or tone, is detected as different in a discrimination task, the amplitude of the N1-P2 component of the AER at the vertex is larger than it is in response to an undetected stimulus difference (Davis, 1964; Karlin, 1970; Sheatz and Chapman, 1969). If the vertex AER responds to the discrimination of speech stimuli in a manner similar to nonspeech, and if the detection of differences between speech stimuli is made task relevant, then the N1-P2 response to discriminably different stop consonant signals should be larger than the corresponding response to signals which are not discriminably different.

The use of the AER technique has another purpose which bears directly on the nature of categorical perception and its interpretation. It is possible that a listener may hear two physically distinct stimuli from within the same phonetic category as slightly different. However, because the listener knows that the two stimuli are both labeled the same in conventional speech and orthography, he may respond that the two stimuli are the same.

In the context of the present study, an estimate of the time necessary to code the acoustic signal into a categorized phonetic description can be made by assessing whether the N1-P2 component of the AER reflects continuous or categorical perception. If the N1-P2 component reflects a categorical response, i.e., a larger response to the stimuli from a different phonemic category than to the stimuli from the same phonemic category, then within 100-200 msec after stimulus onset the acoustic signal has been recoded into a phonetic representation. This would suggest that a categorized phonetic coding is an immediate and obligatory transformation of the acoustic signal.

METHOD

Subjects. Fifty undergraduate students at the University of Connecticut served as Ss. No S had previously participated in research involving synthetic speech or electroencephalographic (EEG) recording.

Apparatus. The Ss sat in a comfortable chair within a dimly lit, electrically shielded room and listened to tape-recorded stimuli presented via stereo headphones (Koss 600A). The sound level at the headphones was 65 db SPL.

Recording of the EEG was made from the scalp using a single silver-disk electrode located at the vertex (Jasper, 1958) which was referenced to the right earlobe. Resistance between electrodes was always less than 5K Ohms.

The EEG signals were transmitted by telemetry (Narco FM-1100-E3) to an AC preamplifier (W-P Instruments DAM 6) and oscilloscope amplifier (Tektronic RM 502A) which also served to monitor the EEG. The frequency response of the system after amplification was flat, between 2.0 and 30 Hz. The amplified EEG was stored for later analysis using a Vetter FM Adapter (FM-3) and a Sony 355 tape deck.

The extraction of the evoked response from the EEG was carried out both on- and off-line by a computer of average transients (Fabri-Tek 1072). The sweep duration was one second. The averaging cycle of the computer was triggered by a pulse from the second channel of the stimulus presentation tape. The onsets of the cuing pulses and the synthetic speech stimuli were simultaneous. The AER records were written out on an X-Y plotter (Hewlett-Packard 7035b).

Stimuli. The three synthetic, stop consonant-vowel syllables used in this study are shown in Figure 1. These stimuli were generated on the Haskins Laboratories computer-controlled parallel-resonance synthesizer (Cooper and Mattingly, 1969).

The three stimuli differed solely along the VOT continuum: 0 msec VOT (0 VOT); 20 msec VOT (20 VOT); and 40 msec VOT (40 VOT). Stimulus duration was 250 msec. For stimulus 0 VOT, the onsets of the first (F1), second (F2), and third (F3) formants were simultaneous; for stimulus 20 VOT, F1 began 20 msec after F2 and F3; for stimulus 40 VOT, F1 began 40 msec after F2 and F3. Aspiration was added to the upper formant frequencies during the period of F1 delay for stimuli 20 and 40 VOT. Thus, each adjacent pair of stimuli along the VOT continuum differed by exactly 20 msec VOT (i.e., 20-0 VOT and 20-40 VOT). Previous identification studies have indicated that stimuli with 0 and 20 VOT are identified as members of the phonetic category [ba] and that stimulus 40 VOT is identified as a member of the phonetic category [pa] (Lisker and Abramson, 1970).[2] Discrimination tests have indicated that the pair 20-40 VOT is discriminated essentially perfectly. The pair 20-0 VOT is discriminated just slightly better than chance (Abramson and Lisker, 1970). In the following account, stimulus 20 VOT will be termed the "standard" stimulus, stimulus 0 VOT the "within-category" shift stimulus, and stimulus 40 VOT the "across-category" shift stimulus.

Preparation of the stimulus tapes. With the aid of the computer-controlled synthesizer four stimulus sequences were recorded on audio tape. Two of the stimulus sequences were composed of varying length runs of standard stimuli (20 VOT), separated by pairs of either within- or across-category shift stimuli. There were a total of 154 standard stimuli and sixteen pairs of shift stimuli in each sequence. The pairs of shift stimuli occurred on the average once every ten successive standard stimuli (range 6-14). In one sequence the pairs of

[2] The three synthetic speech stimuli used in this study were slight modifications of the stimuli used by Lisker and Abramson (1970). Informal listening tests by the author and his colleagues indicated that the 20 VOT stimulus used in the present study was labeled more consistently as a [ba] than the 20 VOT stimulus used by Lisker and Abramson. These tests also indicated that the 20 VOT stimulus was discriminated less often from the 0 VOT stimulus than was the corresponding stimulus used by Lisker and Abramson.

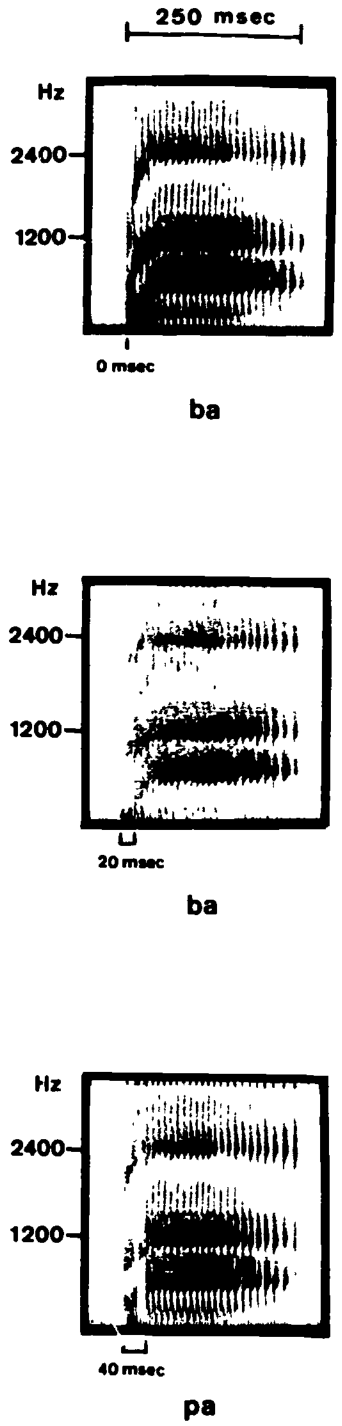


Fig. I

Figure 1: Sound spectrograms of the speech stimuli 0 VOT [ba], 20 VOT [ba], and 40 VOT [pa].

shift stimuli were within-category stimuli; in the other, across-category stimuli. A third stimulus tape consisted of a single sequence of 186 standard stimuli.

The fourth stimulus sequence contained an alternating sequence of blocks of ten within-category stimuli and ten across-category stimuli separated by 30-sec interblock intervals. There were three blocks of each shift category. The interstimulus interval (onset to onset) for all sequences was 2 sec.

Design. The Ss were assigned to five groups (ten Ss per group). The groups were run successively. The experimental task for the Ss in Groups 1, 2, 3, and 4 was to detect the occurrence of shift stimuli embedded in the sequence of standard stimuli.

The Ss in Group 1 listened first to the within-category shift sequence (20-0 VOT), then, on the following day, to the across-category shift sequence (20-40 VOT). The Ss in Group 2 also listened to both sequences on successive days, but in the reverse order.

Group 3 was given twenty practice trials with both the standard and within-category stimuli before listening to the within-category shift sequence. Pretraining consisted of twenty presentations of a group of four stimuli; two standard stimuli followed by two within-category stimuli. The interval between the groups was 5 sec. The Ss were told the order of the different stimuli and were instructed to try to detect any difference between the sounds. The within-category shift sequence was begun immediately after pretraining. These Ss were given pretraining to determine whether increased familiarity with the "unfamiliar" nonphonemic distinction would improve performance.

In a no-shift condition (Group 4) the Ss listened to the tape which contained all standard stimuli. The purpose of this control was to establish a baseline from which to assess the effects of the different shift conditions. In the other control condition (Group 5) the Ss listened to the randomized sequence of blocks of within- and across-category stimuli (the fourth stimulus sequence). The purpose of this control was to determine the amplitude of the AER to the across- and within-category stimuli in a setting unrelated to the discrimination tasks and thus to assess the "inherent" amplitude of the AERs to the 0 and 40 VOT stimuli.

Groups 3, 4, and 5 were tested in a single session. The session duration was approximately 7 minutes.

Analysis of the evoked potentials. The amplitude differences between the N1 and P2 responses was determined from the X-Y plots by measuring the difference in millimeters between the maximum wave of negativity between 75 and 125 msec after stimulus onset (N1) and the maximum wave of positivity between 175 and 225 msec (P2).

Each AER was the sum of sixteen individual responses. Responses to the standard and shift stimuli were averaged separately in all conditions. A separate AER was accumulated for each member of the shift pairs. The AER to the last standard stimulus before the shift pair was designated as the AER to the standard stimulus. In the no-shift condition (Group 4) evoked responses were accumulated for the stimuli which occurred in the same positions as the standard and shift stimuli in the shift conditions. For the stimulus control condition (Group 5) separate evoked responses were accumulated for the within- and across-category stimuli by summing over blocks of trials.

Procedure. All Ss were instructed to remain as motionless as possible during the experiment and to fixate on a point in front of them. The Ss in Groups 1, 2, 3, and 4 were instructed to "listen for" the occurrence of "any change" from the standard stimuli. The Ss were not told which pair of shift stimuli would occur in a given test sequence. The Ss in Group 3, after practice with the within-category and standard stimuli, were told to "listen for" the same changes in the test sequence as they had listened to in the practice sessions. The Ss in Group 5 were told that they would hear separate blocks of [pa] and [ba] and were instructed simply to listen to the stimuli.

RESULTS

Amplitude of N1-P2. For each S, the amplitude scores for both shift stimuli were expressed as the ratio of the shift stimulus applitude to the standard stimulus amplitude. A ratio score of 1.0 indicated that the amplitudes of the standard and shift stimuli were identical. A ratio score greater than 1.0 indicated a larger shift stimulus amplitude than standard stimulus amplitude. For the Ss in Group 1 (across shift then within shift) and Group 2 (within shift then across shift) separate ratio scores were computed for the within- and across-category shift conditions. The ratio scores for Groups 1-4 collapsed across Ss are shown in Table 1.

TABLE 1

Average Ratio of the Standard Stimulus N1-P2 Amplitude to the N1-P2 Amplitude of the Shift Stimuli

Shift Category	Position in Shift Pair	
	1st	2nd
Group 1		
Across	1.36	1.60
Within	0.92	0.83
Group 2		
Within	0.95	0.90
Across	1.35	1.51
Group 3		
Pretrained Within	0.92	0.90
Group 4		
No Shift	0.95	0.90

For Groups 1 and 2, the effects of presentation order (within shift then across shift vs. across shift then within shift), shift type (within vs. across), and location in the shift pair (first vs. second) were compared in an analysis of variance. A reliable main effect due to shift type was obtained

[$F_{1,18}$ = 25.00, $p < .01$; $\bar{X}$ (within shift) = .91, $\bar{X}$ (between shift) = 1.45]. No other main effects were significant. A shift type x location interaction was also obtained ($F_{1,18}$ = 4.66, $p < .05$).

The difference in N1-P2 amplitude to the within- and across-category shifts is illustrated for a representative S in Figure 2. In the across-category shift, the amplitude of both members of the shift pair (BS1 and BS2) exceeded that of the standard stimulus (S). For the within-category shift, neither member of the shift pair was larger than the standard stimulus.

Since the analysis of variance showed no significant effect due to presentation order, the data for the within- and across-category shifts were pooled over Groups 1 and 2. Two additional analyses of variance were then computed with the pooled data.

The first analysis compared the pooled across-category shift condition from Groups 1 and 2 with Group 3 (pretrained within-category shift) and Group 4 (no shift). In the groups x location analysis of variance only the groups effect was significant ($F_{2,37}$ = 13.16, $p < .01$). Post hoc comparisons according to Scheffe revealed that the pooled across-category shift condition (X = 1.46) differed from both the pretrained within-category condition (X = .91) and the no-shift condition (X = .92) at the .05 level. A second analysis of variance compared the pooled within-category shift condition from Groups 1 and 2 with Groups 3 and 4. The analysis of variance showed no reliable effects.

For Group 5, the absolute (N1-P2) amplitude difference of the AERs to the within-category stimulus (0 VOT) and to the across-category stimulus (40 VOT) were compared by a correlated t-test. The amplitudes of the two stimuli were not significantly different (T_9 = 1.01, n.s.).

DISCUSSION

The comparison of the within- and across-category shift conditions demonstrated that the across-category shift (20-40 VOT) elicited a larger N1-P2 response than the within-category shift (20-0 VOT). The difference in N1-P2 amplitude in the two shift conditions cannot be attributed to an "inherently" larger N1-P2 response to the across-category stimulus (40 VOT) than to the within-category stimulus (0 VOT), since in the stimulus control condition (Group 5) the amplitude of the N1-P2 response to 0 VOT [ba] and to 40 VOT [pa] did not differ. This outcome suggests that the difference in N1-P2 amplitude in the within- and across-category shift conditions was due to the difference in discriminability of the two types of shift.

The comparison involving the within-category shift group and the no-shift control (Group 4) revealed that the N1-P2 response in the two conditions did not differ. Furthermore, pretraining with the within-category and standard stimuli (Group 3) did not alter the amplitude of the N1-P2 response in the within-category shift situation.

Thus, the behavior of the N1-P2 component of the AER, under the conditions of the present study, mirrored the relative discriminability of the stop consonant pairs.

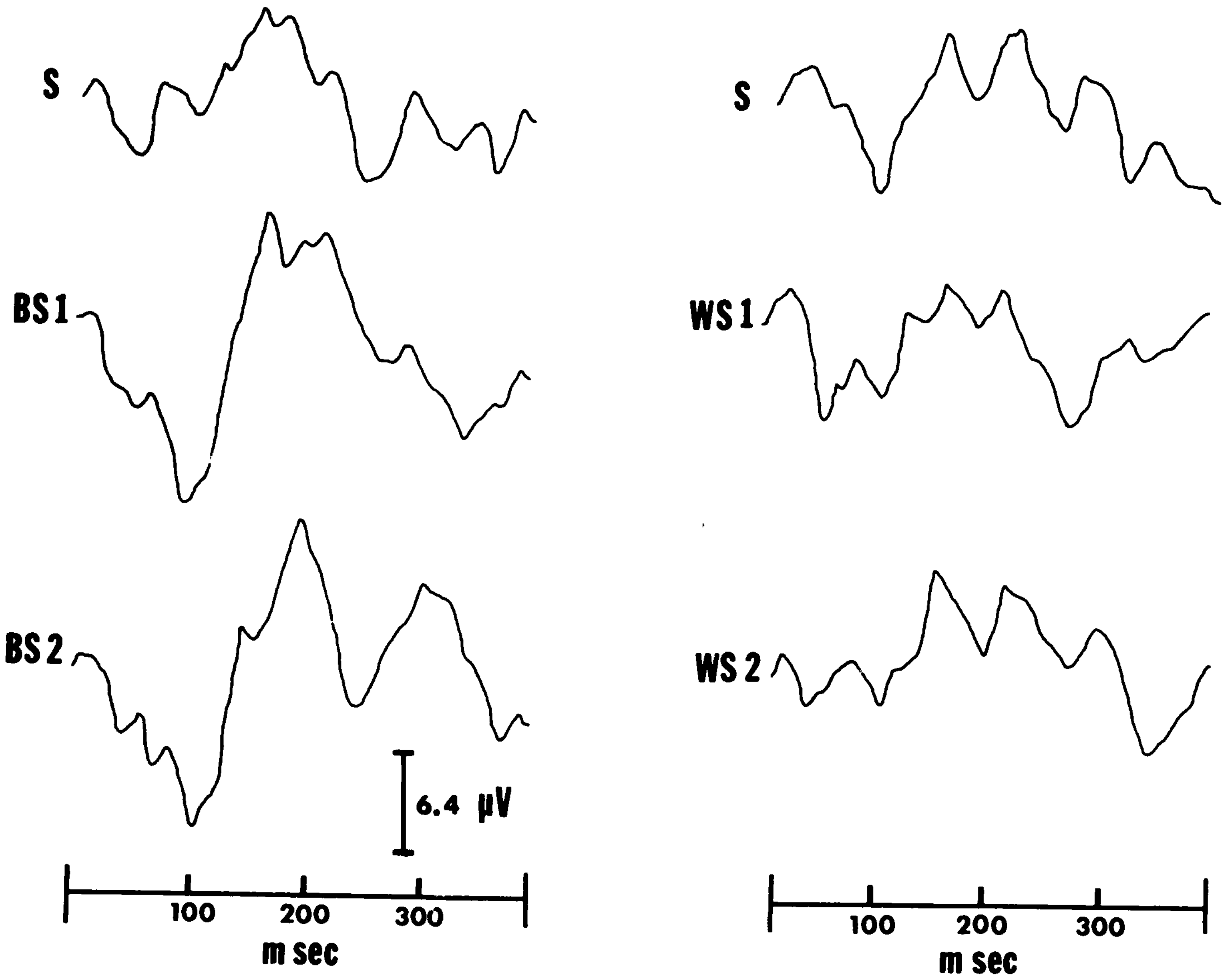


Fig.2

Figure 2: Auditory evoked responses to the within- and across-category shifts for a representative S. The standard stimuli are labeled S. The within- and across-category shift pairs are labeled WS and AS respectively.

Auditory to phonetic recoding. The "categorical" response of the N1-P2 component of the AER suggests that within 100-200 msec after the onset of a stop consonant, the finely detailed acoustic stimulus has been recoded into a categorized phonetic representation. The data from the present study do not support the suggestion that a categorical response is generated at a "long" interval after stimulus onset as a function of an arbitrary labeling of two discriminably different stimuli as belonging to the same phonetic category.

This interpretation of the data bears directly on the nature of the processing of the highly encoded stop consonants. After a stop consonant has been recoded into a categorized phonetic representation, a listener knows very little about the detailed acoustic structure of the auditory signal (e.g., VOT). The processing mechanism for the stop consonants appears to act like a "digitizing" device, accepting as input a highly variable and finely detailed auditory signal and then rapidly recoding it into a quantized phonetic representation (Mattingly et al., 1971). After recoding, the detailed auditory information does not seem to be stored in any accessible form.

This interpretation of the data is supported by two recent studies exploring differences in the processing of stop consonants and steady-state vowels. Crowder (1971) using a serial recall task found that if the vowel portions of CV syllables were varied in a serial list, then a large recency effect was obtained during recall. If, however, the consonant portions of the syllables were varied in the lists, then no recency effect was obtained. If the recency effect is contingent upon an "echoic" or "precategorical" acoustic memory store of 2-3 sec duration, as Crowder and Morton (1969) have suggested, then the representation of a stop consonant does not persist 2-3 sec in "precategorical" auditory memory.

The life span of auditory memory for stop consonants has also been studied using recognition memory tasks. In one of a series of studies, Pisoni (1971) varied the interval (0, .25, .50, 1.0, 2.0 sec) between vowel pairs and stop consonant-vowel pairs in an A-X discrimination paradigm. The discrimination of vowel stimuli was markedly affected by the A-X interval, with longer intervals producing poorer discrimination. Stop consonant discrimination, however, was relatively unaffected by A-X interval. Pisoni concluded that "information other than a binding phonetic categorization is unavailable for use in discrimination [of stop consonants]." The results of the present study are in complete agreement with those of Pisoni (1971) and Crowder (1971) and further reinforce the notion of a special mode of processing for the stop consonants characterized by the absence of a persistant noncategorical auditory image.

REFERENCES

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

Cohen, R. (1971) Differential cerebral processing of noise and speech stimuli. Science 172, 599-601.

Cooper, F. and I. Mattingly. (1969) Computer controlled PCM system for investigation of dichotic speech perception. J. acoust. Soc. Amer. 46, 115 (A).

Crowder, R. (1971) The sound of vowels and consonants in immediate memory. J. verb.Learning verb. Behav. 10, 587-596.

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

Davis, H. (1964) Enhancement of evoked cortical potentials in humans related to a task requiring a decision. Science 145, 182-183.

Greenberg, H. and J. Graham. (1970) EEG changes during learning of speech and nonspeech stimuli. J. verb. Learning verb. Behav. 9, 274-281.

Jasper, H. H. (1958) The ten-twenty electrode system of the International Federation. Electroenceph. clin. Neurophysiol. 10, 371-375.

Karlin, L. (1970) Cognition, preparation and sensory-evoked potentials. Psycho. Bull. 73, 122-136.

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

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

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

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

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

Pisoni, D. (1971) On the nature of categorical perception of speech sounds. Supplement to Haskins Laboratories Status Report on Speech Research.

Pollack, I. (1952) The information of elementary auditory displays. J. acoust. Soc. Amer. 24, 745-749.

Sheatz, G. C. and R. M. Chapman. (1969) Task relevance and auditory evoked responses. Electroenceph. clin. Neurophysiol. 26, 468-475.

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

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

Short-Term Habituation of the Infant Auditory Evoked Response[*]

Michael F. Dorman[+] and Robert Hoffmann[++]

In adults, the amplitude of the auditory evoked response (AER) at the vertex decreases as a negative exponential function of the number of stimulus presentations; it decreases faster, the faster the stimulus presentation rate, and recovers spontaneously when stimuli are withheld (Fruhstorfer, Soveri, and Jarvilehto, 1970). The decrease in AER amplitude reaches asymptote by the third to fifth presentation of a stimulus in a train (Ritter, Vaughn, and Costa, 1968; Fruhstorfer et al., 1970). Fruhstorfer (1971) has argued that the observed short-term reduction in AER amplitude over the first three to five presentations of a stimulus in a train is an instance of habituation (Thompson and Spencer, 1966).

In infants, habituation to stimuli in the auditory modality has been difficult to demonstrate (Jeffrey and Cohen, in press). The present study used a short-term habituation paradigm similar to that of Fruhstorfer et al. (1970) to investigate the effects of repeated stimulus presentation on the amplitude of the infant vertex AER. At the same time, the study served to establish an efficient methodology for collecting reliable AERs from awake infants.

METHOD

Subjects. A total of nine infants completed all of the conditions of the study. Artifact-free AERs were obtained from six (five male, one female) of the infants. All of these Ss were between 10 and 14 weeks old.

Apparatus. Recording of the electroencephalogram (EEG) was made from the scalp using a single silver-disk electrode located at the vertex (Jasper, 1958) which was referenced to the right earlobe. Electrodes were attached to the

[*] This paper reports a portion of the research carried out for a Ph.D. dissertation accepted by the University of Connecticut in 1971.

[+] Haskins Laboratories, New Haven. Currently Herbert Lehman College of the City University of New York.

[++] University of Connecticut, Storrs.

Acknowledgment: The authors wish to exprers their deepest appreciation to Dr. B. Z. Karmel for making available the funds and laboratory facilities necessary for the conduct of this study. The authors also wish to thank Drs. Cooper and Liberman of the Haskins Laboratories for making available the facilities necessary for the generation of the stimulus material.

scalp by styrofoam adhesive pads and an elastic headband. Electrode impedence was less than 6K Ohms.

The EEG signals were transmitted by telemetry (Narco FM-1100-E3) to an AC preamplifier (W-P Instruments DAM 6) and an oscilloscope amplifier (Tektronic RM 502A) which also served as a monitor. The frequency response curve after amplification was flat, between 2.0 Hz and 30 Hz. The amplified EEG was stored on tape for later analysis using a Vetter FM-3 Recording Adapter and Sony 355 tape deck.

The extraction of the evoked response from the EEG was carried out on- and off-line by a computer of average transients (Fabri-Tek 1072). The sweep duration was 1 sec. The averaging cycle of the computer was triggered by a pulse from the second channel of the stimulus presentation tape. The onsets of the cuing pulses and the synthetic speech sounds were simultaneous. The AER records were written out on an X-Y plotter (Hewlett-Packard 7035b).

Stimuli. The stimuli used in this study were trains of the stop consonant-vowel syllable [ba]. The duration of the syllable was 250 msec; the rise time, 25 msec; the intensity, 65 db SPL. The stimuli were generated on the Haskins Laboratories computer-controlled speech synthesizer (Mattingly, 1968).

Design and procedure. During a session, fifteen trains of four stimuli were presented at a rate of 1 train/30 sec from an AR 4-X loudspeaker placed two feet in front of the *S*.[1] The repetition rate of the stimuli was 1 stimulus/2 sec.

The *S*s were held in their mother's lap and were either bottle or breast fed during the test session. The mothers were instructed to hold the infants as quietly as possible and not to move the infant's bottle during presentation of the stimuli.

Analysis of the AERs. The amplitude of the N1-P2 response was determined from the X-Y plots by measuring the difference in millimeters between the maximum peak of negativity between 75 and 150 msec after stimulus onset (N1) and the maximum peak of positivity between 175 and 275 msec (P2). The responses to each member of the stimulus train were averaged separately. Ten good responses (i.e., those with no movement artifacts) were accumulated for each average.

RESULTS

The amplitude of the N1-P2 response as a function of the position of the stimulus in the train is shown in Figure 1. The amplitudes of the second, third, and fourth stimuli in the train are expressed as a percentage of the first stimulus amplitude. The mean amplitudes of the second, third, and fourth stimuli in the train were 36.0%, 41.0%, and 21.7% of the first stimulus amplitude. All amplitude reductions were significantly different from the first stimulus amplitude ($p < 0.01$) using a rank sum test.

[1] The stimuli were never presented when an infant was active or fussing. Thus, on a number of trials for all *S*s, an intertrain interval of greater than 30 sec was used.

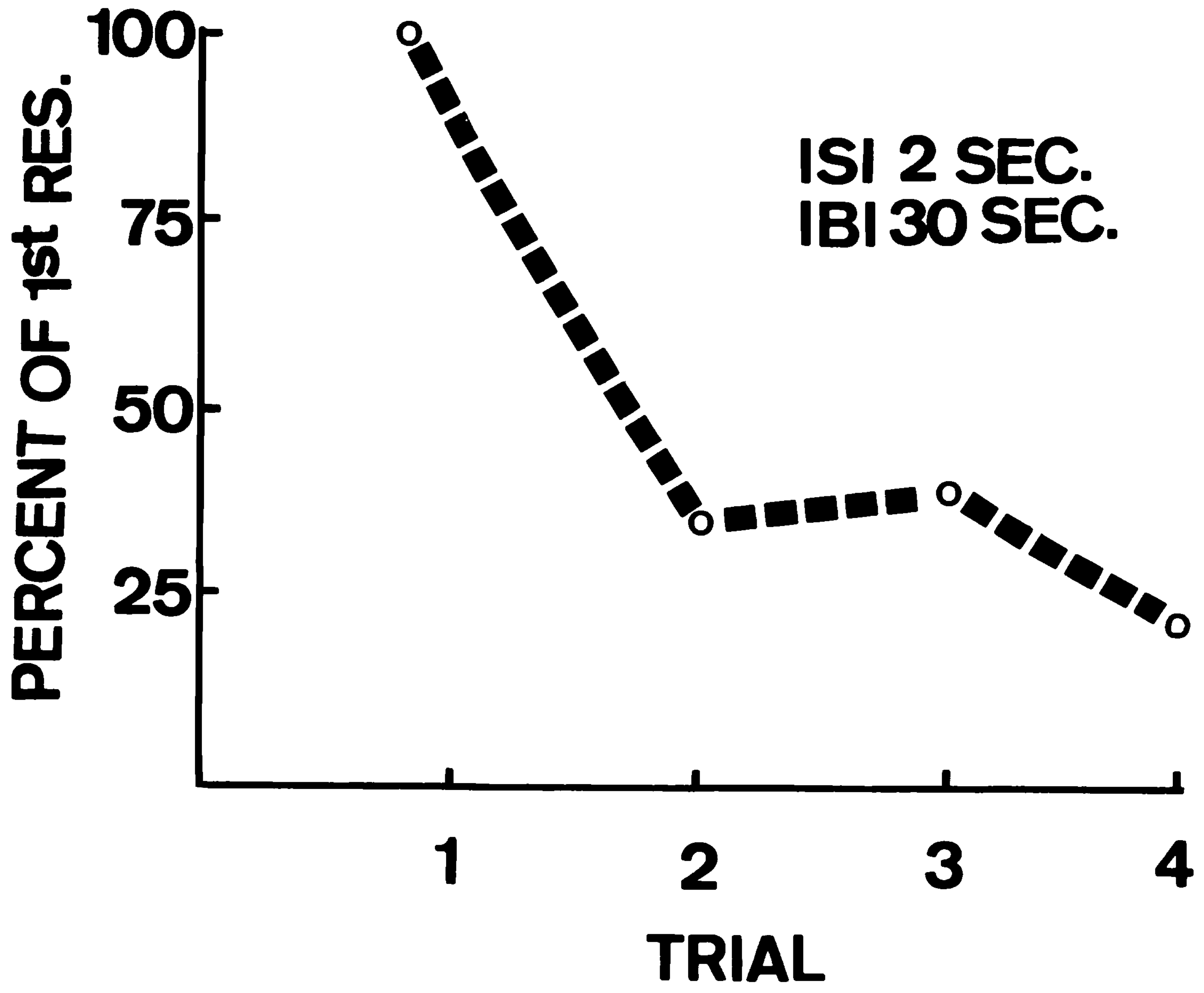


Fig. I

Figure 1: The amplitude of the N1-P2 response as a function of the position of the stimulus in the train, expressed as a percentage of the first response. ISI 2 sec refers to a 2-sec interstimulus interval; IBI 30 sec refers ' a 30-sec interblock interval.

DISCUSSION

Little difficulty was encountered in collecting artifact-free AERs from the awake infants. As long as the infants were brought into the laboratory hungry and were fed during the recording session, artifacts due to infant movement were nil. The use of FM telemetry rather than long cables to convey the EEG data to the recording apparatus also helped minimize movement artifacts.

The N1-P2 amplitude of the vertex AER in the awake infants decreased rapidly as a function of the repeated presentation of the syllable [ba] in a stimulus train. The magnitude and time course of the decrease in N1-P2 amplitude of the infant AER is consistent with the findings of both Ritter et al. (1968) and Fruhstorfer et al. (1970) on the short-term habituation of the adult AER. However, because of the differences in the interstimulus and intertrai: intervals between the present study and the previously cited studies with adults, the rates of habituation of the infant and adult AERs cannot be directly compared.

When the stimulus train was withheld during the 30-sec intertrain interval, the N1-P2 amplitude recovered spontaneously. This was evidenced in the absolute N1-P2 amplitude to the first and fourth members of the stimulus train. Thus, the decrease in the amplitude of the N1-P2 components of the infant AER in response to the repeated presentation of the syllable [ba] satisfies two of the characteristics of short-term AER habituation (Fruhstorfer, 1971).

In adults, a habituated AER to the syllable [ba] can be at least partially dishabituated by the presentation of a novel syllable [pa] (Dorman, in preparation). The results of the present study suggest that the AER could serve as a useful dependent variable in studying the perceptual abilities of awake infants.

REFERENCES

Fruhstorfer, H. (1971) Habituation and dishabituation of the human vertex response. Electroenceph. clin. Neurophysiol. 30, 306-312.

Fruhstorfer, H., P. Soveri, and T. Jarvilehto. (1970) Short-term habituation of the auditory evoked response in man. Electroenceph. clin. Neurophysiol. 28, 153-161.

Jasper, H. H. (1958) The ten-twenty electrode system of the International Federation. Electroenceph. clin. Neurophysiol. 10, 371-375.

Jeffrey, W. and L. Cohen. (in press) Habituation in the human infant. In H. Reese, ed., Advances in Child Development and Behavior (New York: Academic Press).

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

Ritter, W., H. Vaughn, and L. Costa. (1968) Orienting and habituation to auditory stimuli: A study of short term changes in average evoked responses. Electroenceph. clin. Neurophysiol. 25, 550-556.

Thompson, R. and W. Spencer. (1966) Habituation: A model phenomenon for the study of neuronal substrates of behavior. Psychol. Rev. 73, 16-43.

in the train, expressed as a percentage of the first response. ISI 2 sec refers to a 2-sec interstimulus interval; IBI 30 sec refers ' . a 30-sec interblock interval.

Early Apical Stop Production: A Voice Onset Time Analysis

Diane Kewley Port[+] and Malcolm S. Preston[++]

ABSTRACT

Voice onset time (VOT) has been shown to effectively differentiate the phonemic categories of stop consonants along the voicing dimension. This study applied the measurement of VOT to the production of apical stops produced by young children acquiring American English. Stops were measured from three children who were recorded regularly between 1 and 2 years of age and from additional children ranging in age from 6 months to 4-1/2 years. Distributions of the percentage of occurrence of apical stops along the VOT continuum are compared longitudinally across subjects as well as with distributions of adult productions of word-initial /d/ and /t/. Drawing on a physiological discussion of the control of timing between the stop release and the onset of vocal fold oscillation, the following pattern of apical stop development is proposed. The earliest instances of stop articulation, around 6 months of age, have uniform distributions along the VOT continuum. At a later stage the distribution of apical stops collapses into an interval corresponding to that of the adult production of /d/. With further development some apical stops are added in the range of adult /t/. The distributions of /d/ and /t/ words for children do not change from 2 to 4-1/2 years, but they do not yet correspond with those of adults.

[+] Haskins Laboratories, New Haven.

[++] The John F. Kennedy Institute, The Johns Hopkins University, Baltimore, Md.

Acknowledgment: This study was begun at the Neurocommunications Laboratory, The Johns Hopkins School of Medicine, where it was supported in part through the National Institute of Child Health and Human Development, Contract No. PH 43-65-637. Completion of the study was supported in part by the National Institute of Child Health and Human Development, Grant No. 5 P01 HD01994 and by the Public Health Service, Health Services and Mental Health Administration, Maternal and Child Health Service, Grant No. 917.

The authors would like to thank Arthur S. Abramson and Leigh Lisker for their advice during the research itself and comments on an earlier draft of this manuscript. We are most grateful to Hajime Hirose, Rachel E. Stark, Grace H. Yeni-Komshian, and Jacqueline Sachs for many fruitful discussions and ideas.

INTRODUCTION

This investigation applies acoustic measurement techniques to a developmental study of the production of stop consonants. The measure selected fcr this project is voice onset time (VOT), which is defined as the time interval between the release of stop occlusion and the onset of vocal fold oscillation. VOT can be easily measured from spectrograms of adult consonant-initial vocalizations. VOT measurements roughly comparable to those of adults can be made from spectrograms of the vocalizations of young infants if criteria for the selection of stop consonants are carefully applied.

Using VOT measurements, the present study investigates the development of stop consonants for three children from 1 to 2 years of age. This report is limited to apical stops because the children in our sample produced apicals almost exclusively during the time period studied. These longitudinal data are supplemented with other data which include a brief study of words for children from 2 to 4-1/2 years of age.

Linguists have claimed that voicing is a primary phonetic dimension for distinguishing among categories of stops produced at the same point of articulation. The voicing dimension for stops has been related .ɔ many different acoustic and articulatory phenomena. Lisker and Abramson (1971:770) have stated that voice onset time is "the single most effective measure" for sorting stops into different phonemic categories with respect to voicing, either productively or perceptually. Their own studies have repeatedly given support to this claim for production (Lisker and Abramson, 1964, 1967, 1970) and perception (Abramson and Lisker, 1965, 1970) across different languages. The measure of VOT is, however, the manifestation of a complex interaction between supralaryngeal and laryngeal musculature used to produce stops. This paper will consider in detail the physiology of the production of stop consonants and its relationship to VOT measurements. Evaluating the data in the context of these discussions, a sequential pattern of the development of apical stops with respect to VOT is proposed covering birth to 4-1/2 years of age.

PROCEDURE

The primary data of this study consisted of three sets of tape-recorded sessions, each set corresponding to one of three normally developing children (E3, E4, and E7) from American English-speaking environments. Tape recordings of E3 were analyzed at 45, 51, 60, 73, 81, 97, and 101 weeks of age. For E4 the ages of analysis were 50, 64, 82, 9b, 111, and 125 weeks. For E7 sessions were analyzed at 34, 40, 51, 64, 75, 83, and 96 weeks. These ages were chosen to correspond across subjects at roughly 12-week (3-month) intervals. Recordings having the greatest amount of vocalization were chosen when more than one recording was available for a time period.

E3 was a male, while E4 and E7 were females. E3 and E4 were the children of medical residents at the Johns Hopkins Hospital while E7 was the child of a senior undergraduate at the Johns Hopkins University. Thus, all three came from educated,middle-class families. Except for occasional colds, the three infants were in good health over the period during which the recordings weie made.

The tape-recording sessions were conducted in a sound-isolated booth (TAC model 1203) with the mother or father and occasionally an experimenter present.

The instructions to the parents were simply to encourage the child to vocalize as much as possible. Quiet toys and objects of interest were present during the recording sessions, which generally lasted about 30 minutes each. The children's vocalizations were recorded at 7-1/2 ips on an Ampex tape recorder (model AG350). A condensor microphone and cathcde follower (Bruel and Kjaer models 4131 and 4133) were connected by cable to the tape recorder outside the booth.

The procedure for analysis involved a transcription of the entire session[1] using a modified version of the Peterson-Shoup articulatory phonetic theory (Peterson and Shoup, 1966). Although phonetic transcriptions of infant vocalizations are obviously necessary to identify the stop consonants appropriate for measurement, the referent of any symbol in that transcription is unclear. Phonetic theories, such as that of Peterson and Shoup (1966), have been developed for the purpose of describing the phone types of the linguistic vocalizations of adults and are based on substantial knowledge of adult acoustics and articulation and of the correspondence between the two. However, far less is known about the articulatory or acoustic properties of the vocalizations of infants, nor is anything known about the reality of the articulatory mechanisms implied by adults ascribing phone types to the vocal sounds of such young children. Hence, at best our phonetic transcriptions must be considered to be a set of adult phone types which seemed most similar to the vocalizations produced by our infant subjects. It is our belief, however, that our phonetic transcriptions are adequate for the purpose of reliably identifying initial stop consonants produced by young children.

Another problem encountered was to select from the children's recordings a set of vocalizations which would be at least roughly comparable to words with initial stop consonants, as spoken by adults. In order to do this, rigorous selection criteria were developed based on the articulatory parameter values of the Peterson-Shoup theory. A vocalization was considered for analysis as long as its initial portion consisted of a stop consonant and a vowel. The transcriber then carefully judged each one as follows.

For the stop, the primary parameters required were plosive, alveolar (apical), and stop. The secondary parameters specified were: pulmonic air mechanism, egressive air direction, nonfrictional airflow, oral airpath (nonnasal), nonlateral lingual air path, open pharynx shape, natural tongue body shape (nonpalatalized and nonverlarized), and nonretroflexed tongue apex. Because infants exhibit a notable lack of control with reference to several secondary parameters, flexible criteria were used. Air pressure, whether lenis, normal, or fortis, was not judged except where it might have contributed to an excessively frictional airflow. The type of release, as relating to aspirated, unaspirated, or phonoaspirated stops, and lip shape were not judged. Laryngeal action was judged only for the vowel.

For the vowel, any horizontal and vertical place of articulation with pulmonic air mechanism, egressive air direction, and nonfrictional airflow was

[1] Two sessions, which had unusually large numbers of infant vocalizations, were only partially transcribed.

accepted. Air pressure, general air path, lingual air path, pharynx shape, tongue shape, apex shape, and lip shape were not judged. Vocal fold oscillation presented a special problem for infants. It would have been impossible to choose as normal any particular kind of oscillation, since vocal fry and falsetto voice were frequently produced by all three infants. Thus the laryngeal actions accepted included breathy, voiced, laryngealized, pulsated, and phonoconstricted; however, voiceless, whispered, constricted, and stopped laryngeal actions were not accepted. Further consideration was not given to portions of a vocalization following the stop and vowel.

Following transcription, wide-band spectrograms were made of the selected vocalizations on a sound spectrograph (Voiceprint model 4691A). To facilitate the measurement process, the vocalizations were always analyzed at half speed. Using the spectrograms, stops were categorized as initial in a small number of ambiguous cases by assuring that the stop was preceded by a pause of at least 50 msec. Stops were discarded where the onset of voicing or release was difficult to identify on the spectrogram.

Measurements of VOT to the nearest 10 msec were taken directly from the spectrograms. VOT is measured as the interval between the first vertical striation representing glottal pulsation and the onset of energy ("burst") representing the release of stop occlusion. When the glottal pulses precede the stop release (voicing lead), the VOT value is given a negative sign; when the stop release precedes the glottal pulses (voicing lag), the VOT value is positive.

A second experimenter checked the VOT measurements and further eliminated any items which in his opinion did not meet the above criteria. Thus, only sounds which were clearly identified as apical stops in initial position and which could be measured for VOT were included in the final analysis. The number of apical stops per session included in the final analysis varies from thirteen to ninety-eight. However, only three of the total twenty sessions had fewer than twenty tokens.

RESULTS

Figure 1 presents the combined data for each of the three subjects in the form of frequency distributions covering the entire period investigated. Comparison distributions for adults borrowed from the work of Lisker and Abramson (1967:13) are also presented in Figure 1. Their two distributions are derived from sentences, some of which contained words starting with the phonemes /d/ or /t/[2], spoken by ten American English speakers. The data for each child are

[2] The adult distribution presented in the figures of this paper combines the VOT values for both stressed and unstressed words and represents what we could consider to be the model of adult /d/ and /t/ distributions presented to the child in normal speech. On the other hand, distributions for words in only the stressed position should correspond more closely to the isolated stop-initial utterances collected from the children. The differences between the two types of distributions are small: in particular, for stressed words there is a better separation between the VOT values for /d/ and /t/, and the mode for /t/ is greater, +50 msec vs. +40 msec.

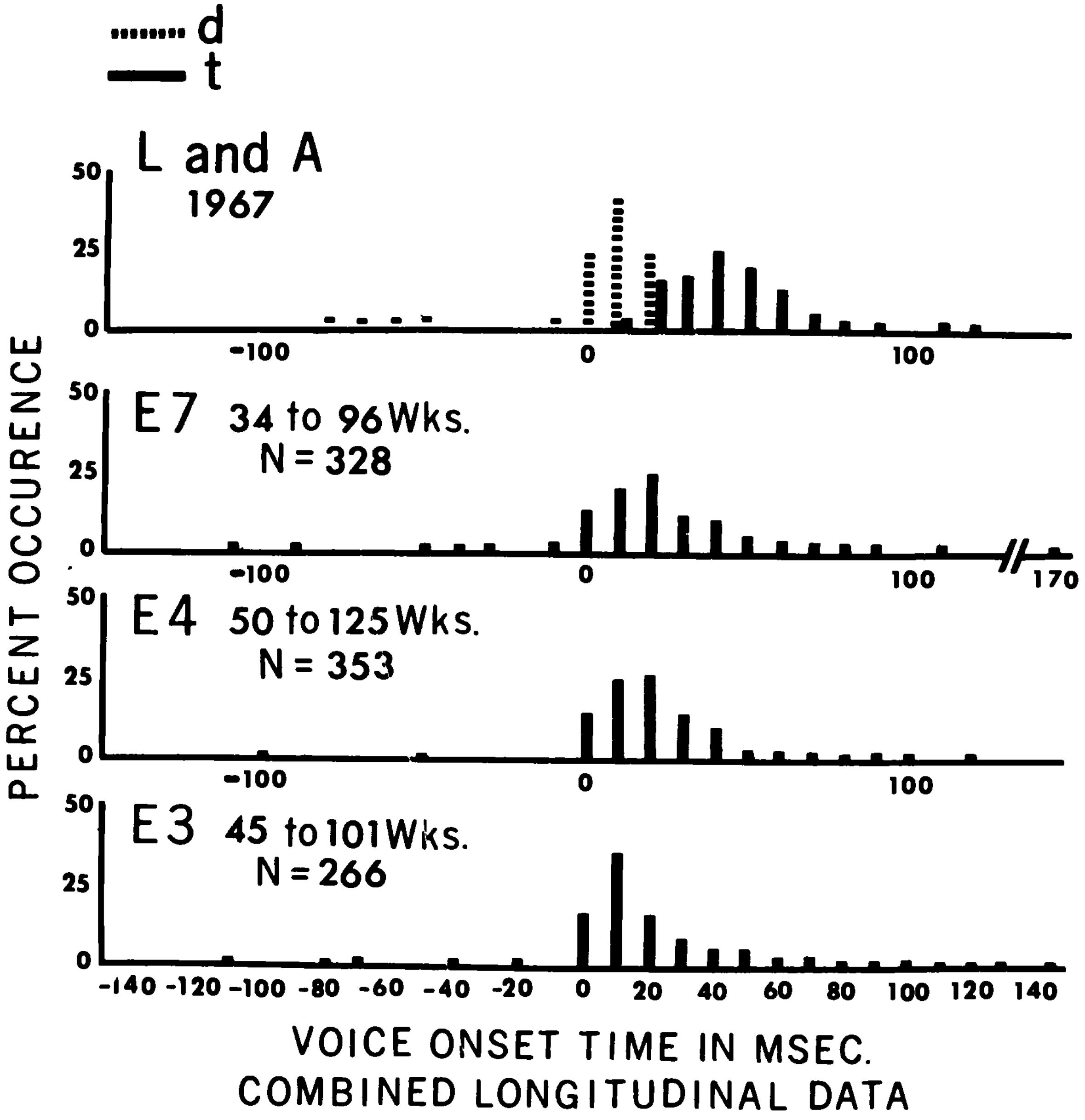

d
t
L and A
1967
E7 34 to 96Wks.
N = 328
E4 50 to 125Wks.
N = 353
E3 45 to 101Wks.
N = 266
PERCENT OCCURENCE
50
25
0
-100
0
100
170
-140 -120 -100 -80 -60 -40 -20 0 20 40 60 80 100 120 140
VOICE ONSET TIME IN MSEC.
COMBINED LONGITUDINAL DATA

Fig. I

presented as a single distribution since there was no way to assign phonemic units to their babbling. Each of the distributions produced by the children should be compared separately with the distributions of /d/ and /t/ for the adults as well as with each other. It is evident that the children's distributions are remarkably similar to one another. Each has a single mode, and the majority of the productions fall in the 0 to +20 msec voicing lag region.

To facilitate comparison of the children's data with those of adults, we introduce some terminology from the studies of Lisker and Abramson (1964, 1970, 1971). In their cross-language studies of initial stop consonants, three categories of stops having a rough correspondence across languages emerge along the voice onset time continuum. The categories are defined as follows: voicing lead, where stops have negative VOT values; short voicing lag, where stops have VOT values from 0 to +20 msec; long voicing lag, where stops have VOT values greater than +40 msec. As Figure 1 shows, measurements of American English apical stops produce two partially overlapping frequency distributions with a boundary between +20 and +30 msec. The majority of VOT values for /d/ lie in the short voicing lag category, although a small percentage occur in the voicing lead category. With respect to American English, it will sometimes be convenient to use the term "d-range" to refer to VOT measurements of +20 msec and less. Similarly, the term "t-range" will refer to VOT values of +30 msec and greater, noting that most values for /t/ lie in the long voicing lag category. The d-range and t-range, as defined, reflect a basic attribute of the voice onset time models which the child will eventually acquire for distinguishing words beginning with /d/ and /t/, namely that values along the voice onset time continuum are divided into two reasonably distinct classes.

A comparison of the children's data with the adult phonemic data suggests that the children reflect the English use of both /d/ and /t/. Only about 5% of the apical stops have voicing lead, whereas 64% are in the short voicing lag category and 31% in the long voicing lag category. Thus, during the period covered for each child, there are productions falling in both the d-range and t-range of VOT with a distinct preference for the d-range at approximately a two-to-one ratio. The children's distributions are unimodal in contrast to the adult data which, if combined into a single distribution, would show two modes, one for each category of apical stop.

Figures 2, 3, and 4 present the data arranged in longitudinal fashion for E3, E4, and E7, respectively. Each distribution in these three figures corresponds to a recording session at a single age going from youngest at the bottom to oldest near the top. The Lisker and Abramson data for adults are again reproduced at the top of each figure.

Inspection of the data for E3 at 45 and 51 weeks shows a concentration of apical stops in the short voicing lag category with only a few tokens in the long voicing lag category. By 101 weeks, E3 shows a considerable number of long voicing lag stops ranging from +30 to +160 msec, with no preference for any particular value. There are almost no stops in the voicing lead range at any age. The mode of all the distributions remains at +10 msec VOT with the exception of 97 weeks where it lies at +20 msec VOT.

For E4 the developmental pattern is much like that of E3. At 50 weeks of age, there is a concentration of short voicing lag stops; although stops do occur

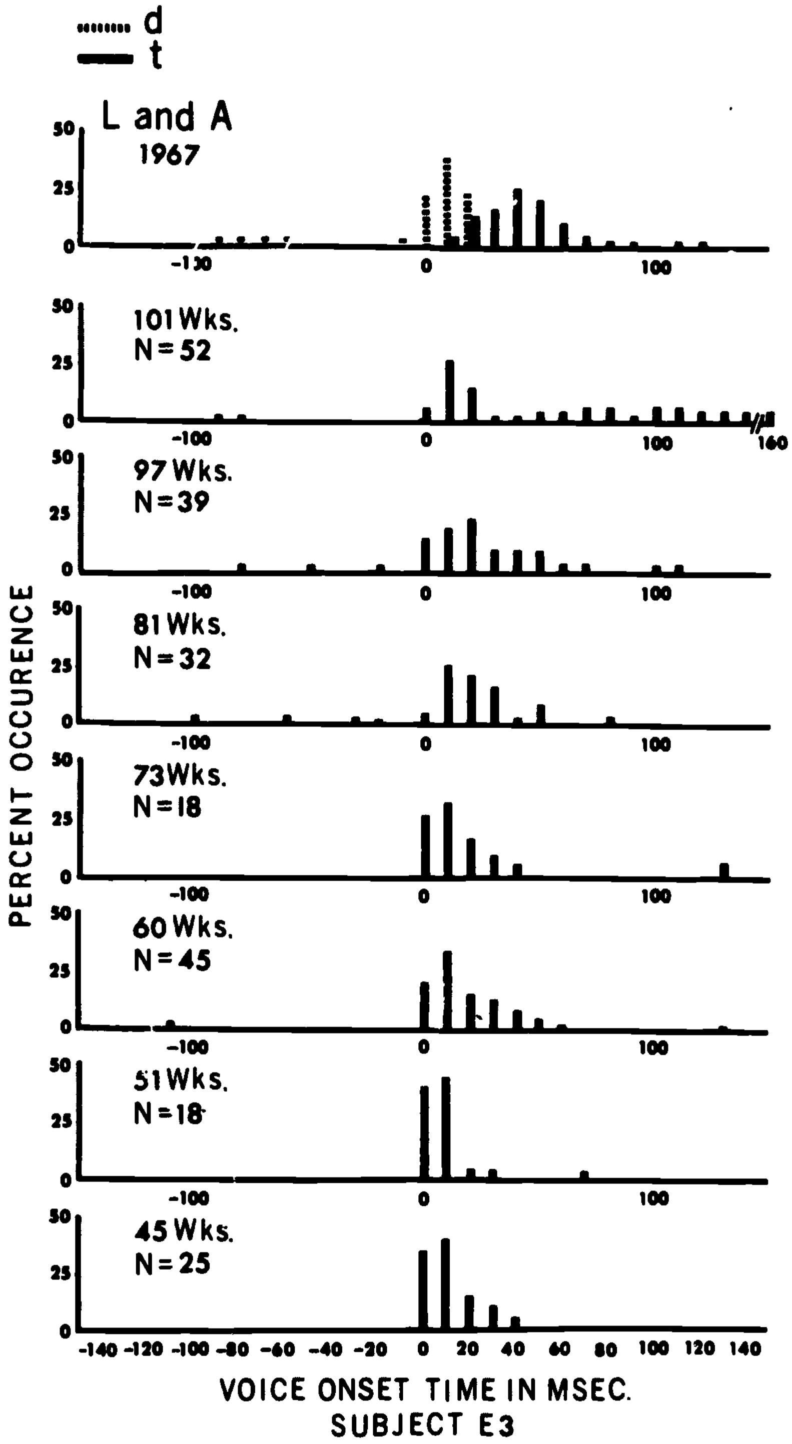
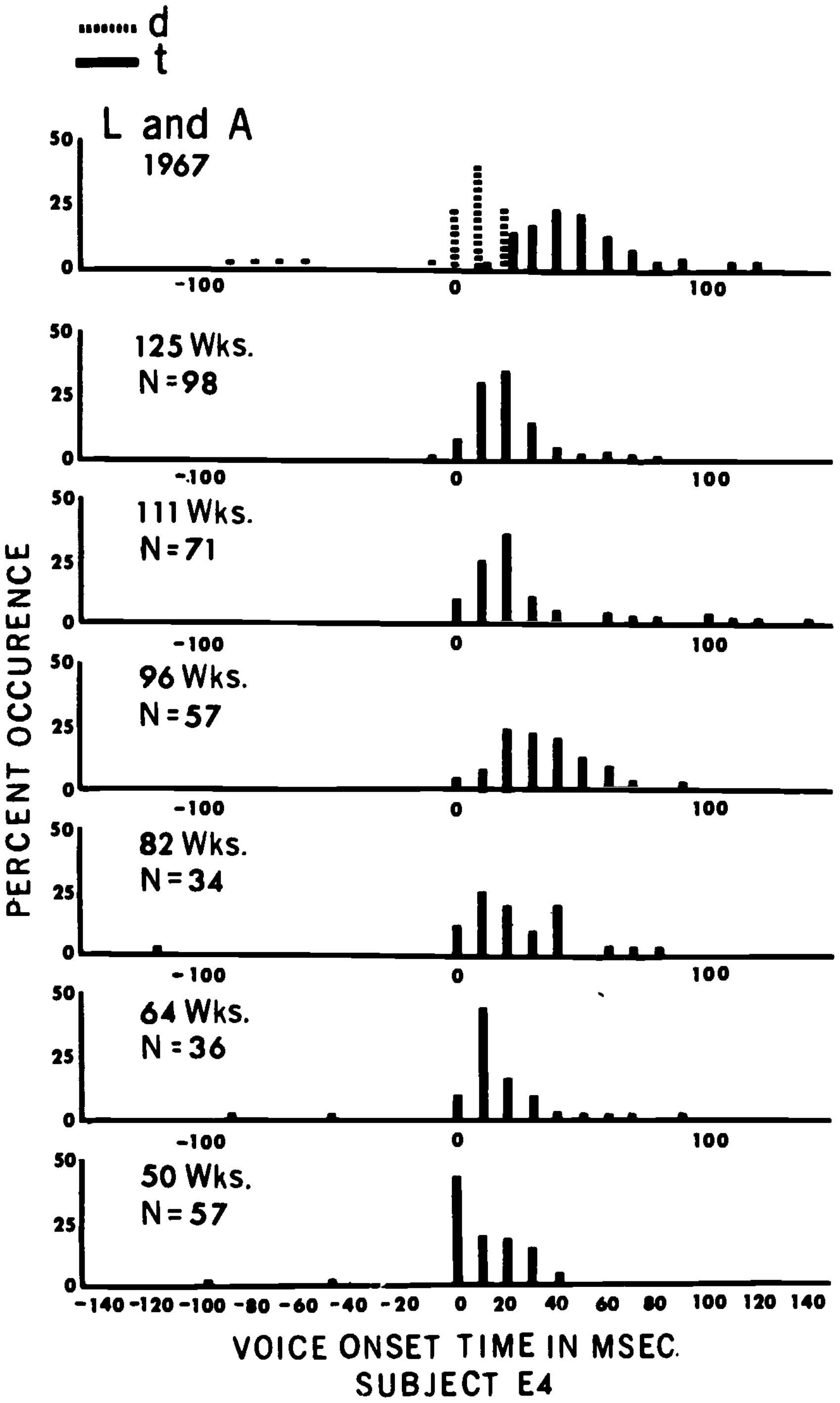
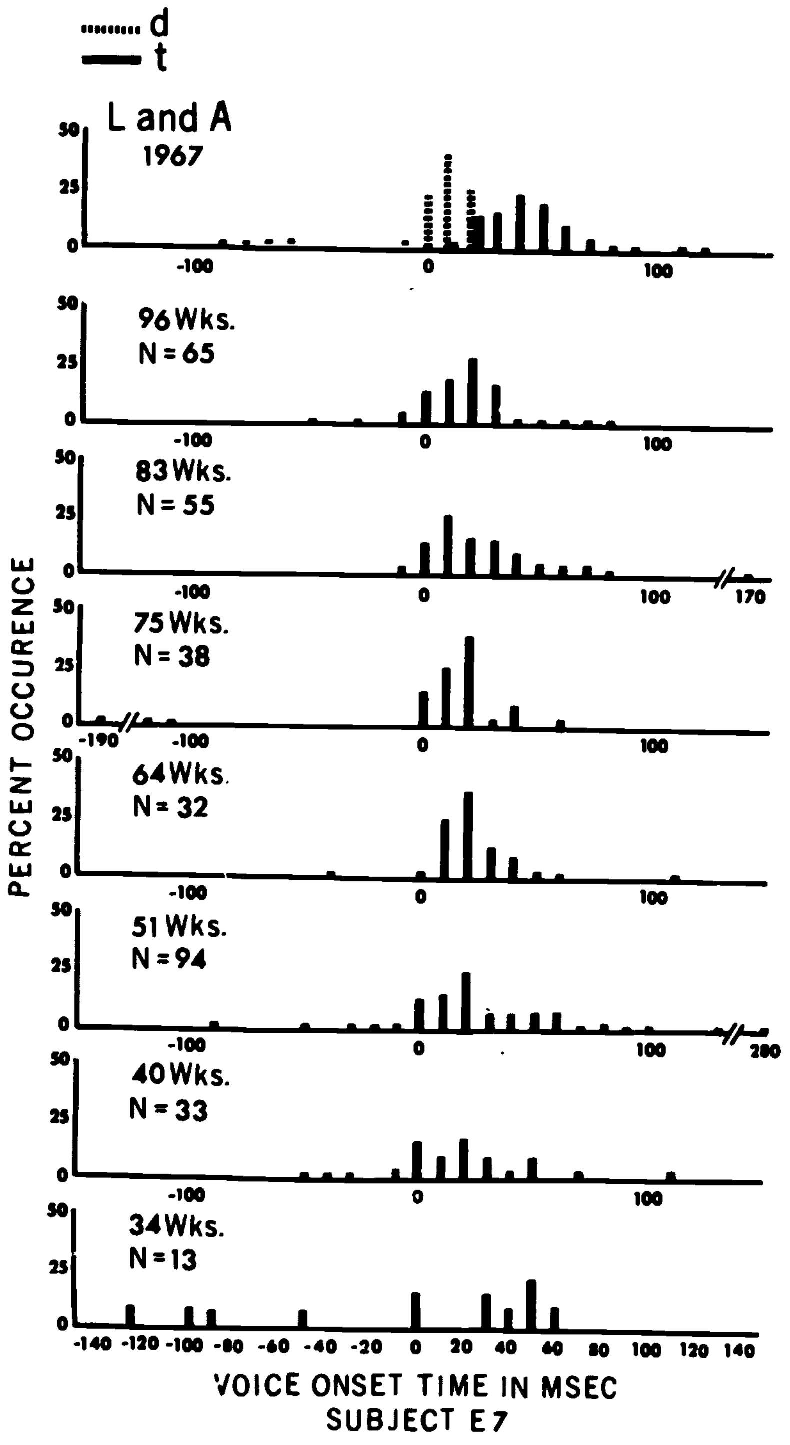

d
t
L and A
1967
101 Wks.
N=52
97 Wks.
N=39
81 Wks.
N=32
73Wks.
N=18
60 Wks.
N=45
51 Wks.
N=18
45 Wks.
N=25
50
25
0
-100
0
100
160
PERCENT OCCURENCE
-140 -120 -100 -80 -60 -40 -20 0 20 40 60 80 100 120 140
VOICE ONSET TIME IN MSEC.
SUBJECT E3

Fig. 2



d
t
L and A
1967
125 Wks.
N=98
111 Wks.
N=71
96 Wks.
N=57
82 Wks.
N=34
64 Wks.
N=36
50 Wks.
N=57
PERCENT OCCURENCE
50
25
0
-100
0
100
-140 -120 -100 -80 -60 -40 -20 0 20 40 60 80 100 120 140
VOICE ONSET TIME IN MSEC.
SUBJECT E4

Fig. 3



d
t
L and A
1967
96 Wks.
N = 65
83 Wks.
N = 55
75 Wks.
N = 38
64 Wks.
N = 32
51 Wks.
N = 94
40 Wks.
N = 33
34 Wks.
N = 13
PERCENT OCCURENCE
50
25
0
-190
-100
0
100
170
280
-140 -120 -100 -80 -60 -40 -20 0 20 40 60 80 100 120 140
VOICE ONSET TIME IN MSEC
SUBJECT E7

Fig. 4

in the other categories. By 96 weeks, E4 produces a considerable number of long voicing lag stops and continues to do so at 111 and 125 weeks. Again few stops with voicing lead occur. Distributions have a single mode that ranges from 0 to +20 msec.

The developmental sequence for E7 contrasts in certain respects with that of E3 and E4. First, E7 was recorded at a much earlier age than E3 or E4. The distributions for the earliest recording sessions (34 and 40 weeks), for which no comparable data exists for E3 or E4, have a wide range of VOT values with no apparent mode. At 51 weeks, unlike E3 and E4, there is still a wide range of VOT values, -90 msec lead to +280 msec lag, although there is now a mode at +20 msec lag. Thereafter up to 96 weeks of age, the mode remains at +10 or +20 msec voicing lag. A concentration of stops in the short voicing lag category does occur at 75 weeks, and then at 83 and 96 weeks long voicing lag stops again appear more frequently. E7 has more stops in the voicing lead range than E3 or E4 up to 51 weeks; after 64 weeks such stops also occur infrequently.

This data can be collapsed by categorizing stops into the d-range or t-range as previously defined. Thus a graphic representation of stops in the t-range as a percentage of the total number of stops at each age characterizes the developmental sequence in which stops representative of the adult models of /d/ and /t/ are observed.

Graphs of this type for E3, E4, and E7 are presented in Figures 5, 6, and 7. Subjects E3 and E4 have a similar developmental pattern from the early sessions (one year) to two years (102 weeks). At about one year, only 15% of stops produced are in the adult t-range. This percentage gradually increases until by two years the percentage is over 50%. The drop in the percentage by E4 for 111 and 125 weeks was the result of a distinct change in vocal behavior. Before two years of age, vocalizations during the half-hour recording sessions were partially babbling and partially recognizable speech with the attention of the child constantly changing. In later sessions, however, almost all vocalizations were recognizable speech and E4's attention was centered through almost the entire session on a single play activity which happened to involve "dishes." Thus the percentage of stops in the t-range for the older sessions is representative of data that is qualitatively different from that of the younger sessions.

Chronologically, the pattern of development for E7 is not similar to that of E3 and E4. The broad distribution of VOT values observed at 34 weeks is divided half into the t-range, half into the d-range. The percentage in the t-range then slowly falls to 12% at 75 weeks. The percentage increases in following sessions, but at almost two years is only 30% compared to over 50% for E3 and E4.

Although there are chronological differences between E7 and the other subjects, we may interpret the data for all three children from another point of view. In particular, by drawing on other developmental and physiological data, we will suggest that there is a single sequential pattern of development which describes the data for all three subjects. According to this interpretation, E7 lags in time behind E3 and E4. It was, in fact, the opinion of the experimenters that the overall language development of E7 lagged considerably behind that of E3 and E4. This includes further observations of E7 until she was 2-1/2 years old--a time period extending beyond that of data collection.

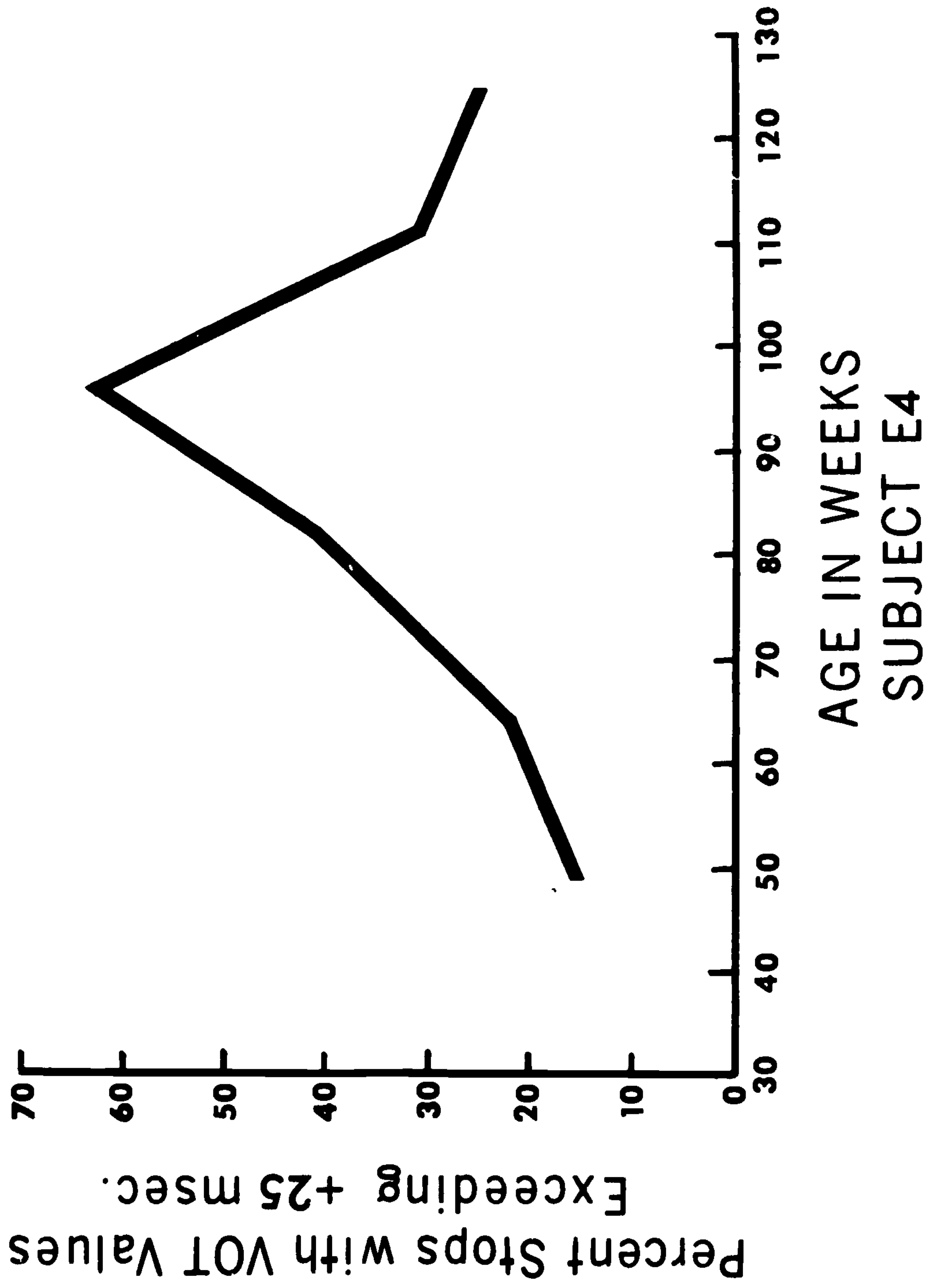

Percent Stops with VOT Values
Exceeding +25 msec.
0
10
20
30
40
50
60
70
30
40
50
60
70
80
90
100
110
120
130
AGE IN WEEKS
SUBJECT E4

Fig. 5

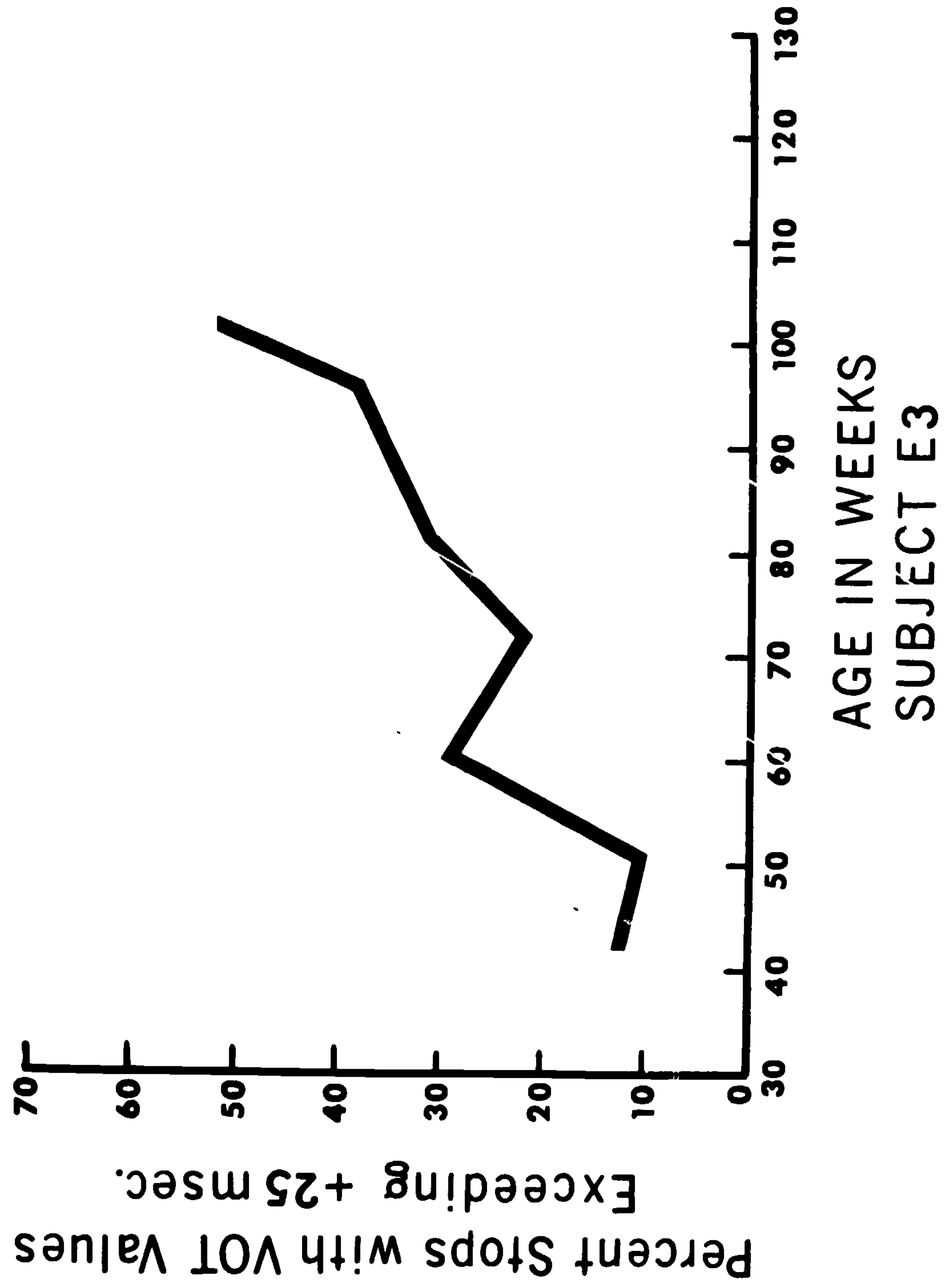

Percent Stops with VOT Values Exceeding +25 msec.
0
10
20
30
40
50
60
70
30
40
50
60
70
80
90
100
110
120
130
AGE IN WEEKS
SUBJECT E3

Fig. 6

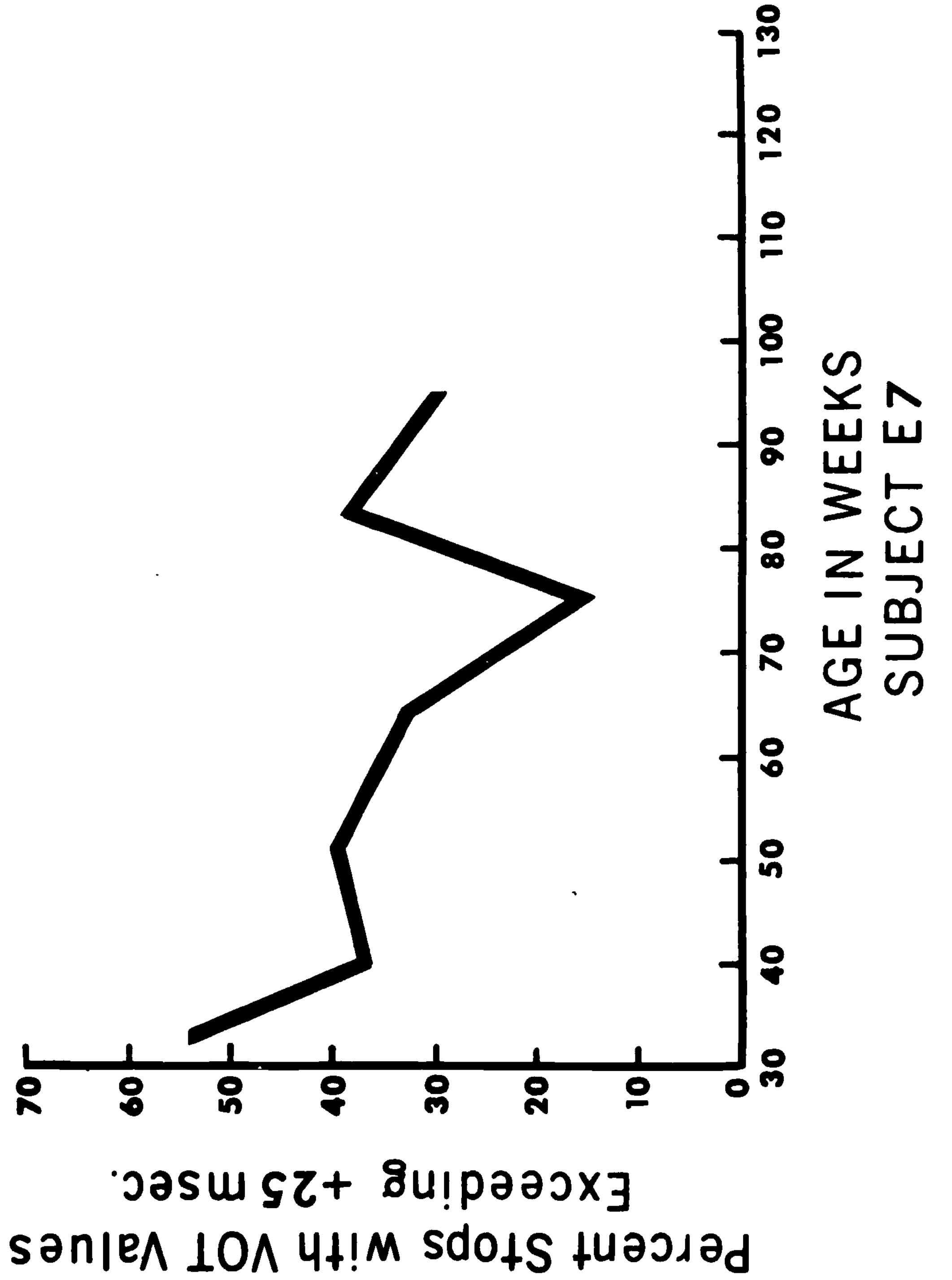

Percent Stops with VOT Values Exceeding +25 msec.
0
10
20
30
40
50
60
70
30
40
50
60
70
80
90
100
110
120
130
AGE IN WEEKS
SUBJECT E7

Fig. 7

DISCUSSION

In this section we will develop two hypotheses based on physiology which will be useful for interpreting the infant data. The hypotheses are that although it is inherently difficult for an infant to control the timing between stop release and the onset of vocal fold oscillation, an infant (learning American English) can produce a short voicing lag stop, like /d/, more easily than a long voicing lag stop, like /t/.

At least three separate articulatory gestures with separate innervations are needed to produce a stop consonant; these are the articulations to permit stop closure and release, to isolate the nasal cavities at the velum, and to initiate vocal fold oscillation. Other articulatory gestures in the vocal tract may also be used by adults to produce stops. However, from the point of view of an infant learning to produce stops, it would appear that control at the point of articulation, the velum, and the larynx must necessarily come first.

The present authors agree with the position of Lisker and Abramson (1964, 1967, 1971) and Rothenberg (1968) that the contrastive differences in the voicing dimension of stops are primarily the result of differences in the timing of glottal articulation relative to supraglottal articulation. We propose that distinct physiological mechanisms underlie the production of stops within each of the three voice onset time categories and, further, that stops in the short voicing lag category are easier to produce than stops in the other two categories.

First we will examine the hypothesis that the infant needs to learn only one type of apical articulatory gesture for the production of apical stops regardless of the VOT category. Studies of adults do not reveal any essential differences in effecting articulatory closure for stops differing with respect to voicing. In a palatography study by Fujii (1970) of the dynamic placement of the tongue against the palate, Japanese /d/ and /t/ were considered to belong to a single articulatory class (compared to other consonants), although there were small, consistent differences between them. In other studies of labial stops, Harris et al. (1965) and Fromkin (1966) investigated electromyographic (EMG) signals from the primary muscle of articulation for labials, the obicularis oris, and found only insignificant differences in peak EMG strength for English /p/ and /b/. Lubker and Parris (1970:632), using simultaneous measurements of EMG and force of labial contact, found the labial gestures for /p/ and /b/ "essentially monotypic." Measurements of closure duration for American English /p/ and /b/ by Lubker and Parris (1970), and Dutch /p/ and /b/ by Slis (1970), found durations to be the same in initial position, varying from 100 to 150 msec depending on context. Although these data concerning articulatory closure is very incomplete, we feel justified in assuming that an infant could learn essentially one type of apical articulation and be able to produce stops in all VOT categories.

The nasal cavities must be isolated from the rest of the vocal tract in order to create the intraoral pressure needed to produce a stop. Muscles attached to the velum and pharyngeal muscles act to close the velum against the pharyngeal wall. Many recent investigations have shown some differential activity in the velopharynx for stops belonging to different VOT categories (Berti and Hirose, 1972; Lubker et al., 1970). The relevance of these studies will be discussed presently.

For stop consonants in initial position, the glottal articulation that must be effected is the adduction of the vocal folds from an open (rest) position to a closed, oscillatory position. Voice onset time measurements reflect the time at which the adduction of the vocal folds is achieved relative to the stop release. For apical stops in the short voicing lag category, VOT measurements range from 0 to +20 msec. Direct observation of the larynx by fiberoptic techniques (Lisker et al., 1970) confirm that the vocal folds have fully adducted and are oscillating at or very near the time of stop release. Thus, articulatory gestures required to produce short voicing lag stops are velopharyngeal closure followed by the complete adduction of the vocal folds at the time of release of the supraglottal articulators, such that vocal fold oscillation begins within 20 msec of release.

In order to initiate vocal fold oscillation, another factor must be considered. Oscillation of adducted vocal folds is the result of airflow through the glottis which in turn occurs when there is a sustained pressure drop across the glottis. When the vocal tract is unobstructed and the vocal folds are adducted, a wide range of transglottal pressure differentials and tensions in the vocal folds will result in some sort of vocal fold oscillation. However, when the vocal tract is obstructed, as during stop closure, and the vocal folds adducted, Rothenberg (1968:91) has argued that oscillation will not occur or be maintained unless special articulatory mechanisms are employed to sustain a transglottal pressure drop. These mechanisms may include active or passive enlargement of the supraglottal cavity, some nasal airflow and heightened subglottal pressure. Thus, if the vocal folds are adducted at any time during apical closure and additional muscle gestures are not made, vocal fold oscillation will not begin until after the stop closure is released.

That is to say, for an infant to successfully produce short lag apical stops in initial position, he may fully close the glottis any time during apical closure providing that the velopharyngeal closure merely isolates the nasal cavities. However, to produce voicing lead stops, the infant must complete glottal closure considerably before oral release and then initiate and sustain vocal fold oscillation by the addition of other articulatory mechanisms (suggested above). These might include velopharyngeal adjustments other than simple velopharyngeal closure.

Stops with long voicing lag are produced with the glottis open at the time of release according to fiberoptic investigations (Lisker et al., 1970). For American English /t/, the onset of vocal oscillations in the Lisker and Abramson data in Figure 1 has a mean of +45 msec VOT. Lisker et al. (1970) show that the vocal folds become fully adducted a short period of time (approximately 30 msec) after oscillation has begun. Kim (1970:111) and other researchers indicate that it takes about 100 to 120 msec to fully adduct the vocal folds from their initial open position. Considering these data, an infant will successfully produce a long voicing lag stop if he leaves the glottis open throughout apical closure and then initiates vocal fold adduction approximately at stop release, having maintained velopharyngeal closure throughout. We note that the gesture for velopharyngeal closure could be approximately the same for the infant to produce short and long voicing lag stops, but it is likely to be different and more complex for the voicing lead stops.

The range along the voice onset time continuum which the different VOT categories cover is also of interest. For the long voicing lag stops, English

/t/ can serve as a representative example; Figure 1 shows 90% of the /t/ stops falling within a 50-msec VOT interval. This indicates that the adult articulation of /t/ involves the very careful control of timing between the supraglottal and glottal articulators, which, we repeat, are separately innervated.

This precise timing constraint is not necessary for the short lag stops. Although the VOT range is ca. 20 msec, adduction of the vocal folds can be achieved any time during apical closure (which duration is about 100 msec) and oscillation will still begin only upon apical release. For those languages containing voicing lead stops in the Lisker and Abramson (1964) study, the range of VOT values was approximately 90 msec. Thus, timing between glottal and supraglottal articulators seems more carefully controlled tor long voicing lag stops than for voicing lead or short voicing lag stops.

On this basis, we are suggesting that short voicing lag stops are easier for the infant to produce than are the other two types. Further support for this argument comes from some recent data collected by Hirose (1971 and personal communication). He studied the EMG signals of the intrinsic muscles of the larynx during stop production. Although he studied primarily intervocalic stops, he did have two contrastive utterances with initial stops, /pʌpə/ and /bʌpə/. Hirose (1971) states that the primary adductors of the vocal folds are the interarytenoid muscles. For one English-speaking subject, the EMG signals, averaged over a set of repetitions of /pʌpə/ and /bʌpə/, show a clear difference in interarytenoid activity for the initial stops. Measuring the time from the beginning of the rise of EMG activity to onset of voicing, he finds that /b/ takes 300 msec while /p/ takes 180 msec. Also, the EMG peak height reached for /b/ was 60 microvolts, but for /p/ it was 90 microvolts. That is, glottal adduction for /b/ takes place more slowly and less forcibly than for /p/. Although too complex to detail here, Hirose's results for other subjects and other adductor muscles support the above statement. Under different conditions of data collection, Hiroto et al. (1967:871) came to a similar conclusion: the time interval from the beginning of change of EMG activity to onset of voicing for the interarytenoids was greater for a set of Japanese short voicing lag consonants (including /b/ and /d/) than for a corresponding set of long voicing lag consonants (including /p/ and /t/).

In other EMG studies, the posterior cricoarytenoid (PCA) muscle, the abductor of the vocal folds, has been shown by Hirose (1971) to act contrastively for /p/ and /b/ in medial position. In initial position in English (Hirose, personal communication), there is also differential activity for /pʌpə/ vs. /bʌpə/. For /b/, there is a suppression of PCA activity during, and at least 250 msec before, labial closure. On the other hand, the PCA is moderately active just prior to the labial closure for /p/. Thus, there is additional muscle activity keeping the glottis open for /p/ which is not present for /b/.

Summing up, short voicing lag stops appear to be easier for the infant to produce than the other two types. Voicing lead stops require muscle gestures in addition to those needed for short voicing lag stops. The long voicing lag stop requires more carefully controlled timing between stop and laryngeal closures and, furthermore, requires a higher level of effort to effect the vocal fold adduction.

It is clear from the foregoing discussion that stops are produced by a complex set of muscle gestures. The gestures of stop release, velopharyngeal closure, and the onset of vocal fold oscillation must be concluded in certain specified orders within a very short period of time if a stop consonant is to be successfully produced. Research on nonhuman primates further suggests that this is not a simple task. Lieberman et al. (1970) state that stops have not been observed spontaneously in the vocalizations of primates, although primates, especially Rhesus monkeys and chimpanzees, are thought to have the gross physiological capability to produce stops. Furthermore, it has proven extremely difficult to train chimpanzees to produce any stops.[3] Research with young deaf children has shown that even when some stops are produced by these children, they are not successful in learning to make stops in more than one voice onset time category (Stark, 1971), typically the short voicing lag category. Thus, we may expect infants to exhibit difficulty in learning to control the timing of the articulatory gestures of stop consonants.

Tracing the development of stops from birth, there is evidence that neonatal humans do not prod·ce stops in their vocalizations (Lieberman et al., 1968). In our own observatiun, it is not until about 6 months of age that infants will produce enough stops to yield even a small number from a half-hour recording. This is about the age of E7's ,oungest data. E7 was recorded weekly from 29 weeks of age, but not until 34 weeks were more than five apical stops recorded in one session. E7's distribution at 34 weeks may be characterized as spanning a wide range of VOT, -120 to +60 msec, and as being approximately uniform over the interval.

Data was not collected for E3 or E4 at these early ages, but another subject was recorded weekly from the even younger age of 26 to 31 weeks of age. This subject produced almost exclusively velar stops, and not very many in most sessions. At 30 weeks, however, twenty-three velar stops were analyzed and gave a distribution which was clearly uniform from -160 msec voicing lead to +160 msec voicing lag.

If we allow a tentative generalization on the basis of these two subjects, the results suggest that the earliest attempts to produce stops give voice onset time measurements randomly distributed over a wide range along the VOT continuum. We note that although the distributions were similar, for one subject the stops were predominantly apical, while for the other subject they were velar. Such a distribution shows no specific patterning after the adult models of /d/ and /t/--or /g/ and /k/ in the other subject--which raises several questions. First, is there reason to expect distributions of stops from their earliest occurrences in infant vocalizations to reflect the adult model? An affirmative answer might be inferred from data which indicate that infants may attend to the adult model of stops prior to producing any of their own. Some recent studies have shown that within the first few months of life infants discriminate sounds which are identified by English-speaking adults as different stop consonants. Specifically, Eimas et al. (1971) showed that infants of 4 weeks and 4 months are able to sort synthetic stop consonants into categories similar to those of the adult phonemic model of voicing. Morse (1971) further

[3] See Emily Hahn's (1971) review of several attempts to teach chimpanzees to talk.

showed that infants of about 7 weeks of age could distinguish between synthetically produced /ba/ and /ga/.

On the other hand, some research by Stark (1968, 1971) on hearing-impaired children suggests that they will not spontaneously produce stop consonants when auditory experience is limited. Stark followed six hearing-impaired children longitudinally from about 2 years of age. These children had severe, and in two cases even profound, hearing losses. Some of the severely deaf children produced labial stops prior to acquiring hearing aids, while none of the profoundly deaf subjects did. After acquiring hearing aids, all six children produced labial stops. Although some severely deaf children also learned to produce apical and velar stops after they were fitted with hearing aids, neither of the profoundly deaf children accomplished this. Furthermore, when the hearing aid of one profoundly deaf child was broken for eight weeks, the child ceased producing any stops.

Thus, there are reasons to expect aspects of the adult models of /d/ and /t/ to be reflected in early apical stop distributions, since relevant perceptual information about the voicing distinction in stops seems already to be available to the infant and because without this (and other linguistic) information, no stops would be produced at all. This leaves us with the question of why no evidence of the adult model is present in the early distributions.

Preceding discussion rules out to a great extent the possibility that stops are produced spontaneously and, therefore, might just as well be randomly distributed with respect to VOT measurements. On the other hand, the inherent difficulty of timing in stop articulations might be great enough that an infant's first attempts to produce stops bear little resemblance to the adult models. This was one of the hypotheses presented in the physiological discussion. Thus, we reason that an infant's earliest attempts to produce stops are uniformly distributed along the VOT continuum because he is unable to control the timing between the muscle gestures at the point of articulation and the larynx.

From the broad distribution of very early stop productions, we propose that a later stage in stop development is a concentration of apical stops in the short voicing lag category of VOT. Two of the subjects, E3 and E4, already had this type of distribution when we first made recordings of them at about 50 weeks of age. E7 did not produce a comparable distribution until 75 weeks, nearly 6 months later. It can be seen from either Figure 4 or Figure 7 that E7 makes steady progress from the broad distribution to a concentration of stops in the short voicing lag category. We checked further on the reliability of the data at 75 weeks using a separate recording of E7 at 74 weeks. Of the forty-six apical stops analyzed, 98% were in the short voicing lag category.

The proposal that a concentration of stops in the short voicing lag category is a stage of apical stop development becomes more reasonable in light of the physiological discussion. The infant initially showed no control over the coordinated timing of gestures needed to differentiate among VOT categories. At some point, however, we assume that the infant does try to produce stops which will match those of the VOT adult model. To succeed he must achieve control over different articulations for apical stops in the d-range than for apical stops in the t-range. It appears that the infant does not acquire control over these articulations simultaneously, but rather acquires the easier one first, i.e., the short voicing lag stop.

Following this stage, stops were added in the t-range, but without a mode characteristic of adult /t/. By this time small numbers of recognizable English words were present in the recordings. Therefore, to carry the investigation further, a small study[4] concerning only words was carried out.

E4 was the child recorded the longest so her data were examined for words beginning with /d/ or /t/. Words were accepted if two adults could easily identify them as English words beginning with "acceptable" infant productions of /d/ or /t/, i.e., not obviously in error with respect to voicing. The entire utterance in which the /d/ or /t/ word occurred was written down in its approximate English equivalent (for example, "Doggie see fish") for presentation to other subjects.

The earliest age for which we could identify a few /d/ and /t/ words was 96 weeks. We also selected words for 111 weeks (2-1/4 years) and 125 weeks (2-1/2 years). For comparison, all of E4's utterances were repeated by four children and three male adults. The children's ages were 3-1/2 years (one child) and 4-1/2 years (three children). For the children, an experimenter or the mother read from the respective lists of E4's utterances which the child was asked to repeat correctly. The adults simply read the utterance lists. Spectrograms were made of all utterances, and the VOT distributions were made separately for the /d/ and /t/ words for each individual subject. This procedure was used in hope that effects of context would be controlled across subjects. Following analysis, the three adult distributions were essentially identical, so only one is reported on here.

The data for all subjects are presented in Figures 8, 9, and 10 corresponding to the utterances for E4 at 96, 111, and 125 weeks of age, respectively.

Consider the distributions for /d/ words. Since there were only two /d/ words at 96 weeks, results are mainly based on distributions at 111 and 125 weeks. The /d/ distributions for all five children for both 111 and 125 weeks are the same; thus, remarks on the /d/ data refer in common to Figures 8, 9, and 10. All children's distributions bear basic similarities to those of the adult, with small differences. Adult JD has a VOT range for /d/ of +10 to +40 msec, with only 10% of the /d/'s falling in the t-range. E4 has a VOT range of 0 to +40 msec, with 25% of the /d/'s falling in the t-range. The four older children have a range for /d/ of 0 to +100 msec, again with 25% of the /d/'s in the t-range. (No /d/'s with voicing lead occurred.) We thus conclude that distributions for /d/ are the same from the earliest word productions to at least 4-1/2 years of age. The /d/ distributions are quite similar to that of the adult model, but children show considerably more error in producing /d/ words with VOT values in the t-range.

For the /t/ words, there is a difference in the distribution for E4 at 96 weeks compared to the two older distributions. E4's /t/ words at 96 weeks have a mean of +40 msec and a range of +20 to +60 msec. These values are significantly smaller than those of adult JD's /t/ distribution, which has a mean of

[4] This study was reported in part at the 79th meeting of the Acoustical Society of America and in Preston and Port (1969).

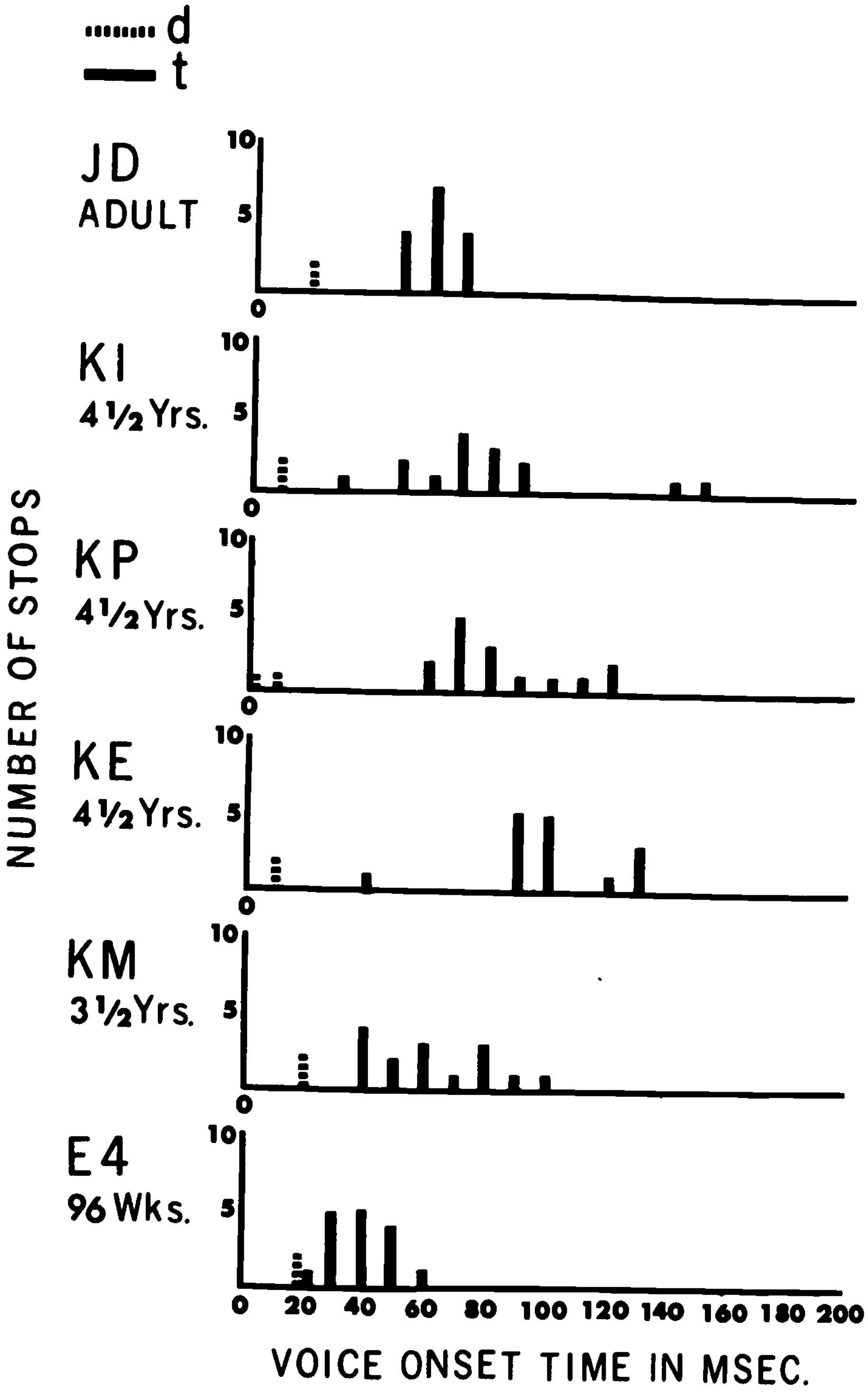


Fig. 8

Stop distributions for each subject corresponding to the repetitions of the /d/ and /t/ words produced by E4 at 96 weeks of age.

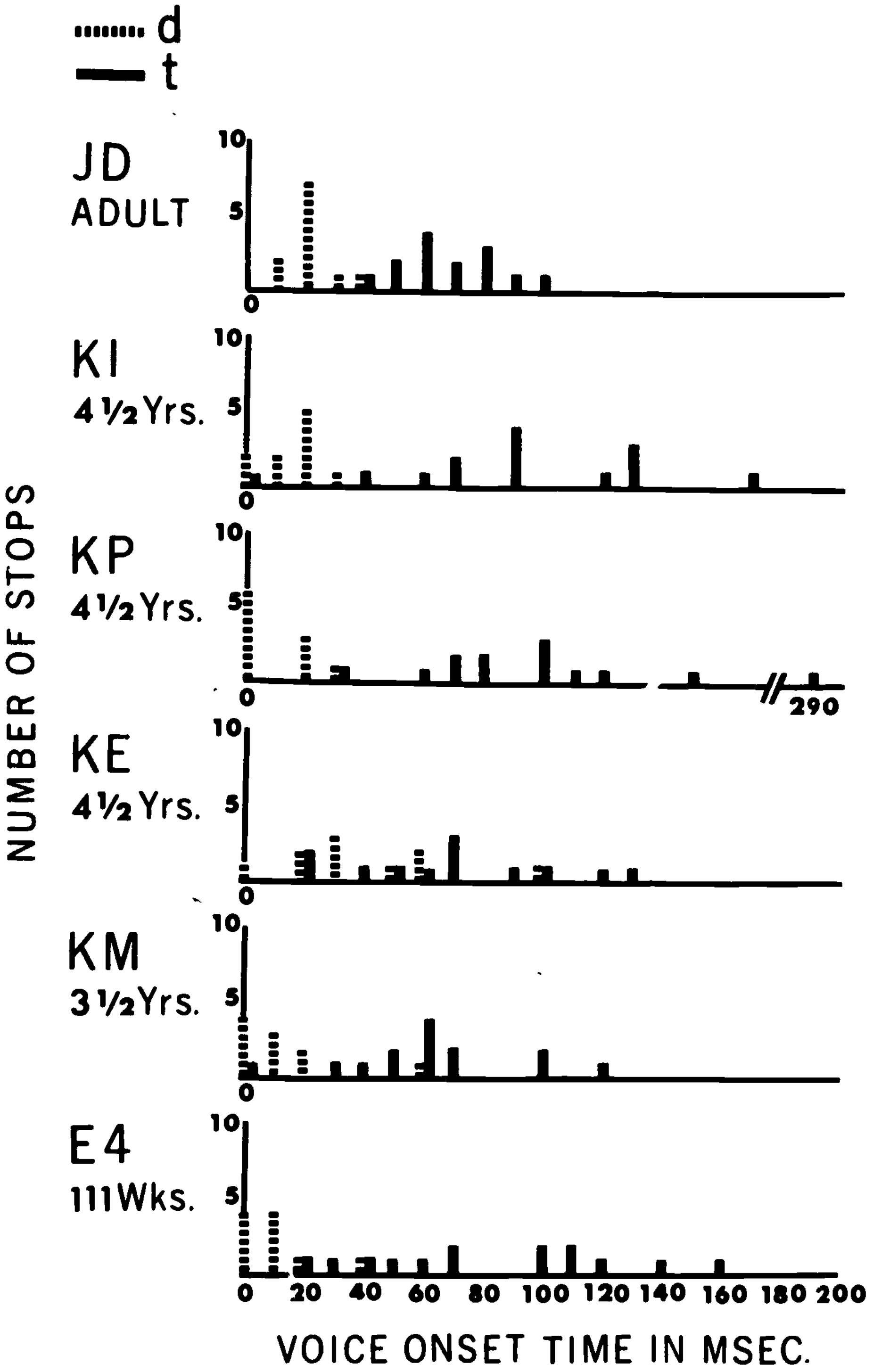


Fig. 9

Stop distributions for each subject corresponding to the repetitions of the /d/ and /t/ words produced by E4 at 111 weeks of age.

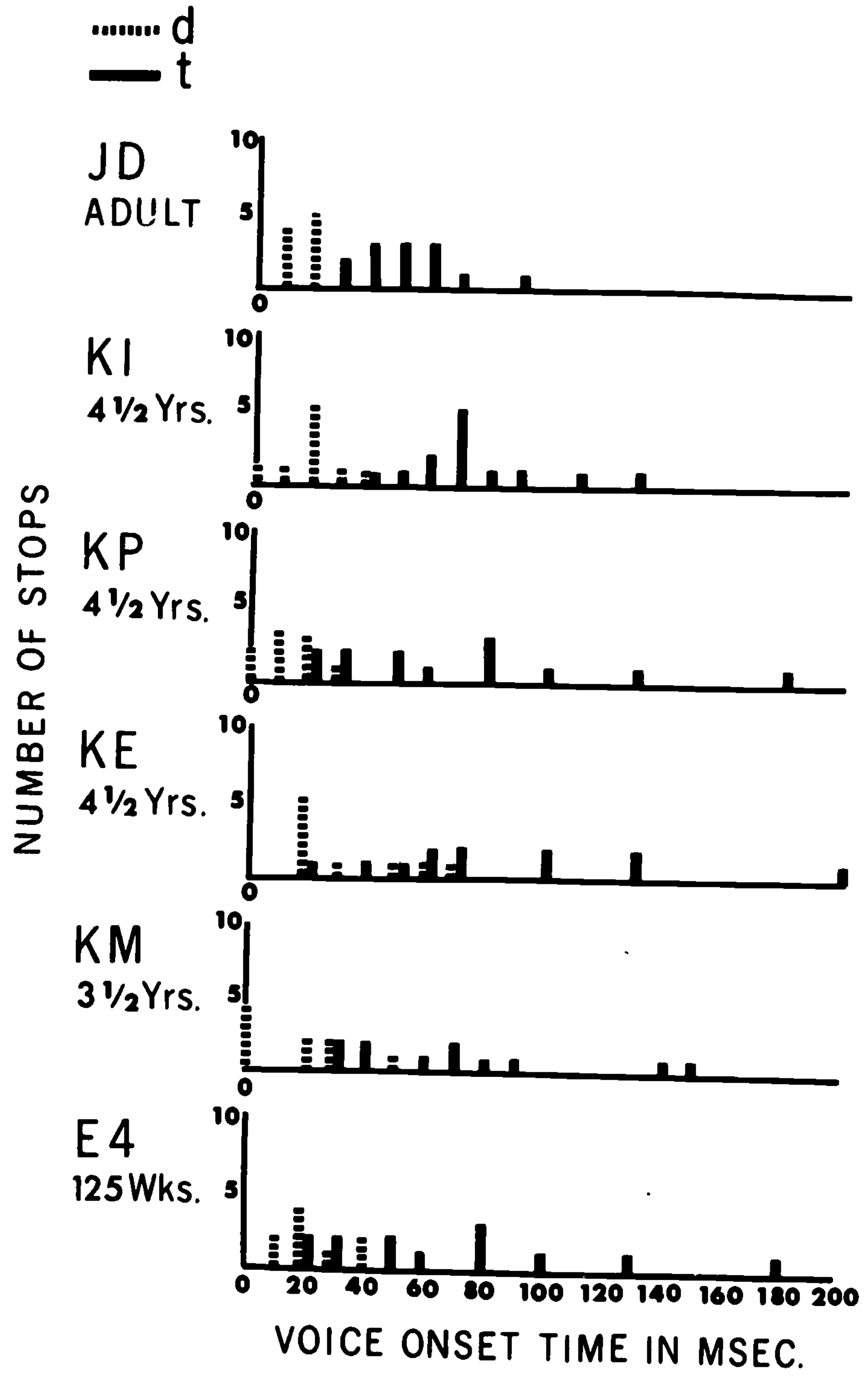


Fig. 10

Stop distributions for each subject corresponding to the repetitions of the /d/ and /t/ words produced by E4 at 125 weeks of age.

+60 msec. On the other hand, few /t/ words have VOT values in the d-range. The /t/ distribution at 96 weeks, in fact, lies midway between adult /d/ and /t/ distributions observed in Figure 1. Therefore, according to listening tests for adults (Abramson and Lisker, 1970), if these stops (for /t/ words) were judge on the basis of voice onset time alone, they would be ambiguously categorized as /d/ or /t/.

E4's /t/ distribution changes radically by 111 weeks of age and is by then similar to all other /t/ distributions for all the children (except E4 at 96 weeks). The characteristic /t/ distribution has a wide range, (0 to +290 msec for all children), but with a mean not significantly different from adult JD's mean of +65 msec for all /t/ distributions. The distributions have no apparent mode, with the possible exception of subject KI (4-1/2 years). For all children, very few of the /t/ words intrude into the d-range. The children's /t/ distributions contrast with that of adult JD which has a narrower range of VOT values, +30 to +100 msec, and a mode at +60 msec. It is also of interest to note that E4's /t/ words actually account for the major portion of all apical stops collected at 96, 111, and 125 weeks that have VOT greater than +30 msec.

A number of conclusions can be drawn from these results. When children begin to use /t/ words, they distinguish their productions functionally from /d/ words along the dimension of voice onset time. However, VOT distributions for /t/ clearly deviate from the adult model. At the earliest age for which we could collect some /t/ words, E4's distribution lies ambiguously across the boundary between the adult /d/ and /t/ distributions, but with VOT values nonetheless clearly larger than those for the majority of adults' or children's /d/ words. We cannot offer an explanation as to why this distribution is so different from the other children's /t/ distributions. It would not appear to be an artifact. The word "toy," or "toys," occurs both at 96 and 111 weeks. For four utterances of "toy(s)" at 96 weeks, the range of VOT is +20 to +40 msec; for ten occurrences of "toy(s)" at 111 weeks, the range is +20 to +140 msec. Thus, at 96 weeks of age E4 has learned one aspect of the adult model for /t/, that /t/ should be produced with VOT greater than +25 msec. But she has not learned to produce apical stops with a large enough delay in the onset of voicing that they would be unambiguously categorized by adults as /t/ on the basis of VOT alone. However, since our adult listeners judged that the /t/ words clearly began with [t], it is possible that some other cue besides voice onset time was being effectively signaled at this age.

By 111 weeks, however, a new pattern of /t/ production occurs which continues to be characteristic of E4's and of our other children's speech through 4-1/2 years of age. This /t/ distribution characteristically has a wide range of VOT values and no mode. At this stage, two of the most important aspects of the adult VOT model have been acquired: the VOT values are greater than +25 msec, and the vast majority of the VOT values are large enough to be unambiguously categorized as /t/ by adult listeners. However, the adult /t/ VOT range and mode are still to be acquired. We have proposed an explanation of why the /t/ range is wide: the production of /t/ is a difficult and complex articulation, wherein the timing evidenced by the adult model just be finely controlled. What is surprising, however, is that at 4-1/2 years, two years after E4's 111-week distribution, no particular change in the /t/ distribution occurs. That is, we know that children will eventually control their /t/ productions within the limits of the adult model, but this control has not been achieved as late as 4-1/2 years. Jacqueline Sachs (personal communication) has some roughly comparable VOT data for /p/ which shows that by 5 years of age, half of her six subjects have a mode and a more restricted range of VOT values. It appears, then, that acquisition of the adult model for /t/ does not occur until after 5 years of age.

Summarizing, we have traced in some detail the development of apical stop consonants with respect to VOT from 1 to 2 years of age and have included supplementary data covering birth to 4-1/2 years of age. From the physiology of stop production, two hypotheses were supported, i.e., that control over timing in stop articulation is inherently difficult and that English /d/ is easier to produce than /t/. A sequential pattern of the development of apical stops with respect to VOT is suggested incorporating these hypotheses.

No stops are observed in neonatal vocalizations. When stops first appear around 6 months of age, the VOT distribution has a wide range of randomly distributed values extending from voicing lead to long voicing lag. This indicates an infant's inability to control timing between the supraglottal and glottal articulatory gestures. Control over timing for the apical stop is achieved first for the short voicing lag category. Apical stops in the long voicing lag category are then gradually added. When words beginning with /d/ and /t/ are first observed--about 2 years of age for subject E4--the characteristics of these distributions remain constant until at least 4-1/2 years. The distribution for /d/ looks similar to the adult's distribution but with more errors into the t-range. The /t/ word distributions bear less resemblance to the adult's, having a wide range of VOT values with no mode, although there are few errors into the d-range.

REFERENCES

Abramson, A. S. and L. Lisker. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. Proc. Fifth Intl. Cong. on Acoustics, ed. by D. E. Commins, A51 (Liège; Imp. G. Thone).

Abramson, A. S. and L. Lisker. (1970) Discriminability along the voicing continuum: Cross-language tests. Proc. Sixth Intl. Cong. Phon. Sci., Prague, 1967 (Prague: Academia) 569-573.

Berti, F. B. and H. Hirose. (1972) Velopharyngeal function in oral/nasal articulation and voicing gestures. Haskins Laboratories Status Report on Speech Research SR-28, 143-156.

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

Fromkin, Victoria. (1966) Neuromuscular specification of linguistic units. Lang. Speech 9, 170-199.

Fujii, I. (1970) Phoneme identification with dynamic palatography. Annual Bulletin, (Research Institute of Logopedrics and Phoniatrics, University of Tokyo), No. 4, 67-73.

Hahn, Emily. (1971) On the side of the apes, I and II. The New Yorker, April 17, 46-97; April 24, 46-91.

Harris, K. S., G. F. Lysaught, and M. M. Schvey. (1965) Some aspects of the production of oral and nasal labial stops. Lang. Speech 8, 135-147.

Hirose, H. (1971) An electromyographic study of laryngeal adjustments during speech articulation: A preliminary report. Haskins Laboratories Status Report on Speech Research, SR-25/26, 107-116.

Hiroto, I., M. Hirano, Y. Toyozumi, and T. Shin. (1967) Electromyographic investigation of the intrinsic laryngeal muscles related to speech sounds. Annual OTOL 76, 861-872.

Kim, C.-W. (1970) A theory of aspiration. Phonetica 21, 107-116.

Lieberman, P., K. S. Harris, P. Wolff, and L. H. Russell. (1968) Newborn infant cry and nonhuman primate vocalizations. Haskins Laboratories Status Report on Speech Research SR-17/18, 23-39. (Also J. Speech Hearing Res., in press.)

Lieberman, P., E. S. Crelin, and D. H. Klatt. (1970) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. Haskins Laboratories Status Report on Speech Research SR-24, 57-90.

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. Proc. Sixth Intl. Cong. Phon. Sci., Prague, 1967 (Prague: Academia) 563-567.

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

Lisker, L., M. Sawashima, and A. S. Abramson. (1970) Cinegraphic observations of the larynx during voiced and voiceless stops. Haskins Laboratories Status Report on Speeeh Research SR-21/22, 201-210.

Lubker, J. F., B. F. Fritzell, and J. Lindqvist. (1970) Velopharyngeal function: An electromyographic study. Royal Institute of Technology, Speech Transmission Laboratory, Quarterly Progress and Status Report 4, 9-20.

Lubker, J. F. and Parris, P. J. (1970) Simultaneous measurements of interoral pressure, force of labial contact, and labial electromyographic activity during production of the stop cognates /p/ and /b/. J. acoust. Soc. Amer. 47, 625-633.

Morse, P. (1971) The discrimination of speech and nonspeech stimuli in early infancy. Ph.D. dissertation, University of Connecticut. (Also J. exp. Child Psychol, in press).

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

Preston, M. S. and D. K. Port. (1969) Further results on the development of voicing in stop-consonants in young children. Haskins Laboratories Status Report on Speech Research SR-19/20, 189-199.

Rothenberg, M. (1968) The breath-stream dynamics of simple-released-polsive production. Bibliotheia Phonetica 6 (Basel: S. Karger).

Slis, I. H. (1970) Articulatory measurements on voiced, voiceless and nasal consonants. Phonetica 21, 193-210.

Stark, R. E. (1968) Voicing in initial stop consonants produced by hearing-impaired children. Annual Report (Neurocommunications Laboratory, Baltimore, Md.), 173-210.

Stark, R. E. (1971) Some features of the vocalizations of young deaf children. Third Symposium on Oral Sensation and Perception: The Mouth in the Infant, ed. by J. F. Bosma (Springfield, Ill.: Charles Thomas & Co.).

ABSTRACT

The Discrimination of Speech and Nonspeech Stimuli in Early Infancy[*]

Philip Allen Morse[+]
Haskins Laboratories, New Haven

The discrimination of synthetic speech and nonspeech stimuli was investigated in infants 40-54 days of age by means of a nonnutritive conjugate sucking procedure. Four groups of Ss were given repeated presentations of one auditory stimulus and, upon habituating to it, were shifted to a second (postshift) stimulus. For Group P (Place) the pre- and postshift stimuli differed according to place of articulation ([ba-] vs. [ga-]). Group I (Intonation) received a stimulus shift consisting of a difference in intonation ([ba-] vs. [ba+], i.e., falling vs. rising intonation). Group C (Control) was presented with the same stimulus during preshift and postshift ([ba-]). For Groups NS (Nonspeech control) the pre- and postshift stimuli consisted of the isolated acoustic cues which differentiate the place stimuli [ba] and [ga]. Changes in hi-amplitude sucking revealed that infants 40-54 days of age can discriminate the acoustic cues for place of articulation and intonation. Furthermore, a comparison of the place and nonspeech control conditions suggested that infants respond to the acoustic cues for place in a linguistically relevant manner.

[*] Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, University of Connecticut. An article based on these data will appear in the Journal of Experimental Child Psychology.

[+] Currently University of Wisconsin, Madison.

ABSTRACT

The Effect of Delayed Channel on the Perception of Dichotically Presented Speech and Nonspeech Sounds[*]

Robert John Porter, Jr.[+]
Haskins Laboratories, New Haven

Previous investigations have found that subjects identify the temporally lagging member of a pair of dichotically asynchronous stop consonant-vowel syllables with greater accuracy than the leading member. This advantage for the lagging syllable has been termed the "lag effect." Three studies are reported which examined the possibility that this effect is a manifestation of the special processes required for the perception of the acoustically encoded stop consonants rather than a general effect to be found for several types of acoustic events.

In the first two experiments, the effects observed for the syllables [bæ, dæ, gæ] were compared to those obtained with nonspeech sounds which were acoustically comparable to the syllables but had been previously shown not to require the same special perceptual processing. Two types of nonspeech sounds were used: (1) the acoustically isolated second formants of the syllables ("bleats"), and (2) acoustically isolated second-formant transitions ("chirps"). Three groups of subjects received dichotically asynchronous pairs of syllables, bleats, or chirps. Twelve stimulus-onset asynchronies, from 0 to 165 msec, were used.

If the lag effect is a general phenomenon of dichotic listening then the nonspeech would be expected to display lag effects similar to those observed for syllables. This did not appear to be the case. Whereas large and reliable lag effects were found for the syllables at asynchronies less than 120 msec (maximal at 60 msec), the lag advantages of the nonspeech controls were very small and variable. The chirps, in some cases, even displayed _lead_ advantages.

The results for the nonspeech signals were interpreted in terms of dichotic masking effects such as are observed in nonspeech auditory masking studies. The considerably larger and more reliable lag effects for the stop-vowel syllables were seen as indicating that the perceptual processing of these signals is particularly sensitive to the conditions of dichotic asynchronous competition. It was argued that this peculiar sensitivity was a manifestation of the special "speech mode" processing known to be required for the perception of the highly acoustically encoded stop consonants.

[*] Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, University of Connecticut.

[+] Currently Kresge Hearing Research Laboratory, Louisiana State University, New Orleans.

In order to examine further the suggested relationship between the lag effect and perception in the speech mode, a third experiment compared the results obtained with stops (in syllables [ba, da, ga]) to those obtained for a liquid and two semi-vowels (in syl!ables [la], [wa], [ja]). The liquids and semi-vowels tend to be acoustically less encoded than the stops and, presumably, require special processing to a lesser degree. Previous studies had demonstrated that steady-state vowels, which can be shown to be less encoded than liquids and semi-vowels, tend not to yield lag effects. If the lag effect is a consequence of the involvement of special decoding processes, the liquids and semi-vowels would be expected to display lag effects to a lesser degree than stops and to a greater degree than vowels.

The procedures and asynchronies used were the same as for the first two experiments. Eleven subjects received both the stop and the liquid and semi-vowel dichotic tests.

The results for the liquid and semi-vowels were consistent with expectation. Five subjects displayed lag effects similar to those they displayed for stops. The results for the six remaining subjects were similar in several respects to those which had been previously observed for chirps and vowels. Apparently, these "intermediately" encoded speech sounds may in some circumstances be perceived like the stops and in other circumstances like vowels and nonspeech.

Taken together, the results of all three experiments suggest that the lag effect is not a general phenomenon of dichotic listening but is specifically associated with the perception of encoded speech sounds. As such, the effect is a possibly valuable source of information concerning the character of these special decoding processes.

ABSTRACT

Phonetic Coding of Kanji[*]

Donna Erickson[+], Ignatius G. Mattingly[++], and Michael Turvey[++]

An experiment in the short-term recall of visually presented Japanese Kanji ideograms suggests that Kanji may, like alphabetic words, be encoded phonetically, despite their lack of phonetic structure. The experiment, based on Kintsch and Buschke's (1969) paradigm, assumed that similarity of items in a list increased errors in recall. Four lists were prepared, each containing sixteen different Kanji. The first included phonetically similar pairs of characters; the second, semantically similar pairs; the third, visually similar pairs; the fourth was a control list containing no similar pairs. The subjects, ten native speakers of Japanese, were presented with randomly ordered versions of each list, at one character per second. After a subject had seen an entire list, he was presented with a cue character selected from the list and asked to recall the character which had been presented immediately before the cue. Confusion in primary memory was significantly greater for the phonetic list than for the other lists. These results strengthen the hypothesis that regardless of structure, visually presented linguistic items are, like speech itself, phonetically processed.

[*] Talk given at the 83rd meeting of the Acoustical Society of America, Buffalo, N. Y., April 1972.

[+] University of Connecticut, Storrs.

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

PUBLICATIONS AND REPORTS[*]

Publications and Manuscripts

Letter Confusions and Reversals of Sequence in the Beginning Reader: Implications for Orton's Theory of Developmental Dyslexia. I. Y. Liberman, D. Shankweiler, C. Orlando, K. S. Harris, and F. B. Berti. Cortex (June 1971) 7, 127-142.

An Auditory Analogue for the Sperling Partial Report Procedure: Evidence for Brief Auditory Storage. C. J. Darwin, M. T. Turvey, and R. G. Crowder. Cognitive Psychology (April 1972) 3, 255-267.

Speech Cues and Sign Stimuli. I. G. Mattingly. American Scientist (May-June 1972) 60, 327-337.

Research on Audible Outputs of Reading Machines for the Blind. F. S. Cooper, J. H. Gaitenby, and I. G. Mattingly. Bulletin of Prosthetics Research (Spring 1971) BPR 10-15, 241-246. Report by same title and authors, Bulletin of Prosthetics Research (Fall 1971) BPR 10-16, 248-252. Report by same title by F. S. Cooper, J. H. Gaitenby, I. G. Mattingly, P. W. Nye, and G. N. Scholes (appearing in this Status Report) Bulletin of Prosthesics Research (Spring 1972) BPR 10-17, in press.

Preceding Vowel Duration as a Cue to the Perception of Word-Final Consonants. L. J. Raphael. Journal of the Acoustical Society of America (April 1972) 51, 1296-1303.

Electromyography of the Intrinsic Muscles During Phonation. T. Gay, H. Hirose, M. Strome, and M. Sawashima. Annals of Otology, Rhinology, and Laryngology (June 1972) 81, 401-409.

On Peripheral and Central Processes in Vision: Inferences from an Information-Processing Analysis of Masking with Patterned Stimuli. M. T. Turvey. Psychological Review, in press.

The Development of Auditory Feedback Monitoring: IV. Delayed Auditory Feedback Studies on the Vocalization of Children Between Six and Nineteen Months. N. F. Belmore, D. K. Port, R. L. Mobley, and V. E. Goodman. Journal of Speech and Hearing Research (June 1972), in press.

Some Aspects of Selective Readout from Iconic Storage. M. T. Turvey.

Some Effects of Oral Anesthesia upon Speech: An Electromyographic Investigation. G. J. Borden.

Laryngeal Control in Vocal Attack: An Electromyographic Study. H. Hirose and T. Gay.

A Parallel Between Encodedness and the Magnitude of the Right Ear Effect. J. E. Cutting.

[*] The contents of this report, SR-29-30, are included in this listing.

The Phi Coefficient as an Index of Ear Differences in Dichotic Listening. G. M. Kuhn.

Auditory Evoked Potential Correlates of Speech Sound Discrimination. M. F. Dorman.

Short-Term Habituation of the Infant Auditory Evoked Response. M. F. Dorman and R. Hoffman.

Early Apical Stop Production: A Voice Onset Time Analysis. D. K. Port and M. S. Preston.

Reports and Oral Presentations

Problems Inherent in Speech and Speaker Recognition. F. S. Cooper. Presented at Yale Department of Engineering and Applied Science Seminar on Adaptive and Learning Systems, New Haven, Conn., 7 March 1972.

Report on "Machines and Speech" sessions of the Conference on Research Trends in Computational Linguistics. F. S. Cooper. Washington, D.C., 14-16 March 1972.

Voice-Timing Perception in Spanish Word-Initial Stops. A. S. Abramson and L. Lisker. Presented at the 83rd Meeting of the Acoustical Society of America, Buffalo, N. Y., 18-21 April 1972.

Mutual Interference Between Two Linguistic Dimensions of the Same Stimuli. R. S. Day and C. C. Wood. Presented at the 83rd Meeting of the Acoustical Society of America, Buffalo, N. Y., 18-21 April 1972.

Phonetic Coding of Kanji. D. Erickson, I. G. Mattingly, and M. Turvey. Presented at the 83rd Meeting of the Acoustical Society of America, Buffalo, N. Y., 18-21 April 1972.

A Backward Look at Speech Research. F. S. Cooper. Luncheon Address: 1972 Conference on Speech Communication and Processing, Newton, Mass., 25 April 1972.

Field Evaluation of an Automated Reading System for the Blind. P. W. Nye, J. D. Hankins, T. Rand, I. G. Mattingly, and F. S. Cooper. Presented at the 1972 Conference on Speech Communication and Processing, Newton, Mass., 24-26 April 1972.

Word and Phrase Stress by Rule for a Reading Machine. J. H. Gaitenby, G. N. Sholes, and G. M. Kuhn. Presented at the 1972 Conference on Speech Communication and Processing, Newton, Mass., 24-26 April 1972.

Conceptual Development in the Speech and Hearing Sciences--panel participant. K. S. Harris. New York Speech and Hearing Association, Ellenville, N. Y., 26 April 1972.

The Relationships Between Speech and Reading. I. G. Mattingly and J. F. Kavanagh Presented at the International Reading Association Convention, Detroit, Mich., May 1972.

Dissertations

Some Effects of Oral Anesthesia upon Speech: A Perceptual and Electromyographic Analysis. Gloria J. Borden. Ph.D. dissertation. City University of New York, 1971.

Auditory Evoked Potential Correlates of Speech Perception. Michael F. Dorman. Ph.D dissertation. University of Connecticut, 1971.

The Discrimination of Speech and Nonspeech Stimuli in Early Infancy. Philip A. Morse. Ph.D. dissertation. University of Connecticut, 1971.

The Effect of Delayed Channel on the Perception of Dichotically Presented Speech and Nonspeech Sounds. Robert J. Porter, Jr. Ph.D. dissertation. University of Connecticut, 1971.