DOCUMENT RESUME

ED 044 679 AL 002 653

TITLE	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 1970.
INSTITUTION	Haskins Labs., New York, N.Y.
SPONS AGENCY	Office of Naval Research, Washington, D.C. Information Systems Research.
REPORT NO	SR-21/22
PUB DATE	Jul 70
NOTE	213p.
EDRS PRICE	EDRS Price MF-$1.00 HC-$10.75
DESCRIPTORS	*Acoustics, Articulation (Speech), Auditory Discrimination, Auditory Perception, Behavioral Science Research, Cerebral Dominance, Consonants, Ears, Phonetics, *Psychoacoustics, *Psycholinguistics, Recall (Psychological), *Speech, Syllables, Vowels

ABSTRACT

This report (for January 1--June 30, 1970) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation and practical application thereof. Extended reports and manuscripts cover the following topics: Hemispheric Specialization for Speech Perception; Ear Differences in the Recall of Fricatives and Vowels; Selective Listening for Temporally Staggered Dichotic CV Syllables; Temporal Order Judgments in Speech; Opposed Effects of a Delayed Channel on Perception of Dichotically and Monotically Presented CV Syllables; Discrimination in Speech and Nonspeech Modes; Effects of Filtering and Vowel Environment on Consonant Perception; A Direct Magnitude Scaling Method to Investigate Categorical Versus Continuous Modes of Speech Perception; On the Speech of Neanderthal Man, Glottal Adjustments for English Obstruents; Cinegraphic Observations of the Larynx During Voiced and Voiceless Stops. (Author/FWB)

SR 21/22 (1970)

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 1970

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

ACKNOWLEDGEMENTS

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

Information Systems Branch, Office of Naval Research
Contract N00014-67-A-0129-0001
Req. No. NR 048-225

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, Veteran Administration
Contract V-1005M-1253

National Institutes of Health
General Research Support Grant FR-5596

CONTENTS

EXTENDED REPORTS
AND
MANUSCRIPTS

Hemispheric Specialization for Speech Perception*

Michael Studdert-Kennedy+ and Donald Shankweiler++
Haskins Laboratories, New Haven

Abstract. Earlier experiments with dichotically presented nonsense syllables had suggested that perception of the sounds of speech depends upon unilateral processors located in the cerebral hemisphere dominant for language. Our aim in this study was to pull the speech signal apart to test its components in order to determine, if possible, which aspects of the perceptual process depend upon the specific language processing machinery of the dominant hemisphere. The stimuli were spoken CVC syllables presented in dichotic pairs which contrasted in only one phone (initial stop consonant, final stop consonant, or vowel). Significant right-ear advantages were found for initial and final stop consonants, nonsignificant right-ear advantages for six medial vowels, and significant right-ear advantages for the articulatory features of voicing and place of production in stop consonants. Analysis of correct responses and errors showed that consonant features are processed independently, in agreement with earlier research employing other methods. Evidence is put forward for the view that specialization of the dominant hemisphere in speech perception is due to its possession of a linguistic device, not to specialized capacities for auditory analysis. We have concluded that, while the general auditory system common to both hemispheres is equipped to extract the auditory parameters of a speech signal, the dominant hemisphere may be specialized for the extraction of linguistic features from those parameters.

---

*This paper appeared in J. Acoust. Soc. Amer. 48, 579-594 (1970).

+Also, Queens College, City University of New York, Flushing.

++Also, University of Connecticut, Storrs.

Acknowledgements. We wish to thank A.M. Liberman, M.P. Haggard, C.J. Darwin, and T. Halwes for many hours of fruitful discussion during the preparation of this paper.

Acknowledgement is due to the Charles E. Merrill Publishing Company for permission to reprint Figures 1 and 2 which are also to appear in Shankweiler (in press).

Introduction

Man is a language-using animal with skeletal structure and brain mechanisms specialized for language. For more than a century, it has been known that language functions are, to a considerable extent, unilaterally represented in one or other of the cerebral hemispheres, most commonly the left. The evidence of cerebral lateralization and localization argues powerfully for the existence of neural machinery specialized for language, but the exact nature of the language function, and characteristics of the neural mechanisms that serve it, remain to be specified. Most studies of the neural basis of language have dealt with higher-level language functions and their dissolution. An alternative approach, which may prove more fruitful, is to investigate the lower-level language functions, that is, to focus on the production and perception of speech sounds.

Study of the evolution of the vocal tract in relation to the physiological requirements for producing the sounds of speech suggests that man has evolved special structures for speech production and has not simply appropriated existing structures designed for eating and breathing (Lieberman, 1968; Lieberman et al., 1969). We may reasonably suppose that he has also evolved matching mechanisms for speech perception. There is, in fact, much evidence that speech perception entails peculiar processes, distinct from those of nonspeech auditory perception (for a review of the evidence, see Liberman et al., 1967). There are also grounds for believing that the sounds of speech are integral to the hierarchical structure of language (Lieberman, 1967; Mattingly and Liberman, 1970). We might, therefore, expect that among the language processes lateralized in the dominant hemisphere are mechanisms for the perception of speech. Evidence of this is not easily gathered from normal subjects with intact nervous systems. But recently a plausible technique has become available and is put to work in the present study.

Kimura (1961a), using a task similar to one described by Broadbent (1954), showed that, if pairs of contrasting digits were presented simultaneously to right and left ears, those presented to the right were more accurately reported. She attributed the effect to functional prepotency of the contralateral pathway from the right ear to language-dominant left hemisphere (Kimura, 1961b). There is evidence for stronger contralateral than ipsilateral auditory pathways in dog (Tunturi, 1946), cat (Rosenzweig, 1951; Hall and Goldstein, 1968), and man (Bocca et al., 1955) and for inhibition of the ipsilateral signal in man during dichotic presentation (Milner et al., 1968; Sparks and Geschwind, 1968).

The right-ear advantage for verbal materials has now been repeatedly confirmed, and attempts to account for it solely in terms of memory, attention, or various response factors have been found inadequate (for reviews, see Bryden, 1967, and Satz, 1968). Kimura's attribution of the effect to cerebral dominance has received support from several other pieces of evidence. She herself (1961b) showed that the effect was reversed--a left-ear advantage appeared--in subjects known to have language dominance in the right hemisphere. She and others (Kimura, 1964; Chaney and Webster, 1965; Curry, 1967) showed that the effect was also reversed for nonspeech materials (melodies, sonar signals, environmental noises). The reversal of the effect for dichotically presented nonspeech fits with other indications that perception of auditory patterns and their attributes typically depends more upon right-hemisphere mechanisms than upon left (Milner, 1962; Spreen et al., 1965; Shankweiler, 1966a, b; Vignolo, 1969).

Kimura's contention that ear advantages in dichotic listening reflect dual cerebral asymmetries of function in perception of verbal and nonverbal materials is thus supported by much evidence from a variety of sources. Dichotic listening techniques, therefore, seem to offer a new way to raise the question of the status of speech (in the narrow sense) and its relation to language. If speech is indeed integral to language, we might expect this fact to be reflected in the neural machinery for its perception. Specifically, we may ask: are the sounds of speech processed by the dominant hemisphere, by the minor hemisphere along with the music, or equally by both hemispheres? All the dichotic speech studies referred to above used meaningful words as stimuli and therefore did not speak to this question. Studies using nonsense syllables have, however, been carried out in order to discover whether the right-ear advantage depends upon the stimuli being meaningful. The results show clearly that it does not (Shankweiler and Studdert-Kennedy, 1966; Curry, 1967; Curry and Rutherford, 1967; Kimura, 1967; Kimura and Folb, 1968; Darwin, 1969; Haggard, 1969). We were therefore encouraged to make further use of dichotic listening experiments as a device for probing in some detail the processes of speech perception. Our general plan was to pull the speech signal apart and to test its components (consonants, vowels, isolated formants, and so on) in order to determine, if possible, which aspects of the perceptual process depend upon lateralized mechanisms and, by looking for information contained in perceptual errors, to guess at some of the characteristics of the processing machinery.

In a study employing synthetic speech (Shankweiler and Studdert-Kennedy, 1967), we compared synthetic CV syllables and steady-state vowels. Our choice of stimuli was dictated by the repeated finding at the Haskins Laboratories

that the identification of stop consonants and vowels engage different perceptual processes, stop consonants being "categorically," vowels "continuously," perceived (for discussion and summary of this evidence, see Liberman et al., 1967; Lane, 1965; Studdert-Kennedy et al., 1970). In our dichotic study of these two classes of phonemes, we found a significant right-ear advantage for the vowels. We also found evidence implicating the articulatory features of voicing and place of production in stop consonant perception and lateralization. The present study[1] was designed to press our analysis of speech perception further by testing the lateralization of "natural" speech rather than synthetic, of final consonants as well as initials, of vowels embedded in CVC syllables rather than steady-state, and of the consonant features of voicing and place.

Method

Test Construction. We wished to study dichotic effects in the perception of initial and final stop consonants followed or preceded by various vowels and of medial vowels followed or preceded by various stop consonants. We constructed four dichotic tests: two consonant and two vowel tests. The stimuli consisted of consonant-vowel-consonant (CVC) syllables formed by pairing each of the six stop consonants, /b,d,g,p,t,k/, with each of the six vowels, /i,ɛ,æ,a,ɔ,u/. In one consonant and one vowel test, all syllables ended with the consonant /p/ [initial-consonant-varying (IC) tests], while in the other pair of tests, all syllables began with the consonant /p/ [final-consonant-varying (FC) tests].

The syllables were spoken by a phonetician. He was given two randomized lists of thirty-six CVC syllables (six consonants X six vowels), one with initial consonants varying, one with final consonants varying. He was asked to read each list once at an even intensity (monitored on a VU meter) and to release the final stop. His utterances were recorded, a spectrogram was made of each syllable, and its duration was measured. The durations averaged around 400 msec., with a range of about 300-500 msec. Most of the variability arose

---

[1]Reports of some of the findings of this study were included in a paper read before the Acoustical Society of America (Shankweiler and Studdert-Kennedy, 1967a), and in a presentation by one of us (D.S.) at the ONR Conference on Perception of Language, University of Pittsburgh, January 1967 (Shankweiler, in press).

from differences in the "natural" length of the vowels and from differences in the delay of the final stop release. For some few syllables, which seemed not perfectly intelligible, the phonetician was asked to make a new recording.

As an example of test construction, we will describe the procedure for the dichotic consonant test in which the initial consonant varied. The thirty-six recorded syllables were dubbed several times with a two-channel tape recorder: half the syllables were assigned to one track of the tape, half to the other, so that each consonant was recorded equally often on each track. The syllables were then spliced into tape loops. Each loop carried a pair of syllables contrasting only in their initial consonants (e.g. /bap/ - /dap/), one on each tape track. There were ninety such loops: each consonant was paired once with every consonant other than itself (fifteen combinations) followed by each of the six vowels.

The next task was to synchronize the onsets of the two syllables on a loop. This was accomplished by playing the loop on a special two-channel tape deck, modified to permit the length of leader tape passing between two playback heads to be varied, until the onsets of the two syllables coincided. Onset was defined on a permanent oscillographic record, obtained from a Honeywell 1508 Visicorder, as the first excursion above noise level that was sustained and followed by clear periodicity. Synchronization of onsets was determined from a three-channel Visicorder record, with two channels displaying the speech waves and the third a 100 Hz sine wave. Figure 1 reproduces the Visicorder record of two syllables with synchronous onsets.

Once the playback of two syllables on a loop had been synchronized, the pair was dubbed on parallel tracks using an Ampex PR-10 recorder. The input channels were matched for peak intensity on the VU meter, and the pair was recorded four times, each syllable going twice to channel 1 and twice to channel 2. In view of the arduous process of construction, this master tape of synchronized, contrasting syllables, distributed evenly over channels, was preserved uncut, as a source of stimuli in possible future experiments. From it, each syllable pair was recorded twice, once in each of its two channel orientations, on an Ampex PR-10. Thus ninety loops, made from dubbings of thirty-six parent recordings, yielded 180 third-generation stimuli in which each consonant was paired with every consonant other than itself followed by each of the six vowels, once on each tape track.

These stimuli were then spliced into a random order with the restriction that each consonant pair should appear once with each vowel in the first half

Temporal Alignment of Syllables for Dichotic Presentation

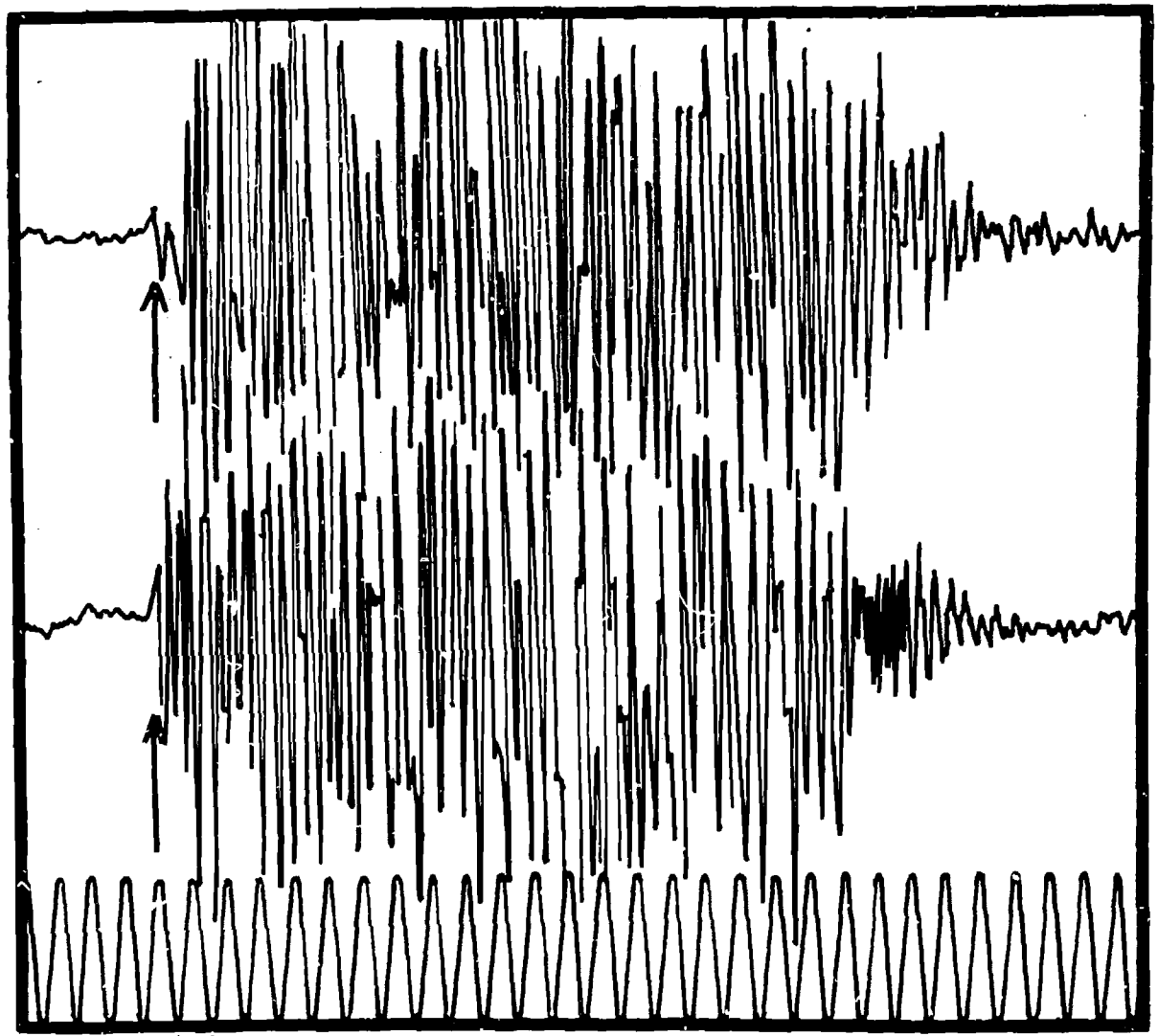


FIG. 1

and once with each vowel in the second half of the test. There was a six-second interval between stimuli, a ten-second interval after every tenth stimulus, a thirty-second interval after the ninetieth.

The IC vowel test was constructed from the original thirty-six recordings in exactly the same way as the IC consonant test, with the single difference that the tape loops were formed from pairs of syllables contrasting only in their vowels.

The FC consonant and vowel tests were constructed in a similar manner. Here the difference was in the alignment procedure: these syllables were synchronized at their final releases. Selecting the exact point of release on an oscillographic record proved a singularly difficult task. Many arbitrary decisions had to be made, and the resulting alignments were almost certainly less precise than those of the corresponding IC pairs.

Subjects. There were twelve subjects: seven women and five men, aged between 18 and 26 years. Audiograms were taken separately on left and right ears. All subjects had normal hearing, considered themselves right-handed and had no left-handed members of their immediate families. They served for four sessions of 45-50 minutes each and were paid for their work.

Procedure. Subjects took the tests individually in a quiet room, listening, over matched PDR-8 earphones, to the output of an Ampex PR-10 two-channel tape recorder.

The order in which the tests were given was counterbalanced. All subjects took a vowel test in their first and fourth sessions: half took the IC, half the FC, on each occasion. All subjects took a consonant test in their second and third sessions: half of those who had taken the IC vowel test in their first session took the FC consonant test in their second and the IC consonant test in their third. The orders for the other subgroups of subjects were appropriately reversed. One subject (BZ) did not come for his final session and so gave no data on the IC vowel test.

The experimenter began a session by playing a steady-state calibrating tone (1000 Hz), spliced to the beginning of each test, on both recorder channels and adjusting the outputs to the voltage equivalent of approximately 70 db SPL. The subject was then given the following, or analogous, instructions to read:

This is an experiment in speech perception. You are going to listen over earphones to a series of monosyllables--consonant-vowel-consonant monosyllables, such as 'pet,' 'bap,' 'doop,' 'pawg,' and so on. They will be presented in simultaneous pairs, one to the left ear, one to the right. In any pair, the two syllables will have the same consonants, but different vowels. The two vowels will always be different, and will be drawn from the set of six given below.

Your task is to identify both vowels. Opposite the appropriate trial number on your answer sheet you should write two of the following:

ee (as in beet)

eh (as in bet)

ae (as in bat)

ah (as in father)

aw (as in bought)

oo (as in boot)

You should always write two vowels, even if you have to guess. Write them in order of confidence. That is to say, write the one you are more sure of first, the one you are less sure of second. There are 180 trials in the first test. You will have a short rest after 90, a longer rest after the 180. Then you will do a second test of the same length.

Each batch of 90 trials takes about ten minutes, and the task may not be easy. But you are asked to give it your fullest possible attention. Don't worry if you think you are missing a lot. Just make careful guesses, and then get ready for the next trial. There are about six seconds between trials.

Any questions? If not, put the earphones on and adjust them so that they fit comfortably on your head.

For the consonant test, the specified responses were: b,d,g,p,t,k. Appropriate changes in instructions were made for the FC tests.

Subjects wrote their responses on two 90-item response sheets, at the top of which the set of letters from which responses were to be selected was displayed. Upon completion of the 180-item test, subjects took a short rest, reversed the orientation of the earphones and took the test again. For each of the four dichotic tests, half the subjects heard channel 1 in their right ear first, half heard it in their left ear first. Channels were switched across ears by phone reversal, rather than electrically, so that bias due to channel and phone characteristics or phone position on the head would not be confounded with ear performance.

Summary. The elaborate procedure of test construction and presentation described above yielded 360 dichotic trials for each subject on each test, that is,

twenty-four judgments on each of the fifteen contrasting phoneme combinations or sixty judgments on each phoneme by ear. Any bias due to neighboring vowel (or consonant), imprecise synchronization of onsets or offsets, recorder channels, earphone characteristics, position of earphones on the head, or sequence of testing was distributed equally over the ears of the entire group of subjects.

Results

Overall Performance. Table I summarizes the raw data and provides percentage bases for subsequent tables. Overall performance on both ears was considerably higher for the IC vowels (82%) than for the IC consonants (68%); FC consonant performance (74%) falls midway.[2] For reasons that will become apparent (see below: an index of the laterality effect) we distinguished between trials on which both syllables were correctly identified and trials on which only one syllable was correctly identified. The distribution of total correct into the two categories is shown in the two right-hand columns of Table I. The difficulty of the IC consonant test as compared with the vowel is again shown by its lower percentage of both-correct trials (43% for consonants, 69% for vowels) and its higher percentages of one-correct trials (25% for consonants, 14% for vowels).

Ear Advantage. Table II presents percentage correct on the three tests, by preference and by ear, for individual subjects and for the group.

On the initial consonant test every subject shows a total right-ear advantage of between 4% (SB, JH) and 22% (AL). The mean total right-ear advantage of 12% is significant on a two-tailed matched pairs t-test ($t=7.19$, $p<0.001$).

For the final consonants, right-ear advantages are smaller and more variable. Ten subjects show a total right-ear advantage of between 2% (JWn) and 15% (LN). Two subjects (MJ, HW) show left-ear advantages of 1% and 3%, respectively. The mean total right-ear advantage of 6% is significant on a two-tailed matched pairs t-test ($t=3.84$, $p<0.01$).

The vowel results are again variable. Seven subjects show right-ear advantages, three (JH, NK, JWn) show small left-ear advantages, one (SB) no advantage. The mean total right-ear advantage of 2% falls short of significance on a two-tailed test at the 0.05 level ($t=2.16$, $p<0.06$).

---

[2]Main results for the FC consonants are presented in Tables I and II. All further consonant data analysis is for IC consonants only, largely due to our dissatisfaction with the FC stimuli. Accordingly, since vowel data were intended for comparison with consonant, only the IC vowel data have been fully analyzed: all reported vowel results are for this test only.

TABLE I

Overall performance: initial consonants, medial vowels, final consonants.

	Test		
	Initial Consonants	Medial Vowels	Final Consonants
Number of syllable combinations	15	15	15
Number of syllable presentations per ear per subject	360	360	360
Number of subjects	12	11	12
Number of syllable presentations per ear for group	4320	3960	4320
Number of syllable presentations for group (both ears)	8640	7920	8640
Total correct (percent)	5858 (68%)*	6516 (82%)	6394 (74%)
Number correct on trials with both correct (percent)	3702 (43%)	5442 (69%)	4505 (52%)
Number correct on trials with only one correct (percent)	2156+ (25%)	1074+ (14%)	1889 (22%)

*All percentages in this table are based on number of syllable presentations for group (both ears).

+Group percentage bases for trials on which only one syllable was correctly identified.

Table 2

PERCENTAGE CORRECT BY PREFERENCE AND BY EAR FOR INDIVIDUAL SUBJECTS

SUBJECT	INITIAL CONSONANTS						MEDIAL VOWELS						FINAL CONSONANTS					
	1st Pref.		2nd Pref.		Total		1st Pref.		2nd Pref.		Total		1st Pref.		2nd Pref.		Total	
EAR	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
SB	42	37	25	35	68	72	42	44	35	33	77	77	40	44	27	26	67	70
JH	40	42	23	25	63	67	45	47	40	36	85	83	36	49	23	23	59	72
MJ	27	46	18	16	45	62	34	48	29	24	63	72	41	45	19	14	60	59
NK	43	45	26	29	69	74	44	44	34	33	78	77	37	44	22	25	59	69
AL	33	21	18	52	51	73	47	10	10	54	57	64	53	14	13	62	66	76
BL	23	59	43	24	67	84	64	34	28	62	92	96	37	57	46	32	83	89
LN	21	65	45	17	65	82	40	59	58	40	98	99	32	62	42	27	74	89
HW	34	44	24	23	58	67	50	46	40	47	90	93	55	38	18	32	73	70
JW	30	50	23	18	53	68	32	37	24	23	56	60	37	54	28	20	65	74
BZ	25	57	42	26	67	83	—	—	—	—	—	—	41	52	43	36	84	88
SZ	33	52	38	28	71	81	58	42	40	57	98	99	37	55	43	30	80	85
JWn	28	53	34	21	62	74	45	54	52	42	97	96	39	52	41	31	81	83
Mean	32	48	30	26	62	74	46	42	35	41	81	83	40	47	31	30	71	77
$\bar{R}-\bar{L}$	16		-4		12		-4		6		2		7		-1		6	

Overall performance is higher on first preferences than on second for all three tests, and for both initial and final consonants, the total right-ear advantage is derived from first preferences (although some subjects--SB and AL on initials, HW and AL on finals--show their larger ear advantage on second preferences). That the right-ear advantage on consonants does not arise from a general tendency to report the right ear first, while the left-ear signal decays in storage, is shown by the fact that the ear advantage on first preferences for the vowels is to the left. Furthermore, the higher overall performance on first preferences is due almost entirely to the right ear on initial consonants, to the left ear on vowels.[3] The tendency to attach greater confidence to correct responses combined with the relatively large number of trials on which both responses were correct leads to nonsignificant reversals of the consonant ear advantages on second preferences.

An Index of the Laterality Effect. The laterality effect has been shown to be a function, under certain circumstances, of task difficulty (Satz et al., 1965; Bartz et al., 1967; Satz, 1968), and a ceiling is necessarily imposed upon it by very high or very low overall performance (Halwes, 1969). Since the vowels evidently set the listeners an easier task than the consonants, we sought a method of data analysis by which the two levels of difficulty might be equated. We found this in trials on which only one of the syllables was correctly identified. All such trials are presumably, in some sense, of equal difficulty, and overall performance on the subset is necessarily equal (50%) for consonants and vowels. No ear advantage can, in any event, be detected on trials for which the syllables are either both correct[4] or both incorrect, so that restriction of a laterality measure to the trials on which only one syllable was correctly identified (see Table I, last column) confines attention to the only occasions on which the effect has an opportunity to appear. Our null hypothesis for these one-correct trials is, then, that the single correct syllables are identified equally often by right and left ears. Deviation from

---

[3]Order of report effects have been shown to be present, but insufficient to account for the entire laterality effect, in many studies. For reviews, see Bryden (1967); Satz (1968); Halwes (1969).

[4]A measure of ear advantage might be derived from both-correct trials by use of preference scores, but these trials may not all be of equal difficulty.

this 50-50 distribution may be expressed as a percentage: (R-L/R+L) 100, where R (or L) is the number of trials on which the correctly identified syllable was delivered to the right (or left) ear. The index will range from 0 (50-50 distribution) to ±100 (0-100 distribution), with negative values indicating a left-ear advantage, positive values a right-ear advantage. Its significance may be tested on the null hypothesis that R/R+L = 0.50, using the normal curve as an approximation to the binomial.

Table III presents values of this index, based on one-correct-only trials for individual subjects on initial consonants, final consonants, and vowels. For initial consonants, the mean percentage laterality effect is 26. Each subject contributes between 150 and 208 trials. For nine subjects, the index is significant; for three subjects (SB, JH, NK), the index is positive but not significant.

For final consonants, the mean percentage laterality effect is 17. Each subject contributes between 89 and 237 trials. For seven subjects, the index is significant; for three subjects (SB, BZ, NWn), the index is positive but not significant; for two subjects (MJ, HW), the index is negative and not significant.

For the vowels, the mean percentage laterality effect is 10, but the reliability of this is low. Subjects vary widely in their indices and in their numbers of one-correct trials. Subject LN, for example, has an index of 50, based on only 8 trials, subject NK an index of -1 based on 143 trials, subject MJ an index of 17 based on 191 trials. For only two subjects (MJ, AL) is the index significant.

Laterality Effect for Individual Stop Consonants and Vowels. Up to this point, we have treated stop consonants and vowels as undifferentiated classes. But do all members of these classes show a laterality effect of the same degree? To answer this question, the group data were broken down by phonemes, and the laterality index was computed for each consonant and vowel. Figure 2 presents the results. The indices are arranged from left to right in order of decreasing magnitude. Consonants and vowels are perfectly segregated by this arrangement. /b/ and /g/ have the highest indices, and the voiced consonant at a given place value is always higher than its unvoiced counterpart. But the right-ear advantage is present for the whole class of initial stop consonants, and all indices are significant with $p < 0.0001$: lateralization is strong and consistent. For the vowels, on the other hand, lateralization is weak and inconsistent: all indices are positive, but only one (for /i/) is significant with $p < 0.01$ and one (for

TABLE III

Individual percentage ear advantages for initial stop consonants, final stop consonants, and medial vowels based on trials containing only one correct response.

	Initial Consonants				Medial Vowels				Final Consonants			
Subject	R-L[+]	R+L	$\frac{R-L}{R+L}$ 100	P	R-L	R+L	$\frac{R-L}{R+L}$ 100	P	R-L	R+L	$\frac{R-L}{R+L}$ 100	P
SB	15	171	9	NS*	2	134	1	NS	10	182	5	NS
JH	18	200	9	NS	-5	99	-5	NS	54	196	28	<.0001
MJ	62	208	30	<.0001	33	191	17	<.02	-3	237	-1	NS
NK	20	178	11	NS	-1	143	-1	NS	37	205	18	<.01
AL	94	204	46	<.0001	26	192	14	<.06	37	177	21	<.01
BL	62	150	41	<.0001	4	32	12	NS	21	89	24	<.05
LN	58	178	33	<.0001	4	8	50	NS	52	122	43	<.0001
HW	32	186	17	<.05	8	50	16	NS	-10	184	-5	NS
JW	55	207	27	<.0001	13	191	7	NS	32	188	17	<.05
BZ	55	153	36	<.0001	-	-	-	-	15	95	16	NS
SZ	36	156	23	<.01	1	9	0	NS	20	106	19	<.06
NWn	43	165	26	<.001	-1	25	-4	NS	12	108	11	NS
TOTAL	550	2156			84	1074			277	1889		
		Mean 26				Mean 10				Mean 17		

[+]R = Number of trials on which only the right ear stimulus was correctly identified.

L = Number of trials on which only the left ear stimulus was correctly identified.

*NS = Not significant at .10 level.

The Right-Ear Advantage for Individual Stop Consonants and Vowels

on Single-Error Trials

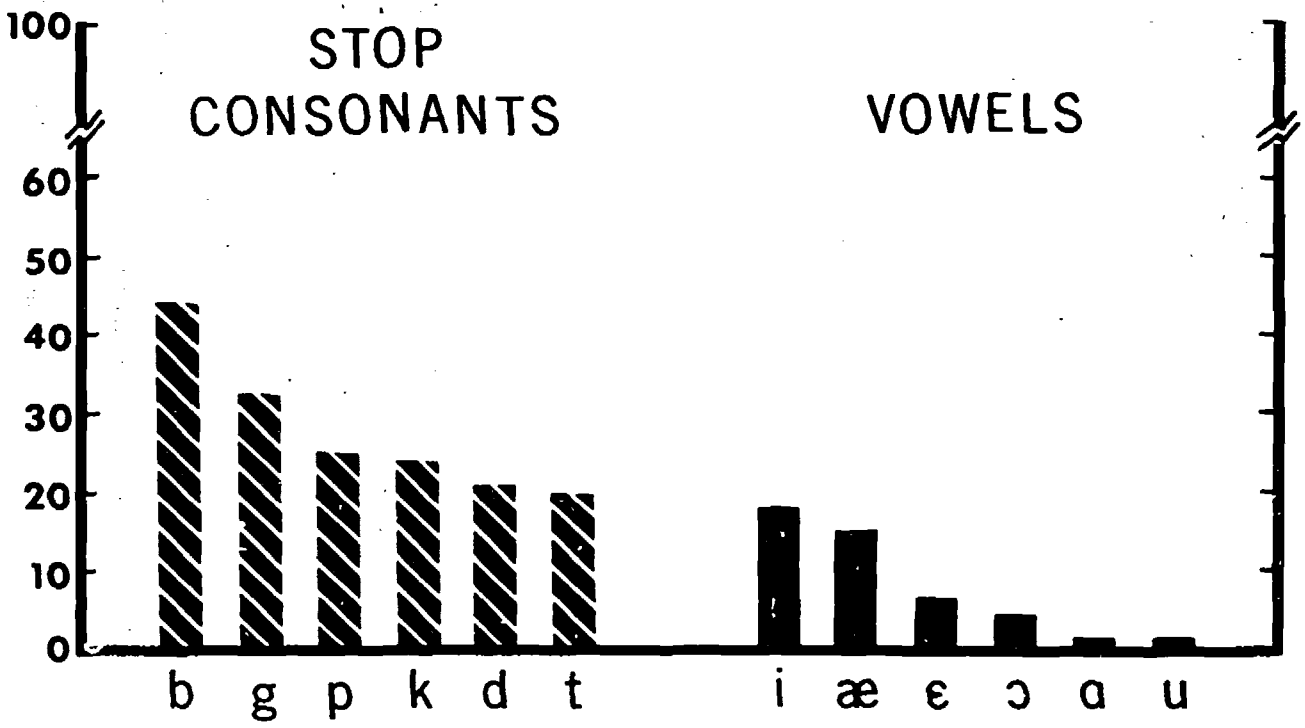


FIG. 2

(For explanation of the index plotted against the ordinate, see text.)

/æ /) with $p < 0.10$.

Laterality Effect and Item Difficulty. We eliminated task difficulty as a variable affecting the apparent lateralization of consonants and vowels by analyzing one-correct trials only. But it would still be possible for differences in the lateralization of individual phonemes on these trials to be linked to item difficulty. Consonants were therefore ranked according to difficulty, measured by total number of errors (order: /k,b,t,g,p,d/) and the value of their indices (order: /b,g,p,k,d,t/). Kendall's tau (Siegel, 1956) was computed and gave a nonsignificant value of 0.20. Vowels ranked according to their levels of difficulty (/æ,ɔ,a,u,ɛ,i/) and indices (/i,æ,ɛ,ɔ,a,u/) yielded a nonsignificant tau of -0.13. There is, therefore, no evidence here for a relation between the observed laterality effect and item difficulty.

The Identification of Consonant Feature Values. Having found that each of the six stop consonants is significantly lateralized, we may now ask whether the same is true of the articulatory features of which they are composed. Logically prior to this, however, is the question of whether these features are even perceived. Their psychological validity is, in fact, attested by the results of scaling the perceived distances among the stop consonants, /b,d,g,p,t,k/ (Greenberg and Jenkins, 1964) and analyses of errors in perception and short-term memory have suggested that the features are separately extracted and stored (Miller and Nicely, 1955; Singh, 1966, 1969; Wickelgren, 1966; Klatt, 1968). Experiments with dichotic listening offer a new approach to study of the perceptual process.

Each of the six stop consonants may be specified in terms of two articulatory features: voicing and place of production. In English, place of production has three values (labial, alveolar, velar), while voicing has only two (voiced, voiceless), so that we can specify each of the stops uniquely within a 2 X 3 matrix. The dichotic pairs may then contrast in voicing (/b,p/, /d,t/, /g,k/), in place (/b,d/, /b,g/, /d,g/, /p,t/, /p,k/, /t,k/), or in voicing and place (/b,t/, /b,k/, /d,p/, /d,k/, /g,p/, /g,t/). In each of these three blocks of trials, each consonant occurs equally often at each ear. If consonants are perceptually irreducible wholes and their component features no more than useful descriptive devices, we would expect performance to display only chance variation across blocks of trials for which articulatory features were the basis of classification. But, in fact, we find, as in our earlier experiment (Shankweiler and Studdert-Kennedy, 1967b), that performance does vary significantly.

Table IV shows that, when a feature value is common to both ears (that is, when the dichotic pair contrasts in only one feature), an error is less likely to be made and both responses are more likely to be correct than when no feature value is common (that is, when the dichotic pair provides a double contrast, a contrast in both voicing and place). Furthermore, performance varies according to which feature is shared: more advantage accrues from shared place than from shared voicing.[5] Or, in opposite terms, the feature more adversely affected by conditions of dichotic competition is place: even when voicing is shared, the contrast in place depresses performance. The outcome confirms the perceptual reality of the features: voicing and place values are indeed separately extracted.

TABLE IV

Percentage of different trial outcomes as a function of feature composition of dichotic pairs.

	Trial Outcomes (Percent)		
Feature Having a Value Shared by the Dichotic Pair	Both Correct	One Correct	Neither Correct
Place	61	37	2
Voice	43	52	5
Neither	33	55	12

The same conclusion is suggested by an analysis of errors. Even if a consonant is wrongly identified, one of its feature values may be correctly identified, and appropriate analysis will permit inferences about the perceptual process. The analysis is confined to trials on which a single error was made, since it is only for these that we can assign an error to its ear and stimulus.

[5]We note here a discrepancy between this result and a finding of our earlier study. There, performance was improved by the sharing of voicing (suggesting the greater difficulty of that feature); here, performance was improved by the sharing of place. Since the inference from Table IV of greater difficulty in the perception of place than voicing is borne out by every other relevant analysis in the present study [as also by the findings of Miller and Nicely (1955) and Singh (1966)], we have discounted the discrepancy in our subsequent discussions.

To ensure that no differential advantage accrues through a shared feature value, the analysis is also confined to trials on which each ear receives a different value of both voicing and place, that is, to double-contrast trials. For these trials, we may then determine the frequency with which each feature was correctly identified on erroneous responses, and we may compare this frequency with that expected by chance. To make the procedure clear, suppose that the stimulus pair is /b,t/ and that the subject correctly identifies /b/, so that we know his error is on /t/. His erroneous response may then be correct on voicing (/p/ or /k/), correct on place (/d/), or correct on neither feature (/g/). Correct guesses, if made on the perceptually unanalyzed phonemes without regard to their component features, would then be distributed in the proportions 2:1:1 for voicing, place, and neither feature correct. Table V shows that, in fact, voicing alone is correctly identified an overwhelmingly large proportion of times. Chi-square for this table equals 200.34, which, with 24 degrees of freedom, is highly significant ($p < 0.001$).

TABLE V

Number and percentage of features correct on single-error responses in double-contrast trials.

Feature Correct	Number	Percent
Voice Alone	678	72
Place Alone	184	19
Neither	83	9
Total	945	100

We may be confident, then, that the features are separately processed and that voicing values are more accurately identified than place. But some advantage may yet accrue to the identification of one feature from the correct identification of the other. In other words, the two perceptual processes may be, at least partially, dependent. The degree of their independence may be estimated by combining correct responses and errors into a single confusion matrix and carrying out an information analysis (Miller and Nicely, 1955; Attneave, 1959). The procedure has the additional advantage of providing a comparison between voicing and place identification in which the unequal guessing

probabilities for the two features may be discounted by expressing, for each feature, the information transmitted as a percentage of the maximum possible transmitted information.

Three confusion matrices were therefore constructed: a 2 X 2 voicing matrix in which stimuli and responses were grouped into labial, alveolar, and velar; a 6 X 6 matrix for the six individual consonants. Entries into these tables could use only those trials on which at least one phoneme was correctly perceived, since when neither phoneme is correct, the erroneous responses cannot be assigned to their appropriate stimuli. This has two consequences for the analysis. First, since all double errors are excluded, it leads to an over-estimate of the transmitted information for the experiment as a whole. But, since the purpose of the analysis is to compare the features and to estimate their degree of independence rather than to make a reliable estimate of information transmission, this need not concern us. A second consequence is that not all phonemes, or classes of phonemes, are equally represented in the trials to be analyzed so that the presented information (and hence the possible transmitted information) is reduced from the value that it would have if the sample were representative of the whole set of stimuli. However, the reduction in presented information proved to be only a few thousandths of a bit for each matrix, so that maximum possible transmitted information remained effectively 1 bit on voicing, 1.58 bits on place, and 2.58 bits on the individual consonants.

The actual information transmitted was computed for each matrix, and the results are displayed on the left side of Table VI. If the features of voicing and place were independently identified, the sum of the information transmitted for voicing and place separately would equal the information transmitted for the individual consonants in which the two features are combined (McGill, 1954; Miller and Nicely, 1955). Table VI shows that the required additivity holds to a close approximation. The independent perception of these features, demonstrated by previous investigators (Miller and Nicely, 1955; Singh, 1966), is again confirmed.

Table VI (right side) also expresses information transmitted as a percentage of maximum possible information transmitted on the two features, thus correcting for their unequal guessing probabilities. We again see the superiority of voicing over place identification: 12% more of the available voicing information is transmitted than of the available place information.

TABLE VI

Information in bits and percentage of maximum possible information transmitted, for each feature separately and for the features combined in individual consonants.

	Absolute Amount of Information Transmitted in Bits				Percentage of Maximum Possible Information Transmitted			
	Voice	Place	(V+P)	Combined	Voice	Place	(V+P) (2)	Combined
	.38	.41	(.79)	.86	33	26	(32)	33
Maximum Possible	1.00	1.58		2.58				

The general superiority of voicing over place identification, shown by the three data analyses described above, may not, of course, hold for all feature values. As a rough test for the homogeneity of the effect, we can compute the percentage correct on each feature value for all trials having at least one correct response (double-error trials again being excluded since responses on these trials cannot be assigned to their stimuli). Table VII shows the results of these computations. There is little difference between performances on the labial and velar place values: both are some 20% lower than performances on either of the two voicing values. The joker in the set is the alveolar performance of 82%, suggesting that perception of this place value is no more affected by dichotic stress than is perception of voicing. However, the result must be viewed with caution, since the data reveal a heavy bias toward alveolar responses: 42% of all place responses on these trials were alveolar, as compared with 29% each for labial and velar responses. A similar, though much smaller, bias appears in the data of Miller and Nicely (1955, Table VIII) for the set of six stop consonants.

TABLE VII

Percentage correct responses on each feature value for trials with at least one correct response.

Feature	Value	Percent Correct
Place	Labial	64
	Alveolar	82
	Velar	63
Voicing	Voiced	85
	Voiceless	83

The bias probably does not reflect listeners' expectations based on their experience with the language. Even though Denes (1963) estimates alveolar stop consonants to be roughly three times as frequent in English as either labial or velar stops, he also estimates voiceless stops to be very nearly twice as frequent as voiced, and no corresponding bias appears in our data (if anything, the reverse: 53% of listeners' responses on these trials were voiced, 47% voiceless). Furthermore, analysis of errors shows that most alveolar responses are made on trials in which at least one of the stimuli carries the alveolar place value. The "bias," therefore, arises when one member of a dichotic pair is alveolar, the other not; the alveolar value, then, "dominates" the contrasting labial or velar value. In other words, our first inference seems to be correct: the "bias" has a perceptual basis, and the alveolar stops in this experiment were less susceptible to dichotic stress than labial or velar stops.

Lateralization of Feature Perception. We may now ask whether the independence of the two features, and the advantage of voicing over place shown in the combined data, holds equally for the two ears. To answer these questions, the data were reanalyzed separately for each ear. We begin with a reanalysis of Table IV. The results are now given in terms of percentage of correct responses for each ear, rather than in terms of trial outcomes, since no difference between the ears can appear on trials for which the responses were either both correct or both incorrect. Table VIII shows the outcome of the reanalysis. For both ears, the ranking is exactly as in Table IV: performance is highest when place is shared, second highest when voicing is shared, lowest when neither feature is shared.

TABLE VIII

Percentage correct responses for the two ears as a function of feature composition of dichotic pairs.

Feature Having a Value Shared by the Dichotic Pair	Percent Correct	
	Left Ear	Right Ear
Place	74	86
Voice	63	75
Neither	54	67

We may notice, furthermore, that the right ear has approximately the same advantage over the left ear (about 12%) for each type of dichotic pair. This suggests that the right-ear advantage is the same for both voicing and place--that one feature is not more heavily lateralized than the other. The same conclusion is suggested by an error analysis along the lines of Table V. Again we make use only of double-contrast trials, and to avoid any bias due to possible interaction between the features (despite their evident independence), we compute for each ear conditional percentages. That is, we compute the percentage correct on voicing, given that place was missed, and the percentage correct on place, given that voicing was missed. Table IX gives the results of these computations: the right-ear advantage is 7% on voicing, 6% on place.

Table IX

Conditional percentages of feature errors for the two ears on single-error responses in double-contrast trials.

Feature in Error	Other Feature	Percent	
		Left	Right
Place	Voicing Correct	86	93
Place	Voicing Incorrect	14	7
Voicing	Place Correct	67	73
Voicing	Place Incorrect	33	27

However, equal lateralization of the two features is not evident in every analysis. Table X shows the breakdown of Table VII by ear. The expected right-ear advantage appears for every value of both features but is somewhat greater for labial and velar place values than for voicing, suggesting stronger lateralization of these place values. [Both ears, incidentally, show a gain in alveolar performance: for the left ear the gain is approximately 20%, as against 13-16% for the right ear, perhaps reflecting a somewhat stronger alveolar preference on the left ear (44% of all left-ear responses, as against 39% of all right-ear responses, were alveolar).]

TABLE X

Percentage correct responses on each feature value for each ear on trials with at least one correct response.

Feature	Value	Percent Correct	
		Left	Right
Place	Labial	59	71
	Alveolar	79	84
	Velar	58	68
Voicing	Voiced	82	89
	Voiceless	80	87

Finally, Table XI displays the results of the information analysis. Both ears transmit a greater percentage of their voicing than of their place information. And for both ears the expected additivity, or independence, of feature information holds quite closely. However, the right-ear advantage is here greater on voicing (18%) than on place (10%). The difference cannot be tested for significance, but the disagreements between Tables VIII and IX (features equivalent in lateralization), Table X (right-ear advantage greater on two place values) and Table XI (right-ear advantage greater on voicing) are obvious.

TABLE XI

Information in bits and percentage of maximum possible information transmitted for each feature, separately and for the features combined in individual consonants, for right and left ears.

	Absolute Amount of Information Transmitted in Bits				Percentage of Maximum Possible Information Transmitted			
	Voice	Place	(V+P)	Combined	Voice	Place	(V+P) (2)	Combined
Right Ear	.49	.50	(.99)	1.06	49	32	40	41
Left Ear	.31	.35	(.66)	0.70	31	22	26	27
Maximum Possible	1.00	1.58	2.58					

There is also disagreement between one particular analysis in this and in our earlier study. In that study, we found differing degrees of laterality effect according to which features were shared (or contrasted) between the ears in a dichotic pair. We took this to indicate some difference in the

degrees of lateralization of the two features. But in the corresponding analysis of the present study (Table VIII), we found no differences in laterality effect.

We therefore conclude that, while both features are clearly and independently lateralized, reliable estimates of their relative degrees of lateralization have eluded us.

Discussion

The results are in general agreement with those of our previous study and of several other investigators (Curry, 1967; Curry and Rutherford, 1967; Kimura, 1967; Darwin, 1969a,b; Haggard, 1969; Halwes, 1969) in demonstrating a laterality effect for the perception of dichotic signals that differ only in their phonetic structure. They show further that the laterality effect extends to the perception of subphonemic features. Before discussing some of the problems that the results present, we will briefly consider a possible mechanism of speech lateralization.

A Mechanism for the Laterality Effect in Speech Perception. As Kimura (1961b, 1964) first suggested, the laterality effect may be accounted for by the assumptions of cerebral dominance and functional prepotency of the contralateral over the ipsilateral auditory pathways. Contralateral prepotency rests upon the greater number of these neurons and upon inhibition of ipsilateral neurons during dichotic stimulation. Strong corroboration of Kimura's argument has come from the work of Milner, Taylor, and Sperry (1968). (See also Sparks and Geschwind, 1968). They studied right-handed patients (presumably left-brained for language) for whom the main commissures linking the cerebral hemispheres had been sectioned to relieve epilepsy. Under dichotic stimulation, these subjects were able to report verbal stimuli presented to the right ear but not those presented to the left; under monaural stimulation, they performed equally well with the two ears. Milner et al. attribute their results to suppression of the ipsilateral pathway from left ear to left (language) hemisphere during dichotic stimulation and, of course, to sectioning of the callosal pathway that should have carried the left ear input from right hemisphere to left. Their data justify the inference that when, under dichotic stimulation, normal, left-brained subjects correctly perceive a left-ear verbal input, the signal has been suppressed ipsilaterally, has traveled the contralateral path to the right hemisphere, and has been transferred across the lateral

commissures to the left hemisphere for processing. Inputs to both ears, therefore, converge on the dominant hemisphere, that from the right ear by the direct contralateral path, and that from the left ear by an indirect path, crossing first to the right hemisphere, then laterally to the left. The right-ear advantage in dichotic studies of speech must then arise because the left-ear input, traveling an indirect path to the left cerebral hemisphere, suffers, on certain trials, a disadvantage or "loss" to which the right-ear input, traveling a direct path, is less susceptible.

The locus of this loss can be broadly specified. We first assume that the two contralateral pathways are equivalent, so that the two signals reach their respective hemispheres in equivalent states; there is, of course, ample opportunity for the signals to interact at subcortical levels, but presumably whatever loss such interaction may induce is induced equally on both signals. If we further assume that the two signals upon arrival in the dominant hemisphere are served by the same set of processors (as evidence, discussed below, suggests), loss in the left-ear signal must occur immediately before, during, or after transfer to the dominant hemisphere.

The nature and source of the left-ear loss are matters of great interest to which we return briefly in a later section of the discussion. Here, we merely remark that a preliminary attack on the problem might be made through careful comparison of error patterns for right- and left-ear inputs. As we have seen in the limited data of the present study, the general pattern of errors is rather similar for the two ears. This suggests that the left-ear input is subject to stress that differs in degree, but not in kind, from that exerted on the right-ear input. The notion of a generalized auditory stress common to both ears, whatever its source, is encouraged by the fact that the error pattern in this experiment is remarkably similar to that found in other studies. The superiority of voicing identification over place, for example, was observed by Miller and Nicely (1955) and by Singh (1966) in studies of speech perception through masking noise.

The Nature of Cerebral Dominance in Speech Perception. To speak of cerebral dominance in speech perception is to imply that at least some portion of the perceptual function is performed more efficiently, or even exclusively, by the dominant hemisphere. The problem is to define that portion. That dichotic inputs must, at some point in their time course, converge on a final, common path is evident from the fact that the two inputs ultimately activate a single articulatory response mechanism. But how early the inputs converge is

the matter of interest. We would like to know, for example, whether convergence occurs before any linguistic analysis of the signal whatever (as would be true if both ears were served by a single set of specialized speech processors in the speech-dominant hemisphere), after partial linguistic analysis (as would be true if, for example, features were separately extracted in the two hemispheres but were recombined in the dominant hemisphere), or after complete linguistic analysis and immediately before response (as would be true if the two hemispheres were equivalent in their capacities to analyze the signal but were served by a single set of specialized output mechanisms in the speech-dominant hemisphere). More generally, is the signal from the nondominant hemisphere transferred to the dominant hemisphere in a linguistic or in an auditory code? Some leverage on this question may be gained from a further analysis of errors in the present study.

Independent processing of subphonemic features requires that, at some point between input and output, a syllable be broken into its component features and that, at some later point, these features be recombined into a unitary response. If convergence of the two inputs occurs before features are recombined, a feature value has an opportunity to lose its local sign, that is, to lose information about its ear of origin. A correctly perceived feature from one ear might then be incorrectly combined with a correctly perceived feature from the opposite ear. The resulting response would be a "blend" of features from opposite ears. However, if convergence of the two inputs occurs <u>after</u> features are recombined, local sign could only be lost for the entire syllable, not for its component features. Blend responses would then occur only by chance. Evidence for greater than chance occurrence of blends is, therefore, evidence for loss of local sign on features and, by inference, for convergence of the inputs before the features are recombined.

Blends cannot be detected on single-contrast trials: even if the error occurs in combining the features, any resulting response will be correct, since one of the crossed feature values is presented to both ears. But on double-contrast trials, blending errors may be detected. For example, if the stimulus pair is /b,t/, the erroneous responses /p/ or /d/ are blends (drawing place values from one ear, voicing values from the other), while the erroneous responses /g/ and /k/ are not blends. Both classes of error would occur equally often if there were no tendency for errors of local sign to occur on the features and if subjects were distributing their errors at random. In fact, blending errors occur with high frequency. Table XII shows that, of

410 errors on double-error, double-contrast trials, 263 (64%) were blends; of 945 errors on single-error, double-contrast trials, 673 (71%) were blends. The overall percentage of blends (69%) is far in excess of chance expectation (50%). For each row of the table, $p<0.0001$ on a test of the chance hypothesis by the normal approximation to the binomial.

TABLE XII

Number and percentage of errors on double-contrast trials that arose by blending or not blending features from opposite ears. Trials affording two errors and trials affording one error are distinguished.

Trial Outcome	Number of "Blend" Errors	Number of "Nonblend" Errors	Total Number of Errors	Percent "Blend" Errors
Double Error	263	147	410	64
Single Error	673	272	945	71
Total	936	419	1355	69

Errors of local sign on the features do then occur in these data, as in those of Kirstein and Shankweiler (1969), with very high frequency. The result is additional evidence for the independent processing of the features. More importantly, it suggests that inputs to left and right ears converge on a common center at some stage before combination of the features into a final unitary response.

We may now ask whether convergence occurs immediately before feature combination or at some later stage. In other words, is the signal that is transferred from right hemisphere to left coded into separate linguistic features, or is it in some form of nonlinguistic auditory code? If the first were true, features of the left-ear syllable and features of the right-ear syllable would be extracted in separate hemispheres, and the feature composition of one syllable should have no effect on the probability of correctly identifying the other. If the second were true, interaction could occur between auditory parameters of the two inputs during the process of feature extraction, and this interaction should be reflected in performance. In fact, we already know from Tables IV and VIII that a response is more likely to be correct if the two inputs have a feature value in common. Furthermore, the advantage of sharing a feature value accrues more frequently if place is shared than if voicing is shared.

We conclude that the inputs converge before rather than after features extraction and that duplication of the auditory information that conveys the shared feature value gives rise to the observed advantage. In other words, we take the systematic relation between performance and the feature composition of dichotic pairs to be evidence consistent with the hypothesis of interaction during, or immediately before, the actual process of feature extraction.

Also consistent with this interpretation are the similar error patterns for left and right ears that we have already reported. As a further example, Table XIII shows the breakdown of Table XII by ear. (Only single-error trials are considered, since double errors cannot be assigned to their ears. An example of a single-error "blend" would be the response /d/ in the response pair /b,d/, given to stimulus pair /b,t/.) While the percentage of "blend" errors is greater for the right ear (75%) than for the left (69%), the difference is not significant at the 0.05 level, and both ears show a heavy preponderance of "blend" over "nonblend" errors.

TABLE XIII

Number and percentage of errors on double-contrast trials that arose by blending or not blending features from opposite ears, for right and left ears. Single-error trials only.

Ear	Number of "Blend" Errors	Number of "Nonblend" Errors	Total Number of Errors	Percent "Blend" Errors
Right	268	91	359	75
Left	405	181	586	69
Total	673	272	945	71

We therefore tentatively conclude that convergence of the two signals in the dominant hemisphere occurs before the extraction of linguistic features and that it is for this process of feature extraction that the dominant hemisphere is specialized. On this hypothesis, we would assign to the dominant hemisphere that portion of the perceptual process which is truly linguistic: the separation and sorting of a complex of auditory parameters into phonological features. Such a specialized "decoding" operation has been shown, on quite other grounds, to be entailed in speech perception (Liberman et al., 1967).

The Role of the General Auditory System in Speech Perception. The foregoing

argument has suggested that the role of the dominant hemisphere is due to its possession of a special linguistic device rather than to superior capacities for auditory analysis. We should, therefore, emphasize the distinction between extraction of the auditory parameters of speech and linguistic "interpretation" of those parameters. It is for the latter that specialized processing is required and for which the dominant hemisphere seems to be equipped, while the former is the domain of the general auditory system common to both hemispheres. In other words, the peculiarity of speech may lie not so much in its acoustic structure as in the phonological information that this structure conveys. There is, therefore, no a priori reason to expect that specialization of the speech perceptual process should extend to the mechanisms by which the acoustic parameters of speech are extracted.

Consider, for example, an acoustic variable underlying the identification of place in stop consonants: the extent and direction of the second formant transition (Liberman et al., 1954). Data bearing on the perception of such frequency transitions in nonspeech have been reported for resonant frequencies (Brady et al., 1961) and, more recently, for tone-bursts (Pollack, 1968; Nabelek and Hirsh, 1969). Nabelek and Hirsh determined the optimal glide durations for the discrimination of frequency change to be, in general, between 20 and 30 msecs. They remark that these values are "close to the durations that were found by Liberman et al. (1956) to be important for the discrimination of speech sounds" (p. 1518). They conclude that this optimum transition duration "is a general property of hearing and...does not only appear in connection with speech sounds" (p. 1518).

Their conclusion does not, of course, imply that there may be no functional differences between the hemispheres in auditory perception. There is, in fact, much evidence that for nonspeech the right, nondominant hemisphere plays a greater role than the left in recognition of auditory patterns and in discrimination of their attributes (Milner, 1962; Kimura, 1964; Benton, 1965; Chaney and Webster, 1965; Shankweiler, 1966 a, b; Curry, 1967; Vignolo, 1969). Whatever the peculiar auditory capabilities of the right hemisphere may be, there is reason to believe that each hemisphere can perform an auditory pattern analysis of the speech signal without the aid of the other. The isolated left hemisphere can, in fact, go further and complete the perceptual process by interpretation of these auditory patterns as sets of linguistic features (as the data of Milner et al., cited above, show).

Whether the right hemisphere can go so far is open to question. Sperry

and Gazzaniga (1967) (see also Smith and Burkland, 1966; Gazzaniga and Sperry, 1967; Sparks and Geschwind, 1968) found that commissurectomized patients, instructed orally to select an object from a concealed tray with the left hand, were able to do so. Since left-hand stereognostic discrimination was known, from other of their tests, to be controlled only by the right hemisphere, it was evident that this hemisphere, in some sense, "perceived" the speech. However, the hemisphere was unaware of what it had "heard": the patients were unable to name the object they had selected and were holding. Similar results have been reported by Milner et al. (1968) for commissurectomized patients to whom instructions had been presented dichotically, thus presumably confining left-hand instructions to the right hemisphere. These authors conclude that "the minor, right hemisphere does show some rudimentary verbal comprehension" (p. 184).

Interpretation of such results is not easy, particularly since these patients had pre-existing epileptogenic lesions in addition to surgical disconnection of the hemispheres. However, it seems possible that the right hemisphere's "rudimentary comprehension" may have rested on auditory analysis which, by repeated association with the outcome of subsequent linguistic processing, had come to control simple discriminative responses. Certainly, a capacity for the auditory analysis of speech would seem to be the least we can attribute to the right hemisphere.

We therefore conclude that the auditory system common to both hemispheres is probably equipped to track formants, register temporal intervals, and in general, extract the auditory parameters of speech. But to the dominant hemisphere may be largely reserved the tasks of linguistic interpretation: for example, selecting from a formant transition the relevant overlapping cues to consonantal place of articulation and to neighboring vowel or selecting from the infinity of temporal intervals automatically registered in the auditory stream the one interval relevant to the perception of voicing (Lisker and Abramson, [illegible]; Abramson and Lisker, 1965). Completion of such tasks is presumably [illegible] equisite to conscious perception of speech.

The interpretation of the laterality effect outlined in preceding [illegible]ions has implications for future work that may best be drawn by first discussing the results for consonants and vowels in the present study.

Consonant Feature Lateralization. Underlying lateralization of cons[illegible]re the independent lateralizations of their component features. Since [illegible] of

consonantal errors are due to the loss of a single feature (see Tables V and IX), any reduction in the laterality effect of one feature would lead to a reduction in the laterality effect of the consonants as a whole. An example of such an effect may have been provided by the final consonants of this study.

The right-ear advantage for the final consonants, though significant, is relatively small. The result is at variance with that of Darwin (1969a,b), who found a strong right-ear advantage for final consonants in dichotically presented synthetic VC syllables.[6] If we accept the difference as genuine and not due to some artefact, such as poor synchronization of the final consonants, in this study, an interesting explanation might be that our reduced effect arose from reduced place lateralization and that place lateralization only occurs for cues carried by a formant transition. A formant transition was the sole source of cues in the unreleased synthetic stops used by Darwin but not in the released "natural" speech stops of the present study, where final bursts may sometimes have provided enough information for clear place identification.

The implication, in light of our previous argument, is that a final burst, standing in relative isolation from the rest of the syllable, may be estimated as well by the minor as by the major hemisphere and that information about its parameters (intensity, duration, frequency band) is liable to relatively little loss during transfer to the dominant hemisphere for feature extraction. A formant transition, on the other hand, in which cues for both vowel and consonant are delicately implicated, even if correctly estimated auditorily by the minor hemisphere, may be subject to degradation during transfer to the dominant hemisphere. The presence of a formant transition was found by Darwin (1969a,b) in an experiment with synthetic (initial) fricatives (/f,s,ʃ,v,z,ʒ/ followed by /ɛ p/) to be a necessary condition of right-ear advantage: fricatives synthesized from friction alone, without transition, were clearly identifiable but gave no right-ear advantage. The likely importance of formant transitions in the laterality effect may also bear on the results for the vowels to which we now turn.

Vocal Lateralization. A main purpose of the present study was to determine whether natural vowels embedded in a consonantal frame would show a greater right-ear advantage than the synthetic, isolated, steady-state vowels of our

---

[6]Trost et al. (1968) report equal right-ear advantages for initial and final consonants in "natural" CVC syllables. But since their test lists included fricatives, liquids, voiced, voiceless, and nasal stops, not all of which occurred equally often in initial and final position, their results are difficult to compare with those of this study.

previous study. They did not. Nonetheless, some tendency toward a right-ear advantage for the vowels is evident. In both studies, the mean advantage, though not significant, was to the right (4%, 2%). Of the twenty-one subjects in the two studies, thirteen gave right-ear advantages (two significant), seven gave left-ear advantages (none significant), one no ear advantage. For the six vowels in the present study, all ear advantages were to the right (one significant). In short, the vowels display a weak, variable right-ear advantage and by this are distinguished from consonants, for which a stronger right-ear advantage is the rule, and also from musical or other nonspeech sounds, for which a left-ear advantage is the rule (Kimura, 1964; Shankweiler, 1966; Chaney and Webster, 1965; Curry, 1967).

The vowels studied up until now seem to occupy a position on the margin of speech. But we should note that the vowels of this experiment, though embedded in CVC syllables, were still of relatively long duration, each syllable lasting between 300 and 500 msecs. Presumably, were they synthetic, we could push them (or isolated, steady-state vowels) toward nonspeech and a left-ear advantage by systematic manipulation of their spectral composition, musicalizing them, perhaps, by reducing the bandwidths of their formants, and increasing their duration. But under what conditions might the tentative right-ear advantage be magnified into a full right-ear advantage comparable with that of the consonants?

If the vowels are isolated and steady-state, merely reducing their duration from 150 msecs. to 40 msecs. has no effect: neither the longer nor the shorter vowels show a significant ear advantage (Darwin, 1969a, b), and reduction of duration much below 40 msecs. is not possible without loss of vowel quality and approach to a nonspeech click. But for vowels placed in CVC syllables, the story may be different. We know that the identification of synthetic CVC vowels may be affected by the rate of articulation (Lindblom and Studdert-Kennedy, 1967). Such vowels may be said to be "encoded" (Liberman et al., 1967) in the sense that cues for their identification are provided simultaneously (in parallel) with cues for the identification of their neighboring consonants. Identification of both vowels and consonants entails a judgment, in some form, of the formant transitions. From the dichotic work of Haggard (1969) we know that synthetic semivowels and laterals (/w,r,l,j/), for which important cues are carried by relatively slow formant transitions, may give a right-ear advantage of the same order as that given by stop consonants. And finally, we have the evidence of Darwin (1969a, b), cited above, on the possible importance of formant transitions in the laterality effect for fricatives.

We may then reasonably hypothesize that reduced, rapidly articulated, "encoded" vowels in CVC syllables, dependent for their recognition on the perception of formant transitions, would show a significant right-ear advantage. Experiments to test this hypothesis are now being planned.

Cerebral Dominance and Information Loss in the Laterality Effect. In the foregoing discussion, we have suggested that differences in right-ear advantage among stops and vowels may be due to differences in the susceptibility of these signal classes to information loss during transmission. In earlier discussions (for example, Shankweiler and Studdert-Kennedy, 1967a; Shankweiler, in press), we have taken such differences in ear advantage to reflect differences in the degree to which consonants and vowels engage the specialized perceptual mechanisms of the dominant hemisphere. We should now make explicit the reasons for this shift in interpretation and, at the same time, summarize our current understanding of the laterality effect.

There are two necessary conditions of an ear advantage in dichotic listening. First, some part of the perceptual process must depend upon unilateral neural machinery; second, the signal from the ipsilateral ear must undergo a significant loss due either to degradation of the signal during transmission to the dominant hemisphere or to its decay during the time it is held before final processing. Wherever a reliable contralateral ear advantage is observed, both these conditions must have been fulfilled. However, Darwin (1969a, b) and Halwes (1969) have independently pointed out that where an ear advantage is not observed, or is small, the outcome is ambiguous: it may indicate either no unilateral processing or no significant information loss in the ipsilateral signal. In other words, the absence of an ear advantage is not inconsistent with complete lateralization of some portion of the perceptual function, since the outcome may simply indicate that the acoustic materials being studied are not susceptible to information loss under certain experimental conditions.

This is the interpretation that the reduced effect for final consonants seems to demand, since, in the interests of parsimony, we must suppose that final consonants require the operation of specialized feature extractors in the dominant hemisphere no less than initials. For the vowels, the situation is not so clear. The "continuous" nature of vowel perception (for a recent discussion, see Studdert-Kennedy et al., 1970) may perhaps be related to vowels not engaging discrete feature extractors in the dominant hemisphere. At the same time, transfer of vowel information to the dominant hemisphere for final perceptual response is unavoidable, and the most parsimonious interpretation

again seems to be that the reduced or null laterality effect for vowels is also due to reduced information loss rather than to absence of cerebral dominance.

We may, finally, distinguish two broad directions that future research with dichotic materials might take. First, there is research of general auditory interest. Much remains to be learned about the experimental and acoustic conditions of ipsilateral transmission loss. Appropriate research may increase our understanding of those features in the design of the auditory system that make it possible to demonstrate laterality effects. Second, there is research directed primarily to the understanding of speech perception. Wherever a laterality effect for speech materials clearly occurs, we may exploit the effect to infer underlying perceptual processes. Here, we should emphasize a point that may easily be missed: the size of the laterality effect is not a measure of its importance or of its value for research. We are not concerned in dichotic experiments to estimate the contribution of a variable to control over perception. We are, rather, exploiting the apparently trivial errors of a system under stress to uncover its functional processes.

Conclusions

This study of dichotically presented, "natural" speech CVC syllables showed: (1) a significant right-ear advantage for initial stop consonants; (2) a significant, though reduced, right-ear advantage for final stop consonants; (3) a nonsignificant right-ear advantage for six medial vowels; (4) significant and independent right-ear advantages for the articulatory features of voicing and place in initial stop consonants.

We have argued, following Kimura (1961b), that the right-ear advantages are to be attributed to left cerebral dominance and functional prepotency of the contralateral pathways during dichotic stimulation. From analysis of the errors made in perception of the initial stop consonants, we have tentatively concluded that, while the general auditory system may be equipped to extract the auditory parameters of a speech signal, the dominant hemisphere is specialized for the extraction of linguistic features from those parameters. The laterality effect would then be due to a loss of auditory information arising from interhemispheric transfer of the ipsilateral signal to the dominant hemisphere for linguistic processing.

References

Abramson, A.S., and Lisker, L. (1965) Voice onset-time in stop consonants: acoustical analysis and synthesis. 5e Congres International d'Acoustique: Rapports (Liege) Ia, A5¹.

Attneave, I. (1959) Applications of Information Theory to Psychology. (Holt, Rhinehart and Winston, New York).

Bartz, W.H., Satz, P., and Fennel, E. (1967) Grouping strategies in dichotic listening: the effects of instructions, rate and ear asymmetry. J. Exp. Psychol. 74, 132-136.

Benton, A.L. (1965) The problem of cerebral dominance. Canad. Psychologist 6a, 332-348.

Bocca, E., Calearo, C., Cassinari, V., and Migliavacca, F. (1955) Testing "cortical" hearing in temporal lobe tumors. Acta Oto-Laryngol. 45, 289-304.

Brady, P.T., House, A.S., and Stevens, K.N. (1961) Perception of sounds characterized by a rapidly changing resonant frequency. J. Acoust. Soc. Amer. 33, 1357-1362.

Branch, C., Milner, B., and Rasmussen, T. (1964) Intracarotid sodium amytal for the lateralization of cerebral speech dominance. J. Neurosurg. 21, 399-405.

Broadbent, D.E. (1954) The roll of auditory localization in attention and memory span. J. Exp. Psychol. 47, 191-196.

Bryden, M.P. (1967) An evaluation of some models of dichotic listening. Acta Oto-Laryngol. 63, 595-604.

Chaney, R.B., and Webster, J.C. Information in certain multidimensional signals. (U.S. Navy Electronics Laboratory Reports, San Diego, Calif.) No. 1339.

Curry, F.K.W. (1967) A comparison of left-handed and right-handed subjects on verbal and non-verbal dichotic listening tasks. Cortex 3, 343-352.

Curry, F.K.W., and Rutherford, D.R. (1967) Recognition and recall of dichotically presented verbal stimuli by right- and left-handed persons. Neuropsychologia 5, 119-126.

Darwin, C.J. (1969a) Auditory Perception and Cerebral Dominance, (Unpublished Ph.D. thesis, University of Cambridge).

Darwin, C.J. (1969b) Laterality effects in the recall of steady-state and transient speech sounds. J. Acoust. Soc. Amer. 46, 114 (Abstract).

Denes, P.B. (1963) On the statistics of spoken English. J. Acoust. Soc. Amer. 35, 892-904.

Gazzaniga, M.S., and Sperry, R.W. (1967) Language after section of the cerebral commissures. Brain 90, 131-148.

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

Haggard, M.P. (1969) Perception of semi-vowels and laterals. J. Acoust. Soc. Amer. 46, 115 (Abstract).

Hall, J.L., and Goldstein, M.H. (1968) Representation of binaural stimuli by single units in primary auditory cortex of unanaesthetized cats. J. Acoust. Soc. Amer. 43, 456-461.

Halwes, T. (1969) Effects of dichotic fusion in the perception of speech. (Unpublished Ph.D. thesis, University of Minnesota).

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

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

Kimura, D. (1964) Left-right differences in the perception of melodies. Quart. J. Exptl. Psychol. 16, 355-358.

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

Kimura, D., and Folb, S. (1968) Neural processing of backwards-speech sounds. Science 161, 395-396.

Kirstein, E., and Shankweiler, D. (1969) Selective listening for dichotically presented consonants and vowels. Paper read before the 40th Annual Meeting of the Eastern Psychological Association, Philadelphia.

Klatt, D.H. (1968) Structure of confusions in short-term memory between English consonants. J. Acoust. Soc. Amer. 44, 401-407.

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

Liberman, A.M., Delattre, P.C., Cooper, F.S., and Gerstman, L.J. (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monographs 68, No. 379.

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

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

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

Lieberman, P. (1967) Intonation, Perception and Language, (M.I.T. Press, Cambridge, Mass.).

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

Lieberman, P., Klatt, D.L., and Wilson, W.A. (1969) Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. Science 164, 1185-1187.

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

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

Mattingly, I., and Liberman, A.M. (1970) The speech code and the physiology of language. In Information Processing in the Nervous System, K.N. Leibovic, Ed. (Springer Verlag, New York) 97-117.

McGill, W.J. (1954) Multivariate information transmission. Psychometrika 19, 97-116.

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

Milner, B., Taylor, L., and Sperry, R.W. (1968) Lateralized suppression of dichotically-presented digits after commissural section in man. Science 161, 184-185.

Nabelek, I., and Hirsh, I.J. (1969). On the discrimination of frequency transitions, J. Acoust. Soc. Amer. 45, 1510-1519.

Pollack, I. (1968) Detection of rate of change of auditory frequency. J. Exptl. Psychol. 77, 535-541.

Rosenzweig, M.R. (1951) Representations of the two ears at the auditory cortex. Amer. J. Physiol. 167, 147-158.

Satz, P. (1968) Laterality effects in dichotic listening. Nature 218, 277-278,

Satz, P., Achenback, K., Pattishall, E., and Fennell, E. (1965) Order of report, ear asymmetry and handedness in dichotic listening. Cortex 1, 377-396.

Satz, P., Fennell, E., and Jones, M.B. (1969) Comments on: a model of the inheritance of handedness and cerebral dominance. Neuropsychologia 7, 101-103.

Shankweiler, D. (1966a) Defects in recognition and reproduction of familiar tunes after unilateral temporal lobectomy. Paper read before the 37th Annual Meeting of the Eastern Psychological Association, New York.

Shankweiler, D. (1966b) Effects of temporal-lobe damage on perception of dichotically presented melodies. J. Comp. Physiol. Psychol. 62, 115-119.

Shankweiler, D. (In press) An analysis of laterality effects in speech perception. In Perception of Language, P. Kjeldergaard, Ed. (Chas. E. Merrill, Columbus, Ohio).

Shankweiler, D., and Studdert-Kennedy, M. (1966). Lateral differences in perception of dichotically presented synthetic consonant-vowel syllables and steady-state vowels. J. Acoust. Soc. Amer. 39, 1256 (Abstract).

Shankweiler, D., and Studdert-Kennedy, M. (1967a) An analysis of perceptual confusions in identification of dichotically presented CVC syllables. J. Acoust. Soc. Amer. 41, 1581 (Abstract).

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

Siegel, S. (1956) Non-parametric Statistics. (McGraw Hill, New York).

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

Singh, S. (1969) Interrelationship of English consonants. In The Proc. 6th International Congress of Phonetic Sciences, (Prague), 542-544.

Smith, A., and Burkland, C.W. (1966) Dominant hemispherectomy: preliminary report on neuropsychological sequelae. Science 153, 1280-1282.

Sparks, R., and Geschwind, N. (1968) Dichotic listening in man after section of neocortical commissures. Cortex 4, 3-16.

Sperry, R.W., and Gazzaniga, M.S. (1967) Language following surgical disconnection of the hemispheres. In Brain Mechanisms Underlying Speech and Language, C.H. Millikan and F.L. Darley, Eds. (Grune and Stratton, New York), 108-121.

Spreen, O., Benton, A.L., and Fincham, R.W. (1965) Auditory agnosia without aphasia. Arch. Neurol. 13, 84-92.

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

Trost, J.E., Shewan, C.M., Nathanson, S.N., and Sant, L.V. (1968) A dichotic study of ear superiority in perception of consonants. Paper read before the 44th Annual Convention of the American Speech and Hearing Association, Denver.

Tunturi, A.R. (1946) A study on the pathway from the medial geniculate body to the acoustic cortex in the dog. Amer. J. Physiol. 147, 311-319.

Vignolo, L.A. (1969) Auditory agnosia: A review and report of recent evidence. In Contributions to Clinical Neuropsychology, A.L. Benton, Ed. (Aldine, Chicago) 172-208.

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

Ear Differences in the Recall of Fricatives and Vowels*

C.J. Darwin+
Haskins Laboratories, New Haven

Summary. Two experiments on the free recall of dichotically presented synthetic speech sounds are reported. The first shows that the right ear advantage for initial fricative consonants is not simply a function of the recognition response class but that it is also a function of the particular acoustic cues used to achieve that response. This is true both for the whole response and for the constituent phonetic features. The second experiment shows that when both the response class and the particular stimuli presented on certain trials are held constant, the right ear advantage for the constant stimuli can be influenced by the range of other stimuli occurring in the experiment. Vowels show a right ear advantage when, within the experiment, there is uncertainty as to vocal tract size, but they show no ear advantage when all the vowels in the experiment are from the same vocal tract. These results are interpreted as demonstrating that there are differences between the ears, and probably between the hemispheres, at some stage between the acoustic analysis of the signal and its identification as a phonetic category.

---

*To be published in Quart. J. Exptl. Psychol. (1971).

+Also, University of Connecticut.

Acknowledgements. The first experiment reported here was performed at the Psychological Laboratory, Cambridge, under the supervision of Dr. D.E. Broadbent, while the author held a Medical Research Council Studentship. The author has also been supported by the Commonwealth Fund.

This paper would not have been possible without the help of Drs. M.P. Haggard, A.M. Liberman, I.G. Mattingly, D.P. Shankweiler, M. Studdert-Kennedy, and T.G. Halwes. Prof. O.L. Zangwill and Dr. F.S. Cooper generously provided research facilities.

Under certain conditions, sounds which enter one ear may subsequently be more efficiently recalled or recognized than similar sounds entering the other ear (Kimura, 1961a, b; 1964). Differences between the ears tend to be obtained more reliably when different sounds enter the two ears simultaneously than when only one ear is stimulated, either with one (Corsi, 1967) or with two simultaneous signals (Shankweiler, in press). Monaural stimulation can give significant ear differences but such experiments have required larger numbers of subjects than the usual dichotic paradigm (Bakker, 1968, 1970).

The type of stimulus material used is probably the only determinant of which ear gives better performance. In similar recognition paradigms, the right ear does better for digit triads (Broadbent and Gregory, 1964) and the left for orchestrated melodies (Kimura, 1964) and simple pitch patterns (Darwin, 1969). In free recall, the right ear again does better for digit triads (Kimura, 1961b) and the left for familiar melodies (Kimura, 1967) and simple pitch sweeps, whether carried on a word or on a nonverbal timbre (Darwin, 1969).

Since patients with vocal speech impaired when their right hemispheres are anaesthetized show an advantage for the left ear in free recall of digit triads (Kimura, 1961a), some link between the ear difference effect and cerebral dominance must be assumed. Authors differ on the nature of this link. Some attribute it to perception (Kimura, 1961b), some to short-term memory (Inglis, 1962), others to attention (Treisman and Geffen, 1968). Some authors have implicitly denied the stimulus specificity of the direction of the effect and claim that there is a general tendency to report material entering the right ear before that entering the left (Oxbury et al., 1967).

One important limitation of free recall experiments was pointed out by Inglis (1962). Serial order effects (see, for example, Broadbent, 1958) could account for ear differences in a free recall paradigm if there were some tendency to report certain types of material from a particular ear first. Bryden (1963) controlled for serial order effects and found a smaller, though still significant, residual advantage for the right ear with digit sequences. Thus, while serial order effects account for some of the ear difference in a free recall paradigm, they do not explain why the sounds from one ear are recalled first or why there is a residual difference. The tendency to report one ear first could derive from whatever causes this residual ear difference.

This residual effect may be due to differences in the efficiency with which material is either perceived or remembered. Making the distinction between perception and memory in terms of the first and second ear reported,

Bryden (1967) summarizes the available data and shows that there is no evidence that the ear difference effect is any smaller on the first than on the second reported ear. Darwin (1969) also failed to find any such evidence for material recalled better either from the right or the left ears.

Treisman and Geffen (1968) suggested that the ear difference effect arises because of an unequal distribution of attention, the left hemisphere finding it easier to attend to the right ear than the left ear. If this were so, we would expect sounds which are more easily separated by selective attention to show a greater ear difference than those which are more difficult to separate. Kirstein and Shankweiler (1969), however, find that, when a subject is asked to report the sounds from a particular ear, he makes fewer errors of attention for vowels than for consonants but that consonants show a greater right ear advantage than vowels. Selective attention may interact with the mechanisms responsible for the ear difference effect, but it is not a basic cause.

Kimura's (1961b) explanation of the ear difference effect as reflecting differences in the efficiency with which material is perceived (in the sense used above) in the two hemispheres can account for all the available data that has been obtained with adequate experimental procedures, provided that we make the assumption that the experimental differences demonstrated between the ears can be attributed to differences between the two hemispheres. No alternative explanation can do so well. What, then, is the nature of this "perceptual" difference? At what stage in the varied processes of perception do differences between the ears and between the hemispheres appear?

The right ear advantage does not depend on the material being meaningful. Significantly greater scores for the right ear than for the left have been detected in free recall paradigms for initial and final stop consonants (Shankweiler and Studdert-Kennedy, 1967a, b) and for laterals and semivowels (Haggard, 1969) in a simple, nonsense syllable context. The right ear advantage for stops remains when order of report is controlled by a suitable method of scoring (Darwin, 1969) or by preinstructing order of report (Kirstein and Shankweiler, 1969). However, these experiments do not tell us whether the difference between the ears occurs before or after the sound has been categorized as a particular phoneme. The failure of vowels to give a right ear advantage in free recall (Shankweiler and Studdert-Kennedy, 1967a; Darwin, 1969) is not relevant here since vowels differ from consonants in both their acoustic structure and their phonological class. Vowels and consonants could have different ear asymmetries at some level <u>after</u> they have been classified

as phonemes. This paper examines whether there are differences between the ears in some perceptual process which occurs before classification of a sound as a phoneme.

Analytically, the sounds of speech form a subset of the sounds of the environment since they are subject to phonetic constraints deriving from the anatomy and physiology of the vocal tract and to phonological and allophonic constraints imposed by particular languages. Maximum efficiency in perception will only be obtained if these constraints are utilized. However, to preserve the efficient perception of sounds not subject to these constraints, some functional division is required in the perceptual system so that one part may deal with the special problems of speech while the other remains free to deal with the remaining sounds.

The phonetic constraints are of two main types, both of which lead to a complex relationship between the perceived phoneme and the acoustic signal. In one case, a complex relation arises because the articulatory specifications for some phonemes are incomplete (for bilabial stops, for example, only a general movement of the lips and jaw is specified); the articulators which are not specified can then assume a wide variety of positions with a correspondingly wide variety of acoustic sequelae. In the second case, the complex relation arises from the variation in size and shape of the vocal tracts producing the sound.

The first set of relations has been extensively studied, and the word encoded has been used (Liberman et al., 1967) to describe this particular lack of acoustic invariance. The second type of variability has received relatively little study. However, the relationship is not likely to be a simple one since, for example, women's vocal tracts are not only smaller than men's but have different relative proportions (Chiba and Kajiyama, 1941). So when a vowel is spoken by two different individuals with the same articulatory gestures, the formant frequencies for one cannot, in general, be obtained by multiplying each formant frequency of the other's by a constant multiple. This multiple varies between speakers, between vowels, and between individual formants (Mattingly, 1966; Fant, 1966). The perceptual system at least partially compensates for these perturbations since it can accommodate some independent variation in the range of the first two formants (Ladefoged and Broadbent, 1957).

These are by no means the only problems for the speech recognition system, but as they are specific to speech, they offer the opportunity of separating speech and nonspeech perceptual mechanisms and of asking whether they are equally

the prerogative of the two ears and of the two hemispheres. The first experiment asks whether the ear advantage is the same for sounds perceived as the same phoneme but requiring to different extents Liberman's "decoder." The second experiment asks the same question of vowel sounds from different sized vocal tracts.

Experiment 1: Fricatives

Fricatives are well suited to the purpose of this experiment since there are two main cues which contribute to their perception. The first, and perceptually most significant, is the spectral peak of the friction itself (Harris, 1958; Heinz and Stevens, 1961); this peak shows relatively little variation with vowel context. The second main cue is the formant transitions to adjacent vowels. These show much more contextual variation with vowels since they depend on the shape of the whole vocal tract. In both voiced and unvoiced fricatives, they assume a major role only in distinguishing /f,v/ from /θ,ð/, although they do contribute to the intelligibility of the other distinctions. Fricatives synthesized with appropriate formant transitions are generally more intelligible than those synthesized without them, although the latter are still highly intelligible provided that the /f,v/ - /θ,ð/ distinction is not required.

Liberman et al. (1967) hypothesized that only those aspects of speech which show appreciable contextual variation give a right ear advantage. This predicts that fricatives containing the appropriate formant transitions will show a right ear advantage, while those without such transitions will not.

Method

The experimental tape was prepared on the Haskins parallel formant synthesizer. Six fricatives, /f,s,ʃ,v,z,ʒ/, were used in the syllabic frame /-ɛp/. The fricatives /θ/ and /ð/ were not used because they are highly confusable with /f/ and /v/, respectively. There were four stimulus conditions:

1) With appropriate friction and appropriate formant transitions.
2) As 1) but with an instantaneous transition into the vowel, which was extended to occupy the time previously allocated to the transition.
3) As 2) but with the vowel deleted, leaving only the steady-state friction.
4) As 1) but without the friction, leaving formant transitions and vowel. This condition sounded like plosives rather than fricatives.

The steady-state friction lasted 45 msec, the transitions 30 msec, and the final syllable 120 msec.

The sounds were assembled into a dichotic tape, using a computer program (Mattingly, 1968) that first laid down marker pulses on the recording tape and then synthesized utterances in a predetermined sequence as the marker pulses were detected. This method allows individual dichotic pairs to be aligned almost perfectly in time, while the use of synthetic speech allows accurate control of the amplitudes and duration of the sounds.

Each sound was paired twice with every other sound in its own stimulus condition to give a basic experimental tape of 240 trials, the second half of which was the same as the first but with the trial order reversed. This whole experimental tape was taken by each subject twice. Prior to the main experiment, the subjects were practiced in identifying the sounds with the following letters: f, s, sh, v, z, j, p, b, d. A pilot experiment showed that the letters p, b, and d were most readily assigned to the quasi-plosives which constituted condition 4. This condition was not basic to the purpose of the experiment but was included in case none of the fricative conditions gave a significant ear advantage. When the subjects were scoring above 75 percent on these single sounds, they were given 10 practice trials with dichotic pairs. They were told to write down the two sounds they heard, putting their more confident choice first. They could write down the same response twice if they wished. They were asked to try to maintain a neutral attention before each trial, rather than to listen for one ear only. After the 10 practice trials, if they had no questions and had not obviously disregarded the instructions, they went on to the main test trials, which came in 16 blocks of 30 trials. Half the subjects started with the headphones reversed, and all subjects reversed their headphones after every 4 blocks.

The experiment was taken by one left-handed and thirteen right-handed undergraduate and graduate subjects. No subject had any hearing defect to the best of his knowledge, and none had a difference of more than 5 db between ears for the threshold at 1500 Hz measured by the method of limits.

Results

Statistical tests are taken from Siegel (1956) and are all two-tailed. Unless otherwise stated, the test used is a Wilcoxon T-test for matched pairs. The overall percents correct for the first and second responses together are given in Table 1.

Table 1

Percents correct for total scores on both responses by stimulus condition.

	Stimulus Condition			
Ear	1 Friction Transition Vowel	2 Friction Vowel	3 Friction	4 Transition Vowel
Left	47.4	44.4	45.4	58.6
Right	53.2	46.6	46.6	63.6
Right-Left	5.8	2.2	1.2	5.0
Right+Left	50.3	45.5	46.0	61.1

A Friedman analysis of variance on total right-minus-left ear scores between the four stimulus conditions is significant ($p < .01$). The total score on the right ear is significantly higher than that on the left for condition 1 ($p < .01$) and condition 4 ($p < .05$) but not for either condition 2 or 3 ($p > .1$). This picture holds both with and without the left-handed subject. Condition 1 gives a significantly greater right ear advantage than either condition 2 ($p < .02$) or condition 3 ($p < .01$). Adding formant transitions thus increases the score more on the right ear than on the left. The left-handed subject shows a large effect in the opposite direction with conditions 2 and 3 showing a greater right ear advantage than condition 1. He is omitted from all remaining statistics.

The total scores show a very significant tendency for the right ear to score higher on condition 1 than on condition 2 ($p < .001$) but only a slight tendency for the left ear to do so ($.1 > p > .05$). A similar pattern prevailed between conditions 1 and 3 but not between conditions 2 and 3. Performance on the right ear is significantly better when formant transitions are added, while that on the left ear is not. Thus, only the right ear can utilize effectively the additional information present in the formant transitions.

Since the preceding analysis has been made in terms of simple percent correct scores, the differences found between the various stimulus conditions may be due partly to changes in preferred order of report, although it is difficult to think of any interesting reason why this should be so. To counter

this objection, however, a scoring system was devised which compensated for order of report effects. These "D scores" are described in the appendix. D_1 scores reflect the first channel reported and D_2 the second. A positive D score indicates a right ear advantage.

Table 2

Mean D scores for fricatives by stimulus condition. Positive D score indicates right ear advantage; subscript denotes order of report.

	Stimulus Condition		
	1 Friction Transition Vowel	2 Friction Vowel	3 Friction
D_1	.253	-.050	.109
D_2	.161	.072	.022
$D_2 - D_1$	-.092	.122	.087

D scores for the three fricative conditions are given in Table 2. A Friedman analysis of variance on the D_1 scores is almost significant ($.1 > p > .05$) but fails significance on the D_2 scores ($p > .1$). The significance level of individual Wilcoxon T-tests on these scores is therefore not reliable. The following significance levels are given, however, as an indication of the pattern of the results. The important differences, those between condition 1 and conditions 2 and 3, respectively, appear large and show apparent significance levels of less than .025 for the D_1 scores. As in the percent correct analysis, there is a large difference between conditions 1 and 2 for the right ear scores ($p < .002$) but a small one for the left ear scores ($p > .1$).

Although the D scores are too variable to allow these significance levels to be accepted, the overall pattern of results is almost identical to that of the percents correct scores. Since the D scores compensate for order of report effects, it is unlikely that the significant patterns seen in the percents correct scores are attributable to a change in order of report preferences. It seems more probable that the D scores are inherently more variable than the simple percents correct from which they are derived.

In summary, a similar pattern of results is obtained with both simple percents

correct scores and a more complicated score which makes some compensation for the order in which the two ears are reported and the overall level of performance. The right ear advantage is greater when appropriate formant transitions are present than when they are absent. The presence of a succeeding vowel in the absence of formant transitions, however, does not appear to influence the ear advantage. The ear difference effect is thus not simply a function of the recognition response class but is also influenced by the particular cues used to achieve a given response. Moreover, the results are as predicted by Liberman et al.'s (1967) encoding hypothesis in that only those sounds with formant transitions show a right ear advantage.

So far in this analysis, we have taken as correct a response which has both the appropriate voicing and place of articulation. It is of some interest to see whether there are ear advantages for these two dimensions independently. There is convincing psychological evidence that the traditional phonetic feature system is implicated in processes of perception (Miller and Nicely, 1955) and short-term memory (Wicklegren, 1966). If the ear difference indeed reflects differences in the perceptual efficacy of the two ears, these differences may be present not only for the perception of the phonemes as a whole but also for the perception of its constituent features.

In a dichotic listening experiment using stop consonants, Halwes (1969) found that a large proportion of errors arose from a failure to combine features correctly rather than from a failure to extract them. Many "incorrect" responses in Halwes's experiment consisted of a feature from one ear combined with a feature from the other ear. Perhaps where, as in this fricatives experiment, a correct response is scored only if both voicing and place of articulation are correct, the ear difference is due to a difference in the efficiency with which the two features are combined into a response rather than to any differences in the efficiency with which they are actually extracted. If this were entirely the case, we would expect there to be no residual ear difference when the ear effects for the two dimensions are assessed separately. On the other hand, it is possible that there <u>are</u> differences between the ears in the efficiency with which the features are actually extracted, in which case we <u>would</u> expect ear differences when we analyze the features separately.

The results of the fricatives experiment were accordingly scored to provide separate analyses of the voicing and place of articulation dimensions. The dimension not under consideration was made irrelevant both in the stimulus and in the response. This procedure is necessary if the analyses of the two dimensions are to be truly independent.

Analysis of place of articulation was carried out in terms of overall percent correct, making voicing irrelevant in both the stimulus and the response. A Friedman analysis of variance gave a significant overall variation over stimulus conditions for right-minus-left ear percent correct scores ($X_r^2 = 7.55$, df=2, $p < .05$). As in the main analysis, the only condition to show a significant right ear advantage was the first, that which had friction and formant transitions (T=4 1/2, n=13, $p < .005$). Neither group 2 nor group 3 showed a significant right ear advantage ($p > .1$). There was a significant difference between the first group and the average of the other two in this respect (T=14 1/2, n=13, $p < .05$). Analysis in terms of D scores was not made because of the large variance with only three response alternatives.

For the voicing dimension, the only trials which contribute differentially to the ear difference are those in which the two stimuli have different voicing but in which the two responses have the same voicing. Only one of the stimuli has then been incorporated into the response. A Friedman analysis of variance on the difference between right and left ear incorporation of voicing for the three fricative conditions is significant ($X_r^2 = 7.0$, df=2, $p < .05$). Individual T-tests show that voicing is incorporated more often from the right ear than from the left in both the first (T=13 1/2, n=12, $p < .05$) and the second (T=11, n=12, $p < .05$) stimulus conditions (the two with the succeeding vowel). There is no significant right ear advantage for the third condition with the isolated friction (T=20 1/2, n=11, $p > .1$). There is a significant difference between groups 2 and 3 in this respect (T=12, n=13, $p < .02$) but not between any of the others. Combining the first two groups gives a highly significant advantage for the right ear (T=1 1/2, n=12, $p < .002$) and a significant difference between their mean and the third group (T=10, n=11, $p < .05$). Thus, the voicing dimension is reported more accurately from the right than from the left ear only when there is a succeeding vowel.

For fricatives, there is thus a dissociation between the stimulus conditions necessary to give a right ear preference for place of articulation and those necessary to give one for voicing. Formant transitions are necessary for the former, but a succeeding vowel suffices for the latter. However, these conclusions must be qualified by their possible contamination with changes in order of report preferences since they are based on an analysis of percent correct scores.

Discussion

The main result of this experiment is that the right ear advantage is not determined solely by the recognition response but is also influenced by the particular sounds used to achieve that response. This appears to be true both for the phonetic response as a whole and for the individual articulatory features which constitute that response. Moreover, the particular acoustic signals that must be present for voicing or for place of articulation to show a right ear advantage are different. For place of articulation, appropriate formant transitions must be present, while for voicing, a succeeding vowel suffices. This dissociation suggests that the difference between the ears is occurring before or during the classification of the sound into features and that it is not simply a consequence of an overall ear difference for the phonemic response. In particular, the presence of a right ear advantage for voicing under condition 2, when there is no overall right ear advantage for the entire phoneme, argues that the ear difference for the individual features is not a consequence of the ear advantage for the entire response but rather that the ear advantage for particular features logically precedes that for the entire response.

If differences between the ears are not simply a function of response class, can the same be said of differences between the hemispheres? Unfortunately, no. An important assumption in the interpretation of ear differences is that there is a functional decussation of the auditory pathways. Although there is electrophysiological evidence that shows that a statistical decussation in subhuman species both for evoked potentials (Tunturi, 1946; Rozensweig, 1951) and for single unit recording (Hall and Goldstein, 1968), the main evidence we have that this decussation is both present in man and sufficient to reveal interhemispheric differences is the result of dichotic listening experiments. The most convincing demonstration occurs in patients with a section of the corpus callosum. These patients can report verbal material equally well from either ear when only one ear is stimulated at a time, but they can report practically nothing from the left ear when similar verbal material is played simultaneously into both ears (Milner et al., 1968). Moreover, this weakening of the left ear response is dependent on the nature of the sounds in the other ear. As the sounds in the right ear are progressively distorted, performance on the left ear improves (Sparks and Geschwind, 1968).

Normal subjects show much smaller ear differences than commissurectomised patients when undistorted digit sequences are played in both ears (Milner et al., 1968; Kimura, 1961b). Normal subjects also show an ear difference effect which is dependent on the nature of the competing stimulus. Initial and final plosive consonants give a reliable right ear advantage when they are opposed by another such consonant (Shankweiler and Studdert-Kennedy, 1967b); however, plosive consonants embedded in a nonsense word and opposed by white noise give no ear difference (Corsi, 1967). An unpublished experiment by the present author showed no ear difference using initial plosives rather than embedded ones in one ear and noise on the other. Thus the ear difference effect is influenced by the nature of the competing stimulus.

The simplest explanation of these effects is that, in normal subjects, considerable information about the sounds on the left ear can be transmitted across the commissures to the left hemisphere. Commissurectomised patients, being deprived of this path, must rely entirely on the direct ipsilateral path. The efficiency of this latter path is critically dependent on the nature of the sounds on the two ears. With no sound on one ear, it can function well, but as progressively less distorted speech is introduced on the other ear, it becomes less and less efficient.

A significant difference between scores from the two ears can be interpreted as showing that there is some difference between the hemispheres and that the sounds on each ear have gone predominantly to their opposite hemispheres. However, if there is no significant difference between the ears, we cannot attribute this failure with any confidence to either an equivalence of the two hemispheres or to a failure of the relevant pathways to decussate sufficiently to reveal an interhemispheric difference. The differences in ear advantage between the various stimulus groups reported in this experiment could then be due either to a difference in the degree to which the two hemispheres are implicated in their processing or to a difference in their abilities to produce a functional decussation of the relevant pathways. We can only conclude that the former is true and, thus, that the hemispheres differ in their ability to classify phonemes if we have independent evidence that those sounds that did not give a right ear advantage were in principle capable of revealing any interhemispheric difference that there might have been.

All the sounds that failed to give an ear advantage for a particular feature in this experiment had a steady state along the physical dimension relevant to that phonetic feature. Thus, place of articulation shows an ear advantage only when it is cued by a moving pattern of formant transitions, while the voicing feature shows an ear advantage only when it is cued by a sound that may be only partially voiced. Perhaps no steady-state discrimination can give an ear difference. The absence of any ear difference for steady-state vowels, whether in CVC context or in isolation (Shankweiler and Studdert-Kennedy, 1967a, b), and of very brief duration (Darwin, 1969) supports this idea. Furthermore, Darwin (1969) found only tenuous evidence for a left ear advantage for recall of steady-state nonverbal timbres similar to those whose discrimination was more impaired after right, rather than left, temporal lobectomy (Milner, 1962). If an ear advantage can be demonstrated for steady-state sounds, we will have more justification for assuming that the steady-state sounds used in this fricatives experiment were, in principle, capable of showing ear differences.

We must now face the logical difficulty that, without further assumptions, we cannot tell whether any change made in the stimulus conditions which produces an ear advantage is having its effect through changing the conditions necessary to reveal differences between the hemispheres or through changing the nature of the task in such a way as to implicate mechanisms for which the hemispheres do, in fact, differ.

One reasonable assumption is that the functional decussation of the auditory pathways is determined only by the particular sounds which are presented on any one trial and is not influenced by the range of sounds which may occur in the experiment. In other words, if we know from the fact that they give an ear advantage that there is good decussation for a particular dichotic pair of sounds in one experiment, we can remove some of the other dichotic pairs from the experiment without changing the functional decussation for that particular pair. In contrast, the number of different dichotic pairs used in an experiment will generally alter the complexity of the task and so perhaps alter the relative contribution of either hemisphere. If, then, we can show that greater ear advantages can be obtained for some sounds when the number of different stimuli used in the experiment is changed, we might assume we are measuring a change in interhemispheric ability rather than a change in the functional decussation of the auditory pathway.

If, then, the steady-state sounds used in this and other experiments have failed to show any ear difference solely because of inadequate functional

auditory decussation, we should not expect such sounds to show an advantage when only the complexity of the perceptual discrimination is changed. The next experiment attempts to demonstrate that the ear advantage is influenced by the complexity of the perceptual discrimination by changing the range of vocal tract sizes that a set of vowels can come from.

Experiment 2: Vowels from Different Sized Vocal Tracts

There is a rough correlation between voice pitch and formant frequencies, since women and children have higher voices and smaller vocal tracts than men. This correlation is utilized in estimating vocal tract size (Fujisaki and Kawashima, 1969). A recent experiment by Haggard (pers. comm.) shows that, when vowel perception depends on the fundamental frequency of the vowel, there is a right ear advantage under free recall conditions. Steady-state sounds show a right ear advantage when there is a difference in pitch between the two ears. Unfortunately for the present argument, this difference in pitch is a reasonable candidate for a factor which changes the conditions necessary to reveal the ear difference effect, as well as one which alters the perceptual complexity of the task. Can we show a right ear advantage for steady-state vowels which have the same pitch on either ear? The most direct way to answer this question is to use sets of vowels from two different sized vocal tracts.

Method

The five vowels /i,ɛ,æ,a,ʌ/ in the context /ən-t/ were synthesized on the Haskins parallel formant synthesizer using only the first two formants. Two sets of these five words were made, the formant frequencies for one set being 25 percent higher than those for the other set. The formant values are given in Table 3.

Two different experimental tapes were then constructed. On one tape, each sound was paired with every other sound except itself and its phonemic homologue from the other vocal tract. On the other tape, only the sounds from the smaller vocal tract were used, and each sound was paired with every other sound except itself. The first tape had 160 trials and the second 40. The order of those trials on the second tape was exactly the same as the order of those trials on the first tape, in which both sounds came from the smaller vocal tract.

Table 3

Formant frequencies for vowels in experiment 2.

	Large vocal tract		Small vocal tract	
Vowel	F1	F2	F1	F2
/ɪ/	386	2078	489	2540
/ɛ/	537	1845	666	2307
/æ/	666	1695	844	2156
/ɑ/	718	1075	894	1312
/ʌ/	640	1232	794	1541

The first tape was taken twice by one group of eighteen subjects, and the second tape was taken twice by a second group of eighteen subjects. All subjects were right-handed, native speakers of American English, who to the best of their knowledge had no hearing defects. The instructions and training they received were similar to those used in the fricatives experiment. The words used to identify the sounds were a nit, a net, a gnat, a knot, a nut, and both groups of subjects used the five letters i, e, a, o, u as their responses. Those who took the first tape had training in identifying the sounds from both vocal tracts, whereas the second group of subjects were only introduced to the sounds from the smaller vocal tract. The usual counterbalancing procedures were observed.

Results

Five stimulus conditions are distinguished in the results. Four come from the first group of subjects and correspond to whether the dichotic pair had sounds from 1) the larger vocal tract only; 2) the smaller vocal tract only; 3) the larger on the left ear and the smaller on the right; 4) the smaller on the left and the larger on the right. The fifth condition corresponds to the second group of subjects who had the smaller vocal tract on both ears all the time. The overall percents correct and the D scores are given in Tables 4 and 5, respectively.

Table 4

Overall percents correct in experiment 2 by dichotic pair composition.

Vocal Tract Size on			Overall Percent Correct on			
Left Ear	Right Ear		Left Ear	Right Ear	Right - Left	p(L=R)
Large	Large	(1)	46.7	50.9	4.2	<.01
Small	Small	(2)	45.8	50.4	4.5	<.01
Large	Small	(3)	45.7	56.2	10.6	<.002
Small	Large	(4)	54.0	51.2	- 2.8	
Total			48.1	52.2	4.1	<.001
Small	Small	(5)	54.0	53.4	- 0.6	>.1

Table 5

D scores for experiment 2 by dichotic pair composition.

Vocal Tract Size on						
Left Ear	Right Ear		D_1	D_2	$p(D_1 = 0)$	
Large	Large	(1)	.083	.078	<.06	<.002
Small	Small	(2)	.103	.067	<.05	
Large	Small	(3)	.226	.206		<.002
Small	Large	(4)	-.062	-.089		
Small	Small	(5)	-.060	-.013	>.1	

The overall superiority for the right ear for the first group of subjects (summing over the first four stimulus conditions) is significant, both on percents correct ($p < .001$) and on D_1 scores ($p < .01$). For the second group of subjects there is no significant right-ear advantage on either score ($p > .1$).

A Friedman analysis of variance over the first four stimulus conditions is significant for differences in percents correct ($p < .01$) and D_1 scores ($p < .02$). The variation in overall level of performance, however, is barely significant ($p < .1$). Individual Wilcoxon T-tests show that ear differences are significant for the first and second stimulus conditions separately on overall percents correct ($p < .01$) and on D_1 ($p < .05$ and $< .06$, respectively).

For conditions 3 and 4 combined, when the two ears had different vocal tracts, the right ear did significantly better than the left ($p < .002$) but there was also a significant tendency for vowels from the smaller vocal tract to be recalled better than those from the larger ($p < .05$). This difference is not present when the two ears receive vowels from the same vocal tract, as in conditions 1 and 2. It is not, then, due to markedly poorer intelligibility for the smaller vocal tract.

There is a significantly greater right ear advantage for the vowels in condition 2 than in condition 5, both for percents correct ($p < .05$) and for D_1 scores ($p < .02$) on Mann-Whitney U-tests. But there is no difference between the averages of conditions 1 and 2 versus conditions 3 and 4 ($p > .1$). In other words, the right ear advantage for vowels in this experiment depends on the nature of the discrimination within the framework of the whole experiment rather than within the individual trial.

A reliable right ear advantage for steady-state vowels, therefore, can be obtained when there is uncertainty within the experiment as to what size vocal tract has produced them. But this right ear advantage is not influenced by whether, on a particular trial, the two alternative sizes of vocal tract are, in fact, present.

Discussion

Vowels can give a right ear advantage. Whether or not the advantage appears in this experiment depends on the complexity of the perceptual discrimination rather than on the particular sounds used on any one trial. On the assumption that the sounds used for the second group of subjects were, in principle, capable of showing a right ear advantage, we can conclude that the hemispheres do differ in their ability to classify vowels from different sized vocal tracts. This assumption seems reasonable since identical sounds did give a right ear advantage when played to the first group of subjects as part of a larger experiment.

The assumption that was necessary to interpret the results of the fricatives experiment in terms of differences between the two hemispheres has received

some justification since the vowels used here are cued mainly by a steady state. More direct confirmation of this could perhaps be obtained by using steady-state friction from different sized vocal tracts.

Can we draw any conclusions about the stage or stages in perception at which ear or hemisphere differences become apparent? The ear difference effect is not solely a function either of the stimulus or of the response but rather of the processes which must mediate between the two. The fricatives experiment showed that it did not depend on the response category alone since whether or not it appeared either for the entire phonetic response or for one of the constituent dimensions of voicing and place of articulation depended on the presence of particular acoustic cues. The vowel experiment described here shows that the effect does not depend solely on either the stimuli presented on a particular trial or on the response category since the same stimuli do or do not show a right ear advantage depending on the complexity of the relationship between the stimuli and the responses.

A similar conclusion has been reached by Studdert-Kennedy and Shankweiler (1970) on the basis of a feature analysis of a dichotic experiment with stop consonants. They, with Halwes (1969), find that a large proportion of errors arise from inappropriate combination of correctly extracted features. They suggest that this arises because acoustic features can be extracted correctly in either hemisphere but that they can only be related to phonemic features and assembled into a phonemic response in the left hemisphere.

More direct evidence that particular acoustic features themselves are not entirely responsible for the ear difference effect comes from an experiment by Haggard (1970). Haggard shows that, when the voicing dimension is cued only by a change in pitch (Haggard et al., 1970) in a dichotic listening paradigm, the recall of this feature shows a right ear advantage. Since Darwin (1969) has shown that simple pitch sweeps give a left ear advantage when carried on a word but do not cue a phonemic distinction, it seems likely that the pitch sweeps which cued voicing in Haggard's experiment would show a left ear advantage in a suitable nonspeech context. Here, then, it is not the extraction of the acoustic cue which is important but its phonetic relevance.

The existence of some stage which mediates between an acoustic representation of the input stimulus and the phonetic output has been suggested by Hiki et al. (1968) on the basis of experiments on a short-term contrast effect in vowel perception (Fry et al., 1962). They suggest that there is some transform which maps acoustic space into a multidimensional phonetic space from which decisions are made about the appropriate phonetic category. The nature of this transform is determined both by the short-term effects that they investigated

and by the longer-term normalization effects demonstrated by Ladefoged and Broadbent (1957).

The arguments put forward here have concentrated on identifying the earliest stage at which differences between the ears become apparent. This is not necessarily the only stage nor that at which the greatest differences may be obtained. Work on temporal lobectomized patients has shown large differences between the two temporal lobes for verbal memory in excess of the short-term memory span (Milner, 1958), but there has been considerably less evidence that verbal perceptual deficits depend on which hemisphere is damaged. Luria (1966) presents some evidence that patients with damage to the left temporal lobe are impaired in their ability to repeat simple nonsense syllables. But this is the only evidence of its kind. The work on commissurectomised patients has given no evidence that there are any perceptual differences between the two hemispheres (Milner et al., 1968), although, of course, recall is largely restricted to one hemisphere only. Perceptual differences may, in fact, exist at the level of phonemic analysis, and these differences may not yet have been revealed because few tests have put strain specifically on the phonetic aspects of speech perception.

That no effects, other than those reported by Luria, have yet appeared does suggest that the lateralization of speech perception is considerably less than that of speech production and verbal memory. This does not necessarily mean that these latter processes are influencing the results of the experiments reported here. It may well be that the dichotic listening technique is particularly sensitive to processes which occur early in the sequence of perception and memory, if only because stimuli are more likely to be differentiated according to ear of arrival immediately after input than at some later time. We must, however, acknowledge the possibility that memory processes may show differential ear effects, although there is as yet little evidence that they do.

Appendix: D Scores

In a free recall dichotic listening experiment, the simple percents correct score is inadequate for two reasons. First, it takes no account of the relative number of times one ear is reported first and the other second, so that errors arising from serial order effects are confounded with those from other sources. Second, differences in percents correct are not strictly comparable between subjects because of varying overall levels of performance; a given difference in detectability gives rise to a wide range of differences in percents correct at different performance levels. The two D scores described here give estimates of the differences in recall between the two ears on the first and second reported channels, respectively. These estimates take into account both the relative number of times each ear is reported first and the absolute probability of being correct on each of these channels.

First and second channels here refer simply to the order of report rather than to any property of the input. The following letter combinations denote the number of trials on which each subject made the corresponding pattern of correct responses.

LR = left ear	correct on first channel,	right ear	correct on second channel.
RL = right	"	left	"
LZ = left	"	neither	"
RZ = right	"	neither	"
ZL = neither	"	left	"
ZR = neither	"	right	"
ZZ = neither	"	neither	"

Then let:

$$p(L_1) = (LR + LZ) / (LR + LZ + ZR)$$
$$p(R_1) = (RL + RZ) / (RL + RZ + ZL)$$
$$p(L_2) = (RL + ZL) / (RL + RZ + ZL)$$
$$p(R_2) = (LR + ZR) / (LR + LZ + ZR)$$

Denoting a normal transformation with a prime we now define

$$D_1 = p'(R_1) - p'(L_1)$$
$$D_2 = p'(R_2) - p'(L_2)$$

This scoring method ignores trials on which neither ear was correct (ZZ) and assumes that making a normal transformation is an adequate compensation for variations in overall performance level (Green and Birdsall, 1964).

References

Bakker, D.J. (1968) Ear-asymmetry with monaural stimulation. Psychonom. Sci. 12, 62.

Bakker, D.J. (1970) Ear-asymmetry with monaural stimulation: relations to lateral dominance and lateral awareness. Neuropsychologia 8, 103-117.

Broadbent, D.E. (1958) Perception and Communication. (Pergamon, London).

Broadbent, D.E., and Gregory, M. (1964) Accuracy of recognition for speech presented to the right and left ears. Quart. J. Exptl. Psychol. 16, 359-360.

Bryden, M.P. (1963) Ear preference in auditory perception. J. Exptl. Psychol. 65, 103-105.

Bryden, M.P. (1967) An evaluation of some models of laterality effects in dichotic listening. Acta oto-laryngol. 63, 595-604.

Chiba, T., and Kajiyama, M. (1941) The Vowel: Its Nature and Structure. (Tokyo-Kaiseikan, Tokyo).

Corsi, P.M. (1967) The effects of contralateral noise upon the perception and immediate recall of monaurally presented verbal material. Unpublished M.A. thesis, McGill University, Montreal.

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

Fant, G. (1966) A note on vocal tract size factors and non-uniform F-pattern scalings. STL-QPSR 4, 22-30.

Fry, D.B., Abramson, A.S., Eimas, P.D., and Liberman, A.M. (1962) The identification and discrimination of synthetic vowels. Language and Speech 4, 171-189.

Fujisaki, H., and Kawashima, T. (1968) The roles of pitch and higher formants in the perception of vowels. IEEE Trans. AV-16, 73-77.

Green, D.M., and Birdsall, T.M. (1964) The effect of vocabulary size on association score. In Signal Detection and Recognition by Human Observers, J.A. Swets, Ed. (Wiley, New York) 609-619.

Haggard, M.P. (1969) Perception of semi-vowels and laterals. J. Acoust. Soc. Amer. 46, 115 (A).

Haggard, M.P. (1970) The use of voicing information. Speech Synthesis and Perception, Progress Report No. 2. (Psychological Laboratory, University of Cambridge).

Haggard, M.P., Ambler, S., and Callow, M. (1969) Pitch as a voicing cue. J. Acoust. Soc. Amer. 47, 613-617.

Hall, J.L., and Goldstein, M.H. (1968) Representation of binaural stimuli by single units in primary auditory cortex of unanaesthetized cats. J. Acoust. Soc. Amer. 43, 456-461.

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

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

Heinz, J.M., and Stevens, K.N. (1961) On the properties of voiceless fricative consonants. J. Acoust. Soc. Amer. 33, 589-596.

Hiki, S., Sato, H., and Oizumi, J. (1968) Dynamic model of vowel perception. Paper presented to 6th International Congress of Acoustics, Tokyo, Japan, August 1968.

Hyde, S.R. (1968) Automatic speech recognition. Post Office Research Dept. Report No. 45.

Inglis, J. (1962) Dichotic stimulation, temporal lobe damage, and the perception and storage of auditory stimuli: A note on Kimura's findings. Canad. J. Psychol. 16, 11-16.

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

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

Kimura, D. (1964) Left-right differences in the perception of melodies. Quart. J. Exptl. Psychol. 14, 355-358.

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

Kirstein, E., and Shankweiler, D.P. (1969) Selective listening for dichotically presented consonants and vowels. Paper read before 40th Annual Meeting of Eastern Psychological Association, Philadelphia.

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

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

Luria, A.R. (1966) Human Brain and Psychological Processes. B. Haigh, Trans. (Harper and Row, New York and London).

Mattingly, I.G. (1966) Speaker variation and vocal-tract size. J. Acoust. Soc. Amer. 39, 1219 (A).

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

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

Milner, B. (1958) Psychological deficits produced by temporal lobe excision. Publ. Ass. Res. Nerv. Ment. Dis. 36, 244-257.

Milner, B. (1962) Laterality effects in audition. In Interhemispheric Relations and Cerebral Dominance, V.B. Mountcastle, Ed. (Johns Hopkins, Baltimore).

Milner, B., Taylor, L.B., and Sperry, R.W. (1968) Lateralized suppression of dichotically-presented digits after commissural section in man. Science 161, 184-185.

Oxbury, S., Oxbury, J., and Gardiner, J. (1967) Laterality effects in dichotic listening. Nature 214, 742-743.

Rosenzweig, M.R. (1951) Representation of the two ears at the auditory cortex. Amer. J. Physiol. 167, 147-158.

Shankweiler, D.P. (in press) An analysis of laterality effects in speech perception. In The Perception of Language, P. Kjeldergaard, Ed. (University of Pittsburgh Press).

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

Shankweiler, D.P., and Studdert-Kennedy, M. (1967b) An analysis of perceptual confusions in identification of dichotically presented CVC syllables. J. Acoust. Soc. Amer. 41, 1581 (A).

Siegel, S. (1956) Non-parametric Statistics. (McGraw Hill, New York).

Sparks, R.W., and Geschwind, N. (1968) Dichotic listening in man after section of neocortical commissures. Cortex 4, 3-16.

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

Treisman, A.M. and Geffen, G. (1968) Selective attention and cerebral dominance in perceiving and responding to speech messages. Quart. J. Exptl. Psychol. 20, 139-150.

Tunturi, A.R. (1946) A study of the pathway from the medial geniculate body to the acoustic cortex in the dog. Amer. J. Physiol. 147, 311-319.

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

Selective Listening for Temporally Staggered Dichotic CV Syllables*

Emily Kirstein+
Haskins Laboratories, New Haven

It is by now well known that when stop-vowel syllables differing only in the initial consonant are delivered simultaneously to opposite ears, recall is more accurate from the right ear than from the left. More recent work on dichotic listening (Lowe et al., 1970; Studdert-Kennedy et al., 1970) has revealed that the maximum suppression of the left ear occurs not when the two syllables are simultaneous but when the syllable delivered to the right ear arrives about 50 msec. after the syllable delivered to the left ear. More generally, an advantage in recall accrues to the lagging consonant regardless of which ear receives that consonant. This lag effect combines with the ear asymmetry to produce the greatest right ear advantage for trials on which the right ear lags in onset behind the left.

The lag effect is as yet poorly understood. We do know that the effect depends on dichotic presentation and that it is, therefore, central in origin and not due to peripheral masking. Moreover, there are indications that, even with dichotic presentation, some types of stimuli do not give a lag advantage. When isolated steady-state vowels are presented dichotically, the leading vowel tends to be heard as clearer than the lagging vowel (Porter et al., 1969). We may speculate that the lag effect is specific to the recall of encoded phones like the stop consonants and that it depends on some fundamental property of the speech decoding mechanisms.

The present study was undertaken primarily to determine whether the lag effect arises during phoneme recognition or whether it arises in the organization of recall after the sounds have been identified. If the lag effect reflects merely a preference for the lagging consonant or a recall strategy in which the syllable arriving second is generally the first to be recalled, then the lagging and leading consonants should be recalled equally well if

---

*Paper presented to the 79th Meeting of the Acoustical Society of America, Atlantic City, N.J., 21-24 April 1970.

+Also, University of Connecticut, Storrs.

the subject were required to listen for and recall only one of the competing consonants. In previous experiments on the lag effect, the subjects had been instructed to report both consonants on each trial. In the present experiment, the listeners were instructed to report only one of the two stimuli. Each subject performed two selective listening tasks. In one task, called the "ear monitoring" task, the subjects were instructed to report one ear and ignore the stimulus at the other ear. Equal periods of time were spent reporting only the right and only the left ear. In the second task, called the "temporal order" task, the subjects were to attend to the order of arrival of consonants within a dichotic pair. On half the trials they were to report only the lagging consonant from the pair and on half the trials only the leading consonant. If the lag effect is truly a robust perceptual phenomenon, the listeners should be more accurate in their report of lagging consonants than of leading consonants, whether they are selecting by ear of arrival or by order of arrival.

The stimuli for the experiment were nine synthetic syllables, /ba/, /da/, /ga/, /be/, /de/, /ge/, /bɔ/, /dɔ/, /gɔ/, each 350 msec. in duration. Pairs of these syllables were recorded on a two-channel tape in such a way that only the consonant differed between channels, resulting in combinations such as /ba/-/ga/ or /de/-/ge/. One of the syllables of a pair was always delayed in onset relative to the other by 10, 30, 50, 70, or 90 msec. There were 6 seconds between pairs. Timing of the recording was under computer control. The frequency of occurrence of individual syllables was completely balanced over ears, tape channels, delays, and recall conditions. The complete counterbalancing required a total of 720 trials for each subject for each task. Testing was conducted in four one-hour sessions, two sessions for each task.

Each testing session was split into four blocks of ninety trials. Selective recall instructions were given at the beginning of each block of trials. For the ear monitoring task, the subjects were instructed as to which ear to report at the beginning of each set of ninety. For the temporal order test, they were told whether to report the lagging or the leading consonant for each block of trials. The subjects were required to give one response on each trial, even if they had to guess and to respond with B, D, or G.

Twelve right-handed students took the two tests. Half of them did the ear monitoring task first and half the temporal order task.

Very similar results were observed for the ear monitoring and temporal order tasks, so the two tasks will be considered together.

Figure 1 shows the percentage of trials on which the subjects correctly recalled the consonant designated by the selective listening instructions, depending on whether the correct consonant was lagging or leading, whether it was arriving at the left or right ear, and the amount of time between syllable onsets. (The abscissa gives the relative onset time of the correct syllable from 90 msec. lag to 90 msec. lead. The solid line represents trials on which the selected syllable was arriving at the right ear, and the dashed line represents trials on which the selected syllable was arriving at the left ear. Each point is based on 36 trials for each of the twelve subjects, 432 trials in all).

For both the ear monitoring and temporal order task, three factors were found to influence recall. The instructions had some effect, and as might be expected, the subjects were more accurate in monitoring either by ear or by temporal order with longer intervals between the onsets of the competing syllables. The most obvious effect is the right ear advantage. Regardless of whether the subjects were recalling by ear or by temporal order, they were more often correct when the selected syllable was arriving at the right ear than at the left. Finally, there was a large advantage in recall for the lagging syllable compared with the leading syllable within a pair. This lag advantage is indicated by the generally negative slopes of the curves.

The errors made by subjects could almost always be interpreted as failures to judge correctly the order of arrival or ear of arrival rather than as incorrect identifications of the consonants. Trials on which subjects selected the stimulus which they should have ignored are termed "intrusions." The difference in frequency between correct responses and intrusions provides a measure of the accuracy of selective recall. Figure 2 gives the accuracy of selection as a function of the length of the delay between syllables. Discrimination of order of arrival was consistently poorer than discrimination of ear of arrival. Although monitoring on either basis improved with longer delays, considerable difficulty was experienced with delays as long as 90 msec.

In view of the fact that the listeners were more accurate in selecting their responses by ear than by temporal order, one might expect that the influence in recall attributable to the lag effect and laterality effect would be diminished for the ear monitoring task. A comparison of the lag effect and laterality effect for the two tasks is shown in Figure 3, which gives the change in magnitude of each of these effects as a function of the delay between syllable onsets for the two tasks. The graphs were constructed

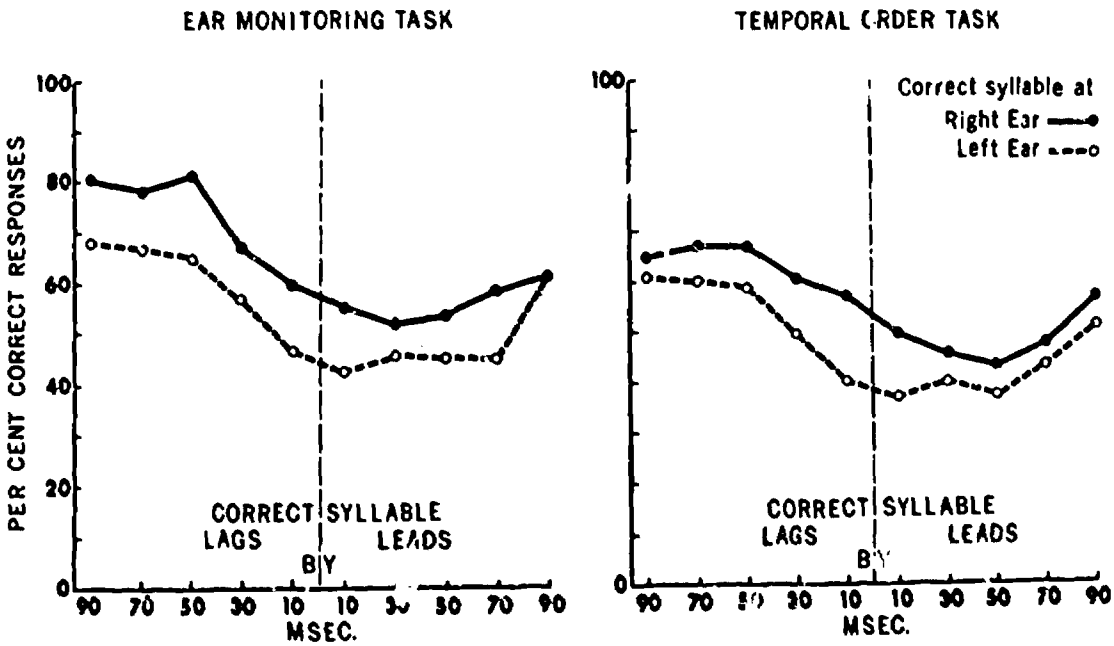

EAR MONITORING TASK
TEMPORAL ORDER TASK
PER CENT CORRECT RESPONSES
100
80
60
40
20
0
CORRECT SYLLABLE
LAGS
LEADS
BY
90 70 50 30 10 10 30 50 70 90
MSEC.
Correct syllable at
Right Ear
Left Ear

FIG. 1

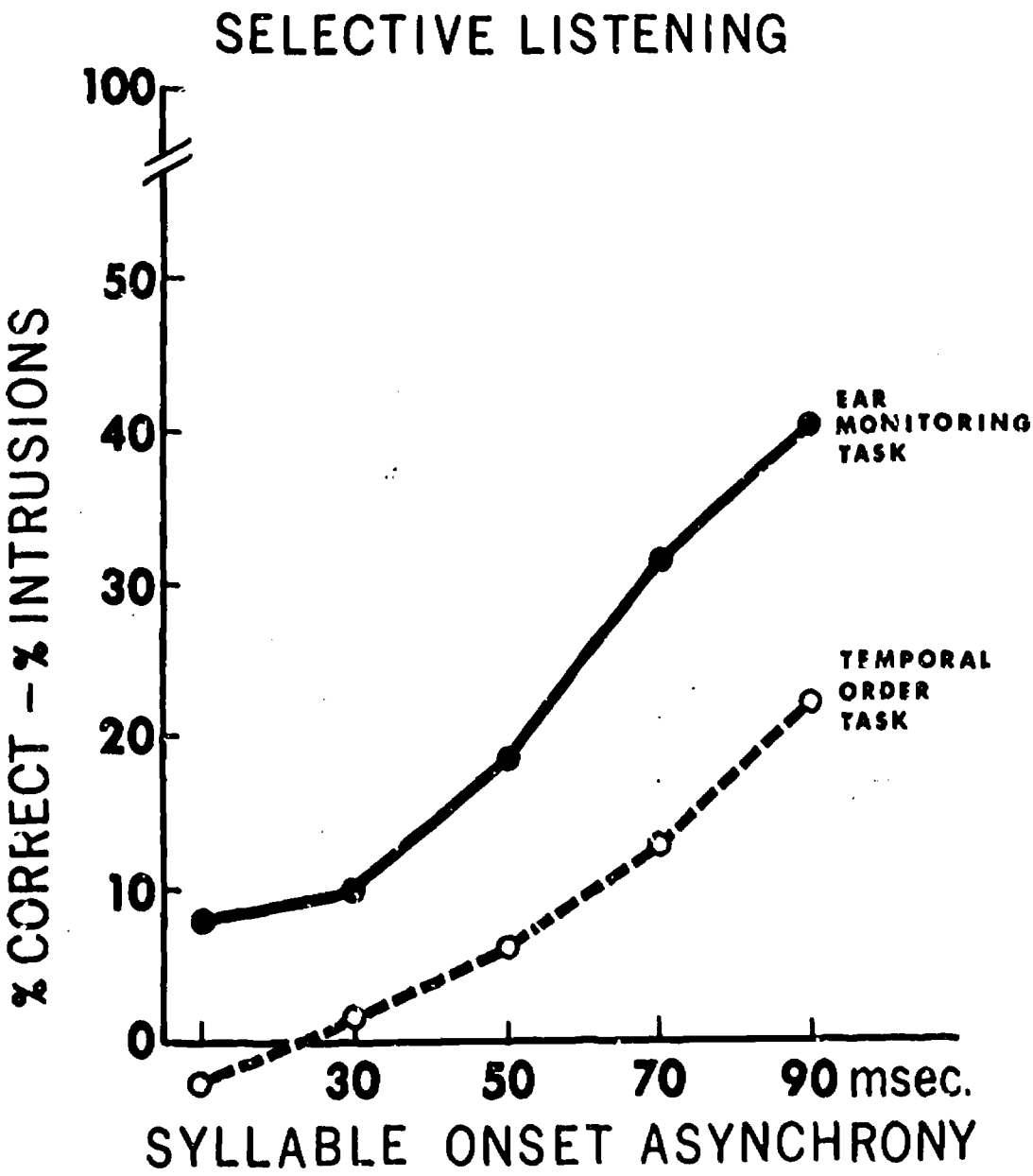

SELECTIVE LISTENING
100
50
40
30
20
10
0
% CORRECT – % INTRUSIONS
EAR MONITORING TASK
TEMPORAL ORDER TASK
30
50
70
90 msec.
SYLLABLE ONSET ASYNCHRONY

FIG. 2

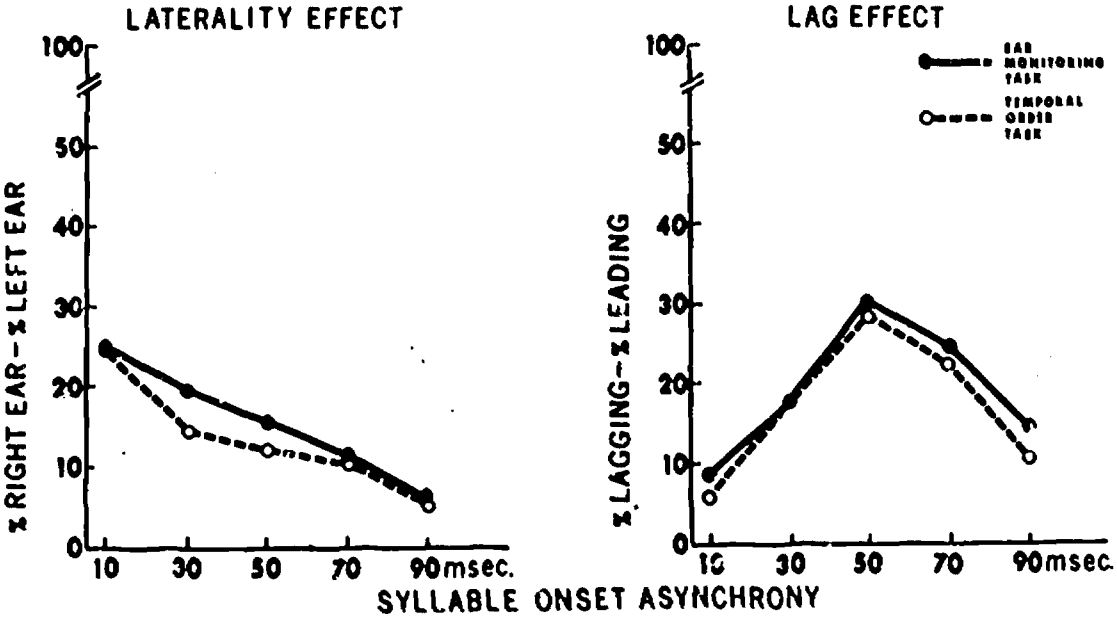

LATERALITY EFFECT
LAG EFFECT
% RIGHT EAR - % LEFT EAR
% LAGGING - % LEADING
100
50
40
30
20
10
0
10
30
50
70
90msec.
SYLLABLE ONSET ASYNCHRONY
EAR MONITORING TASK
TEMPORAL ORDER TASK

FIG. 3

as follows: each response was classified as to whether it corresponded to a stimulus presented at the right ear, the left ear, or neither and also whether it corresponded to a lagging stimulus, a leading stimulus, or neither. The percentages of all responses falling into each of these categories was computed. The laterality effect chart gives the percentage by which right ear report exceeded left ear report. The lag effect chart gives the percentage by which recall of lagging syllables exceeded recall of leading syllables.

For both the ear monitoring and temporal order tasks, the preference of the right ear was strongest when the two syllables were most nearly simultaneous, that is, with delays of 10 msec. Longer delays between ears reduced the ear effect. The magnitude of the ear advantage did not differ reliably between the two tasks. This decrease in the magnitude of the ear advantage with increases in interaural delay is further evidence that the laterality effect depends on competition between ears. Although the effect does not depend critically on simultaneity of onsets of the competing syllables, it is clear that longer delays reduce the advantage for the right ear over the left.

Turning to the lag effect, it can be seen that the magnitude of the advantage for the lagging ear did not differ between the two tasks. For both the ear monitoring task and the temporal order task, the maximum advantage for the lagging ear was achieved when syllables were separated in onset time by 50 msec. The peaking of the lag effect at 50 msec. is seen more clearly here than in previous experiments. It is possible that the location of the peak depends on the stimuli used. The synthetic syllables used in this experiment had second formant transition lasting from 35 to 50 msec. The peak at 50 msec. may reflect a critical time for processing these transitions.

The results of this experiment suggest that the lag effect and the laterality effect exert an advantage in recall independent of each other and that the magnitude of both these effects is independent of the recall strategy. It seems likely that the lag effect and ear effect do not account for the confusions in selection by ear or temporal order; such selection errors probably occur often in any case because of the fact that the acoustic features which differentiate the competing consonants last no more than 50 msec. Rather, consonants which are lagging or which arrive at the right ear appear to gain in saliency for the listener, while those arriving first or arriving at the left ear appear to lose by the same amount without altering the overall level of performance on the selection task.

In conclusion, the lag advantage and right ear advantage in recall of dichotically presented CV syllables are extremely robust phenomena which cannot be eliminated by manipulation of recall strategy. These asymmetries in recall attributable to ear or order of arrival most likely arise during the identification of consonants and not during the organization of responses for recall. Further research should be undertaken to elucidate the mechanisms underlying these effects.

References

Lowe, S.S., J.K. Cullen, C. Thompson, C.I. Berlin, L.L. Kirkpatrick, and J.T. Ryan. (1970) Dichotic and monotic simultaneous and time-staggered speech. J. Acoust. Soc. Amer. 47, 76 (A).

Porter, R., D. Shankweiler, and A.M. Liberman. (1969) Differential effects of binaural time differences in perception of stop consonants and vowels. Proc. 77th Annual Convention of the Amer. Psychol. Assn.

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

Temporal Order Judgments in Speech: Are Individuals Language-Bound or Stimulus-Bound?*

Ruth S. Day+
Haskins Laboratories, New Haven

Abstract. Speech stimuli, such as BANKET and LANKET, were presented dichotically, with relative onset time varying over trials from 0 to ±100 msec. When asked to report which phoneme led, subjects fell into two groups: those who performed well, and those who were misled by the temporal constraints on English. A tentative model of temporal order judgment is presented that suggests that there are two modes of listening: a linguistic mode and a nonlinguistic mode.

Introduction

Previous studies of dichotic listening have emphasized the rivalry between the two ears. For example, when the digit TWO is presented to one ear over earphones, while at the same time the digit THREE is presented to the other ear, subjects typically report hearing TWO, or THREE, or both TWO and THREE. Fusion does not occur: no one reports hearing THRU or TEE. However, a study in the present series (Day, 1968) has shown that fusions can occur when the proper psycholinguistic variables are taken into account. For example, given BACK to one ear and LACK to the other, subjects typically report hearing the fusion, BLACK.

Method

The present experiment was designed to study the role of time cues in facilitating or retarding the fusion effect. Figure 1 shows a dual-beam oscilloscope photograph of a sample item. The top channel represents BANKET and the bottom channel represents LANKET. Both are real-speech samples that have been edited using the pulse code modulation system at the Haskins Laboratories (Cooper and Mattingly, 1969). This system permits the experimenter to do four things: 1) he can determine where an item begins and discard all that

---

*Paper presented at the Ninth Annual Meeting of the Psychonomic Society, St. Louis, November 1969.

+Also, Yale University, New Haven.

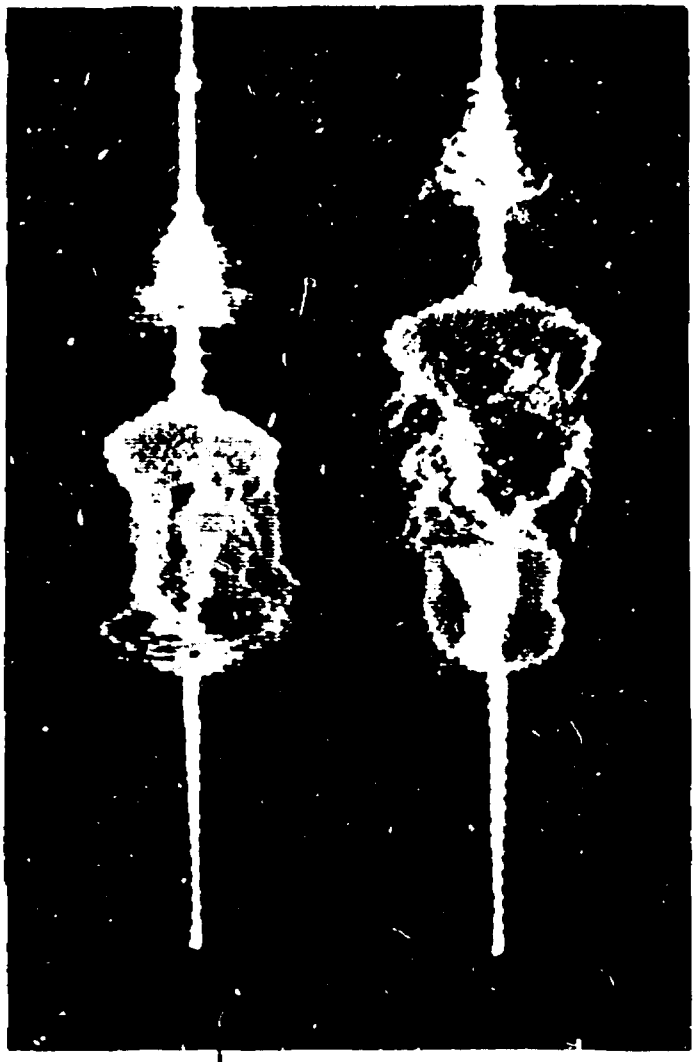


FIG. 1

precedes that point; 2) he can determine where an item ends and discard all that follows; 3) he can equalize the over-all intensities of the two items so that they are equally loud; 4) finally, he can line up the onsets of the two utterances with accuracy on the order of 500 microseconds. Note that in this particular example of BANKET/LANKET both utterances begin at the same point in time. We will refer to this situation as the simultaneous onset case, or the 0-lead time case.

Figure 2 shows the general paradigm of the experiment. On one set of trials, BANKET began first by 25, 50, 75, or 100 msec. On another set of trials, LANKET began first by the same intervals. And on the final set, both utterances began at the same point in time. There were ten items, as shown in Figure 3. All involved initial stop and liquid consonants. Each stop consonant (/p,t,k, b,d,g/) was paired with the liquids /r,l/.[1] Thus, for the /pr/ cluster, the inputs PAHDUCT/RAHDUCT can be fused to yield PRODUCT; for the /pl/ cluster, PANET/LANET yields PLANET; for the /tr/ cluster, TEETMENT/REETMENT yields TREATMENT; and so on. All possible fusion responses were acceptable English words, although other experiments (e.g., Day, 1968, Exp. II) have shown that subjects will report fusions that are nonwords, e.g., GORIGIN/LORIGIN yields GLORIGIN. In addition, all inputs were nonwords. (While "wordness" does correlate with meaningfulness, the notion should not be taken too seriously: although BANKET is not an acceptable word, it does answer the question, "What shall I do with the money?" and hence, it is in some sense meaningful.)

Results

The Effect of Relative Onset Time on Fusion Probability. We want to determine what the probability of fusion response is for each of the lead-time conditions. Will fusions occur more readily when the stop consonant (e.g., /b/) leads than when the liquid (e.g., /l/) leads? If so, we would expect that fusion response probability will be higher on the left side of the display in Figure 2 than on the right side. As shown in Figure 4, the obtained function was more or less a straight line. Perhaps fusion was somewhat more probable when the stop led by 75 msec, but in any event, it looks as if time cues per se were not affecting fusion levels. People were about as likely to hear BLANKET when LANKET led.

[1] /tl-/ and /dl-/ were excluded since these clusters do not occur in initial position in English.

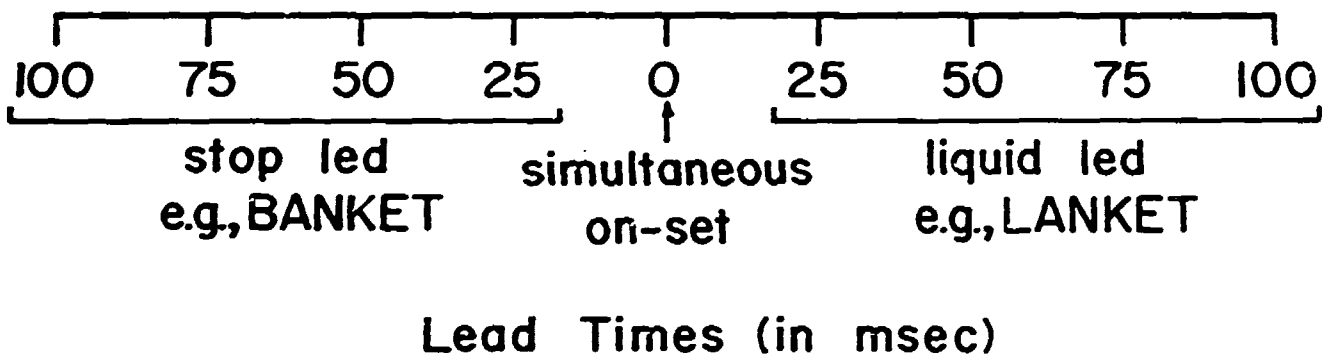

100
75
50
25
0
25
50
75
100
stop led
e.g., BANKET
simultaneous
on-set
liquid led
e.g., LANKET
Lead Times (in msec)

FIG. 2

Experimental Items

	r	l
p	PRODUCT	PLANET
t	TREATMENT	
k	CRACKER	CLOSET
b	BREAKFAST	BLANKET
d	DREADFUL	
g	GREEDY	GLEAMING

FIG. 3

Fusion Response Probability
.10
.20
.30
.40
.50
.60
.70
.80
.90
1.00
100
75
50
25
0
25
50
75
100
stop led
simultaneous on-set
liquid led
Phoneme Lead Times (in msec)

FIG. 4

Fusion Rates over Subjects. A surprising set of findings emerged from the data based on the performance of individual subjects. Each subject was given a score that reflected how often he fused. This was simply the proportion of times he fused over all trials (180). Contrary to scores on most psychological tasks, fusion rates were not normally distributed (Figure 5). Instead, subjects fell into two groups: those who fused most of the time ("high fusers") and those who fused relatively infrequently ("low fusers"). This experiment with sixteen right-handed subjects has been repeated with sixteen left-handers, and the fusion results are comparable. So the addition of more subjects makes the bimodal distribution even more striking.

Temporal Order Judgments (TOJ). Up to this point, we have been discussing the fusion task. In this task, subjects were asked to report out loud whatever they heard: one word, two words, real words, or nonsense words. There was a second task: temporal order judgment (TOJ). Here, subjects were asked to write down the first sound they heard on every trial. For example, if the first sound they heard was /b/, as in BOY, they wrote down the letter B. In this task we wanted to determine how well subjects could determine which phoneme led as a function of the lead conditions. Consider an individual subject as shown in the top display of Figure 6. When the stop (e.g., /b/) led by 100 msec, he performed perfectly: he always said that the stop led. As the stops lead decreased down to 25 msec, he always said that the stop led. Now consider trials where the liquid (e.g., /l/) led. When the liquid led by 25 msec, the subject performed miserably, that is, he always said that the stop led. As the liquid's lead increased, this subject's performance did not improve at all. He simply reported hearing the stop first, independent of the stimulus conditions. There were twenty observations per point, so the data for each subject are fairly stable. I should also point out that the point representing the 0-lead case is a special case: since neither item led, there is no "correct" temporal order judgment. The open circle at 0-lead indicates the percent of stop consonant responses so that we can assess the overall level of the subject's bias.

Now let's look at another subject as shown in the lower part of Figure 6. He looks very much like the first subject. Note that neither subject improved with increased lead time on either side of the continuum. There were about four more subjects who performed like those of Figure 6. There were other subjects who showed the same over-all effects, but did show a slight increase in performance at the longer leads; nevertheless, their performance on liquid leads never rose above chance.

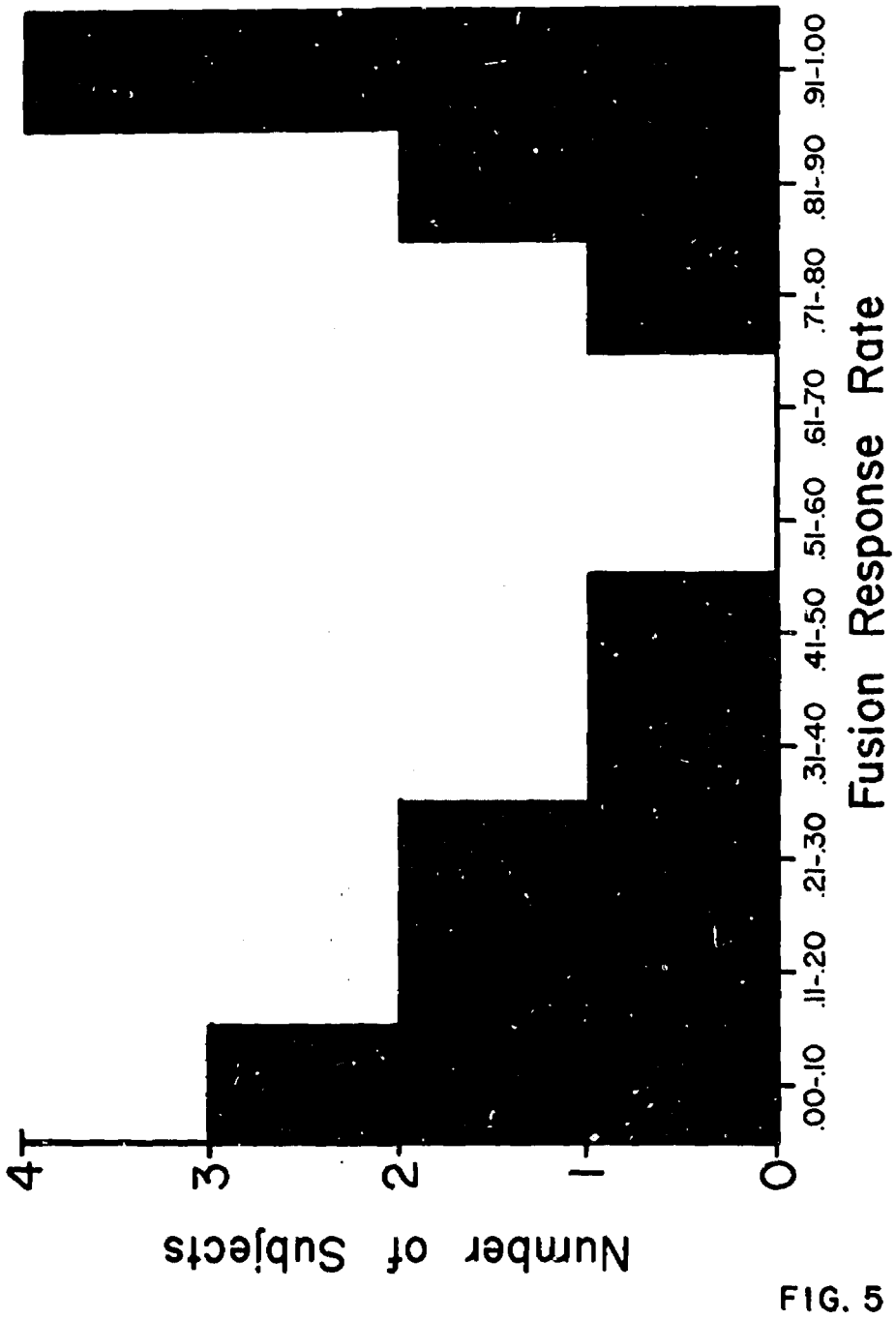

Number of Subjects
4
3
2
1
0
.00-.10
.11-.20
.21-.30
.31-.40
.41-.50
.51-.60
.61-.70
.71-.80
.81-.90
.91-1.00
Fusion Response Rate

FIG. 5

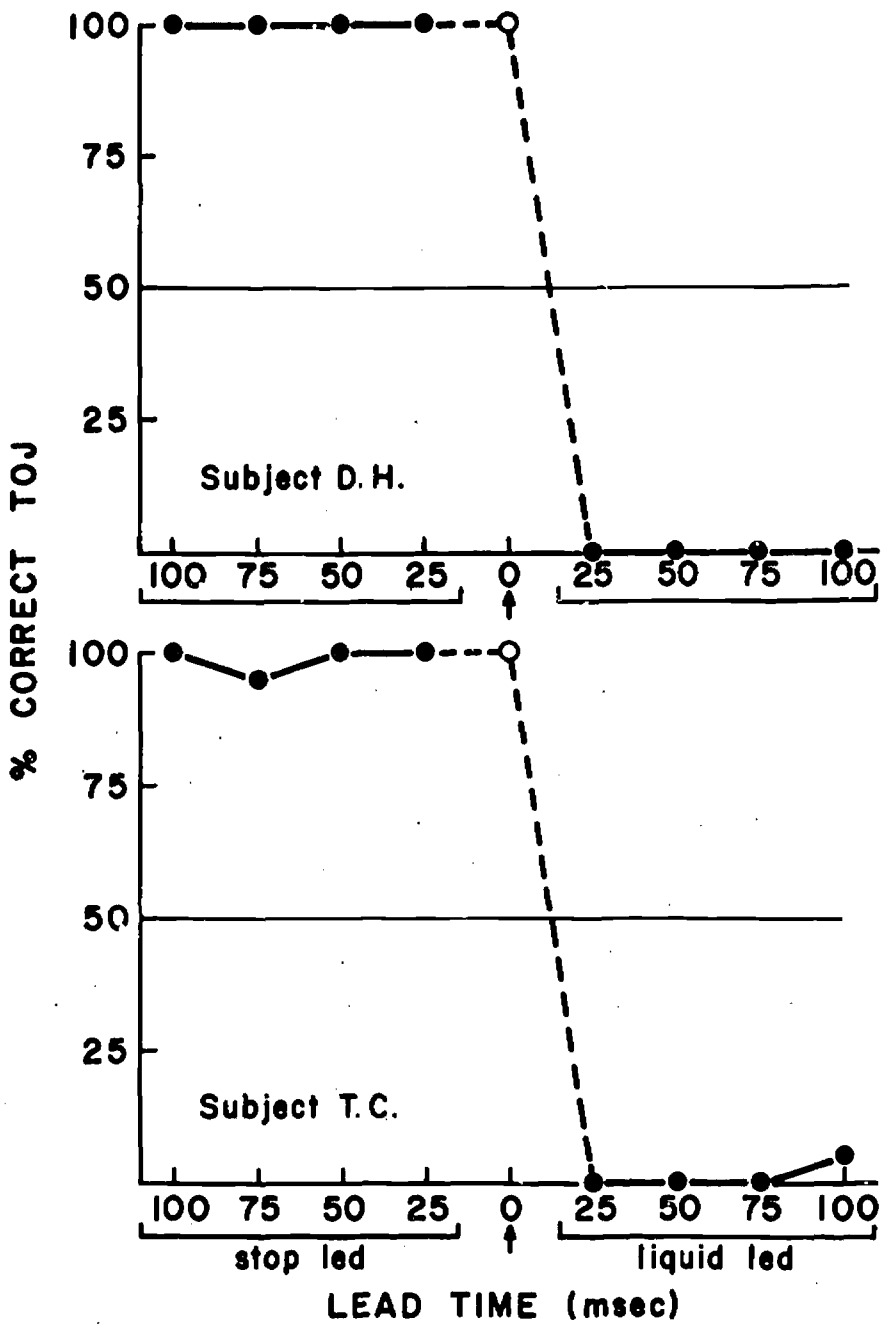

100
75
50
25
% CORRECT TOJ
Subject D. H.
100 75 50 25 0 25 50 75 100
100
75
50
25
Subject T. C.
100 75 50 25 0 25 50 75 100
stop led
liquid led
LEAD TIME (msec)

FIG. 6

There were some radically different subjects. Consider those shown in Figure 7. When the stop led, they correctly identified it as leading. When the liquid led, they also correctly identified it as leading. Note that both subjects were sensitive to increased lead times on both sides of the continuum.

Figure 8 shows a schematic diagram of the two types of TOJ performance. The top display shows overall performance that is poor. Subjects here perceive the stop as leading, independent of the stimulus conditions. Further, they show no improvement with increased time leads. These subjects are wholly bound by the facts of the language: in English, (stop + liquid) can occur in initial position, but (liquid + stop) cannot. These facts bias subjects against hearing the phonemes in their given order. We will call them "language-bound" subjects (for lack of a better term). In the bottom display of Figure 8, overall performance is good. Subjects here can tell which stimulus led, and they are sensitive to increased time leads. Since their responses do reflect the stimulus condition, we will call them "stimulus-bound."

Relation of the Fusion and TOJ Tasks. A brief review is in order. On the fusion task, we found two groups of subjects: high fusers and low fusers. On the TOJ task with the same subjects, we found two groups of subjects: those who performed poorly (language-bound) and those who performed well (stimulus-bound). Question: Is there any way to predict how a subject will do on the TOJ task given that we know he is a high fuser or low fuser? Thus we want to correlate performance on the two tasks for the same subjects. Figure 9 shows the scatter diagram for this relationship. Along the ordinate is each subject's TOJ accuracy.[2] Not only is there a negative correlation, but subjects tend to cluster into two groups: those who are high fusers and poor temporal order judges and those who are low fusers and good temporal order judges.

Discussion

Ordinarily, when we talk about "individual differences," we are trying to account for noisy data. Here, however, the individual difference data suggest that there may be two different types of language perceivers. We have

[2]The score used here was percent correct on liquid-leading items. Several other scores have been used, and all give essentially the same results.

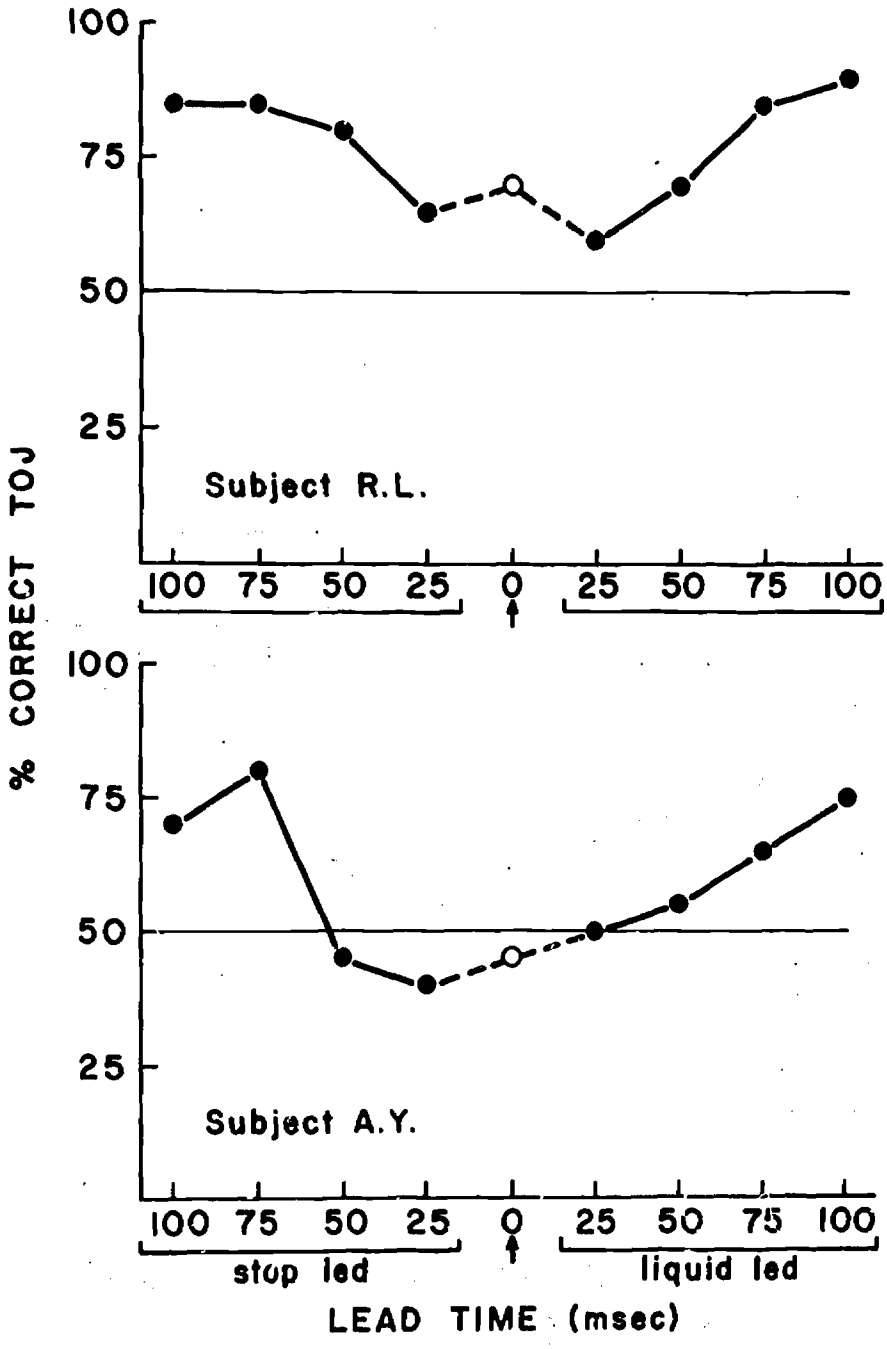

100
75
50
25
Subject R.L.
100 75 50 25 0 25 50 75 100
100
75
50
25
Subject A.Y.
100 75 50 25 0 25 50 75 100
stop led
liquid led
% CORRECT TOJ
LEAD TIME (msec)

FIG. 7

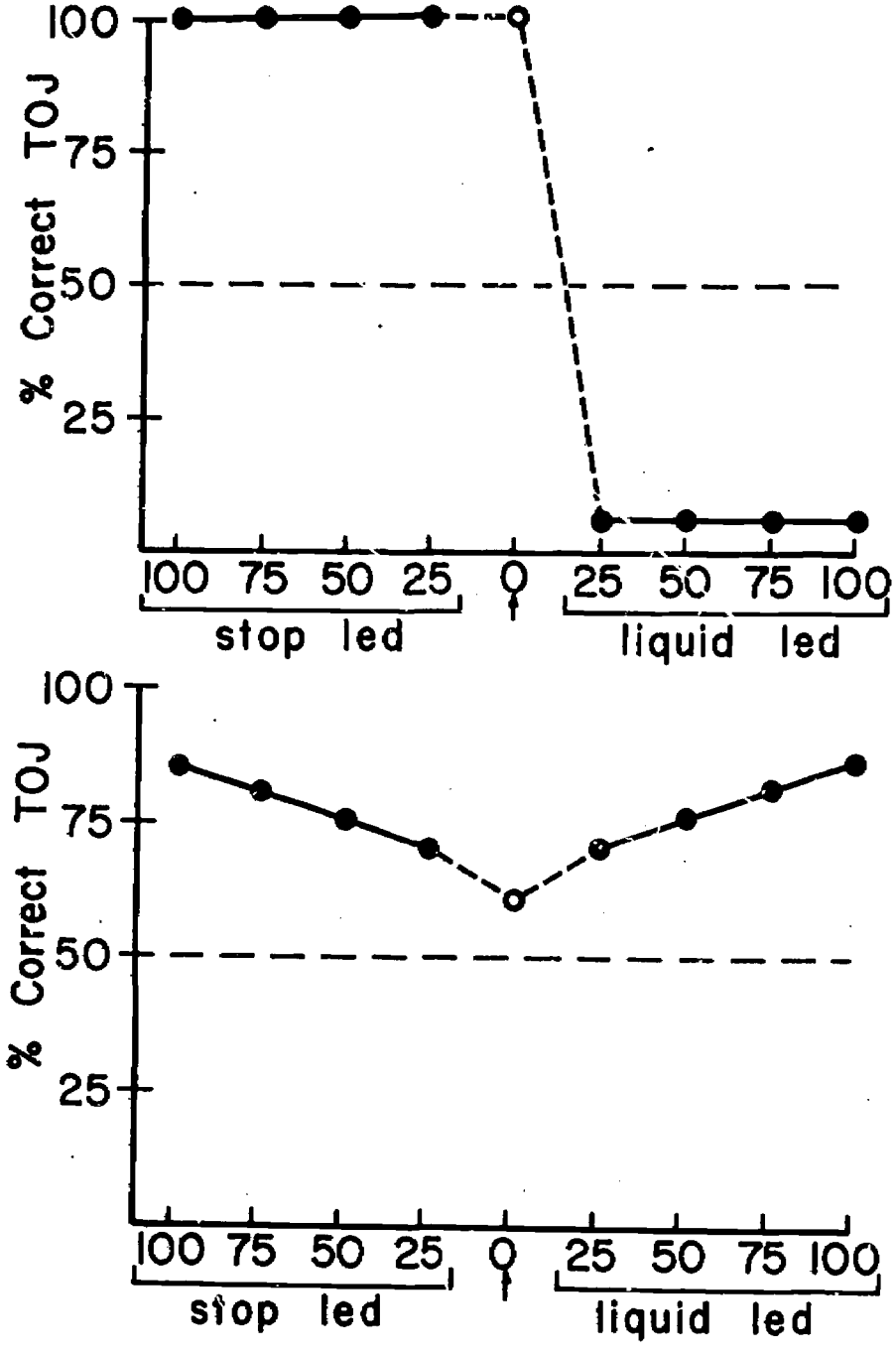

100
75
50
25
% Correct TOJ
100 75 50 25 0 25 50 75 100
stop led
liquid led
100
75
50
25
% Correct TOJ
100 75 50 25 0 25 50 75 100
stop led
liquid led

FIG. 8

Fusion Rate (by Subjects)
1.00
.80
.60
.40
.20
TOJ Accuracy (by Subjects)
.20
.40
.60
.80
1.00

FIG. 9

noticed other differences about the two groups. At the end of the experiment, language-bound subjects are often surprised to learn about the nature of the stimuli and still hear fused clusters even when they are told that there are none on the tape. On the other hand, stimulus-bound subjects can usually tell the experimenter exactly what is on the tape. There are further questions to ask: Will these two groups retain their identities on other speech perception tasks? What about auditory short-term memory tests? We are currently identifying subjects in each category and giving them a battery of relevant tests.

A preliminary and tentative model to describe how subjects make temporal judgments in the present experiment is given in Figure 10. The model is to be used only as a point of departure: I do not know how many boxes there should be, nor how they should be arranged, nor which way all the arrows should go. However, the model does serve as a way to begin thinking about the processes involved. Consider first the analysis stage. At some point, subjects do analyze the stimuli into phonemes. They know that they are deciding between /b/ and /l/, or between /p/ and /r/. At a later stage, synthesis work must be done. That is, the phonemes must be arranged into some order. Before a subject can give a response, the results must be related to past experience with the language, perhaps by way of a linguistic filter or similar device. The filter operates on the basis of the sequential dependencies of phonemes in the language.[3] For example, if LBANKET emerges from the synthesis stage, it has difficulty in passing through the linguistic filter and is therefore returned to synthesis for new ordering. If the output is BLANKET, it can pass through the filter, and hence the subject reports hearing /b/ first. The filtering system may have different bias levels across subjects, which would account for the obtained individual differences.

The discussion of temporal order judgment thus far has involved processing of a linguistic nature. However, before the analysis work is done, certain acoustic decisions can be made. For example, there may be a simple detection, a decision that a signal is on, and further, a decision concerning which ear received the signal. A subsequent experiment has been performed

[3]Perhaps the linguistic filter does not come after synthesis is completed; instead, the synthesis stage itself may have preset probabilities for various sequential dependencies. But this distinction is not crucial for the present discussion.

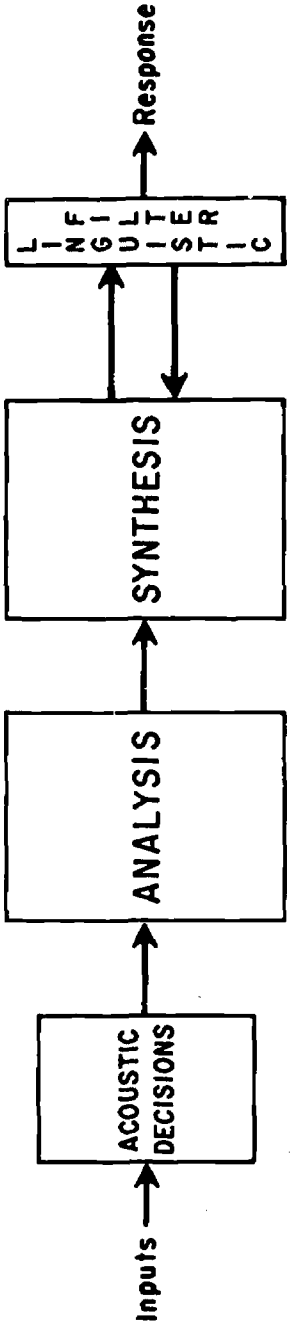

Response
LINGUISTIC
SYNTHESIS
ANALYSIS
ACOUSTIC DECISIONS
Inputs

FIG. 10

(Day and Cutting, 1970) in which subjects indicated which ear led. Performance on the ear task was much better: subjects were highly accurate, even though they were language-bound on the phoneme task.

At present, we cannot be sure that acoustic decisions of this sort are made primarily at an early stage. Nor can we assert that they necessarily require less information processing. The claim at this point is simply that they do not require linguistic processing. When subjects judge temporal order at this nonlinguistic level, they perform well. It is only when they must do some linguistic processing, that is, analysis into phonemes, that they get into trouble. Thus, the model suggests that there are two types of processing that speech signals can undergo: linguistic processing and nonlinguistic processing. Further, given that a subject's performance does not reflect the stimulus events when asked to identify the leading phoneme, the model suggests where the information is lost: namely during linguistic processing.

Two complementary research strategies have emerged from this work. The first, as described above, involves presenting speech stimuli and asking for temporal order judgments that require linguistic vs. nonlinguistic processing. The second involves presenting speech stimuli and analogous nonspeech stimuli, such as complex tones, and asking for temporal order judgments in the two situations. The results thus far are promising: the data support the notion that there are two general modes of auditory perception, a linguistic mode and a nonlinguistic mode.

Another approach involves investigation of critical cases within the linguistic system. Reversible clusters are of particular interest here (Day, 1970). Given the dichotic item TASS/TACK, subjects can easily determine which phoneme came last since both orders are permissible: TASK and TACKS.

Conclusion

In conclusion, we have seen that: 1) the effect of time cues on fusion and temporal order judgment is surprisingly negligible, and 2) individuals perform in two very different ways, some appear to be language-bound, while others accurately reflect the stimulus conditions. These results suggest that there may be two types of language perceivers in the population at large. A preliminary and tentative model of temporal order judgment was presented. It suggests that there are two modes of listening: a linguistic mode and a nonlinguistic mode.

The dichotic fusion technique is also useful in studying the role of the two cerebral hemispheres in the perception of speech. Therefore, we have extended these studies to other populations: left-handers, temporal lobe patients, and split-brain patients. But those accounts can wait for another occasion.

References

Cooper, F.S. and Mattingly, I.G. (in press) Computer-controlled PCM system for investigation of dichotic speech perception. Paper presented at the 77th meeting of the Acoustical Society of America, Philadelphia, April, 1969. J. Acoust. Soc. Amer.

Day, R.S. (1968) Fusion in dichotic listening. Unpublished Ph.D. thesis, Stanford University.

Day, R.S. (1970) Temporal order perception of reversible phoneme cluster. Paper presented at the 79th meeting of the Acoustical Society of America, Atlantic City, April.

Day, R.S. and Cutting, James E. (1970) Levels of processing in speech perception. Paper presented to the Tenth Annual Meeting of the Psychonomic Society, San Antonio, November.

Opposed Effects of a Delayed Channel on Perception of Dichotically and Monotically Presented CV Syllables*

M. Studdert-Kennedy,+ D. Shankweiler,++ and S. Schulman
Haskins Laboratories, New Haven

We wish to report a new phenomenon in binaural speech listening that we have termed the "lag effect." The effect is seen in the greater accuracy with which subjects identify the lagging member of a pair of temporally overlapped syllables presented to opposite ears. Earlier experiments had shown that if CV or CVC syllables, differing only in their initial or final consonants, were presented in dichotic competition, those presented to the right ear were correctly reported significantly more often than those presented to the left (Shankweiler and Studdert-Kennedy, 1967; Studdert-Kennedy and Shankweiler, 1970). As part of a general program of research into the conditions of this right ear advantage for consonants, we undertook to titrate the effect in temporal units: our plan was to estimate the number of milliseconds by which the left ear syllable should lead the right ear syllable for the right ear advantage to be abolished. In the event, we found that the right ear advantage was more readily abolished by a left ear lag than by a left ear lead. The effect has now been repeatedly confirmed both at our own laboratory and elsewhere (Berlin et al., 1970; Lowe et al., 1970). Here we wish simply to report some of its conditions as uncovered in the original experiment.

The stimuli were formed from the syllables /ba, da, ga, pa, ta, ka/, synthesized on the Haskins Laboratories Parallel Formant Synthesizer and each 250 msec long. Syllables were recorded in pairs, one on each channel of a balanced two-track tape recorder. By means of a computer-aided routine, two 240-pair random order tapes were prepared: the onset of one member of each pair was made to lead (or lag) the onset of the other by 0, 5, 10, 20, 25, 50, 70, or 120 msec. The two tapes provided a fully balanced 480-item

*This paper appeared in J. Acoust. Soc. Amer. 48, 599-602 (1970).
+Also, Queens College, City University of New York, Flushing.
++Also, University of Connecticut, Storrs.

test in which each syllable occurred equally often on each channel, paired with each syllable other than itself, for a total of thirty presentations at each lead and lag value other than zero, at which it occurred sixty times. These tapes were intended for dichotic presentation. A second pair of tapes was prepared for monotic presentation by mixing two channels of the dichotic tape electronically and recording the output on a single track.

The subjects were sixteen, right-handed, undergraduate women, all of whom had scored 95% or better on both ears in monaural identification tests of the synthetic syllables. As in previous dichotic studies (Shankweiler and Studdert-Kennedy, 1967; Studdert-Kennedy and Shankweiler, 1970), appropriate counterbalancing procedures distributed all effects due to recorder channels, earphone characteristics, positions of earphones on the head, or sequence of testing equally over the ears of the entire group of subjects. Subjects were instructed to record two from the set of six consonants on each trial, writing on an answer sheet and guessing if necessary.

As a baseline against which the dichotic data may be judged, we first present the group monotic data: in Figure 1, mean percent correct is plotted as a function of temporal lag (negative) and lead (positive) in milliseconds, for right and left ears. Each point is based on 480 judgments (960 at 0 msec). The two ears give essentially identical results: performance is at chance level for syllables with onsets that lag by 25 msec or more but then rises steadily to virtually perfect performance for syllables that lead by 50 msec or more. The functions were similar for all subjects: every one of the sixteen reached at least 95% correct for a lead of 50 msec.

The results seem open to a straightforward peripheral masking interpretation. Although each syllable was approximately 250 msec long, the important cues for the identification of its initial consonant occur in the first 50 msec, during which the syllable is also rising to its maximum amplitude. As lead time is increased from zero, more and more of the crucial portion of the syllable is presented without interference from the lagging syllable until, at 50 msec, all needed consonantal information in the leading syllable is freely available, and performance is almost perfect. On the other hand, as lag time is increased from zero, more and more of the crucial portion of the lagging syllable occurs during the period of maximum amplitude of the leading syllable until, at -50 msec, the important cues in the lagging syllable are fully masked and performance drops to chance. This account squares with the subjective impression of the monotic pairs at the longer lead/lag values: one hears a single

Mean percent correct by ear on monotically presented CV syllable pairs as a function of temporal lead in milliseconds. For sixteen subjects.

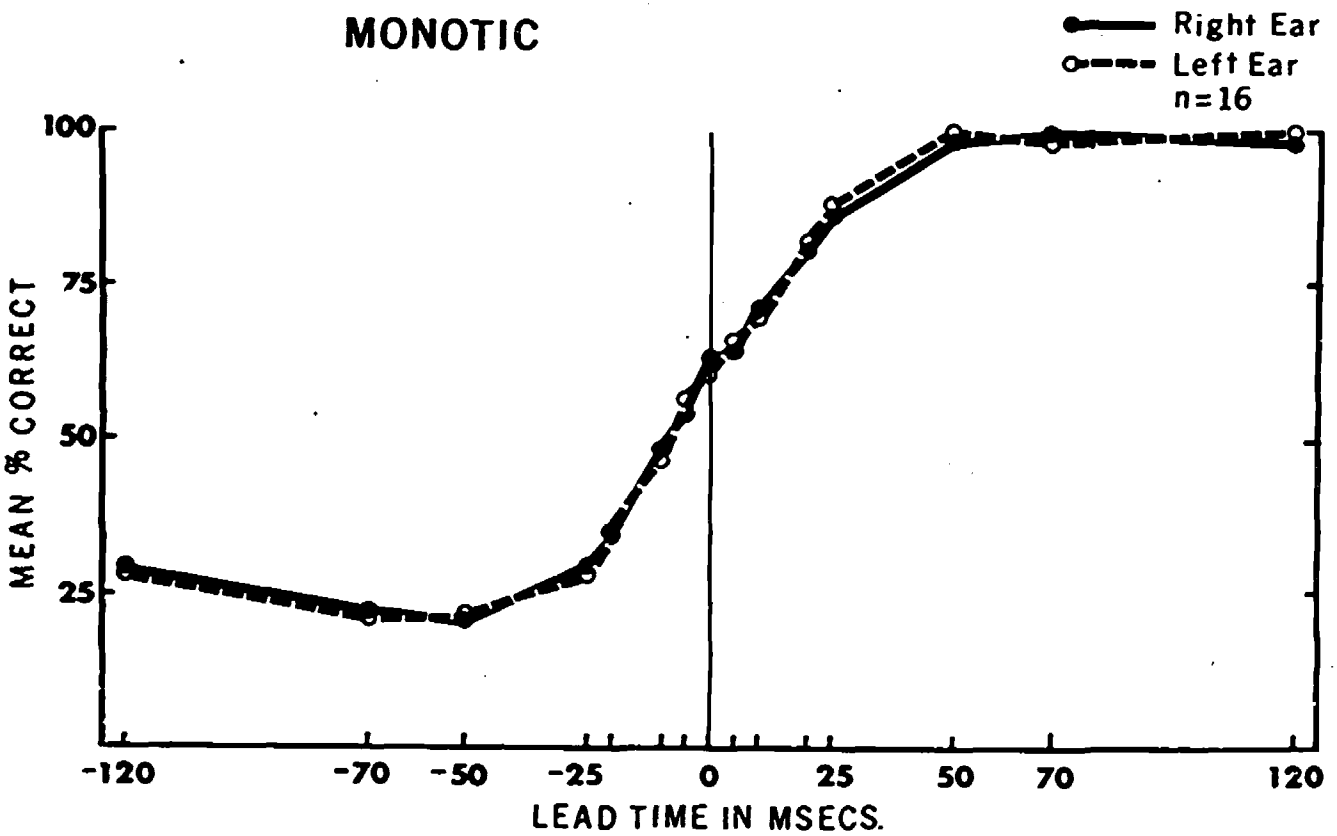


FIG. 1

Mean percent correct by ear on dichotically presented CV syllable pairs as a function of temporal lead in milliseconds. For sixteen subjects.

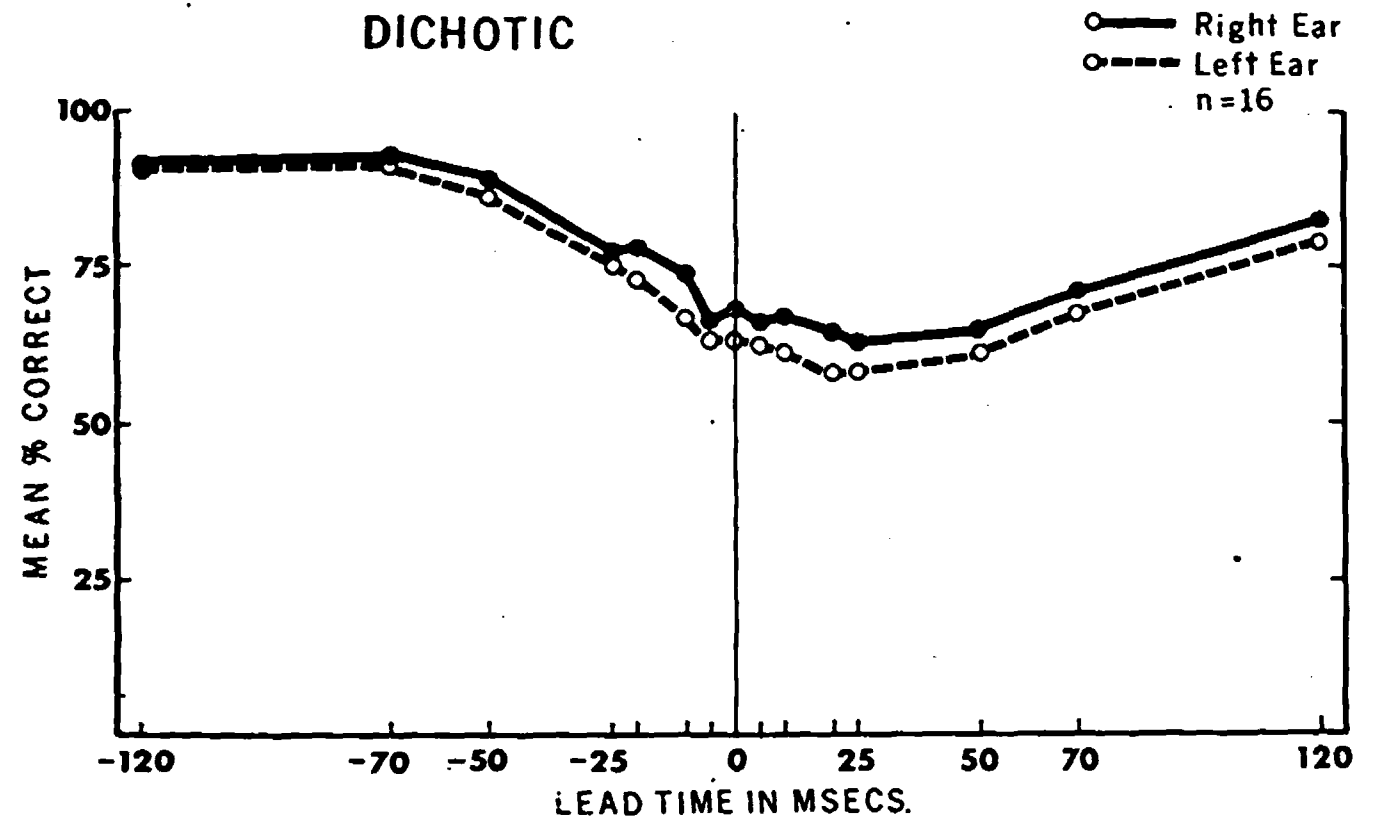


FIG. 2

syllable with a superimposed click.

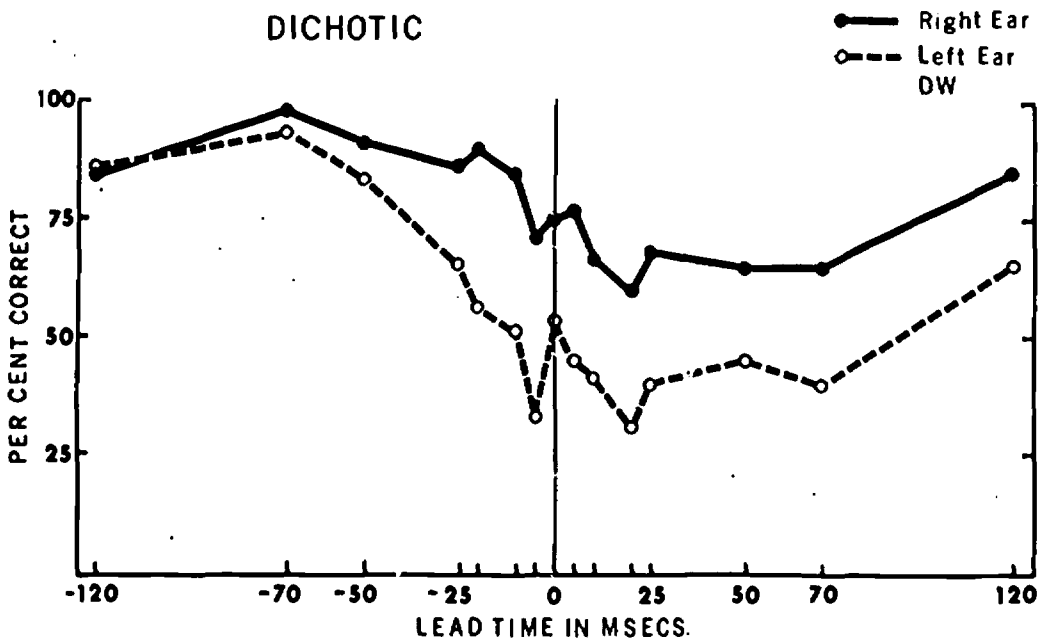
The dichotic results present a quite different picture. Figure 2 displays the group dichotic results plotted on the same coordinates as Figure 1. On this plot, the difference between levels for left and right ears is a measure of the ear advantage (laterality effect), and the slopes of the functions from their minima measure the advantages accruing from changes in lead or lag time. Where the two functions are parallel, laterality effect and temporal effects are additive; significant deviations from the parallel indicate some interaction between the two effects.

We note first a clear laterality effect: right ear performance is superior to left at every lag/lead value other than -120 msec. Ten of the sixteen subjects show significant right ear advantages by matched pair t-tests over the lag/lead range; four show no significant ear advantage; two show significant left ear advantages. Subject by ear interaction is significant by analysis of variance, hence the overall ear effect is not significant. Individual differences of this order are common in dichotic experiments and may be related to differences in cerebral language dominance. Figure 3 gives some idea of the variability: examples of a clearly right-eared subject (above) and of a subject showing no significant ear advantage (below).

Second, we note that increases in the amount of lag yield, for both ears, increases rather than decreases in performance. Furthermore, the functions are not symmetrical: they reach their minima at lead values of 20 or 25 msec, rather than at zero; they reach their maxima at a lag value of -70 msec, where performance is superior by some 20% to performance at the corresponding lead value. In other words, the functions climb more rapidly over the lag than over the lead range. And this is true of every subject, despite considerably greater intersubject variability in the dichotic than in the monotic data. The overall effect of temporal offset is highly significant by analysis of variance, and there is no significant subject by temporal offset interaction.

The advantage of the lagging over the leading syllable may be more clearly seen if we replot the data of Figure 2 so that each pair of points shows the mean percent correct by ear for all trials of a given type. For example, the pair of points at the extreme left in Figure 4 gives performance for trials on which the left ear lagged by 120 msec, and the corresponding pair at the extreme right gives performance for trials on which the right ear lagged by 120 msec. (Figure 4 may be generated by rotating the right ear function of

Percent correct by ear for two subjects on dichotically presented CV syllable pairs as a function of temporal lead in milliseconds.



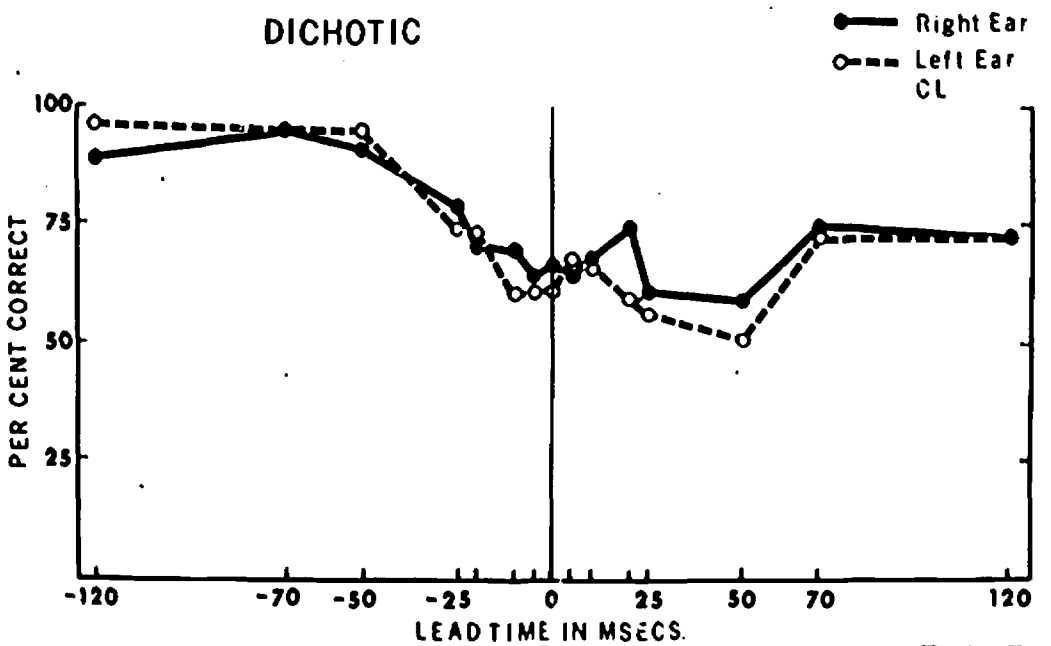


FIG. 3

Mean percent correct by ear on left lag and right lag CV syllable pairs, dichotically presented. For sixteen subjects.

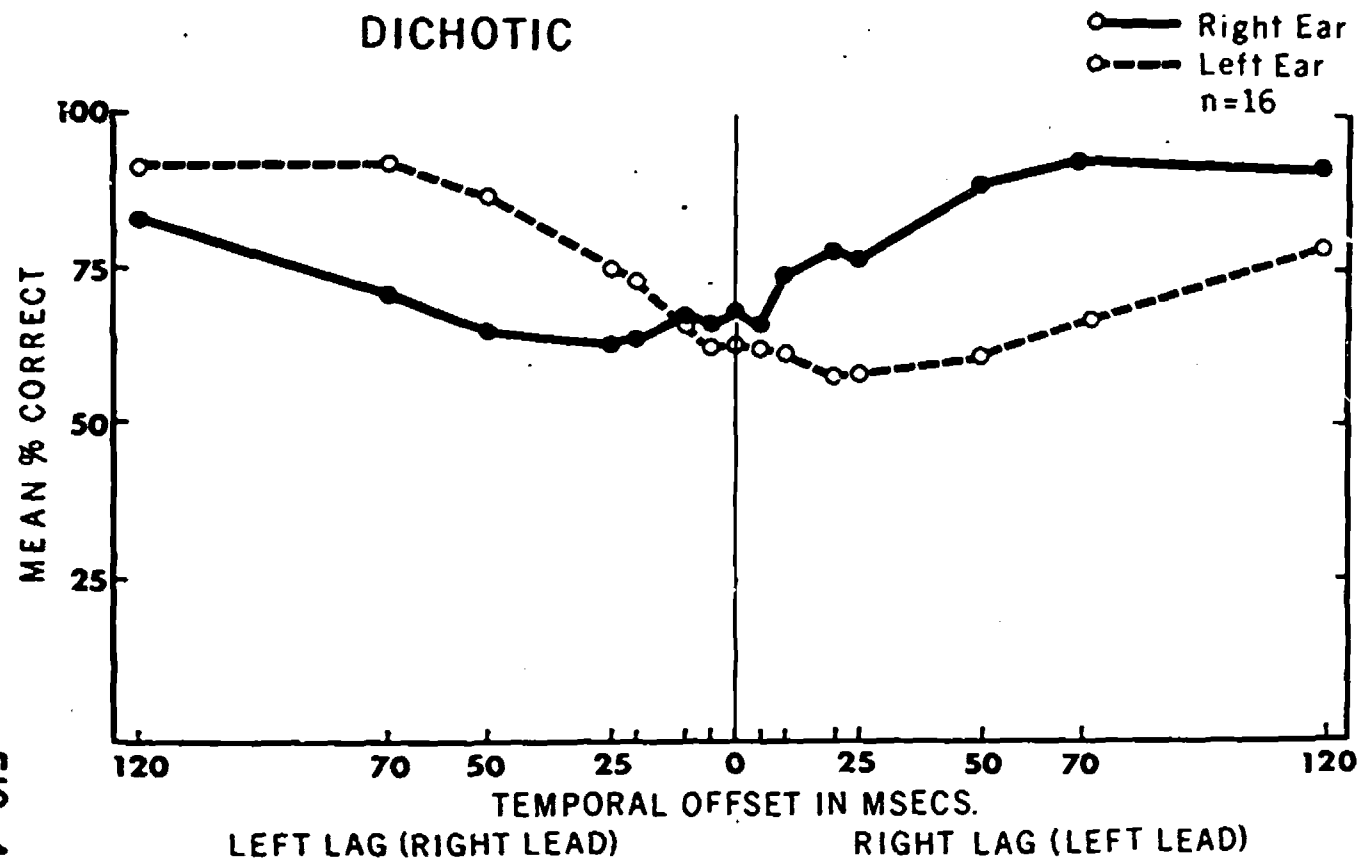


FIG. 4

Figure 2 through 180 degrees in a plane vertical to the page.) We see immediately that the ear to which the lagging syllable is presented almost invariably has the advantage over the leading ear. The exception is over the short, left ear lags (0-10 msec), where the right ear advantage under dichotic stimulation is sufficient to cancel the left ear advantage from lag. In fact, for these group data, 10 msec is the titration value that we originally sought, that is, the temporal advantage to the left ear necessary to cancel the dichotic advantage to the right. However, the value is not reliable across subjects.

Cancellation of the right ear advantage by an appropriate left ear lag suggests that the laterality and lag effects are independent phenomena. The same conclusion is suggested by the asymmetry of Figure 4: the wider separation of the two curves over the right lag range than over the left is due to the fact that the right ear, whether leading or lagging, has an overall higher level of performance than the left. The generally parallel courses of the two curves in Figure 2 makes the same point, and analysis of variance shows no significant interaction between ear and temporal offset.

We may now pose the problem raised by the dichotic lag effect fairly precisely. From Figure 2 it is evident that there is little variation in performance between -5 and +50 msec; within this range, the functions for both ears reach a broad minimum. For fifteen out of sixteen subjects this is the range within which both ears reach their minima; the sixteenth subject gives her minima at +70 msec. Thus, for every subject, dichotic performance is at its worst in the very range of lead values over which monotic performance is at, or rising to, its peak. The paradox sharpens when we recall that the conditions of presentation for the leading portion of the leading syllable are identical under monotic and dichotic presentation. For example, under both conditions, the initial 50 msec of a syllable leading by that amount are presented without interruption to a single ear. These 50 msec carry all the information needed for identification of the initial consonant, and under monotic conditions, virtually perfect identification is achieved by every subject, while performance on the syllable that lags by 50 msec drops to chance. Under dichotic conditions, on the other hand, performance on the leading syllable is, for every subject, close to her function minimum and on the lagging syllable close to her function maximum.

What gives rise to this reversal of the direction of the effect under dichotic conditions? The question is of interest for the light that its answer may throw on the processes of speech perception. For while the monotic lead

effect may be interpreted as an instance of peripheral, simultaneous masking, the dichotic lag effect seems to be of central origin, possibly analogous to metacontrast effects in vision. Werner (1935) showed that perception of a disc flashed on a screen might be blocked if rapidly followed by presentation of a ring having the same internal diameter as the disc. He attributed the effect to interference by the ring with development of the disc's contour. Later work (for example, Kolers and Rosner, 1960) showed that a similar effect might be obtained dichoptically and hence, that it involved central mechanisms.

Interpretation of the dichotic lag effect along analogous lines would assume processing of the important cues in the leading stimulus to be incomplete at the time the lagging stimulus arrived along a different channel to compete for, and frequently capture, the processors. Occlusion of the leading syllable by a switch in channels just as the crucial information in that syllable is being processed recalls the finding of Huggins (1964) that the rate of across-ears switching most disruptive to speech perception is roughly equal to the syllable rate. A similar disruption does not occur when the lagging syllable is presented along the same channel in wake of the first, presumably because it is masked at a peripheral point in the pathway.

The notion that the lag effect reflects interruption of speech processing is further suggested by control data. Studies with nonspeech have not yet been completed, but Porter, Shankweiler, and Liberman (1969) have reported that, if the stimuli are steady-state synthetic vowels, the advantage tends to the leading, rather than to the lagging, stimulus. Given that such stimuli have been found, under other experimental conditions, to be perceived in the manner more of nonspeech than of speech (Liberman et al., 1967; Shankweiler and Studdert-Kennedy, 1967; Studdert-Kennedy and Shankweiler, 1970; Studdert-Kennedy et al., 1970), we may reasonably suspect that the lag effect is tied to speech and, specifically, to those components of the speech stream for which a relatively complex decoding operation is necessary.

However, an adequate account of the effect and of its implications for speech perception calls for much further study. Several experiments are already under way at Haskins Laboratories. These include studies of individual differences, nonspeech controls, attention switching, channel tracking, and consonant feature errors.

References

Berlin, C.I., Willett, M.E., Thompson, C., Cullen, J.K., and Lowe, S.S. (1970) Voiceless versus voiced CV perception in dichotic and monotic listening. J. Acoust. Soc. Amer. 47, 75 (A).

Huggins, A.W.F. (1964) Distortion of the temporal pattern of speech: interruption and alteration. J. Acoust. Soc. Amer. 36, 1055-1064.

Kolers, P. and Rosner, B.S. (1960) On visual masking (metacontrast): dichoptic observation. Amer. J. Psychol. 73, 2-21.

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

Lowe, S.S., Cullen, J.K., Thompson, C., Berlin, C.I., Kirkpatrick, L.L., and Ryan, J.T. (1970) Dichotic and monotic simultaneous and time-staggered speech. J. Acoust. Soc. Amer. 47, 76 (A).

Porter, R., Shankweiler, D., and Liberman, A.M. (1969) Differential effects of binaural time differences in perception of stop consonants and vowels. Proc. 77th Annual Convention of the Amer. Psychol. Assn.

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

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

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

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

Discrimination in Speech and Nonspeech Modes*

Ignatius G. Mattingly,+ Alvin M. Liberman,++ Ann K. Syrdal+++ and Terry Halwes+
Haskins Laboratories, New Haven

Abstract. Discrimination of second-formant transitions was measured under two conditions: when, as the only variation in two-formant patterns, these transitions were responsible for the perceived distinctions among the stop-vowel syllables [bæ], [dæ], and [gæ]; and when, in isolation, they were heard, not as speech, but as bird-like chirps. The discrimination functions obtained with the synthetic syllables showed high peaks at phoneme boundaries and deep troughs within phoneme classes; those of the nonspeech chirps did not. Reversal of the stimulus patterns, producing vowel-stop syllables in the speech context and mirror-image chirps in isolation, affected the speech and nonspeech functions differently. An additional nonspeech condition, presentation of the transitions plus the second-formant steady state, yielded data similar to those obtained with the transitions in isolation. These results support the conclusion that there is a speech processor different from that for other sounds.

For many years, the authors and their colleagues have been interested in the differences in perception between speech and other sounds. That a difference exists is suggested first by the nature of the relation between the perceived phonetic message and the acoustic signal that conveys it: message and signal are linked by a complex code for which there is no parallel in any class of nonspeech sounds; we therefore infer that speech perception is accomplished by a special decoder (Liberman et al., 1967). This complex speech code is not unique but is, rather, similar in form to the grammatical codes at the higher levels of language: syntax and phonology (Mattingly and Liberman, 1969).

These inferences are supported by experimental results that point more directly to a special mode of perception for speech and suggest that this mode is related to a still broader one that characterizes perception of language in general. Numerous experiments on dichotic listening indicate that the encoded sounds of speech (like the higher levels of language) are normally processed

---

*This paper incorporates data, some of which have been reported earlier by Mattingly et al. (1969), Syrdal et al. (1970) and Liberman (in press).

+Also, University of Connecticut, Storrs.

++Also, University of Connecticut, Storrs, and Yale University, New Haven.

+++Also, University of Minnesota, Minneapolis.

primarily in the left hemisphere of the brain, while nonspeech sounds and the relatively unencoded aspects of speech (such as steady-state vowels) are either processed in the right hemisphere or are not lateralized at all (Kimura, 1964, 1967; Kirstein and Shankweiler, 1969; Shankweiler and Studdert-Kennedy, 1967; Studdert-Kennedy and Shankweiler, 1970).

Other experimental observations imply additional differences in perception between speech and nonspeech. One such observation, which is particularly relevant to the experiments to be reported here, is that the encoded acoustic cues sound very different in and out of speech context. Though the difference has not been precisely measured, its existence is clear enough. When transitions of the second formant, which are sufficient cues for the place distinctions among stop consonants, are presented in isolation, we hear them as we should expect to--that is, as pitch glides or as differently pitched "chirps." But when they are embedded in synthetic syllables, we hear unique linguistic events, [bɑ], [dɑ], [gɑ], which cannot be analyzed in auditory terms. Thus, speech perception cannot be straightforwardly mapped onto the physical dimensions of the speech signal.

There is a more specific sense in which speech perception does not correspond to acoustic reality. If asked to discriminate physically continuous variations in a speech cue, a listener does not hear a continuum of sounds but, rather, quantal jumps from one sound to another. His discrimination function displays high peaks at phonetic boundaries. These high peaks (and the adjacent troughs) reflect a kind of perception in which the listener hears phonetic units but not intraphonetic variations. In the extreme case, he discriminates no more stimuli than he can absolutely identify. Perception of this sort has been called "categorical" (Liberman et al., 1957; Studdert-Kennedy et al., 1970); it is unusual, if not unique, since, in the perception of nonspeech sounds, many more stimuli can be discriminated than can be identified. Of course, categoricalness is a property of language generally: active/passive and singular/plural, for example, do not admit of degree.

In this paper we shall make use of categorical perception to study the difference between speech and nonspeech. To capture the difference as directly as possible, we will compare listeners' discrimination of the same acoustic variable, once in speech context, where it serves as a cue for a phonetic distinction, and once in nonspeech, where it does not.

Several such comparisons have already been made. In one of these studies (Liberman et al., 1961b), the acoustic variable was the "cutback" or delay of onset of the first formant, which in initial position is a major cue to the

voiced/voiceless distinction. The speech-like stimuli were made on the Haskins Pattern Playback from a series of spectrographic patterns with increasing delay in the onset of the first formant (F1) relative to the onsets of the second and third formants (F2 and F3). Stimuli for which the delay was sufficiently long were heard as [to], other stimuli as [do]. The nonspeech control stimuli were synthesized from inverted versions of these same spectrographic patterns. Thus, the same inf mation was presented in both speech-like and control stimuli, but the control stimuli did not sound like speech. The inversion, however, affected the acoustic variable itself; as the authors point out,

> in the control stimuli the formant whose time of onset varied was at a higher frequency than the other two formants, while in the speech stimuli it lay at a lower frequency than the other formants.

Subjects were asked to identify the speech stimuli as [to] or [do] and, in the case of both speech and control stimuli, to discriminate between neighbors along the acoustic series. For a typical subject, the speech discrimination function showed a peak at a delay of 20-30 msec, corresponding to the phonetic boundary predicted by the cross-over point of the two identification functions, while the control discrimination function showed no such peak and, in fact, never rose very far above the chance level.

In the other study (Liberman et al., 1961a), the acoustic variable was the length of the silent interval associated with stop consonants; in intervocalic position, this length is a cue to voicing. The speech-like stimuli were synthesized from a series of spectrographic patterns representing a word containing a medial stop, with a silent interval of increasing length: stimuli for which the interval was sufficiently long were heard as rapid, other stimuli as rabid. Each control stimulus consisted of two bursts of band-limited white noise, with the same durations and energy envelopes as the two syllables of a speech stimulus and separated by a silent interval. The silent intervals matched those of the speech stimuli. As in the [to]/[do] study, subjects were asked to identify the speech stimuli and to discriminate the speech from the control stimuli. The speech discrimination functions showed no peaks and were, in general, lower than the speech functions but subtantially higher than chance.

Both of these studies indicated that perception of the relative timing of two acoustic events was different if the difference in timing cued a distinction between two speech sounds. In the case of speech, there were peaks in the discrimination functions at the phonetic boundaries; in the case of nonspeech, there were not. Moreover, the results indicated that the peaks in the speech

represented, by comparison with nonspeech, a sharpening of discrimination at phonetic boundaries, not a reduction of discrimination within the phonetic category. Although these results are suggestive, their interpretation is complicated by the fact that the acoustic variable was the same for speech and nonspeech only in a derived sense: the time intervals between two sounds were identical, but the sounds themselves were different.[1]

The purpose of the two experiments reported here was to provide a more appropriate nonspeech context for comparison with speech. To that end, we examined the perception of the second-formant transition. Unlike the timing cues of the earlier studies, the second-formant transition is itself an actual acoustic event. The problem of devising an appropriate nonspeech control context thus becomes much more straightforward. In fact, it is possible to use the simplest context of all: isolation. As we have noted, second-formant transitions distinguish [b], [d], and [g] in speech context but sound in isolation like chirps.[2]

EXPERIMENT I

The purpose of the first experiment was to compare the discrimination of F2 transitions in stop-vowel syllables and in isolation.

---

[1] An interesting and somewhat relevant experiment, in which the speech and nonspeech context were determined not by the stimuli but by the experimenters' instructions to the subjects, has been carried out by Cross and Lane (1964). Presented with synthetic speech stimuli of marginal realism, one group of subjects was told that they were being tested in speech/sound discrimination, while another group was told that the test had to do with discrimination of tones. The discrimination functions obtained with the first group showed peaks at the phonetic boundaries; the discrimination functions for the other group did not.

[2] This method of comparing speech and nonspeech was suggested by Kirstein's (1966) pilot study, in which she used isolated second formants with an initial transition and a following steady state--what we have called "bleats" in this paper. In an experiment applying detection theory to categorical perception, Popper (1967) included a same/different discrimination using bleats with final transitions in the [b]/[d] range. The d' functions obtained can be compared with that for a test with the same subjects using speech stimuli. The results of both Kirstein and Popper are consistent with the results reported here.

Method

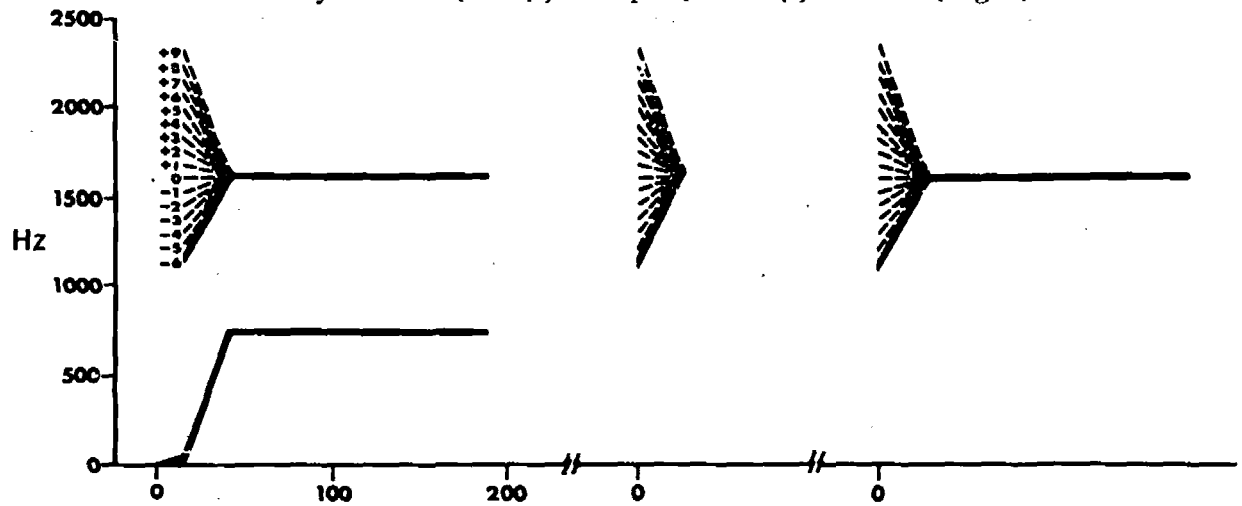
Stimuli. The Haskins Laboratories computer-controlled synthesizer (Mattingly, 1968) was used to produce the stimuli of the experiment. A stimulus to be synthesized is specified by time functions for each of the several parameters of the synthesizer (e.g., F0, the fundamental frequency; F1, the first-formant frequency; and so on). Each of these functions is represented by a series of digital values stored in computer memory. To produce the stimulus, a set of values, one for each parameter, is transmitted every five msec by the computer to the synthesizer.

The two sets of stimuli used in Experiment I are shown in Figure 1, top left and center. (The other stimuli shown in Figure 1 were used only in Experiment II.) The set at top left are speech stimuli and consist of sixteen syllables, each beginning with a voiced stop and ending with the vowel [œ]. In all the syllables, the fundamental frequency is constant at 90 Hz, and only the first and second formants of the synthesizer are used. A 15-msec period of closure voicing, represented by a low-amplitude F1 at 150 Hz, is followed by a 40-msec transitional period during which the two formants move toward the steady-state frequencies appropriate to [œ]: F1 = 740 Hz, F2 = 1620 Hz. The steady-state period of the stimulus is 190 msec long. Throughout the stimulus, the two formants are of equal amplitude. The F1 transition always starts at 150 Hz. The experimental variable is the starting point of the F2 transition. This is varied in fifteen approximately equal steps from 1150 to 2310 Hz. In Figure 1, top left, the level transition, for which the starting point is 1620 Hz, is labeled 0; transitions with higher (or lower) starting points are labeled positively (or negatively) with reference to the level transition. Depending on the starting point (and therefore the slope) of the F2 transition, these stimuli are heard as [bœ], [dœ], or [gœ].

The second set of stimuli (Figure 1, top center) are the nonspeech controls. They consist simply of transitions identical to those of the first set but with the closure voicing, the steady state of F2, and all of F1 absent. In the first set--that is, in the syllables--the transitions were the only cues to the point of articulation. In the second set, the transitions have been removed from their speech contexts and do not sound at all like speech. To most listeners they sound like chirps, and it is not hard, at least in the case of the more extreme members of the set, to tell whether a chirp is rising to a higher or falling to a lower frequency.

Stimuli for Forward Condition, with Initial Transitions

Syllables (left), Chirps (center), Bleats (right)



Similar Stimuli for Backward Condition, with Final Transitions

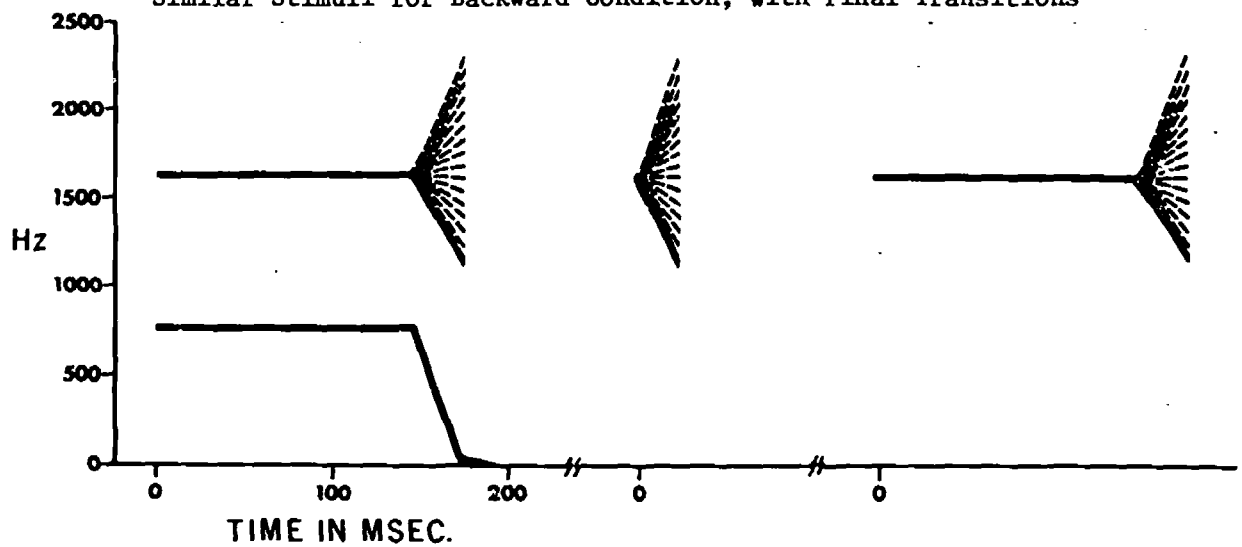


FIG. 1

Procedure. With the aid of the synthesis system, the digital parametric representations of all the stimuli were stored on a disc file, and the tests required for the various experiments were then automatically compiled and recorded The test included an identification test for the syllables and discrimination test for the syllables and for the chirps.

The purpose of the identification test was to determine where, and how reliably, the subject placed the phonetic boundaries. It consisted of 160 syllables in ten groups of 16. Each of the different syllables occurred once in each group, and each group was differently randomized. The subject's task was to identify each of the 160 syllables as beginning with [b], [d], or [g].

To find out how well the subjects could discriminate the stimuli, we used an oddity method: each item in the test consisted of a triad in which one member of a pair of stimuli to be discriminated occurred once and the other, twice; the subject's task was to select the odd stimulus. For each pair there are six ways in which a triad can be ordered. Pairs of stimuli two steps apart along the continuum of Figure 1 were to be discriminated; for each set there are fourteen such pairs. Each test consisted of the eighty-four possible triads in six groups of fourteen. Each stimulus pair was used to form one triad in each group. The assignment of the six triad orderings to the six groups was separately randomized for each group. There were four differently randomized forms of the discrimination test. The tests for syllables and chirps were made in the same way.

The tests were presented to the subjects over headphones. The gain on the tape recorder was set so that subjects could listen to the syllable stimuli comfortably; this same gain setting was used for the chirps.

For each subject, there were five experimental sessions on five separate days. On each day, the subject was given different forms of the discrimination test for the syllables and different forms of the discrimination test for chirps, in random order. Altogether, he received all four forms of the syllable discrimination test twice and all four forms of the chirp discrimination test twice. Thus, for each stimulus comparison, each subject gave forty-eight judgments. The identification test was given once on each of the first, second, fourth, and fifth days. Each stimulus was presented for judgment ten times on each identification test; there was then a total of forty judgments per stimulus.

Subjects. There were seven subjects, all undergraduate students at the University of Minnesota and all paid volunteers. None was told the purpose of the experiment.

Results

In Figure 2 are the results for two of the seven subjects, chosen on a basis to be described later. The upper portion of the block for each subject plots his identification functions for [b], [d], and [g]: the abscissa represents the stimuli ordered according to the series of F2 starting points; the ordinate represents the percentage of responses for each of the three stops.

Both of the subjects shown sorted the stimuli cleanly into the three phonetic categories. The areas of uncertainty are small by comparison with those where the subjects apply the phonetic labels with consistency. Of the seven subjects, six yielded identification functions approximately as reliable as those shown; moreover, agreement among the subjects in the location of the phonetic boundaries is almost perfect. One subject did very poorly on the identification; he labeled stimuli inconsistently, and there was substantial overlap in the identification functions for the three stops. We have rejected all the data from this subject because we suspect that he did not hear the synthetic patterns very well as speech; if he did not, then a comparison of the way he perceived speech and nonspeech stimuli, which is the purpose of this experiment, becomes meaningless.

The lower portion of each block in Figure 2 plots the subject's discrimination functions for syllables (solid line) and for chirps (dashed line). Each point along the abscissa corresponds to the stimulus pair whose members are the stimuli one step higher and one stop lower in the series than the stimulus represented by the corresponding point in the abscissa of the identification test plot. The ordinate is the percentage of correct discriminations for each pair. The horizontal broken line at 33% represents the level of discrimination expected by chance.

For the syllables, the discrimination function shows peaks near the phonetic boundaries indicated by the identification functions for each subject. Since the boundaries are constant from subject to subject, the locations of the peaks are likewise constant. The peaks for [b]-[d] boundaries are generally somewhat higher than those for [d]-[g] boundaries. Away from phonetic boundaries, the discrimination functions are at or near chance.

The chirp discrimination functions are quite different. There are no peaks in discrimination at points corresponding to the phonetic boundaries. Both subjects have a peak at +6, but we believe that this is to be attributed to an artifact resulting from a previously unremarked shortcoming of the synthesizer: its pitch generator was free-running, so that the occurrence of the first pitch pulse of a chirp (or indeed, of any other stimulus) could lag by as much as half

Identification and Discrimination: Two Subjects

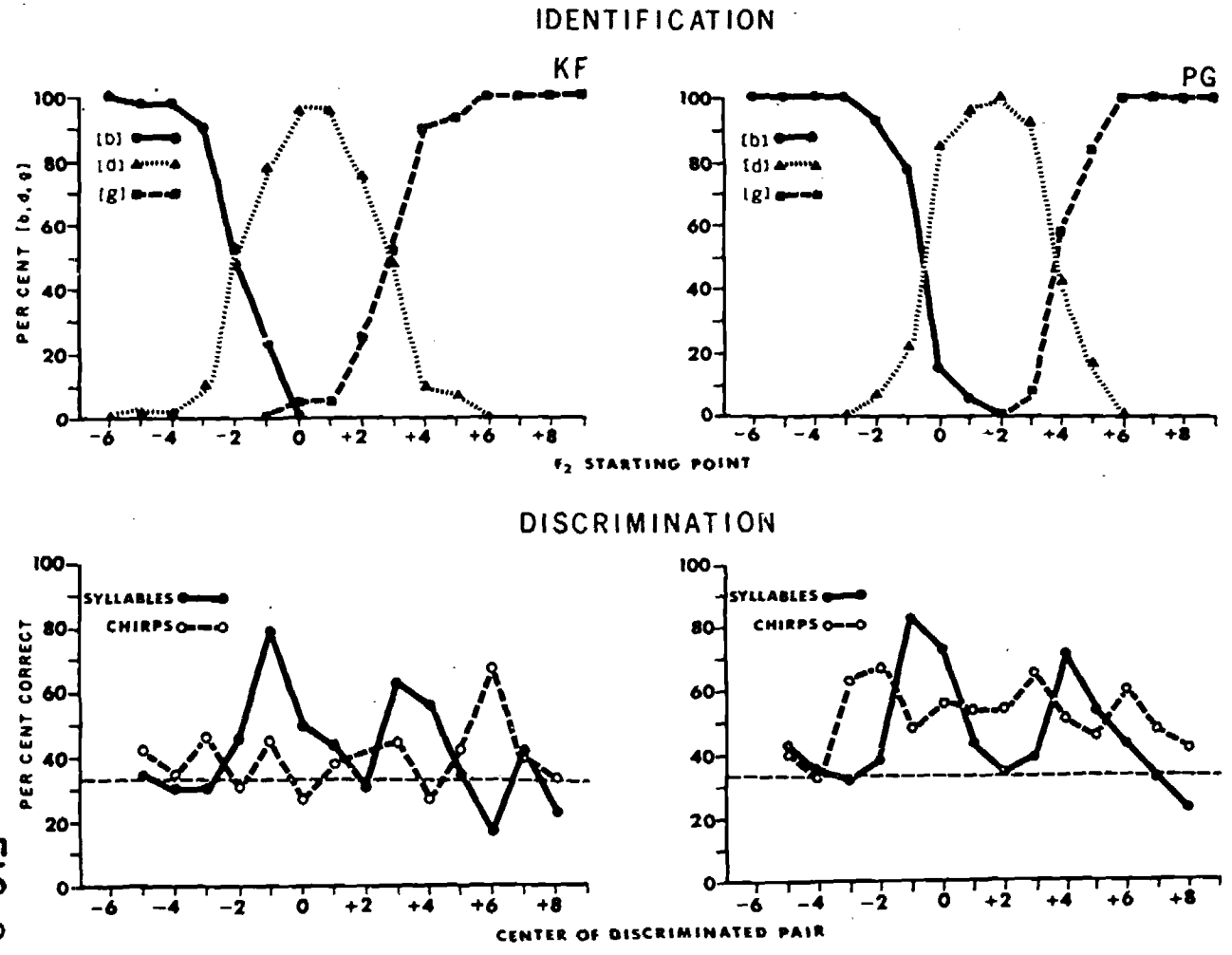


FIG. 2

a pitch period (6.5 msec) behind the nominal starting point. The synthesizer parameter values change stepwise; for the more extreme stimuli, for which F2 moves rapidly, there would, therefore, be substantial variation in the actual initial frequency, as well as in the duration, of the transition in the different tokens of the "same" stimulus. Such variation was, of course, randomized across these several tokens. However, inspection of the tokens for Stimuli +5 and +7 (discrimination of which produced the peak at +6) reveals that the variations were unbalanced in such a way that careful listeners could discriminate accurately on the basis of differences in duration or exaggerated differences in F2 starting point. That this is, in fact, the cause of the peak is indicated by the results of later experiments in which we synchronized the pitch pulses and the peak at +6 disappeared.

The two subjects whose results are shown in Figure 2 were chosen to illustrate the extremes in the general level at which the chirps were discriminated. One of them (KF) discriminates the chirps at a level only slightly above chance, except at +6; the other (PG) does considerably better. In general, the variation among subjects in level of discrimination, and also in the shape of the function, was greater for the chirps than for speech.

In Figure 3 is a plot of the pooled discrimination data of the six (out of seven) subjects who identified the syllables well. The chirp discrimination function and the syllable discrimination function are clearly different. The chirp function is low (except for the peak at +6) but above chance. The syllable function shows peaks near phonetic boundaries and is at or near chance away from phonetic boundaries. The subjects' perception of the second-formant transition apparently depends on whether they are listening in the speech mode.

EXPERIMENT II

The second experiment was prompted by the observation, made in one of the studies with synthetic speech, that the F2 transition is a less powerful cue to place of articulation in final position than in initial position (Liberman et al., 1954). Though this difference reflected directly only the relative difficulty of identifying the transitions, it is reasonable to suppose that discrimination might also be different in final than in initial position. Preliminary experiments have since suggested that this is so. As with so many findings in speech perception, the question arises whether this difference is to be accounted for psychoacoustically or whether it is, rather, a consequence

Pooled Discrimination Function Data from Experiment 1

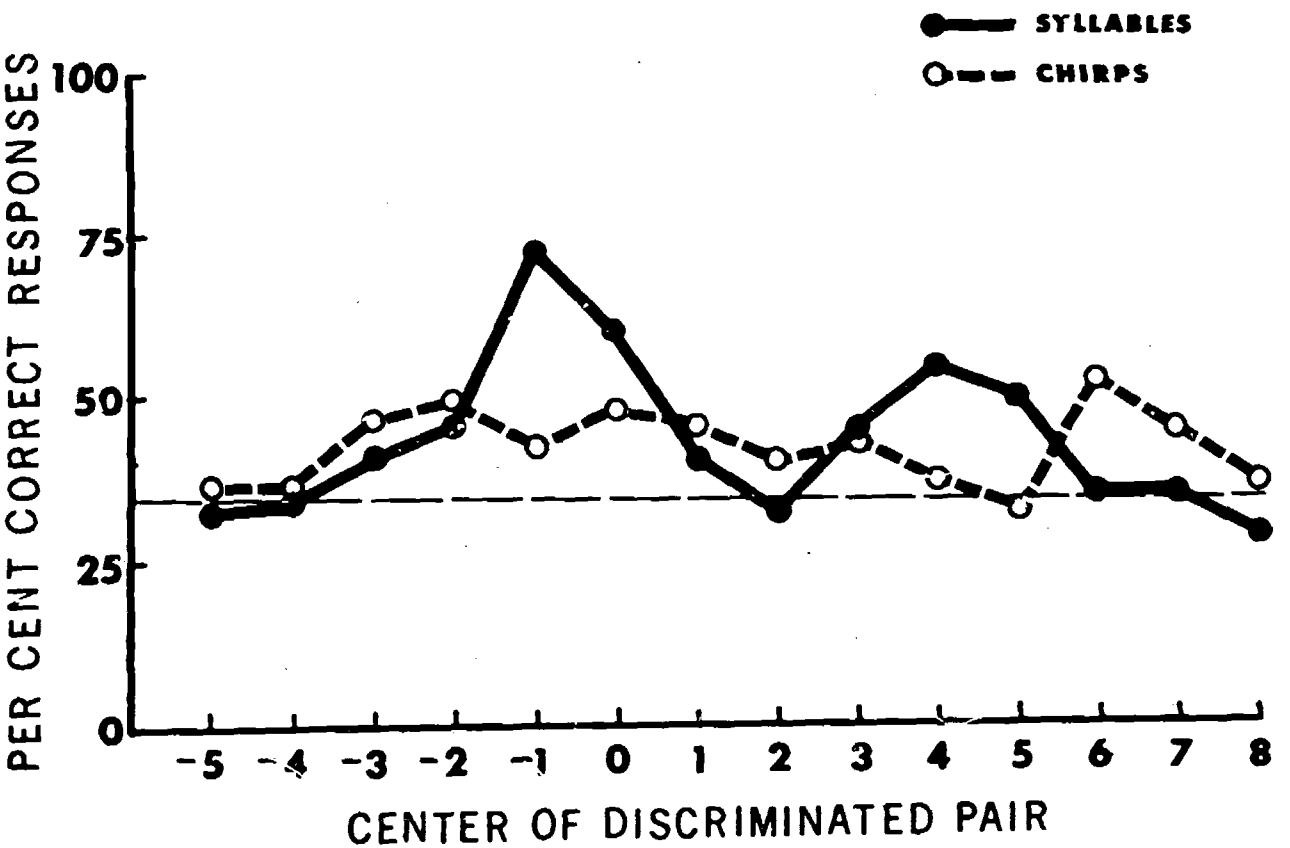


FIG. 3

of the special processing that the speech signal undergoes. If the explanation is psychoacoustic, then we should expect that the F2 transitions in nonspeech context--that is, the chirps--would also be differently discriminated in final position. The second experiment was designed to provide relevant data. In it we have compared the discriminability of the F2 transitions in initial and final positions when they are, in one condition, cues for speech and, in another, not.

The second experiment was intended also to determine whether possible reservations about the chirp control are justified. It might be argued that this control is faulty: when the F2 is in initial position in the syllable, the vowel steady state may provide a reference that is, of course, absent in the chirp. When F2 is in final position in the syllable, as in this experiment, the steady state may provide a reference and, conceivably, a fatigue effect. Therefore, we introduced in Experiment II an additional set of nonspeech control stimuli (Figure 1, top right). These stimuli have not only the various second-formant transitions, as do the chirps, but also the second-formant steady state. Naive subjects do not commonly hear these as speech. We have called them "bleats."

Six sets of stimuli were required for the experiment: F2 transitions in initial and final positions in two-formant syllables; F2 transitions in isolation in "initial" and "final" positions (chirps); and F2 transitions attached to steady-state second formants in initial and final positions (bleats). The syllables and chirps with initial F2 transitions were produced as in Experiment I; the bleats with initial transitions were produced by synthesizing two-formant syllables with F1 turned off. The production of the stimuli was better controlled than in Experiment I. The synthesizer was made to produce its first pulse at the start of every stimulus, instead of randomly, so that each token of a stimulus had exactly the same duration and frequency excursion, thus eliminating the basis for the pile-up of correct discriminations at +6 in the first experiment. It was not necessary to produce separate sets of stimuli with final F2 transitions, since these stimuli (Figure 1, bottom) were equivalent to the available stimuli in reverse temporal order. Thus, tests requiring stimuli with initial transitions were run by playing the test tapes <u>forward</u>; tests requiring stimuli with final transitions were run by playing these same tapes <u>backward</u>.

Procedure. The formants of the identification test (for the syllables) and the discrimination test (for the syllables, chirps, and bleats) were the same as in Experiment I. The subject's task included the oddity judgment (selecting the one stimulus of each triad that he thought different from the other two) used in Experiment I and, in addition, a confidence rating. For the purposes of the confidence rating, the subject was asked to estimate the correctness of each discrimination judgment on a three-point scale. These estimates were then treated according to a method developed by Strange and Halwes (in press) and successfully applied by them to increase the sensitivity of discrimination measures of the voiced/voiceless distinction. By their method, the confidence-rating score for each discriminated pair is determined by multiplying the number of correct responses for which the subject used a particular confidence rating by a weight assigned to this rating, summing these products over all ratings, and dividing by the number of trials per pair to give a number between 0 and 1. The weight is equal to $\frac{3p-1}{2}$, where p is the ratio, for all pairs in a given testing condition, of the number of correct responses for which a particular confidence rating was used to the total number of responses for which this rating was used. Thus the weight for a rating is 0 when the level of discrimination over all pairs is at chance ($p = 1/3$); and 1 when discrimination is perfect ($p = 1$). The advantage of the confidence rating is that is permits a reliable approximation of a subject's discrimination function with fewer responses per stimulus pair than if only the correctness or incorrectness of his responses is considered.

All subjects were given (1) the syllable identification test, once in the forward and once in the backward condition; (2) the syllable discrimination test, three forms forward and three forms backward; and (3) one of the two nonspeech discrimination tests, three forms forward and three forms backward. The chirps served as nonspeech controls for half the subjects, the bleats for the other half. For each subject, there were three separate test sessions on three separate days. Each chirp subject took a different form of each of the four discrimination tests (forward and backward, syllables and chirps) each day in a different random order. In the case of the bleat subjects, however, since the bleats were more like syllables than the chirps, we thought it wiser to protect the subjects' naivete by presenting all the bleat tests before all the syllable tests. During the first day and a half, therefore, each bleat subject took three forms of the forward and backward bleat tests; during the remaining

day and a half, he took three forms of each of the two speech tests. Thus, for each discrimination test, there were eighteen judgments per stimulus pair for each subject. The syllable identification test in the forward condition was given to all subjects at the end of the second day and the identification test in the backward condition at the end of the third day.

Subjects. There were eleven subjects, all undergraduate students at the University of Minnesota and all paid volunteers. None were told the purpose of the experiment. Three subjects were eliminated because of their inability to identify the syllables accurately. Data were provided, then, by eight subjects, four in each of the two experimental subgroups (chirps and bleats).

Results

In Figure 4 are the results for one typical subject in the syllable-chirp half of Experiment II. In the left-hand column are the results for the forward condition and in the right-hand column the results for the backward condition. The topmost graphs show his identification functions; the middle graphs, his discrimination functions without regard to his confidence ratings; and the lowest graphs, his discrimination functions, taking into account the confidence ratings.

Figure 5 shows syllable and chirp discrimination functions based on pooled data for all four subjects. The upper portion of the figure shows the forward condition; the lower portion, the backward condition.

For the forward condition, the results are consistent with the first experiment. Discrimination functions for syllables peak at the phonetic boundaries implied by the identification function but tend toward random elsewhere. Discrimination functions for chirps appear to have no relation to discrimination functions for syllables. The characteristic peaks and troughs of syllable discrimination are even more pronounced in confidence-rating analyses; on the other hand, the adventitious peaks of the chirp functions tend to be leveled. Still, chirp discrimination levels for all four subjects are clearly above random. One exceptional subject has a much higher overall level of chirp discrimination than that of the subject shown in Figure 4 (or, indeed, of any of the other subjects). That subject also has chirp peaks at the same points as his speech peaks, 0 and +3; in his confidence-rating analysis the peak at 0 becomes more pronounced by comparison with the one at +3.

Syllables, as expected, are much less consistently identified in the

Identification Function for Syllables and Discrimination Functions for Syllables and Chirps for One Subject

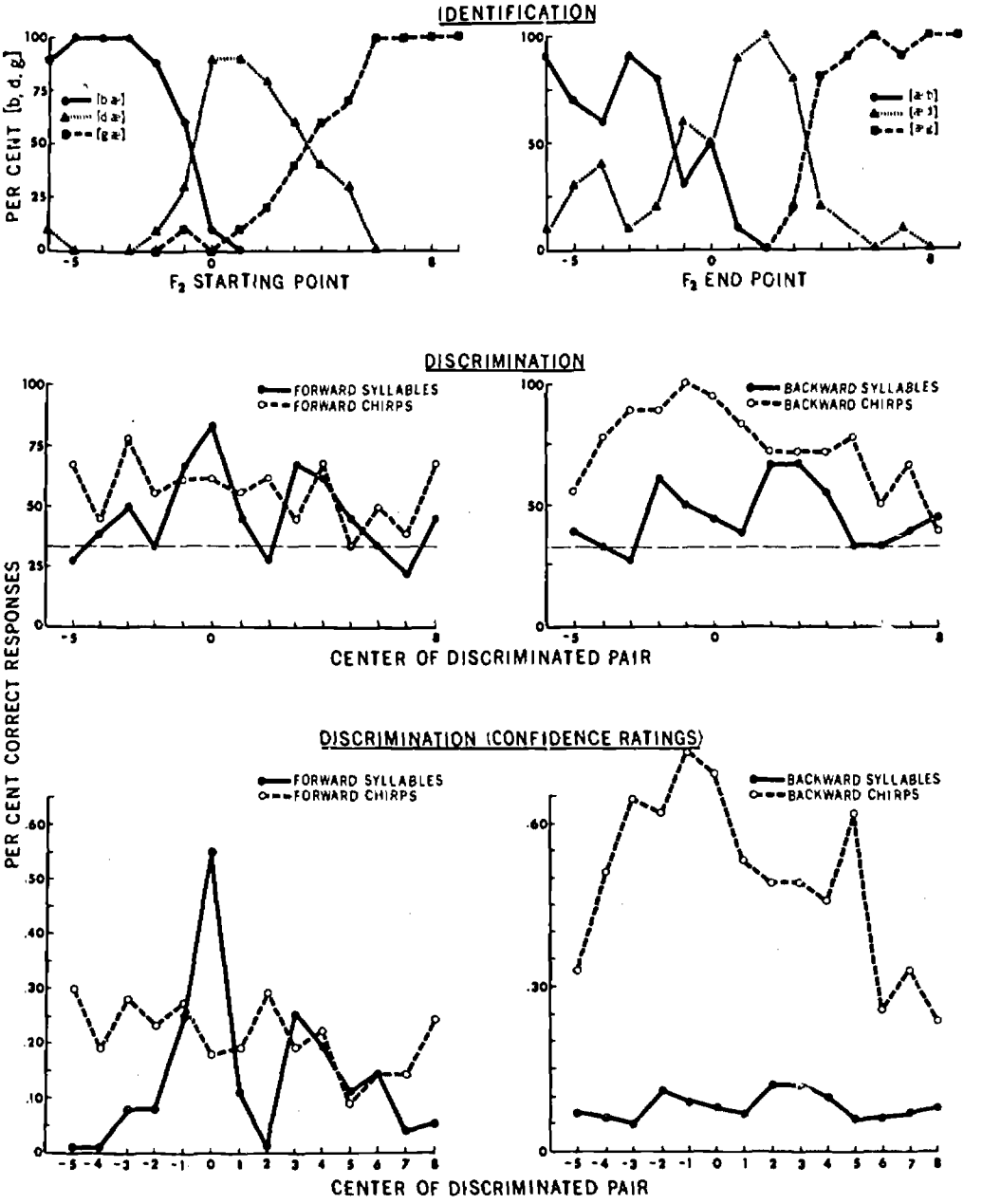


FIG. 4

Pooled Discrimination Function Data
for Four Subjects

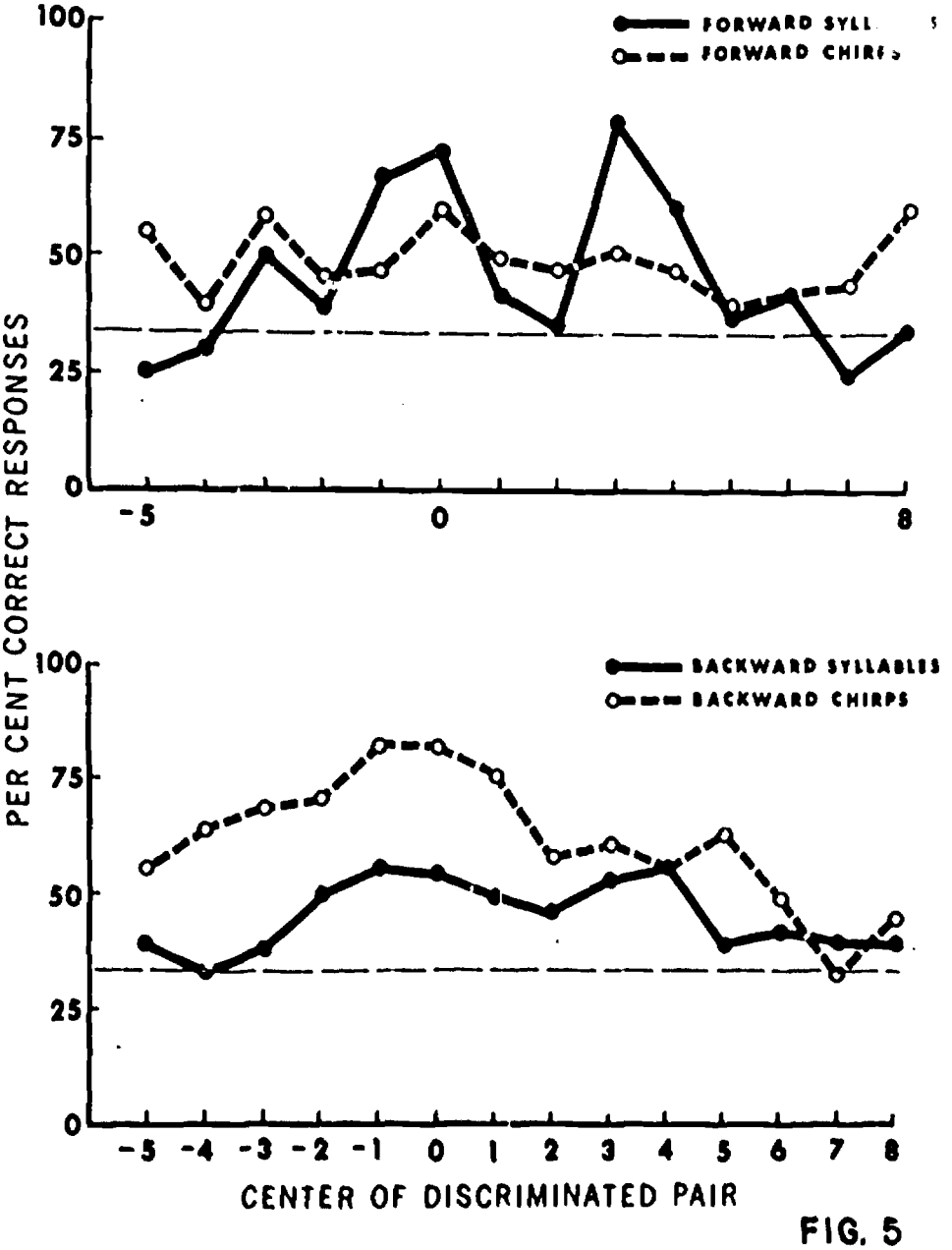


FIG. 5

backward than in the forward condition. There is also a certain tendency, shown by all subjects, for the cross-over point for [d]-[g] to move to the right, increasing the range over which subjects tended to hear [d]. Proper identification functions predicted lower peaks in the discrimination functions, and indeed, for the subject shown in Figure 4 and for all other subjects, syllable discrimination peaks are lower in the backward condition. The confidence-rating analysis accentuates this difference between the two conditions. But while the peaks are lower, the troughs are not as deep. The difference in both peaks and troughs is obvious in the pooled data of Figure 5.

Unlike the syllables, chirps are clearly much better discriminated in the backward than in the forward condition. This is true of all subjects, although the absolute level of performance varies among subjects just as in the forward condition. The discrimination functions for the backward chirps for two subjects are as good as their backward syllable discrimination functions, and for the two other subjects, including the one for whom data are given in Figure 4, the chirp functions are substantially better than the syllable functions at every point along the abscissa. All four backward chirp functions have their highest peak in the -1, 0, +1 range, but subjects tend to have idiosyncratic peaks elsewhere. The confidence-rating analyses emphasize the difference between forward and backward chirps and between backward chirps and backward syllables, and they accentuate the peaks near 0. The improved discrimination of chirps in the backward condition and the tendency to peak in the -1, 0, +1 range are apparent from comparison of the forward and backward chirp functions in Figure 5. In short, perception of chirps differs greatly from perception of syllables in the backward, as well as in the forward, condition; and the increase in discrimination induced by reversing the chirps does not appear to parallel the similarly induced change in perception of syllables.

The results for the bleat subjects are quite similar to those for the chirp subjects. In Figure 6 are the data obtained from a typical subject, arranged as in Figure 4. Pooled data for all four subjects, showing discrimination functions for syllables and bleats, are shown in Figure 7 (cf., Figure 5). Discrimination of syllables is high at phonetic boundaries, near random elsewhere; identification is more consistent and discrimination of the boundaries better in the forward condition. In fact, for these subjects, the backward syllable discrimination function has lost its bimodal shape and its characteristic troughs and looks not unlike the backward chirp function.

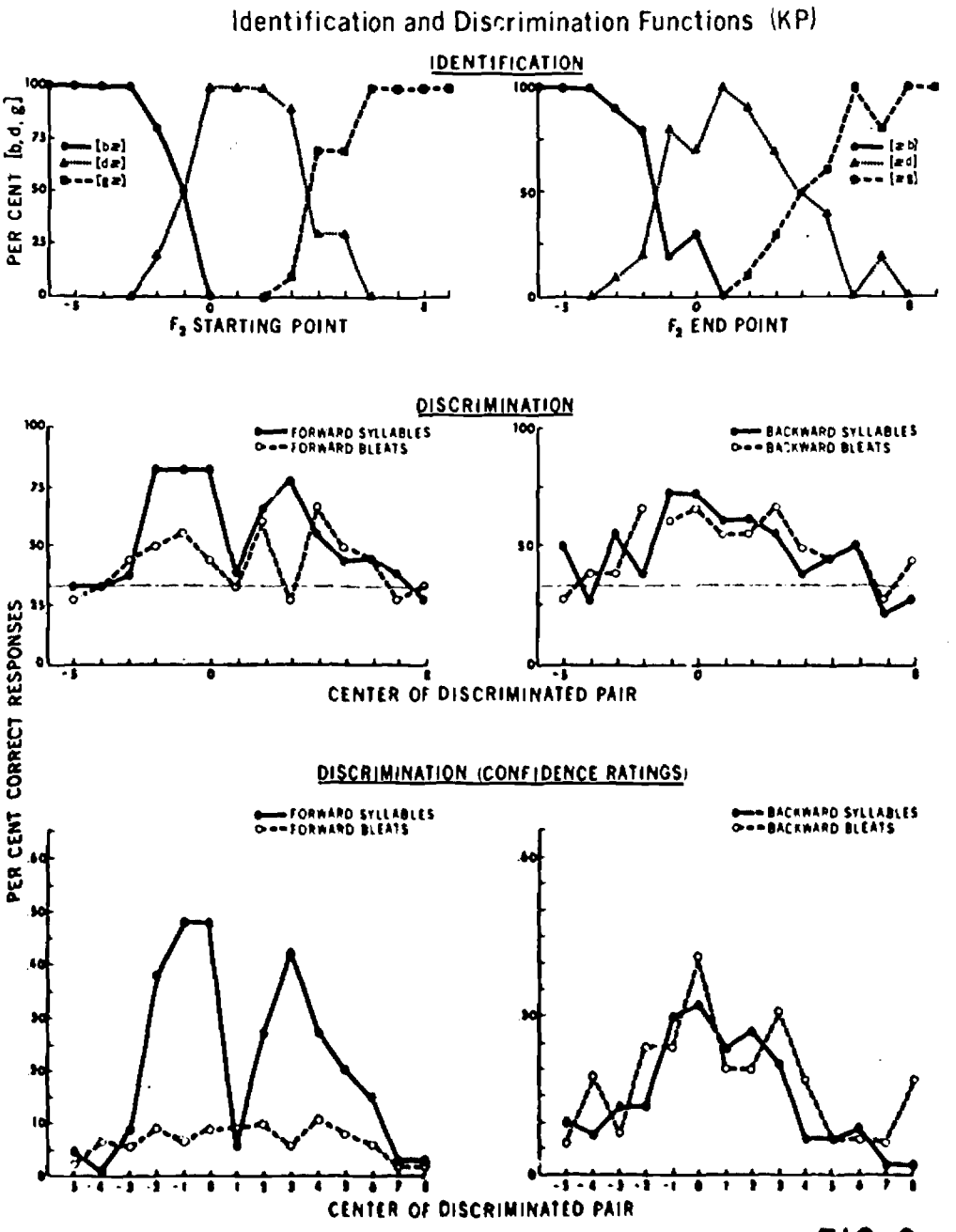

Identification and Discrimination Functions (KP)
IDENTIFICATION
PER CENT [b, d, g]
[ba]
[da]
[ga]
F2 STARTING POINT
F2 END POINT
DISCRIMINATION
FORWARD SYLLABLES
FORWARD BLEATS
BACKWARD SYLLABLES
BACKWARD BLEATS
PER CENT CORRECT RESPONSES
CENTER OF DISCRIMINATED PAIR
DISCRIMINATION (CONFIDENCE RATINGS)
CENTER OF DISCRIMINATED PAIR

FIG. 6

Pooled Discrimination Function Data
for Four Subjects

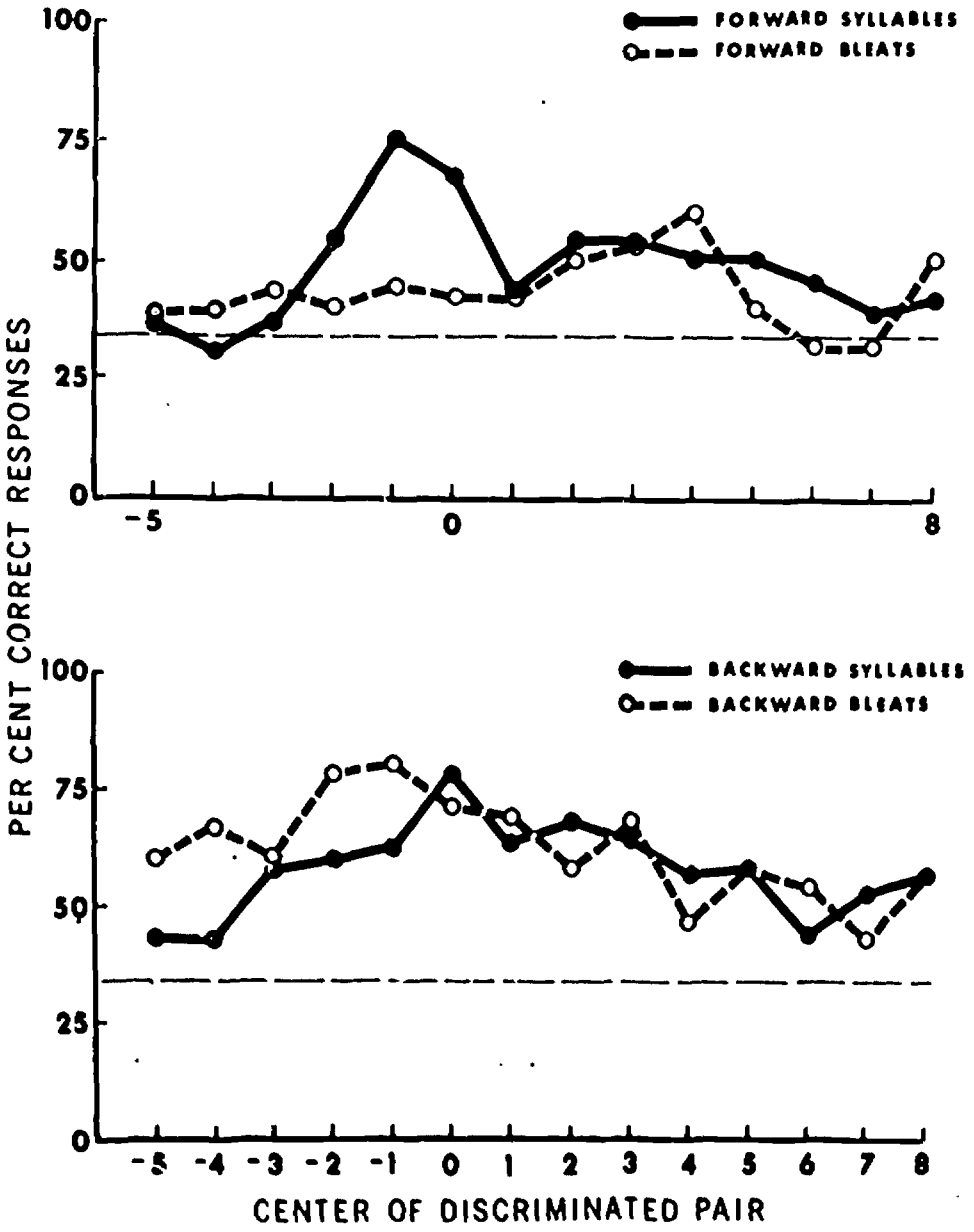


FIG. 7

To facilitate comparison of the results with chirps and bleats, we have presented together in Figure 8 the pooled data for these two nonspeech controls. The discrimination functions for the bleats parallel those for the chirps: the functions in the forward condition are above random, low, and irregular, while the functions in the backward condition are considerably higher and show peaks in the -1, 0, +1 range. As with the backward chirps, individual subjects (including the subject shown in Figure 6) show idiosyncratic peaks in their backward bleat functions, but there is no sign in either forward or backward chirp or bleat functions of an artifact such as gave trouble in Experiment I. However, in the forward condition, discrimination of the chirps is somewhat better than discrimination of the bleats.

At this point, we must consider whether there is any difference between the discrimination functions for the chirps and those for the bleats which would lend plausibility to the argument that the comparison between chirps and speech is in one respect or another unfair. Had we found that bleats were discriminated better than chirps in either forward or backward conditions, we might have supposed that the absence of a steady-state second formant at a constant frequency in the chirp stimuli made them more difficult to perceive than the syllable stimuli. No such result was obtained; in fact, forward bleats are not discriminated quite as well as forward chirps. (This is probably attributable to the fact that bleat subjects took all the nonspeech discrimination tests first). Had we found that backward chirps were discriminated better than backward bleats, with no comparable improvement in the forward condition, we might have supposed that the absence of a fatiguing steady state in the chirps made them easier to perceive than the syllables. Although our bleat control was imperfect since it is still possible to argue that fatigue might be induced by the presence of the steady states of both first and second formants, the least that can be said is that the outcome of the bleat experiment does not encourage such an argument. Chirps are discriminated at the same level as bleats in the backward condition. Since the shapes of the corresponding chirp and bleat functions are similar, the effect of the second-formant steady state can probably be ignored; and it will be convenient for purposes of our discussion to pool the results for the two groups of subjects in Experiment II, as in Figures 9 and 10.

Let us sum up the results of Experiment II, referring to Figures 9 and 10. In forward condition, the speech discrimination function shows peaks at phonetic boundaries and troughs within phonetic categories. The nonspeech function shows

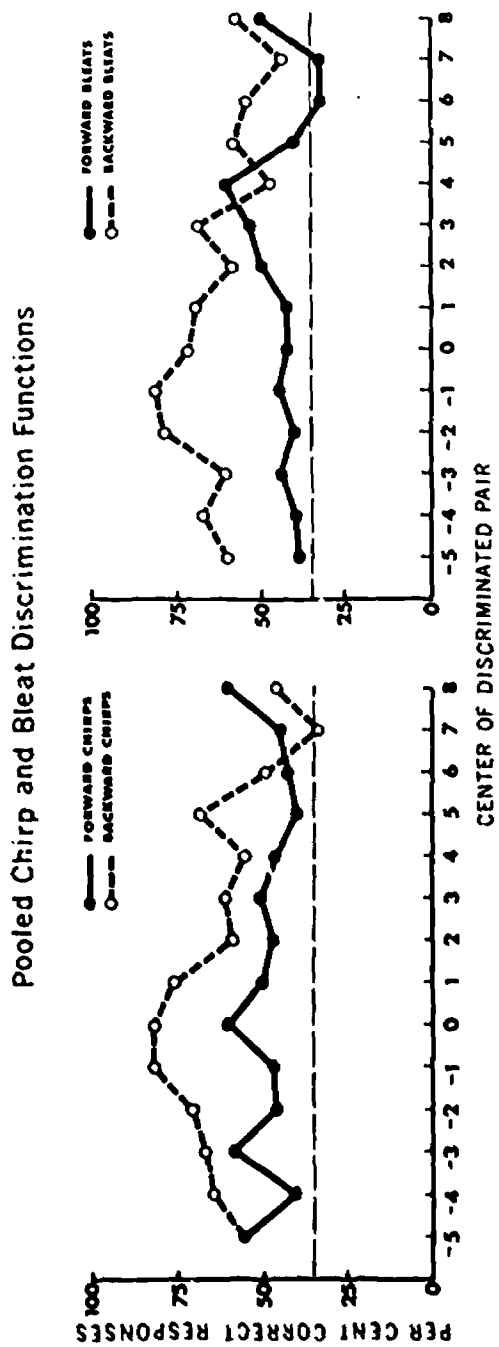

Pooled Chirp and Bleat Discrimination Functions
FORWARD CHIRPS
BACKWARD CHIRPS
FORWARD BLEATS
BACKWARD BLEATS
PER CENT CORRECT RESPONSES
100
75
50
25
0
-5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8
CENTER OF DISCRIMINATED PAIR

FIG. 8

Speech and Non-speech Discrimination Functions

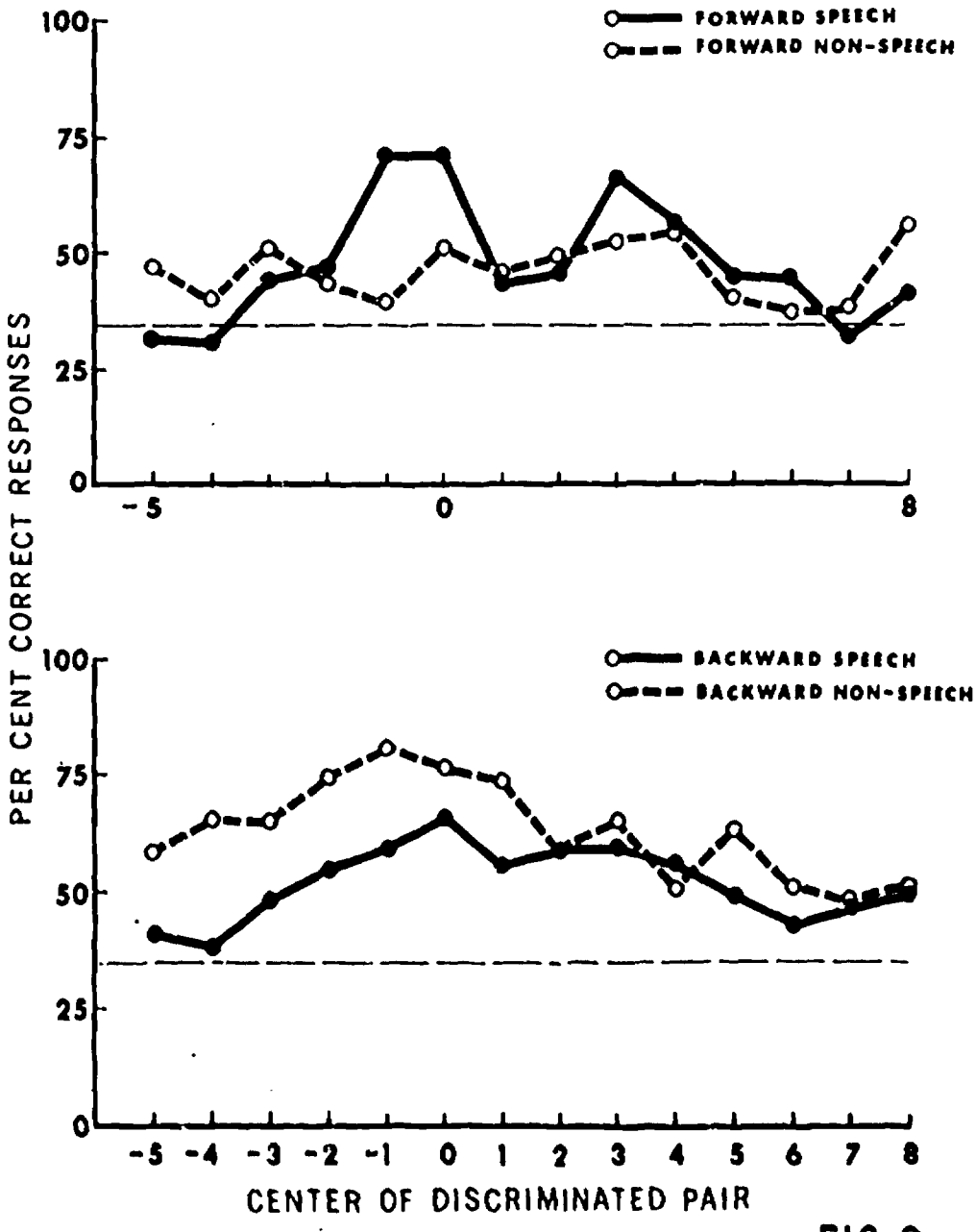


FIG. 9

Forward and Backward Discrimination Functions

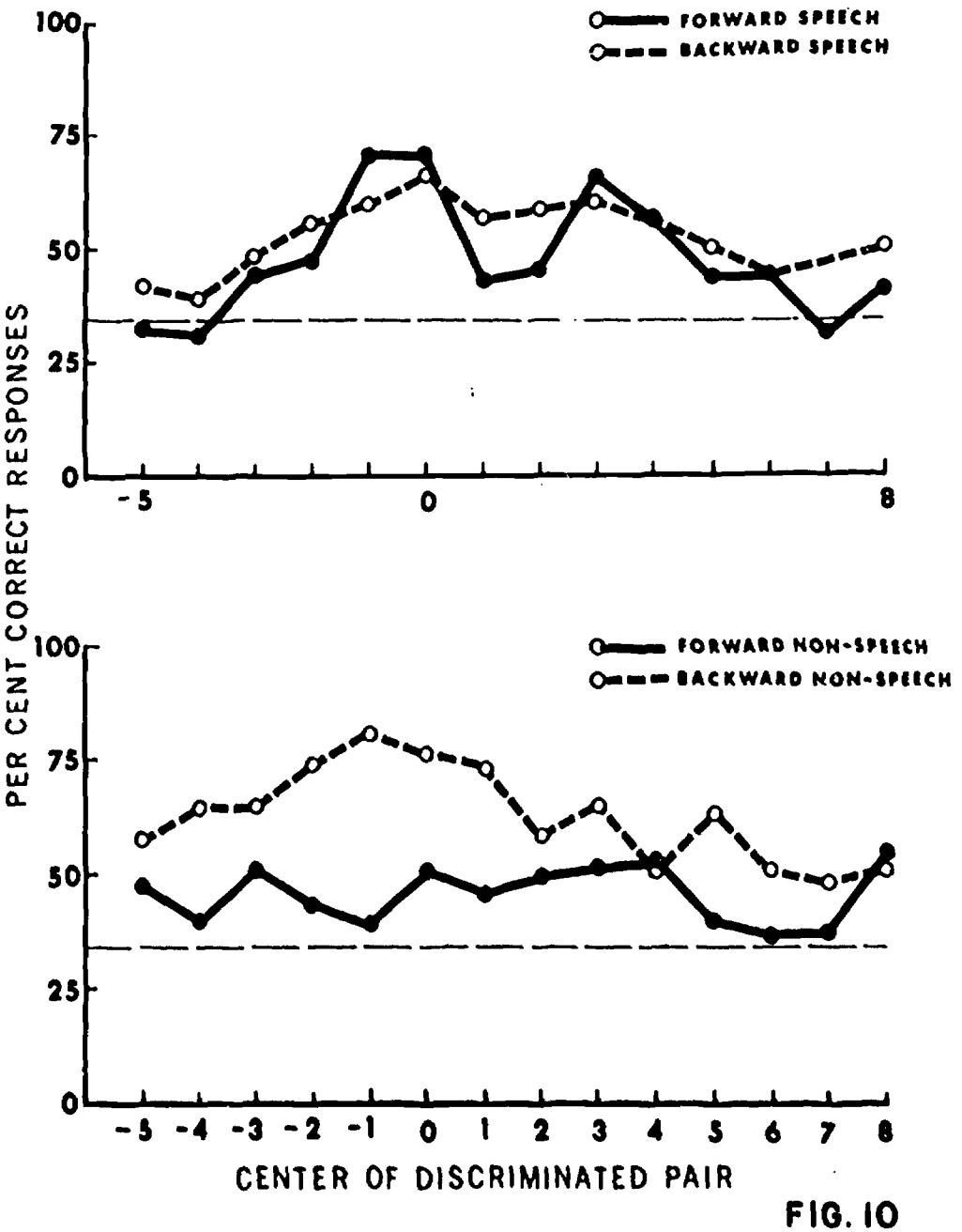


FIG. 10

no such peaks or troughs; it is irregular and low, though above random. In backward condition, the level of discrimination for speech is about the same as in forward condition, but the function has all but lost the peaks and troughs. The nonspeech function peaks near zero; it is higher than the speech function and much higher than the nonspeech function in forward condition. Thus, speech and nonspeech differ in each condition (Figure 9), and the change of conditions affects speech in one way and nonspeech in another (Figure 10).

GENERAL DISCUSSION

There are three different classes of phenomena to be accounted for: the responses of subjects to the chirps and bleats which served as nonspeech control stimuli; the responses to the speech-like stimuli; and the differences in the responses to the corresponding speech-like and nonspeech stimuli.

We must first attempt to interpret the results for the nonspeech stimuli. For convenience, we will speak of chirps, but it will be seen that the argument applies just as well to the bleats. Since, surprisingly, we have been able to find only one psychoacoustic study of dynamically varied resonances (Brady et al., 1961) against which we could check our conclusions, this interpretation must be considered as highly tentative.

For each of several stimuli similar to our chirps, with various durations and initial and final frequencies, Brady et al. asked their subjects to adjust the frequency of a steady-state resonance until it sounded most like the test stimulus. The subjects showed a very pronounced tendency to select a steady-state frequency approximately equal to the final frequency of the chirp. It seems plausible to infer that, for some reason, subjects find it easier to estimate the final frequency of a chirp than its frequency at some earlier moment. If so, we should expect to find, as we do in the present experiment, that a discrimination task in which the stimuli differed most in their final frequencies and not at all in their initial frequencies (the backward condition) would be easier than a task for which the reverse was true (the forward condition).

But we cannot go on to assume that, in our experiment, subjects discriminate simply by comparing the three estimated frequencies of each oddity triad. Suppose that subjects were given a chirp discrimination test in which both the initial and final frequencies of the chirps were varied. Before a subject could compare the three chirps in a triad, it would be necessary for him (1) to estimate the frequency at some fixed time during each of the three chirps and (2)

to determine the slope of each chirp. However, in the special case where either the initial or the final frequencies of all chirps are constant throughout a test, either (1) or (2) would give the subject sufficient information to discriminate the stimuli one from another.

Which of these two methods are the subjects using? In the case of the backward chirps, it seems clear that subjects are using method (2). They discriminate best those pairs of stimuli straddling values -1, 0, +1, i.e., pairs having negative and zero, negative and positive, and zero and positive slopes, respectively. It is not surprising that these three special cases of slope comparison should prove easy. On the other hand, these pairs of stimuli have no particular significance in terms of method (1), comparison of frequencies.

With respect to the forward chirps, no similar conclusion can be drawn. Performance was, in general, too poor to reveal any significant pattern, although one of the four subjects has a peak at +1 and another at 0, and the highest peak of the pooled data is at 0. But if we make the assumption that subjects are comparing slopes in the case of forward as well as backward chirps, a further inference, about the way subjects determine slopes, is possible.

Conceivably, a subject might estimate the slope directly. Alternatively, he might estimate the frequency at two different moments during the signal t and $t + \Delta t$ (or possibly just the difference between these two frequencies) and compute

$$\frac{f_t - f_{t + \Delta t}}{\Delta t}$$

Computing the slope in this way does not, of course, involve the kind of frequency estimation required for method (1): it is not necessary to hold t constant for estimates for all three members of a triad. Moreover, in the case of the backward chirps, he could then let $t = 0$ and take advantage of the fact that, for this value of t, f_t is a constant; in the case of the forward chirps, similarly, he could let $t + \Delta t = 40$ msec.

Now if the subject estimates the slope directly, he should do as well with forward as with backward chirps. If he computes the slope, this will not necessarily be the case since the computational process is not the same. For the backward chirps, the subject can choose Δt freely (his optimal choice is 40 msec), and he knows its value at t. For the forward chirps, on the other hand, either the subject must compute $\Delta t = 40$ msec $- t$, or he must, before t, choose Δt

and compute t = 40 msec - Δt, or he must wait until he hears $t + \Delta t$ to measure Δt. If these constraints make it more difficult for the subject to evaluate t or Δt, his slope computations would in turn be affected. Thus, there is a second reason why we should expect the backward chirps to be better discriminated. Not only is the final frequency of chirp apparently easier to estimate than its initial frequency, but also, the time estimation required to compute the slope is easier when the initial frequency is known to be constant and the final frequency is varied than in the reverse case.

Brady et al. point out the conflict between their result and the much greater cue value of the second-formant transition in initial than in final position in speech context, and they conclude that speech perception cannot be accounted for on the same basis as their experimental result. We face a similar question. Can we account for the discrimination function for the speech stimuli on a strictly psychoacoustic basis? To do so requires either that we point out resemblances to the corresponding nonspeech functions or that we propose some convincing explanation for the differences.

We recall first that the forward speech functions have characteristic peaks and troughs; these peaks and troughs occur consistently for all subjects and are obvious in the pooled data of Figure 9. The same peaks and troughs, much less pronounced, appear in the speech function for the backward condition. Nothing corresponding to these peaks and troughs occurs for the nonspeech stimuli, except that the nonspeech functions, like the speech functions, have peaks near or at 0. As we shall see shortly, this is probably a coincidence, and there is no obvious parallel in the nonspeech function for the other peak of the forward speech function or for its troughs. Furthermore, we note that performance is consistently better for nonspeech stimuli in backward condition than in forward condition, while for speech stimuli, there is no corresponding consistent improvement (Figure 10).

As we have seen, the perception of the speech stimuli tends to be categorical: the peaks are found near phonetic boundaries while the troughs correspond to zones inside these boundaries. It has been noted before (Liberman, 1957) that there is an obvious articulatory reference for such perception. When there is articulatory continuity, as in several tokens of [t], each with somewhat different second-formant transitions, such as might, in a human speaker, have resulted from different varieties of apical closure, the listener finds it difficult or impossible to discriminate. When, on the other hand, the difference

in the formant transition, though physically no greater, is at a point in the continuum such that it could only have resulted from one sound having been made with labial closure and the other by apical closure, there is a discontinuity in articulation and the listener discriminates quite readily. Because of the particular vowel used in the stimuli, this point of discontinuity happened to fall at stimulus 0. For a vowel with a higher (or lower) second-formant steady state, the boundary would have been lower (or higher) relative to this steady state.

The articulatory basis for the fact that initial transitions result in better phonetic separation than final transitions is less clear, but a study by Öhman (1966) suggests a possible answer. He found that consonants tend to be coarticulated much more with a following vowel than with a preceding vowel. In production of V_1CV_2 syllables, the character of the transition from V_1 to C depends not merely on V_1 and C but quite considerably on V_2, whereas the transition from C to V_2 is only slightly affected by V_1. Thus, an initial transition (CV) is apt to be a better consonantal cue than a final transition (VC). And in fact, in natural speech, final stops are often followed by a release, consisting of a burst (itself a supplementary cue to point of articulation) and low-amplitude transitions toward [ə]; unreleased stops, on the other hand, are notoriously ambiguous. The stops in the backward speech stimuli used in this experiment were, of course, unreleased.

In previous experiments comparing perception of speech and nonspeech, the nonspeech results were interpreted as representing the discrimination of an acoustic variable before the acquisition by the subjects of this articulatory knowledge. Differences between the discrimination of speech as opposed to nonspeech could then be assigned to "acquired distinctiveness" or "acquired similarity." The results of the [to]/[do] and rapid/rabid experiments were taken as evidence of acquired distinctiveness. A more conservative and, we now think, more proper view would have taken the results of those experiments, just as we take the results of our own present experiment, to be evidence for the existence of a speech mode that differs in interesting ways from the auditory mode. Questions about the role of learning in the development of the speech mode stand apart from questions about its existence and are answered by experiments different from those of the kind we have been considering here. Thus, to see the effects of experience, we should look to the cross-language studies of Lisker and Abramson (1970; also, Abramson and Lisker, 1970) on the perception

of the distinction between voiced and voiceless stops. These studies have shown that peaks in discrimination similar to those of our experiment are present or absent depending on the linguistic background of the listener. It does not follow, however, that the peaks are simply a consequence of differential reinforcement or of the mediational processes usually associated with the concepts of acquired distinctiveness and acquired similarity. In that connection, we should take note of other results obtained by the same investigators which show that the location of the voiced/voiceless boundaries is very much the same in a number of unrelated languages. When we consider, in addition, that the voicing distinction is universal, or very nearly so, we see that learning does not, in any case, exert its effect in the arbitrary way that Lane (1965), for example, or Quine (1960:85-90) suppose. The biologically given constraints are important and must surely be of the greatest interest to anyone who is concerned to understand the development on consonant perception and the peaks that characterize consonant discrimination. This view is strengthened by the findings of recent experiments on infants by Moffitt (1969) and Eimas et al. (1970), which show that consonant discrimination is present at a very early age. In the study by Eimas et al. it was found that one-month-old infants discriminate synthetic [ba] and [pa]. Of even greater interest is the fact that, given a fixed physical difference in the relevant acoustic cue, these infants discriminate better across a phonetic boundary than within a phonetic category. Thus, like our adult subjects, they show a discontinuity in discrimination of the voiced/voiceless distinction just as our adult subjects do for the place distinction. It is most likely that the infants' perception of the voicing distinction was, like so many deeply biological processes, not entirely uninfluenced by their experience. If they had been reared in a soundless environment, they would conceivably not have been able to discriminate [ba] from [pa] as they did. Indeed, it is possible that the experience of having heard speech was a necessary condition for the performance that Eimas et al. found. But it is hardly conceivable that the effects were produced at the age of one month by the simple processes of differential reinforcement or by the more complex mediational mechanisms implied by the concepts of acquired distinctiveness and acquired similarity.

The outcome of our present study also raises other doubts about the applicability of acquired distinctiveness and similarity. In the forward condition, for some distance on either side of the peaks corresponding to the phone

boundaries, the speech function is well above the nonspeech function. This, therefore, we would have to attribute to acquired distinctiveness. For portions of the continuum well within phonetic boundaries, the speech function is at or near random and usually well below the nonspeech function. This we would have to attribute to acquired similarity. So far, nothing is seriously amiss, though it would be more parsimonious if it were possible to invoke only one of these processes.

In the case of the backward functions, however, our embarrassment is of a different character. The nonspeech function is higher than the speech functio at almost every point. We are, therefore, compelled to invoke acquired similari ty to account for the peaks as well as the troughs of the speech function. But why should there be any acquired similarity for stimuli on opposite sides of a phonetic boundary--that is, for stimuli which the listener has learned to call by different names?

Although there are surely ways out of this difficulty that yet preserve concepts like acquired distinctiveness and acquired similarity, it seems to us preferable to conclude, rather, that we are dealing with two basically different modes of perception. One of these modes is the psychoacoustic. The results of discrimination studies in this mode require an interpretation of the kind we advanced in trying to account for the chirp and bleat data. The other mode is the speech mode. Its characteristics are the consequence of the special processor that decodes the complexly encoded speech signal and recovers the phonetic message. The results of perceptual experiments on the stop consonants do not yield to an interpretation in terms of psychoacoustic perception, with or without such modification as might have been produced by discrimination learning.

In connection with the conclusion that speech and nonspeech are processed differently, we should note that speech and nonspeech functions differ not only in their shape and level but in their reliability. The nonspeech functions vary not only from subject to subject but also for a single subject from one session to the next. Such factors as the relative naiveté, the alertness, and the motivation of the subject and the strategy he adopts for the task of discrimination may make a very substantial difference. In informal tests, in which two of the authors served as subjects, higher levels of chirp discrimination in the forward condition were attained than for any of the subjects for which data have been presented here. The remarkable thing about the perception

of the speech-like stimuli, on the other hand, is precisely its insensitivity to all such factors. Within wide limits, the performance of a subject is relatively stable and predictable, provided only that he hears the synthetic stimuli as speech. Even subjects who are quite familiar with the stimuli--for example, the authors--do little better than naive subjects away from phonetic boundaries, while naive subjects do little worse than the authors hear phonetic boundaries. The speech mode appears to act like some digitizing device which, accepting a signal of quite variable quality and much fine detail, converts it to a perceptual response that is coarsely but reliably quantized.

The backward speech discrimination functions at first appear to contradict what has just been said, since these functions are variable and unstable. In the backward speech test, the subjects were confronted with a confusing task. They were given speech-like stimuli which, as the identification function showed, were difficult to perceive as speech. One might have expected them, in such a situation, to discriminate speech poorly: that is, to produce a discrimination function in which the peaks corresponding to those observed in the forward condition were lower and the troughs--near random in the forward condition--remained near random. Such an outcome, however, would have suggested that there was, after all, considerable variability in the level of speech discrimination and that, for some kinds of speech, discrimination is much less reliable than we have just suggested. What actually happens, however, is that, while the peaks are indeed lower, the troughs are higher (Figure 10). The function appears to be a combination of the forward speech function and the backward chirp function. Our interpretation is that the subjects tried to respond to the stimuli as speech. When they found this too difficult, they reverted to the nonspeech mode. But whenever they did respond to the stimuli as speech, they did so, we suspect, as reliably as in the forward condition.

This interpretation of the data bears on an important and difficult question: what conditions must be presented to insure perception in the speech mode? The very fact that perceptual experimentation with very simple synthetic speech patterns has been possible shows that a high degree of naturalness is not an important factor, though it seems reasonable to suppose that, at a minimum, some representation of the first two formants may be essential. However the subjects' response to the backwards speech, where formants were present but speech cues were weak and few in number, suggests that a requirement for perception in the speech mode is that the cues for the distinctions among phonetic

segments be present in sufficient strength and number to keep the perceptual machinery active. If this requirement is not met, the listener may slip into the nonspeech mode. Thus, the apparently exceptional backward speech results offer an interesting and, to us, unexpected insight into the nature of the special mode of perception which, our experiments suggest, is required for speech.

References

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

Brady, P.T., House, A.S.,and Stevens, K.N. (1961) Perception of sounds characterized by a rapidly changing resonant frequency. J. Acoust. Soc. Amer. 33, 1357-1362.

Cross, D.V. and Lane, H.L. (1964) An Analysis of the Relations Between Identification and Discrimination Functions for Speech and Nonspeech Continua. Report No. 05613-3-P. (Ann Arbor: Behavior Analysis Laboratory, University of Michigan).

Eimas, P., Siqueland, E.R., Jusczyk, P., and Vigorito, J. (1970) Speech perception in early infancy. Paper presented to the Eastern Psychological Association, April 1970.

Kimura, D. (1964) Left-right differences in the perception of melodies. Quart. J. Exptl. Psychol. 16, 335-358.

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

Kirstein, E. (1966) Perception of second-formant transitions in non-speech patterns. Status Report on Speech Research 7/8 (New York: Haskins Laboratories) paper 9.

Kirstein, E., and Shankweiler, D. (1969) Selective listening for dichotically presented consonants and vowels. Paper presented to the Eastern Psychological Association, Philadelphia.

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

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

Liberman, A.M. (in press) Some characteristics of perception in the speech mode. Proc. Association for Research in Nervous and Mental Diseases. (Baltimore: Williams and Wilkins).

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

Liberman, A.M., Delattre, P.C., Cooper, F.S., and Gerstman, L.J. (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monog. No. 8, 1-13.

Liberman, A.M., Harris, K.S., Eimas, P., Lisker, L., and Bastian, J. (1961a) An effect of learning on speech perception: the discrimination of durations of silence with and without phonemic significance. Lang. & Speech 4, 175-195.

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

Liberman, A.M., Harris, K.S., Kinney, J.A., and Lane, H. (1961b) The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. J. Exptl. Psychol. 61, 379-388.

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

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

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

Mattingly, I.G., and Liberman, A.M. (1969) The speech code and the physiology of language. In Information Processing and the Nervous System, K.N. Leibovic, ed. (Berlin: Springer Verlag) 97-117.

Mattingly, I.G., Liberman, A.M., Syrdal, A.K., and Halwes, T. (1969) Discrimination of F2 transitions in speech context and in isolation. J. Acoust. Soc. Amer. 45, 314-315.

Moffitt, A.R. (1969) Speech perception by 20-24 week old infants. Paper presented to the Society for Research in Child Development, Santa Monica, Calif., March 1969.

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

Popper, R. (1967) Linguistic determinism and the perception of synthetic voiced stops. Unpublished Ph.D. thesis, U.C.L.A.

Quine, W. v. O. (1960) Word and Object. (Cambridge, Mass.: M.I.T. Press).

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

Strange, W., and Halwes, T. (in press) Confidence ratings in speech research: Experimental evaluation of an efficient technique for discrimination testing. Perception and Psychophysics.

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

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

Syrdal, A.K., Mattingly, I.G., Liberman, A.M., and Halwes, T. (1970) Discrimination of F2 trnasitions in speech and nonspeech contexts. J. Acoust. Soc. Amer. 48, 94.

Effects of Filtering and Vowel Environment on Consonant Perception*

Thomas Gay+
Haskins Laboratories, New Haven

Abstract. The purpose of this experiment was to determine the effects of filtering and vowel environment on consonant perception. Sixteen consonants in CV combination with seven vowels were recorded on tape, low-pass filtered, and played back to a group of listeners. In general, the results indicated that /t,k,b,d,g,s,f,z,w,r,n/ are affected by filter cut-off points, /k,g,f,v, m/ show multivowel effects, and /p,b,d,j,n/ show consistently lower scores only when followed by /i/. As expected, error types were predominantly "place," with "manner," "voicing," and "nasality" errors occurring only at the less favorable cut-off frequencies. The results are discussed in terms of the predictability of the effects as a function of CV transition characteristics and the suitability of small sample PB lists for assessing speech discrimination of individuals with high frequency hearing loss.

---

*The paper appeared in J. Acoust. Soc. Amer. 48, 4, Part 2 (October 1970) 993-998.

+Also, University of Connecticut School of Dental Medicine, Storrs.

Acknowledgements. The author wishes to thank Mrs. Frieda Toback and Mrs. Celia Dorrow for their assistance in the collection and analysis of the data.

Since their introduction in the 1940s, phonetically balanced word lists have been used extensively for testing speech discrimination in both the clinic and research laboratory. The lists' main features are that the words are common, familiar, easy to administer, and of course, are in "phonetic balance." Originally, the aim of phonetic balancing was to provide a list of words whose phonemic content occurred with the same frequency of occurrence as the phonemes found in everyday speech. This was accomplished simply by assigning a certain overall proportion to each phoneme in the list, without regard for the internal phonemic make-up of the words. It has since been recognized, however, that coupling effects exist for different consonant and vowel sequences, with the articulatory and acoustic properties of a given phoneme often depending on those of its neighbor. In this sense, then, it is not unreasonable to suspect that conditions may exist where the perception of a given sound might be either enhanced or degraded by the coarticulation effects of the adjacent phoneme. The most likely conditions, of course, would be one in which the spectral characteristics of the phoneme are either altered or eliminated, as in filtering, or, on a physiological level, a hearing impairment. In both cases, important cue information provided by the CV transition might be reduced by varying degrees, depending on the amount of the transition eliminated by the distortion.

The experiment reported here attempts to describe some of these effects, specifically, the extent to which various vowel environments influence the identification of consonants in CV syllables heard under conditions of low-pass filtering. Although coarticulation effects in real speech extend beyond simple CV sequences, the data obtained from this experiment can be considered a first step in determining the extent of these effects. These data will be examined in two ways: first, as strictly normative and second, since low-pass filtering somewhat resembles a high frequency hearing loss, as a basis for speculating on certain clinical speech discrimination problems.

Procedures

The general procedure was to construct lists of various consonant-vowel syllables and record, filter, and play back these lists to a group of listeners.

The stimuli consisted of the sixteen consonants, /p,t,k,b,d,g,s,f,z,v,w, j,r,l,m,n/, each in CV combination with the seven vowels, /i,ɛ,æ,a,ʌ,ɔ,u/. The total number of syllables was 112. These items, each repeated three times, were randomized into a master list. Three such randomizations were made, one

for each of the three speakers. The speakers were three adult males whose speech was typical of the New York City dialect area. Recordings were made on one track of an Ampex Model AG-500 two-track tape, recorded through an Electrovoice Model 654 microphone. The items were recorded at approximately three-second intervals, with longer rest periods occurring after groups of ten. The carrier word write preceded each utterance. Gain levels for each speaker were adjusted so that the vowel / ɔ / peaked at zero on the tape recorder's VU meter. Other than that, no attempt was made to equalize within-list gain levels. This meant, of course, that, due to normal vowel-level differences, relative intensities among the tokens differed by as much as 8 db. The master tape, then, contained all stimuli, each repeated three times by each of three speakers for a total of 1008 items (112 x 3 x 3).

This tape was edited into five different randomizations, one for each of five low-pass filter conditions. Filter cut-off points were 800, 1000, 1200, 1400, and 1600 Hz. Exploratory work showed these settings to cover the range between apparent chance responses and unmeaningfully high scores. The filtering was accomplished by playing the tapes back on one Ampex AG-500 through two Allison Model 2B variable filters connected in series and re-recording the tapes on a second Ampex AG-500. The filters provided a roll-off of approximately 60 db/octave.

Listeners were seven normal-hearing, undergraduate and graduate college students. Each was told about the make-up of the lists only in general terms. The response mode was open-set, with the listeners free to choose any of the phonemically permissible CV combinations. Twenty-five practice items preceded each filter condition. The tapes were played back to the subjects (random-order presentation) binaurally through Telephonics TDH-39 earphones in a quiet but not fully sound-treated room. Playback levels for all lists were adjusted to approximately 80 db, overall SPL, as measured on a B&K audiometer calibration unit.

Results

As would be expected, vowels were highly intelligible under all filter conditions and, except for /i,u/, exceeded 95 percent in all cases. Not unexpectedly, /i,u/ were sometimes confused with each other (consistently more /u/ confusion for /i/ than vice versa) but with no observable consonant influence. Also, although evidence of occasional speaker influences for a small number of tokens existed, there were no consistent trends, and thus, all data were averaged over the three speakers. As expected, then, the major effects

Mean Correct Scores for Six Stops and Five Filtering Conditions

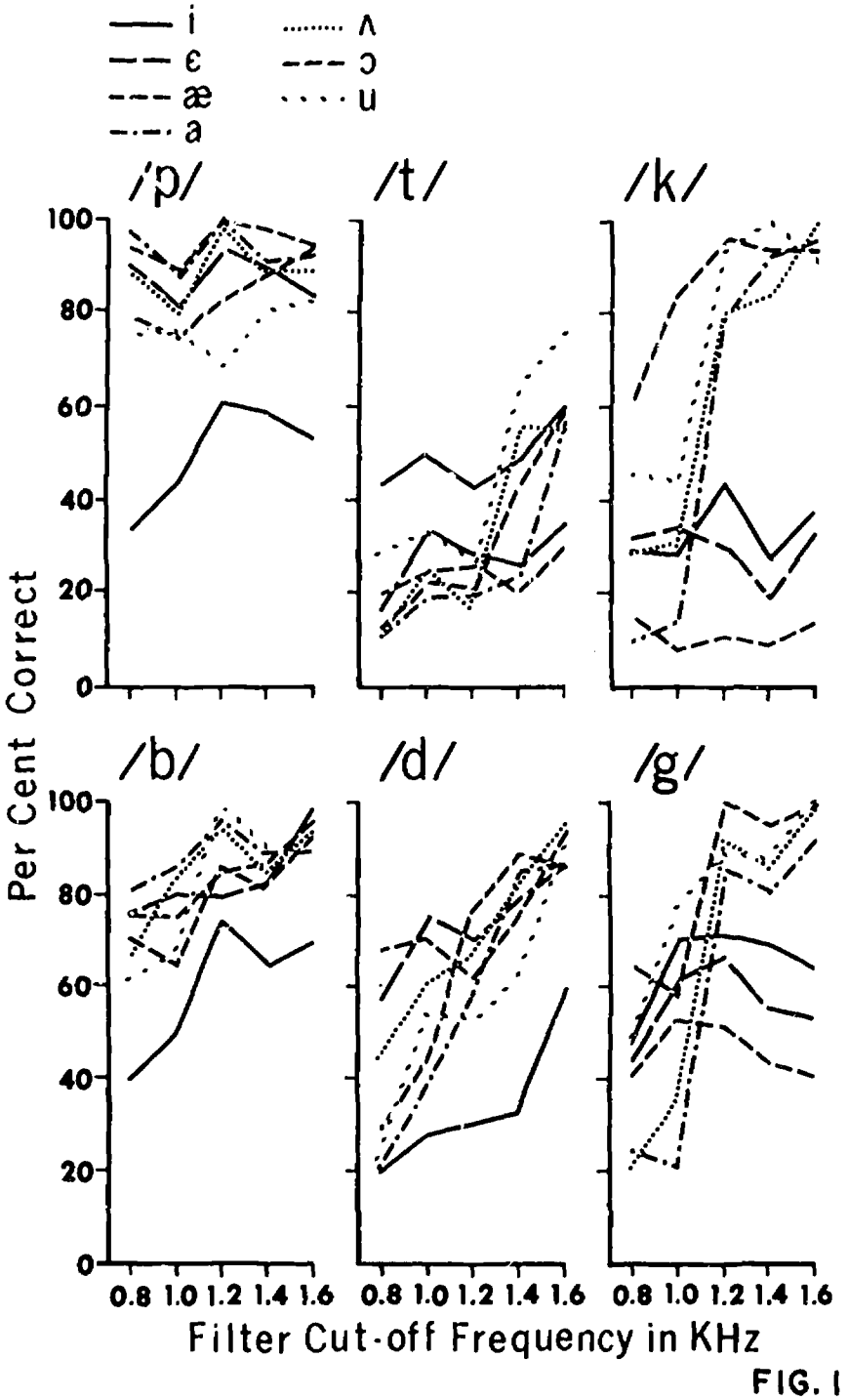


FIG. 1

are those of filter condition and vowel environment.

The percent correct scores for all consonants under each filter condition are plotted separately for each consonant category in Figures 1-4.

A. Stops

Figure 1 shows the mean percent scores for the group of stop consonants, /p,t,k,b,d,g/. As the graphs show, each consonant is somewhat differently affected by filter cut-off point and vowel environment. For /p/, there is no consistent filter cut-off effect, as the curves run moderately flat across the five filter conditions. The most conspicuous vowel effect is for /i/, where scores are consistently lowest. This might be explained somewhat by the fact that, since the second formant for /i/ is somewhere in the vicinity of 2200 Hz, much of the information-bearing second formant transition rising to this level is probably eliminated by the filtering. (Similar /i/ effects occur for four of the remaining fifteen consonants.) Like /p/, /b/ shows no real vowel effect (except for /i/), but scores generally increase with the more favorable filter conditions.

Unlike their labial counterparts, /t,d/ bear little similarity to each other. Scores for /t/ followed by /i/ are clearly higher except at the two highest cut-off points. A cut-off effect exists only at 1400 Hz for three of the seven vowels. /d/, on the other hand, is characterized by a sharp increase across the cut-off points along with the deleterious effect of a following /i/.

The most interesting of the stops are /k,g/. Here, both pronounced vowel and cut-off effects occur. For both consonants, back vowel combinations show an increase in intelligibility at cut-off points of [illegible] Hz and higher. A ready explanation of this occurrence can be found in synthetic speech work of Delattre, Liberman, and Cooper (1955), who found that the course of the formant transitions for /k,g/ originate at two different starting points in frequency. The theoretical starting point, or locus, of a /k,g/ transition for a back vowel was found to be approximately 1200 Hz, while the locus for a front vowel transition was at about 3000 Hz. The filtering effects found here, then, can be explained by the fact that information for /k,g/ preceding a back vowel does not appear below frequencies of 1200 Hz (hence the lower scores for filter points below 1200 Hz) and that significant information for /k,g/ preceding a front vowel does not appear at frequencies below 3000 Hz (with lower scores expected for all cut-off points).

Mean Correct Scores for Four Fricatives

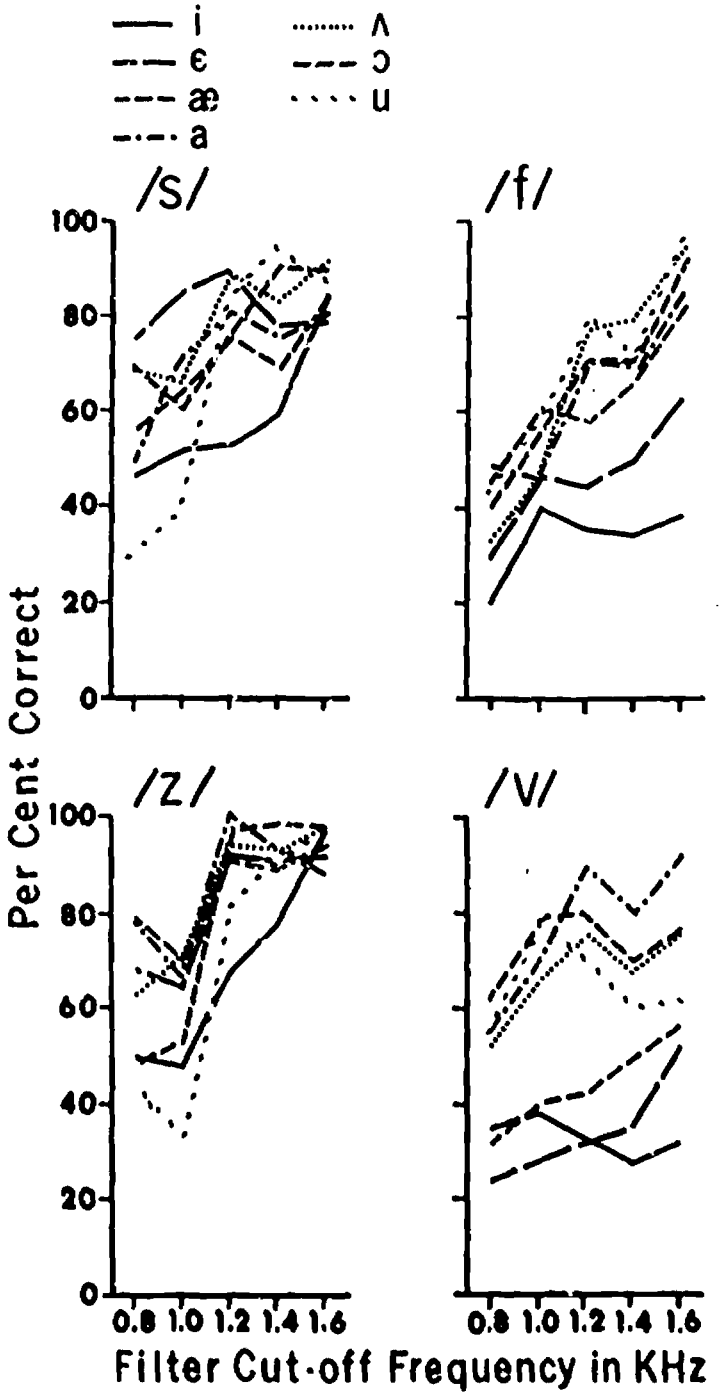


FIG. 2

B. Fricatives

The fricatives, /s,z,f,v/, like the stops, show certain individual peculiarities (Figure 2). For /s/, the vowel effects are inconsistent but the curves generally rise across the filter cut-off points. In general, /s/ followed by /i/ shows the lowest scores. On the other hand, /z/, although showing no real vowel effects (except for /i/), shows a sharp increase in intelligibility beginning at the 1200 Hz position.

Both cut-off frequency and vowel environment affect /f,v/ identification. For both consonants, but especially /v/, back vowel combinations are more intelligible than front vowel combinations. These effects are superimposed upon the increases across cut-off points. The behavior of the fricatives might be explained by the fact that, while /s,z/ are identified primarily by their noise characteristics, /f,v/ are cued more by their second formant transitions (Harris, 1958; Heinz and Stevens, 1961). The assumption here is that, as the transitions extend down to lower frequencies, more transition information remains intact for back vowel combinations.

C. Semivowels

The results for the semivowels, /w,r,l,j/, are shown in Figure 3. Filter cut-off effects occur for all consonants except /l/, whose intelligibility is highest of all consonants, regardless of filter cut-off conditions. The greatest cut-off effects occur for /w,r/. The vowel effects for /w/ are somewhat unusual in that higher intelligibility generally accompanies front vowels, especially at the lowest cut-off points. For /r/, there are also vowel effects at the lowest cut-off point. There are no real vowel effects for /l/ (except for a slight decrease in scores for /i/). The /i/ effect for /j/ is the most marked of any consonant.

D. Nasals

The results for /m,n/ are plotted in Figure 4. Both sounds show marked (though complicated) vowel and cut-off effects, with strong vowel-filter interactions most evident for /m/. In general though, front vowel curves are somewhat lower than back vowel curves. Except for /i/, the only significant vowel and cut-off effects for /n/ occur at 1200 Hz. However, no special vowel group preference emerges.

E. Error Types

Figure 5 summarizes the types of confusions that occurred for each of the

Mean Correct Scores for Four Semivowels

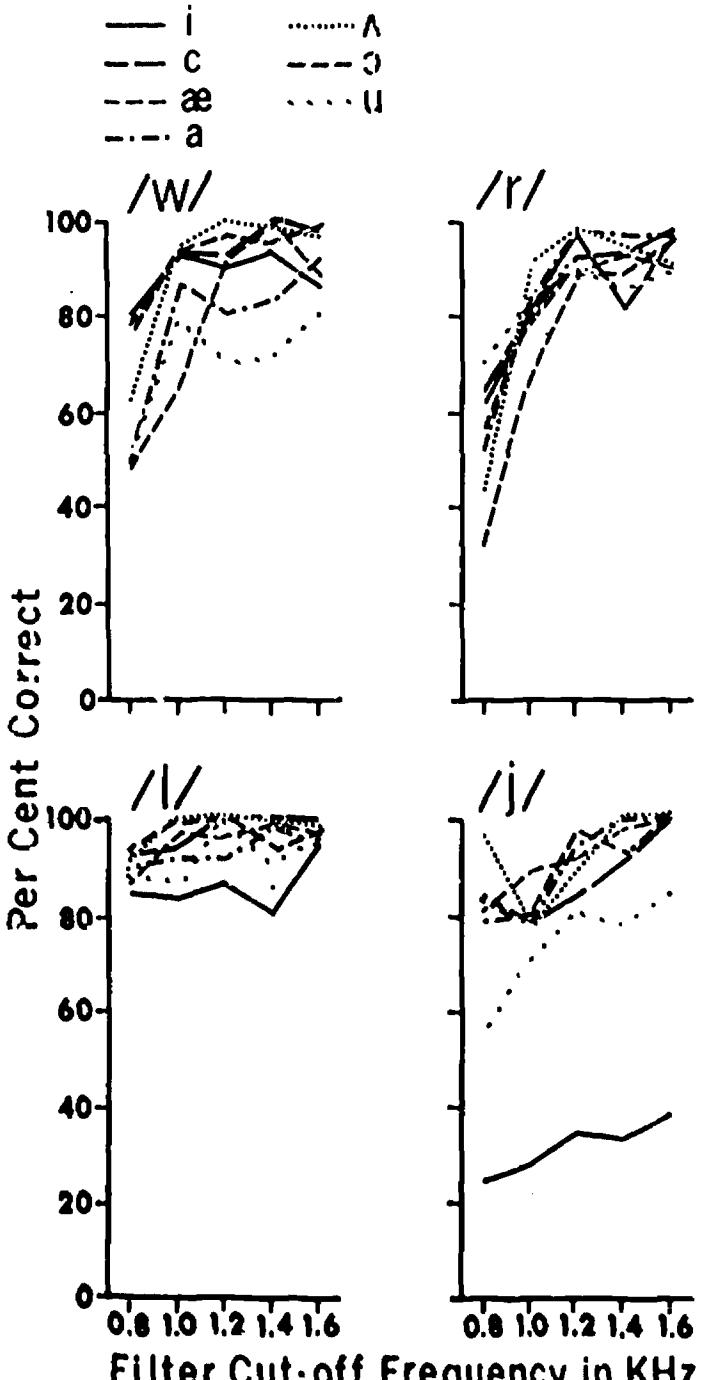


FIG. 3

Mean Correct Scores for Two Nasals

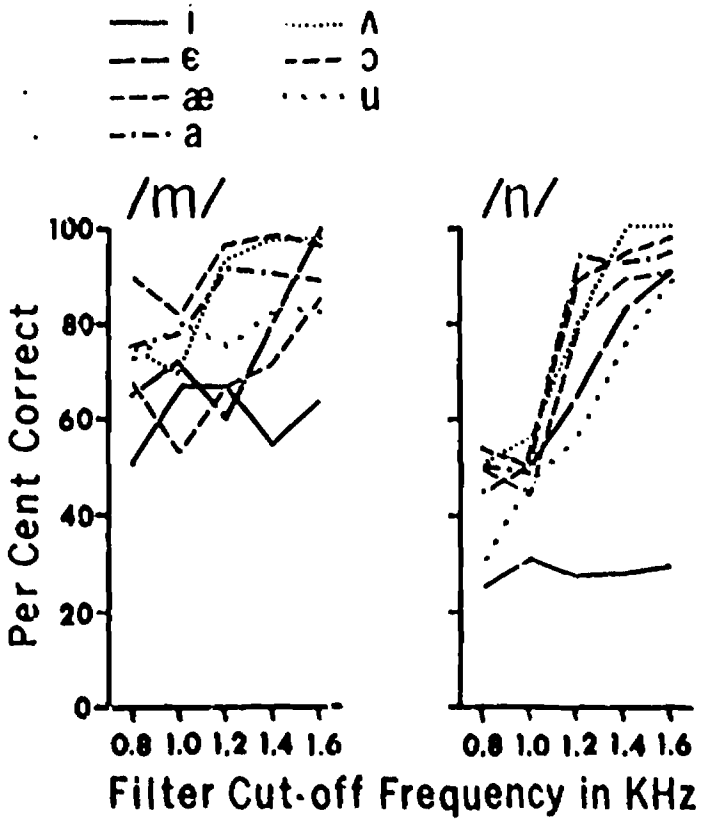
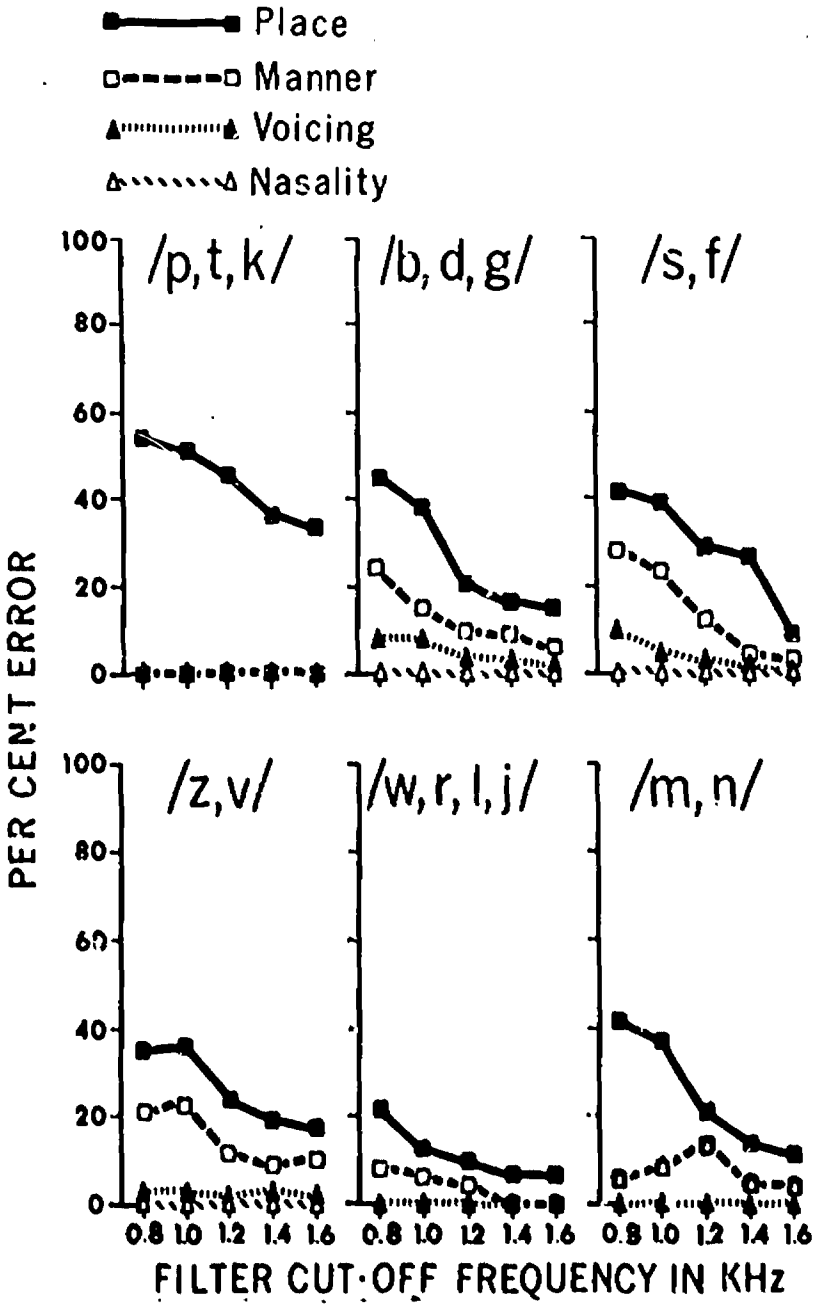


FIG. 4

Mean Error Scores Showing Type of Error for
Six Groups of Consonants and Five Filtering Conditions



Note: All manner errors for /m,n/ are also, by definition, nasality errors.

FIG. 5

six different consonant categories and the five low-pass filter conditions.[1] As would be expected, place errors accounted for most of the confusions, regardless of filter cut-off frequency (Miller and Nicely, 1955). At the lower cut-off frequencies, however, additional error types occur, generally in the order of manner, voicing, nasality. As can be seen from the graphs, only the voiceless stops, /p,t,k/, are characterized almost wholly by place errors.

To summarize the above results briefly: the effects of filter cut-off frequency and vowel environment on consonant perception are complicated. Some consonants are affected by filter cut-off points, others are not. Those affected are /t,k,b,d,g,s,f,z,w,r,n/. Likewise, some consonants are affected by vowel environment, while others are not. The greatest multivowel effects occur for /k,g,f,v,m/. Of the sixteen consonants, /p,b,d,j,n/ show consistently lower scores when followed by the vowel /i/. Error types were predominantly "place," with "manner," "voicing," and "nasality" errors occurring only at the less favorable cut-off points.

Discussion

A. Filter-Transition Relationships

As was mentioned at the outset, a reasonable basis exists for predicting the perceptual effects of certain consonants heard under conditions of low-pass filtering. This, of course, is based on the cue information provided by the CV transition and the extent to which it is eliminated by the filtering. These cut-off and vowel effects were most clearly demonstrated in the /k,g/ data, which supported Delattre, Liberman, and Cooper's (1955) notion of a variable locus for these phonemes. The perception of some of the other sounds, however, including the remaining four stops, is not so easily explained. If a fixed locus for the labials (720 Hz) and dentals (1800 Hz) is assumed, then a lower level of intelligibility would be expected for those stimuli containing vowels with a higher frequency F2, as more of the transition is eliminated by the filtering. (It is assumed that virtually all F3 information is missing under these conditions.[2]) This, however, is not always the case.

---

[1] Error types were classified as place, manner, voicing, and nasality. Multiple-type errors were counted in each appropriate category, e.g., if a /p/ was heard as a /d/, the error would be classed as both a place and a voicing error.

[2] The overall higher intelligibility of /p,b/ over the rest of the stops can be interpreted in much the same way; that is, since the labials are characterized by a lower frequency transition starting point, they are less vulnerable to missing higher frequency components.

For both /p/ and /b/, a following /i/ might be expected to degrade the consonant's intelligibility, and this, indeed, is borne out by the data. The remaining vowels, on the other hand, do not follow in this order. The overall picture is one rather of a grouping of the remaining curves without any hierarchical vowel preference. The data for /t,d/ are perhaps even more unusual. For /t/, at all but the two highest cut-offs a following /i/ provides the highest intelligibility levels, whereas for /d/, a following /i/ is accompanied by the lowest intelligibility levels at all cut-off points.

There are perhaps three explanations for the variability found for both sets of stops. First, certain unfiltered segments of the transition might, in one way or another, provide the necessary place cues; second, supplementary cue information might be contained in the burst segment of the phoneme; and third, perceptually significant variations might exist in the transition starting points, or even loci, of these phonemes. Support for this last possibility can be found in a recent experiment by Fant (1969), whose measurements for Swedish stops in CV syllables showed some large variation in F2 and F3 transition starting points, depending on the following vowel.

The behavior of the fricatives is generally straightforward, with a minimal vowel effect for /s,z/ and an important, predictable one for /f,v/. As was mentioned earlier, this can be explained by the fact that /f/ and /v/ are cued primarily by their transitions, which remain more intact when extending down to the lower F2 back vowels. The consonants /s,z/, on the other hand, are cued more by their noise segments, the major portions of which are located above the filter cut-off points. Cut-off effects occur for all consonants, with those for /s,z/ apparently due to the increased presence of the frication. The /f,v/ filter effects, like the vowel effects, are more consistent, with increases for all vowels occurring with each increase in cut-off frequency.

Except for /ji/ (and perhaps /li/), the semivowels show few consistent vowel effects. This is not unusual, as these phonemes are distinguished from one another by the onset frequencies of their F1, F2, and F3 transitions. What is somewhat unusual, however, are the cut-off effects for /w,r/. This is especially true of /w/, which is presumed to be cued by low frequency F1 and F2, in contrast, for example, to the higher frequency starting points of /l/, which shows no cut-off effects (O'Connor et al., 1957).

The place cues for /m,n/ are generally considered to be identical to those of the stops and thus might be expected to behave somewhat like their labial and dental counterparts. Unfortunately, the data for /m/ are not very clear, although it might be suggested that the front vowel stimuli, as a whole, are

less intelligible than the back vowel stimuli. The curves for /n/ seem to be similar to those of /d/ but with sharper slopes.

In summary, then, the filtering effects for four of the six stop consonants (/p,t,b,d/) cannot be related clearly to the course and extent of their CV transitions. Fricative behavior is generally straightforward, but unexplainable are the cut-off effects for the semivowels, /w,r/, and perhaps, the lack of them for /l/.

B. Clinical Implications

Although the results of this experiment are essentially normative, they can be applied to certain speech discrimination problems of the hearing-impaired. This is not to say, however, that low-pass filtering produces the same effects as a high frequency hearing loss.[3] The comparisons made here are based only on the fact that similar portions of the spectrum are eliminated by the two conditions and that this might produce some similar perceptual effects. In this sense, then, if these or similar vowel and cut-off effects exist for the hearing-impaired, then the use of a small sample word list, such as the W-22's, for testing speech discrimination would suggest the possibility of certain perceptual biases caused by the presence or absence of a given phoneme sequence. This assumes, of course, that common phoneme sequences are not adequately represented in the W-22 distributions. As was mentioned earlier, the W-22 frequencies, originally based on those of Dewey (1923), involved only overall frequencies of occurrence. Not until 1963, with the publication of Denes's data, was there any detailed information available on CV, VC, or CC syllable frequencies. When the present W-22 lists are analyzed according to these frequencies, however, the following can be noted: first, many familiar CV and VC syllables are not represented in the W-22 lists, and second, between twenty and twenty-five percent of the W-22 words contain consonant clusters, most of which are hardly common in everyday speech. The significance, especially of the latter, is that the acoustical characteristics and, consequently, perceptual cues of many consonants are quite different when in CC or CV position. Specifically, the first element of a cluster is no longer characterized by its often perceptually significant second formant transition.

Apparently, then, the internal phonemic make-up of the present PB words is not adequate. Although adequate representation can be built into a list,

[3]Beside the lack of evidence supporting a comparison of filtering with the pure tone audiogram, filtering does not take into account factors such as recruitment, equal loudness contour effects, or other nonlinear distortion that might accompany a high frequency hearing loss.

the job would be difficult and the results cumbersome. Indeed, it could also be argued that, in the clinical sense, such representation might not even be necessary. Since certain consonants and vowels are highly resistant or even insensitive to most hearing loss conditions, these phonemes might be replaced in a list by those that show more complicated effects or interactions. This approach would probably provide a more detailed, less redundant account of an individual's speech discrimination ability. Although lists of this nature are not as yet available, some existing lists can be adapted. For example, both Fairbanks's Rhyme Test (1958) and House, Williams, Hecker, and Kryter's closed response CVC lists (1965) control phoneme environment and, in addition, have the advantage of allowing an inventory of specific phoneme errors to be easily made.

Finally, it might be mentioned that the data of this experiment can also be applied to the selection and use of speech materials for clinical auditory training. They might be useful in providing a basis for determining the degree of difficulty for various syllables and words used in clinical sessions, especially beginning ones.

References

Delattre, P.C., A.M. Liberman, and F.S. Cooper. (1955) Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Amer. 27, 769-773.

Denes, P.B. (1963) On the statistics of spoken English. J. Acoust. Soc. Amer. 35, 892-904.

Dewey, G. (1923) Relative Frequency of English Speech Sounds. (Harvard University Press, Cambridge, Mass.).

Fairbanks, G. (1958) Test of phonemic differentiation: The rhyme test. J. Acoust. Soc. Amer. 30, 596-600.

Fant, G. (1969) Stops in CV syllables. STL-QPSR 4, 1-25.

Harris, K.S. (1958) Cues for the discrimination of American English fricatives in spoken syllables. Language and Speech 1, 1-7.

Heinz, J.M., and K.N. Stevens. (1961) On the properties of voiceless fricative consonants. J. Acoust. Soc. Amer. 33, 589-596.

House, A.S., C.E. Williams, M.H.L. Hecker, and K.D. Kryter. (1965) Articulation testing methods: Consonantal differences with a closed-response set. J. Acoust. Soc. Amer. 37, 158-166.

Miller, G.A., and P.E. Nicely. (1955) An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Amer. 27, 338-353.

O'Connor, J., L.J. Gerstman, A.M. Liberman, P.C. Delattre, and F.S. Cooper. (1957) Acoustic cues for the perception of initial /w,j,r,l/ in English. Word 13, 24-43.

A Direct Magnitude Scaling Method to Investigate Categorical Versus Continuous Modes of Speech Perception

M.D. Vinegrad[+]
Haskins Laboratories, New Haven

Summary. The perception of synthetic steady-state vowels, synthetic consonant-vowel syllables, and pure tones was investigated using a psychophysical scaling procedure involving direct magnitude estimation. In each of the three stimulus classes, there were thirteen members equally spaced along a physical continuum; the investigation was designed to measure the degree to which the stimulus members appeared to be evenly spaced along a corresponding perceptual continuum. The experimental technique required the subject to judge the members of each set of stimuli in terms of their similarity to each other. The results suggested that for stops, the perceptual spacing depended upon phoneme identification but that for vowels, the spacing was relatively independent of phoneme identity. The vowel data, in fact, approximated to the tone data (included to provide a nonspeech comparison). The results are interpreted in terms of the notion of categorical versus continuous modes of speech perception.

[+]Currently, Goldsmith's College, University of London

Acknowledgement. The author is indebted to Dr. A.M. Liberman for his encouragement, suggestions, and helpful criticism through the course of this research.

The experiment reported here is a psychophysical study of synthetic speech perception. The primary aim was to investigate quantitatively the tendency for stop consonants to be perceived categorically and steady-state vowels to be perceived continuously. The method used was a stimulus scaling technique that has proved useful in other areas of perceptual research. In the experiment, a sequence of three stimuli was presented to a subject; the stimuli were spaced along a physical continuum and the subject was required to judge how the stimuli appeared to be spaced along a perceptual continuum. The subject indicated his judgment by spacing points along a line.

The technique is both simple and direct. A variety of studies using nonspeech stimuli have shown the reliability and usefulness of this approach (Torgerson, 1958). Speech stimuli have not previously been studied in this way, partly because of their complexity and multidimensionality. Most of the studies found in the scaling literature are limited to situations where changes in both stimulus and perception are unidimensional. One aim of this investigation, therefore, was to test the suitability of the method for synthetic speech stimuli. At a practical level, scaling has the advantage of being relatively easy to carry out and, in addition, directly reflects the way stimuli are perceived by a subject. Other psychophysical procedures (for example, discrimination measurements) lead only to inferences about the mode of perception. Results from investigations using discrimination techniques (Liberman et al., 1967; Stevens et al., 1969) suggest that there may be different modes of perception for vowels and stop consonants. The present experiment is intended to be a further examination of this question. The stimuli used were steady-state vowels, stop consonants, and pure tones; the tones were included to provide a nonspeech comparison.

Description of Stimuli

The speech sounds were thirteen steady-state vowels and thirteen consonant-vowel syllables. They were synthesized by Stevens, Liberman, Studdert-Kennedy, and Öhman (1969) on the OVE II speech synthesizer at the speech transmission laboratory at the Royal Institute of Technology at Stockholm. A full description of the stimuli is given in Stevens et al. (1969). In each set, the stimuli (numbered 1 through 13, for reference) were evenly spaced along a physical continuum. Listened to in order, the vowels appeared to form a smooth series running through the American English vowels /i/, /I/, and /ɛ/. The physical

spacing of the thirteen vowels corresponded to changes in the frequencies of the first three formants of a five-formant pattern. Moving along the continuum from stimulus 1 to stimulus 13, the frequency of the first formant rose in approximately equal steps while the frequencies of the second and third formants decreased in approximately equal steps. For stimulus 1, the respective values of the first three formants were 270.5, 2300, and 3019 cps and for stimulus 13, the respective values were 530.5, 1858, and 2492 cps. Small deviations from even spacing existed because formant frequencies could only be set within a few cps. The bandwidths of the first three formants were fixed at 60, 80, and 100 cps. The frequencies of the fourth and fifth formants were constant throughout. The duration of each stimulus was 300 msec. The consonant-vowel syllables consisted of a stop consonant followed by the vowel /ɛ/. The stimuli (again labeled 1 through 13), when listened to in order, seemed to form a series broken into three segments each characterized by a change in stop. To most North American English listeners, the transition seemed to be /g/---/d/---/b/. Each of the thirteen stimuli was 300 msec in duration. The final 260 msec corresponded to the vowel portion of the syllable; it was a steady-state vowel with the first three formants fixed at 700, 1550, and 2600 cps. Before reaching these fixed values, the three formants underwent transitions along parabolic contours. The transitions for the second and third formants started at different points for different stimuli, and it was in terms of these starting points that the stimuli were spaced along the physical continuum. Going along the continuum from stimulus 1 to 13, the starting frequency for the second formant decreased in equal steps, while the starting frequency for the third formant increased in equal steps as far as stimulus 7 and then decreased in equal steps over the remainder of the range. This variation was designed to parallel the change in speech sound to be expected if the place of consonantal articulation were to be moved in thirteen equal stages from velar to alveolar to labial position. The set of thirteen tones was recorded directly from an audiogenerator. The frequencies of stimuli 1 and 13 were 250 and 298 cps, respectively, with the eleven intermediate stimuli evenly spaced in steps of 4 cps. The output of the audiogenerator was constant and was recorded on a Roberts tape recorder.

Experimental Procedure

In preparing stimuli for the experiment, multiple copies of the

original recordings of the speech stimuli were made on the Roberts recorder. Magnetic tape segments of each vowel and consonant-vowel syllable were then cut and spliced to form sequences of stimuli suitable for scaling. Similar sequences of tones were made by splicing magnetic tape segments of each frequency recorded from the audiogenerator.

The experiment was divided into three parts: (1) vowels, (2) stops, (3) tones. A single scaling procedure was used throughout. All subjects did the three parts in the same order and each part was completed before the subject had any experience with the stimuli of a later part.

The scaling procedure required the subject to listen to two sequences of stimuli and then to make a judgment. The first sequence always contained the same seven stimuli, viz., 1-3-5-7-9-11-13. The second sequence always contained three stimuli, viz., 1-x-13, where x is any of the thirteen stimuli. The subject was required to make a judgment about stimulus x; he was required to indicate how similar or close stimulus x seemed to be to stimulus 1 (or stimulus 13). He was not required to identify x in absolute terms but merely to indicate the position of x by marking a point on a line such that the point bore the same position in relation to the ends of the line as stimulus x bore to stimuli 1 and 13. The stimulus presentation is more cumbersome than is usual for direct magnitude estimation. The procedure was devised by trial and error and was designed to eliminate context effects. These are discussed in more detail below.

The subject was given a straight line, seven inches in length, and told that the left-hand end of the line was to be taken as representing the first stimulus and the right-hand end as representing the third stimulus. If the middle stimulus sounded exactly like the first, the subject was instructed to place a point at the extreme left-hand end of the line; if the middle stimulus sounded exactly like the third stimulus, the subject was instructed to place a point at the extreme right-hand end; if the middle stimulus sounded slightly different from the first, the instruction was to place a point slightly in from the left; and so on. Demonstrations and practice were provided until the subject had mastered the task. In addition, the subject was told that, although there were a number of different stimuli that might occur in the middle position, any one of them might be repeated. Subjects were discouraged from trying to identify the middle stimulus; they were urged to respond according to how similar the stimuli sounded. The two sequences were presented repeatedly but with random rotation in the choice of stimulus to fill the x position.

The subject was provided with a sheet with six seven-inch lines horizontally spaced out between two verticals. There were no labels or other marks to guide the subject in his judgment. The subject was told to use one line per judgment and to turn over to new sheets as necessary. No description of the stimuli was given; subjects were left to form their own frames of reference.

For part one of the experiment, a tape was made containing thirty-nine replications of the two sequences of stimuli; the middle position was filled by each of the thirteen stimuli three times; the order was random. There were two versions of the tape, one a partial rerandomization of the other. For parts two and three, each tape contained fifty-two replications of the stimulus sequences; the middle position was filled by each of the thirteen stimuli four times. The randomization was restricted so that each of the thirteen stimuli occurred twice in the first twenty-six presentations and twice in the second set of twenty-six. (The second set of twenty-six was, in fact, a partial rerandomization of the first twenty-six.) The vowel data were collected over eight experimental sessions, and the stop and tone data were each collected over six sessions. In an experimental session, the subject listened to a tape once through, making 39 judgments in the case of the vowels and 52 in the case of the stops and tones. In total, each subject made 312 judgments on each of the three kinds of stimulus.

The timing of the stimulus presentations was as follows: there was a three-second pause between the end of the first sequence and the beginning of the second and a five-second period for the subject to make his judgment, and then the cycle began again with the repetition of the first stimulus sequence. In each sequence, there was a half-second pause between one stimulus and the next. The five-second judgment period appeared to be optimum; longer periods left too much time for doubt, and subjects found the pace helped them form a suitable set for responding.

The tapes were played on a Hewlett Packard tape deck through a loudspeaker in a language laboratory. Subjects worked simultaneously but were not able to see each other's responses. Practice trials were given at the beginning of each session.

Context Effects

The experimental procedure required the subject to make a succession of judgments, and because there was no absolute standard to judge by, the subject

tended to make judgments in relation to each other. The practical consequence was that one judgment was partly determined by the preceding one. In the preliminary experiments, this was a serious source of error. With the vowels and tones, different sets of judgments were obtained for the same triplets presented in different orders, but with the stops, the effect was either negligible or absent. It was to overcome the context effect that each triplet was preceded by the longer (constant) sequence of stimuli. The constant sequence seemed to reset the subject's frame of reference and to eliminate, or drastically reduce, interference from the preceding judgment. In order to provide a uniform procedure throughout the experiment, the constant sequence was also used when scaling the stops. The fact that the context effect occurred with the vowels and tones but not with the stop consonants is an important difference between these classes of stimuli.

Subjects

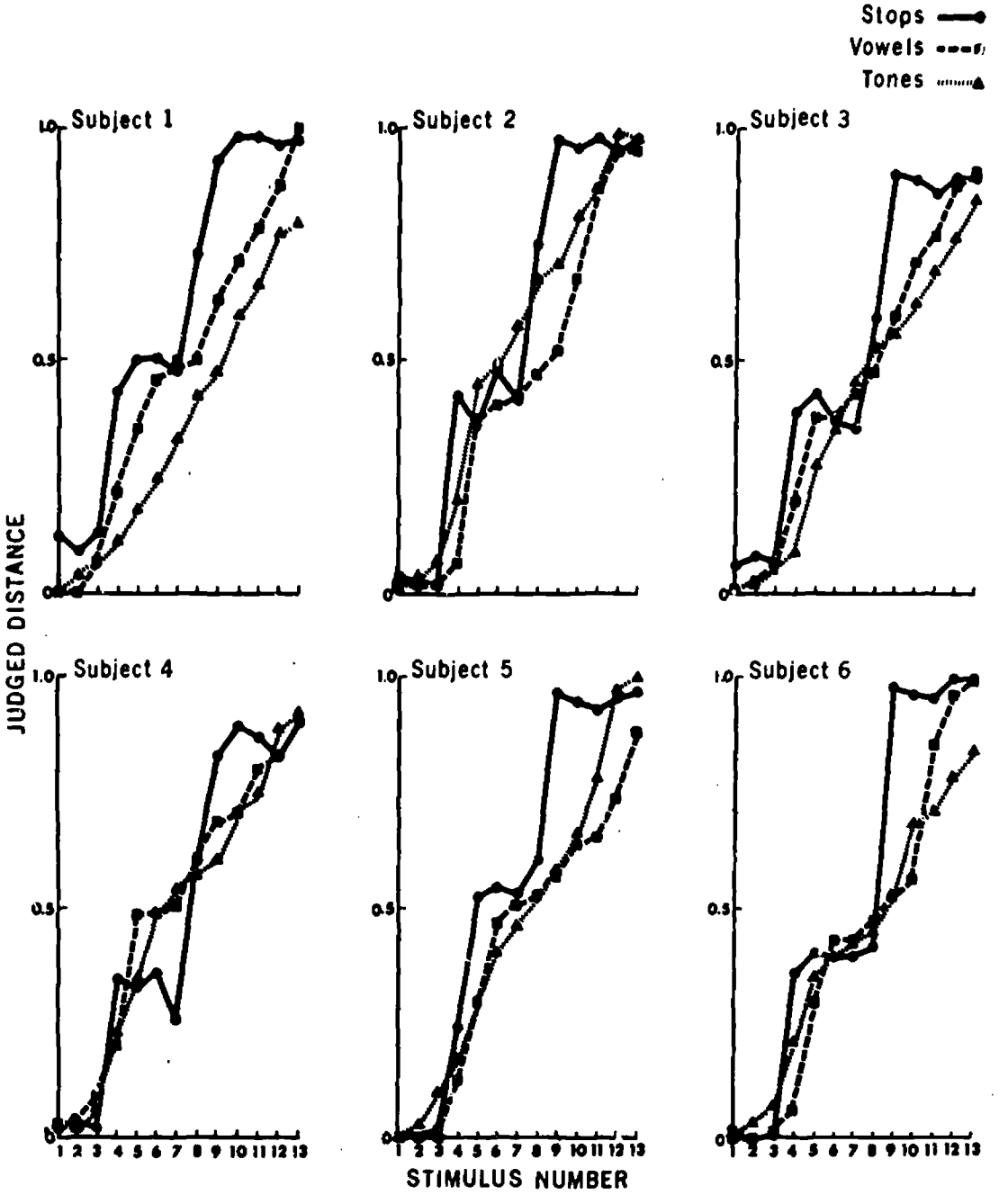
Six Canadian English-speaking undergraduates at a university in Ontario served as subjects. Their ages were between 18 and 21; three were male and three female.

Results and Discussion

In each part of the experiment, the subject judged each of the thirteen stimuli twenty-four times. The judgments were tabulated by measuring the distance of each point marked by the subject from the left-hand end of the line. The mean distance (i.e., judgment) for each stimulus is shown on the ordinates in Figure 1. Results are shown separately for each subject. The ordinates are calibrated so that 0 corresponds to a point marked at the extreme left-hand end of the line and 1.0 to a point marked at the extreme right-hand end of the line. There are clear individual differences, but the curves tend to exhibit certain common features from subject to subject. For stops, the curves seem to be broken into three segments; for tones, they seem essentially continuous; while for vowels, there is a tendency for the curves to be intermediate in form to the other two. Figure 2 shows mean curves calculated over the six subjects.

The clear trends and the consistency from subject to subject seem an adequate answer to one question posed by the study: the speech stimuli are scalable by the method used.

The judged distance along a perceptual continuum of each member of a series of thirteen stimuli.



Note: 0 and 1.0 represent points on the continuum corresponding to the first and last members of the series. There were three series of stimuli: vowels, stops, and pure tones.

FIG. 1

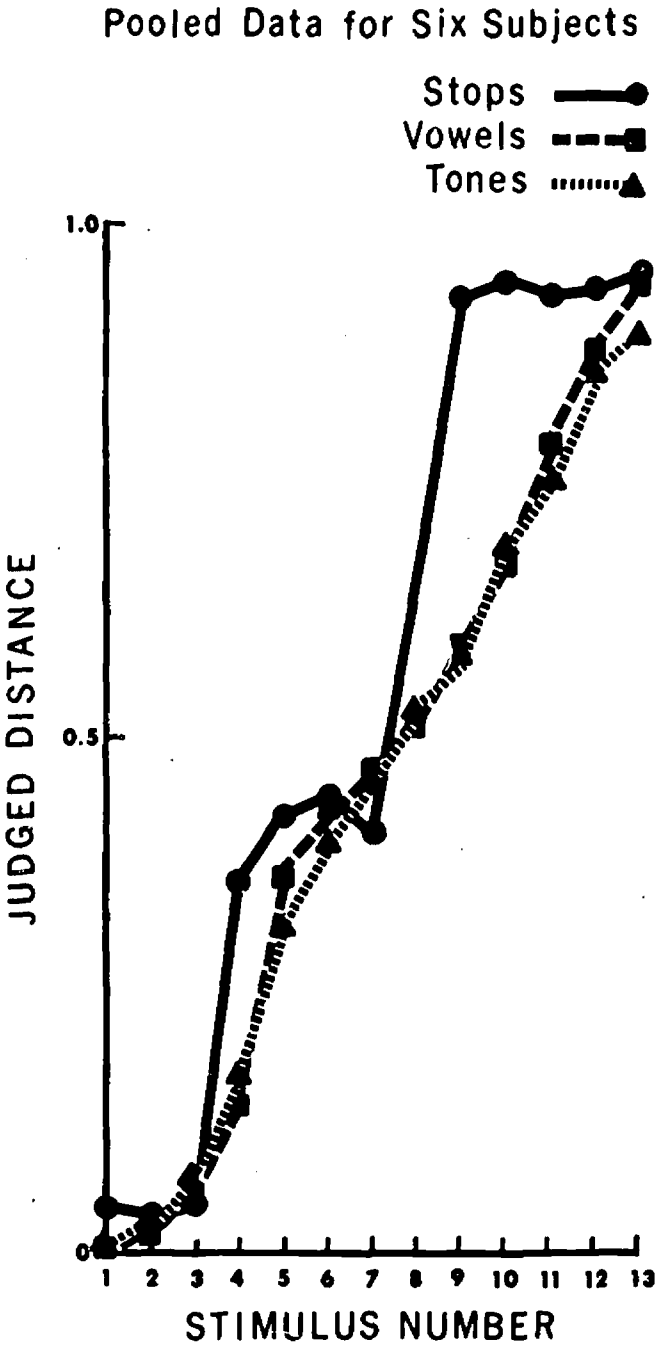

Pooled Data for Six Subjects
Stops
Vowels
Tones
1.0
0.5
0
JUDGED DISTANCE
1 2 3 4 5 6 7 8 9 10 11 12 13
STIMULUS NUMBER

FIG. 2

The speech stimuli were spaced along a physical continuum covering three phonemes so that each sequence cut across phoneme boundaries. Adjacent stimuli either fell within a single phoneme category or fell in a region of transition from one phoneme category to the next. How did these phoneme boundaries affect the judgments? The effect of phoneme boundaries on perception has been reported by a number of workers: Eimas (1963); Griffith (1958); Liberman et al. (1957); Studdert-Kennedy et al. (1963, 1964); Fry et al. (1962); Stevens et al. (1963). The synthetic speech stimuli of the present investigation were previously used in a study by Stevens, Liberman, Studdert-Kennedy and Öhman (1969). These workers established phoneme boundaries for both the vowels and stops. The stimuli in each set showed some overlapping, but the general picture was clear. For stops, the preponderance of identification responses placed stimuli 1, 2, and 3 in phoneme category /g/; stimuli 4, 5, 6, and 7 in phoneme category /d/; and stimuli 9, 10, 11, 12, and 13 in phoneme category /b/. For vowels, a preponderance of responses placed stimuli 1, 2, 3, and 4 in phoneme category /i/; stimuli 5, 6, 7, 8, and 9 in phoneme category /I/; and stimuli 10, 11, 12, and 13 in phoneme category /ɛ/.

In the present experiment, these phoneme boundaries can be seen to have a marked effect upon the judgment of the stops and a small, perhaps negligible, effect upon the judgment of the vowels. In general, for stops, there was little change in judgment from one stimulus to the next when the stimuli fell in the same phoneme category but a marked change in judgment when the stimuli crossed a phonema boundary. For vowels, the change in judgment seemed to be more a continuous function of stimulus variation along a continuum and was little affected by the phoneme boundaries. The nonspeech stimuli, the tones, gave results like the vowels, only the curves are somewhat smoother.

These results tend to confirm the findings of Stevens et al. (1969), who also obtained discrimination functions for the stimuli. They showed that discrimination of adjacent members of the stimulus series was poorest when the pair fell at the center of a phoneme category and best when the pair fell in different phoneme categories. This phenomenon was far more marked in the case of the stops than of the vowels. In the case of the stops, discrimination was little better than would have been the case if the subject had been able to discriminate about as well as he could identify. In the case of the vowels, however, discrimination was considerably

better than could be predicted from the identification data; as in the present experiment, the vowel data tended to bear more of a continuous relation to the stimulus variation.

Thus, the two experiments, using very different psychophysical procedures, agree in making a distinction between steady-state vowels and stop consonants that may amount to a difference in mode of perception. The stimuli for the stop consonants tend to be perceived in terms of category of identification to a much greater extent than the stimuli for the steady-state vowels; the difference amounts to a greater limitation on the perception of stimulus differences for stops than for vowels. The vowels seem to be closer to the tones in mode of perception, but it is possible that vowels presented in a context of other speech sounds would behave more like the consonants. The present scaling technique seems to provide a fairly suitable method of comparing the perceptual characteristics of different classes of speech sound.

References

Eimas, P.D. (1963) The relation between identification and discrimination along speech and non-speech continua. Language and Speech 6, 206.

Fry, D.B., A.S. Abramson, P.D. Eimas, and A.M. Liberman. (1962) The identification and discrimination of synthetic vowels. Language and Speech 5, 171.

Griffith, B.C. (1958) A study of the relation between phoneme labelling and discriminability in the perception of synthetic stop consonants. Unpublished Ph.D. dissertation, University of Connecticut.

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

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

Stevens, K.N., S.E.G. Öhman, and A.M. Liberman. (1963) Identification and discrimination of rounded and unrounded vowels. J. Acoust. Soc. Amer. 35, 1900 (Abstract).

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

Studdert-Kennedy, M., A.M. Liberman, and K.N. Stevens. (1963) Reaction time to synthetic stop consonants and vowels at phoneme centers and at phoneme boundaries. J. Acoust. Soc. Amer. 35, 1900 (Abstract).

Studdert-Kennedy, M., A.M. Liberman, and K.N. Stevens. (1964) Reaction time during the discrimination of synthetic stop consonants. J. Acoust. Soc. Amer. 36, 1989 (Abstract).

Torgerson, W.S. (1958) Theory and Methods of Scaling. (John Wiley & Sons, New York).

On the Speech of Neanderthal Man*

Philip Lieberman+ and Edmund S. Crelin++

Language is undoubtedly the most important factor that differentiates man from other animals. It is, in itself, a system of abstract logic, allowing man to extend his rational ability. Indeed, it has often been virtually equated with man's abstract logical ability (Chomsky, 1966). It is therefore of great interest to know when a linguistic ability similar to that of modern man evolved. One of the most significant determinants of the form of man's linguistic ability is his use of "articulate" speech. We will discuss the speech ability of an example of Neanderthal man, the La Chapelle-aux-Saints fossil, in the light of its similarity to certain skeletal features in Newborn humans. We herein use the term "Neanderthal" as referring to the so-called classic Neanderthal man of the Würm or last glacial period.[1]

Our discussion involves essentially two factors. We have previously determined by means of acoustic analysis that Newborn humans, like nonhuman primates, lack the anatomical mechanism that is necessary to produce articulate speech (Lieberman, 1968; Lieberman et al., 1968, 1969), that is, they cannot produce the range of sounds that characterizes human speech. We can now demonstrate that the skeletal features of Neanderthal man show that his

---

*To be published in Linguistic Inquiry 2, No. 2, March 1971.

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

++Yale University School of Medicine, New Haven.

Acknowledgements. We thank Professors W. Henke and D.H. Klatt for providing the computer program and suggesting some of the supralaryngeal area functions in the speech synthesis procedure. We also would like to thank Professors H.V. Vallois, J.E. Pfeiffer, D. Pilbeam, W.S. Laughlin, W.W. Howells, and F. Bordes and Dr. K.P. Oakley for many helpful comments, as well as Drs. Y. Coppens and J.L. Heim of the Musée de L'Homme for making the La Chapelle-aux-Saints and La Ferrassie fossils available.

[1]The La Chapelle-aux-Saints fossil as described by Boule (1911-13) is perhaps the archetypal example of "classic" Neanderthal man. As Howells (1968) notes, there is a class of classic Neanderthal fossils that can be quantitatively differentiated from other fossil hominids. We recognize that some of these other fossil hominids exhibit characteristics that are intermediate between classic Neanderthal man and modern Man. These fossils may have possessed intermediate degrees of phonetic ability, but we will limit our discussion to the La Chapelle-aux-Saints fossil in this paper.

supralaryngeal vocal apparatus was similar to that of a Newborn human. We will also discuss the status of Neanderthal man in human evolution.

The Anatomical Basis of Speech

Human speech is essentially the produce of a source (the larynx for vowels) and a supralaryngeal vocal tract transfer function. The supralaryngeal vocal tract, which extends from the larynx to the lips, in effect filters the source (Chiba and Kajiyama, 1958; Fant, 1960). The activity of the larynx determines the fundamental frequency of the vowel, whereas its formant frequencies are the resonant modes of the supralaryngeal vocal tract transfer function. The formant frequencies are determined by the area function of the supralaryngeal vocal tract. The vowels /a/ and /i/, for example, have different formant frequencies although they may have the same fundamental frequency. Sounds like the consonants /b/ and /d/ may also be characterized in terms of their formant frequencies. Consonants, however, typically involve transitions or rapid changes in their formant frequencies, which reflect rapid changes in the area function of the supralaryngeal tract. The source for many consonants like /p/ or /s/ may be air turbulence generated at constrictions in the vocal tract.

A useful mechanical analog to this aspect of speech production is a pipe organ. The musical quality of each note is determined by the length and shape of each pipe. (The pipes have different lengths and may be open at one end or closed at both ends.) The pipes are all excited by the same source. The resonant modes of each pipe determine the pipe's "filter" function. In human speech, the phonetic qualities that differentiate vowels like /i/ and /a/ are determined by the resonant modes of the supralaryngeal vocal tract.

The acoustic theory of speech production, which we have briefly outlined, thus relates an acoustic signal to a supralaryngeal area function and a source. It is therefore possible to calculate the range of sounds that an animal can produce if the range of supralaryngeal vocal tract area function variation is known. The phonetic repertoire of the animal can be further expanded if different sources are used with similar supralaryngeal vocal tract area functions. We can, however, isolate the constraints that the range of supralaryngeal vocal tract variation will impose on the phonetic repertoire by studying the effects of different source functions. In short, we can see what limits would be imposed on the Neanderthal phonetic repertoire by studying his supralaryngeal vocal tract even though we cannot reconstruct his larynx.

Skeletal Structure

The human Newborn specimens used in this study were six skulls and six heads and necks completely divided in the midsagittal plane plus all of the cadavers dissected by the coauthor (E.S.C.) for his book on newborn anatomy (Crelin, 1969). The specimens of adult Man were fifty skulls, six heads and necks completely divided in the midsagittal plane, and the knowledge derived from dissections of adult cadavers made by the coauthor and his students during twenty continuous years of teaching human anatomy. The Neanderthal specimens were casts of two skulls with mandibles and an additional mandible of the fossil man from La Chapelle-aux-Saints described by Boule (1911-13). The casts were purchased from the Museum of the University of Pennsylvania. Detailed measurements were made on the casts and from photographs of this fossil. The original fossil was also examined at the Musée de L'Homme in Paris by one of the authors (P.L.). Skulls of a chimpanzee and of an adult female gorilla were also studied.

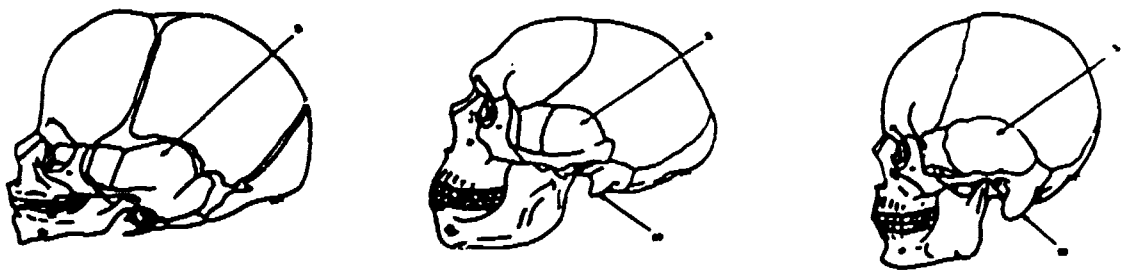
When the skulls of Newborn and adult Man are placed beside the cast of the Neanderthal skull, there appears to be little similarity among them, especially from an anterior view (Fig. 1). Much of this is due to the disparity in size: when they are all made to appear nearly equal in size and are viewed laterally, the Newborn skull more closely resembles the Neanderthal skull than that of adult Man (Fig. 2). The Newborn and Neanderthal skulls are relatively more elongated from front to back and relatively more flattened from top to bottom than that of adult Man. The squamous part of the temporal bone is similar in Newborn and Neanderthal (Fig. 2). The fact that the mastoid process is absent in Newborn and relatively small in Neanderthal adds to their similarity when compared with the skull of adult Man shown in Figure 2. However, the size of the mastoid process varies greatly in adult Man. It is not unusual to find mastoid processes in normal adult Man as small as those of Neanderthal, especially in females. The mastoid process is absent in the chimpanzee and relatively small in the gorilla. Other features that make the Newborn and Neanderthal skulls appear similar from a lateral view are the shape of the mandible and the morphology of the base of the skull.

Newborn and Neanderthal lack a chin; thus they share a pongid characteristic (Fig. 2). The body of the Newborn and Neanderthal mandible is longer than the ramus, whereas they are nearly equal in adult Man (Fig. 3). The posterior border of the Newborn and Neanderthal mandibular ramus is more inclined away from the vertical plane than is that of adult Man. In

SKULL OF NEWBORN
CAST OF NEANDERTHAL SKULL
SKULL OF ADULT MAN
A
B
C

FIG. I

LATERAL VIEWS OF SKULLS



A

NEWBORN

B

NEANDERTHAL

C

ADULT MAN

M - Mastoid Process
S - Squamous Portion of Temporal Bone

FIG. 2

LATERAL VIEWS OF SKULLS

NEWBORN NEANDERTHAL ADULT MAN

L - Angle of Pterygoid Lamina
S - Angle of Styloid Process
P - Coronoid Process
N - Notch
R - Ramus
M - Body

FIG. 3

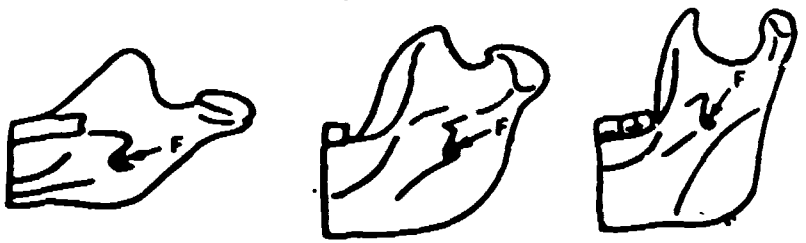
Newborn and Neanderthal, there is a similar inclination of the mandibular foramen leading to the mandibular canal through which the inferior alveolar artery and nerve pass (Fig. 4). The mandibular coronoid process is broad and the mandibular notch is relatively shallow in Newborn and Neanderthal (Fig. 3).

The pterygoid process of the sphenoid bone is relatively short and the posterior border of its lateral lamina is more inclined away from the vertical plane in Newborn and Neanderthal when compared with adult Man (Fig. 3). The styloid process is also more inclined away from the vertical plane in Newborn and Neanderthal than in adult Man (Fig. 3). There are sufficient fossil remains of the Neanderthal left styloid process to determine accurately its original approximate size and inclination.

The dental arch of the Newborn and Neanderthal maxillas is U-shaped, a pongid feature, whereas it is more V-shaped in adult Man (Fig. 5).

In the Newborn skull the anteroposterior length of the palate is less than the distance between the posterior border of the palate and the anterior border of the foramen magnum, i.e., 2.1 cm average (range 2.0-2.2 cm) and 2.6 cm average (range 2.5-2.7 cm) respectively (Fig. 5). In Neanderthal, the length of the palate is equal to the distance between the palate and the foramen magnum, i.e., 6.2 cm. In the skull of adult Man, the length of the palate is greater than the distance between the palate and the foramen magnum, i.e., 5.1 cm average (range 4.6-5.7 cm) and 4.1 cm (range 3.6-4.9 cm) respectively. Only two of the fifty skulls of modern, adult Man studied were exceptions. In one, the distance between the palate and the foramen magnum was 0.4 cm greater than the length of the palate, and in the other, the distances were the same (4.6 cm). Note the great absolute distance between the palate and the foramen magnum in Neanderthal man compared to adult Man. The greater distance between the palate and the foramen magnum in Newborn and Neanderthal when compared with adult Man is related to the similar relative size and shape of the roof of the nasopharynx in Newborn and Neanderthal. The basilar part of the occipital bone, between the foramen magnum and the sphenoid bone, is only slightly inclined away from the horizontal toward the vertical plane in these specimens (Fig. 5). Therefore, the roof of the nasopharynx is a relatively shallow and elongated arch, whereas in adult Man it forms a relatively deep, short arch (Figs. 8 and 9). In adult Man, without exception, the basilar part of the occipital bone is inclined more toward the vertical plane than the horizontal plane. The vomer bone in Newborn and Neanderthal is relatively shorter in its vertical height than is that in Man, and its posterior border is inclined away from the vertical plane to a greater degree, thus affecting the shape of the roof of the nasopharynx

DEEP SURFACE OF RAMUS MANDIBLE



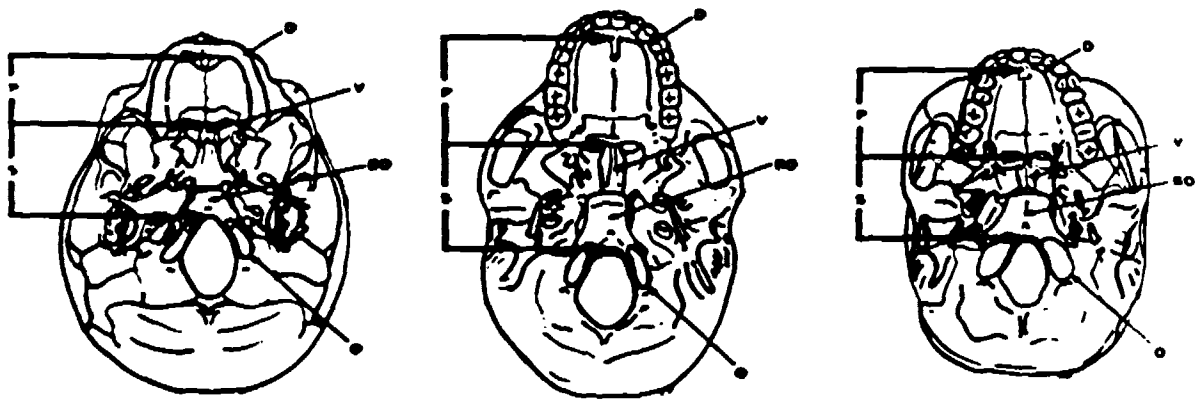
A B C

NEWBORN NEANDERTHAL ADULT MAN

FIG. 4

F - Mandibular Foramen

INFERIOR VIEWS OF BASE OF SKULL



A — NEWBORN

B — NEANDERTHAL

C — ADULT MAN

D - Dental Arch
P - Palate
S - Distance Between Palate and Foramen Magnum

V - Vomer Bone
BO - Basilar Part of Occipital
O - Occipital Condyle

FIG. 5

(Figs. 5 and 9).

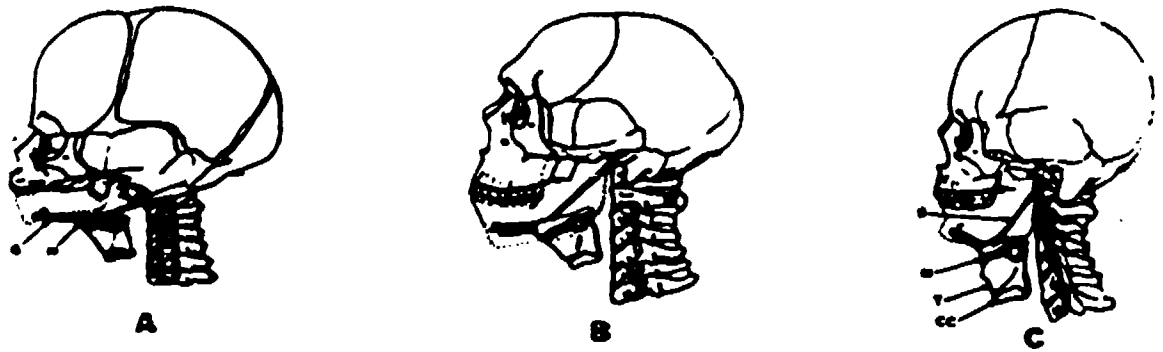
In Figure 5 the foramen magnum is shown to be elongated in the antero-posterior plane in the Newborn, Neanderthal, and adult Man. Its shape is variable in both Newborn and adult Man, where it is frequently more circular. The occipital condyles of Neanderthal are similar to those of the Newborn and the gorilla in that they are relatively small and elongated. Since the second, third, and fourth cervical vertebrae of the man from La Chapelle-aux-Saints are lacking, they were reconstructed to conform with those of adult Man (Fig. 6). The Neanderthal skull is placed on top of an erect cervical vertebral column instead of on one sloping forward as depicted by Boule (1911-13) and Keith (1925). This is in agreement with Straus and Cave (1957). In addition, the spinous processes of the lower cervical vertebrae shown for adult Man in Figure 6 are curved slightly upward. They are from a normal vertebral column and were purposely chosen to show that those of Neanderthal were not necessarily pongid in form. In fact, the cervical vertebral column of Neanderthal also resembles that of Newborn (Fig. 6).

Reconstruction of the Supralaryngeal Vocal Tract

In order to reconstruct the supralaryngeal vocal tract of Neanderthal, it was essential to locate the larynx properly. Because of the many similarities of the base of the skull and the mandible between Newborn and Neanderthal, coupled with the known detailed anatomy of Newborn, of adult Man, and of apes, it was possible to do this with a high degree of confidence (Fig. 6). Although the larynx was judged to be positioned as high in Neanderthal man as in Newborn and apes, it was, in this model, dropped to a slightly lower level to give the Neanderthal every possible advantage in his ability to speak.

Once the position of the larynx in Neanderthal was determined, it was a rather straightforward process to reconstruct his tongue and pharyngeal musculature (Fig. 7). The next step was to reconstruct the vocal tract of Neanderthal by building his laryngeal, pharyngeal, and oral cavities with modeling clay in direct contact with the skull cast. After this was done, a silicone-rubber cast of the air passages, including the nasal cavity, was made from the clay mold. At the same time, similar casts were made of the air passages, including the nasal cavity, of Newborn and adult Man. This was done by filling each side of the split air passages separately in the sagittally-sectioned Newborn and adult Man heads and necks to ensure perfect filling

SKULL, VERTEBRAL COLUMN, AND LARYNX



NEWBORN

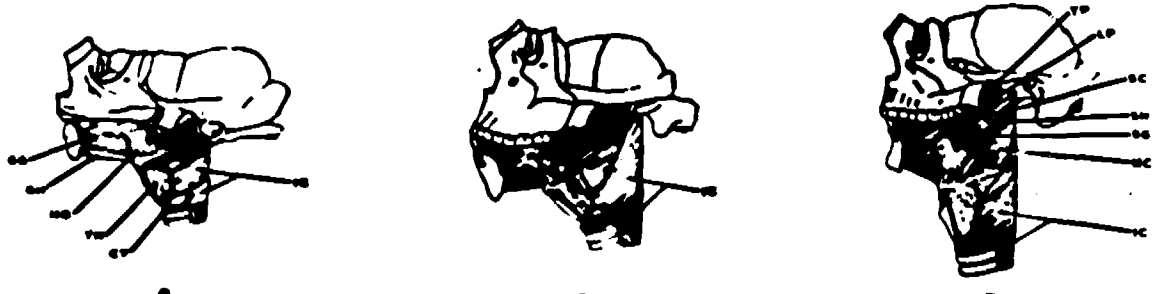
RECONSTRUCTION OF NEANDERTHAL

ADULT MAN

FIG. 6

G - Genichyoid Muscle
H - Hyoid Bone
S - Stylohyoid Ligament
M - Thyrohyoid Membrane
T - Thyroid Cartilage
CC - Cricoid Cartilage

TONGUE AND PHARYNGEAL MUSCULATURE



A — NEWBORN

B — RECONSTRUCTION OF NEANDERTHAL

C — ADULT MAN

GG - Genioglossus
GH - Geniohyoid
HG - Hyoglossus
TH - Thyrohyoid
CT - Cricothyroid
TP - Tensor Veli Palatini
LP - Levator Veli Palatini
SC - Superior Pharyngeal Constrictor
MC - Middle Pharyngeal Constrictor
IC - Inferior Pharyngeal Constrictor
SH - Stylohyoid
SC - Styloglossus

FIG. 7

of the cavities. The casts from each side of a head and neck were then fused together to make a complete cast of the air passages.

Even though the cast of the Newborn air passages is much smaller than those of Neanderthal and adult Man, it is apparent (Fig. 8) that the casts of the Newborn and Neanderthal are quite similar and have pongid characteristics (Negus, 1949). When an outline of the air passages from all three are made nearly equal in size, one can more readily recognize the basic differences and similarities (Fig. 9). Although the nasal and oral cavities of Neanderthal are actually larger than those of adult Man, they are quite similar in shape to those of Newborn, being very elongated. The high position of the opening of the larynx into the pharynx in Newborn and apes is directly related to the high position of the hyoid bone; the opening of the larynx into the pharynx is, therefore, in a high position (Fig. 9). The development of the Newborn pharynx into the adult type is primarily a shift in the location of the opening of the larynx into it from a high to a low position. This is probably the result of differential growth where the posterior third of the tongue, between the foramen cecum and the epiglottis, shifts from a horizontal resting position within the oral cavity to a vertical resting position to form the anterior wall of the oral part of the pharynx (Fig. 9). In this shift, the epiglottis becomes widely separated from the soft palate. Also, the large, posterior portion of the pharynx below the opening of the larynx in the Newborn is lost as it, in large part, becomes part of the acquired supralaryngeal portion.

Supralaryngeal Vocal Tract Limits on the Neanderthal Phonetic Inventory

We cannot say much about either the laryngeal source or the dynamic control of Neanderthal man's vocal apparatus. We can, however, determine some of the limits on the range of sounds that Neanderthal man could have produced by modeling the reconstruction of his supralaryngeal vocal tract.

We measured the cross-sectional area of the Neanderthal and Newborn vocal tracts shown in Figure 8 at 0.5 cm intervals. These measurements gave us "neutral" area functions which we perturbed toward area functions that would be reasonable if a Newborn or a Neanderthal vocal tract attempted to produce the full range of human vowels. This can be conveniently done by attempting to produce vowels that are as near as possible to /u/, /a/, and /i/ (the vowels in the words boot, father, and feet). These three vowels delimit the human vowel space (Fant, 1960). We also investigated vocal tract area functions for various consonants. In all of these area functions, we made

CASTS OF AIR PASSAGES

NEWBORN | RECONSTRUCTION OF NEANDERTHAL | ADULT MAN

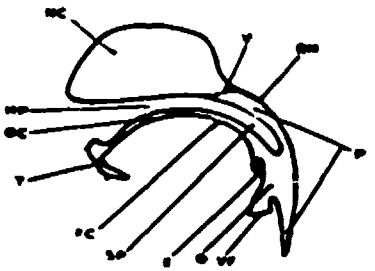
A B C

FIG. 8

Note: The nasal oral and pharyngeal air passages are shown.

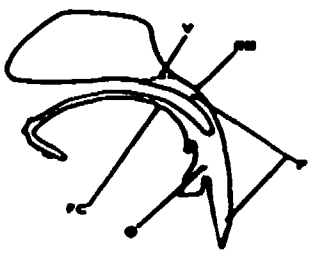
SUPRALARYNGEAL AIR PASSAGES

NEWBORN



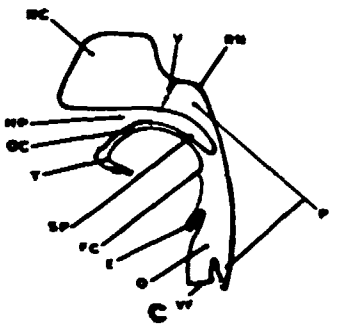
A

RECONSTRUCTION OF NEANDERTHAL



B

ADULT MAN



C

NC - Nasal Cavity
V - Vomer Bone
RN - Roof of Nasopharynx
P - Pharynx

HP - Hard Palate
SP - Soft Palate
OC - Oral Cavity
T - Tip of Tongue

E - Epiglottis
O - Opening of Larynx into Pharynx
VF - Level of Vocal Folds

FIG. 9

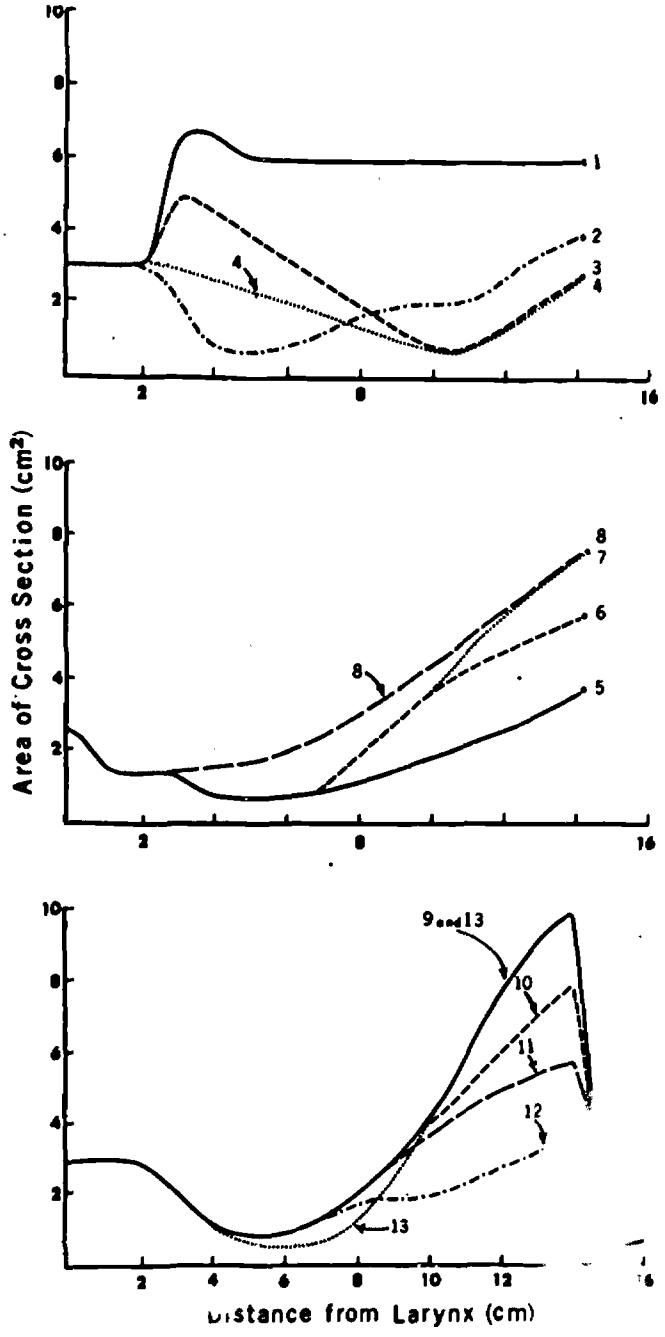
use of our knowledge of the skull and muscle geometry of adult Man and Newborn and the Neanderthal skull as well as cineradiographic data on vocalization in adult Man (Perkell, 1969) and Newborn (Truby et al., 1965). When we were in doubt as, for example, with respect to the range of variation in the area of the larynx, we used data derived from adult Man that would enhance the phonetic ability of the Neanderthal vocal tract (Fant, 1960).

Typical supralaryngeal area functions for the nonnasal portion of the Neanderthal vocal tract are plotted in Figure 10. We were able to determine what sounds would result from these area functions by using them to control a computer-implemented analog of the supralaryngeal vocal tract.

The computer program represented the supralaryngeal vocal tract by means of a series of contiguous cylindrical sections, each of fixed area. Each section can be described by a characteristic impedance and a complex propagation constant, both of which are well-known quantities for uniform cylindrical tubes. Junctions between sections satisfy the constraints of continuity of pressure and conservation of volume velocity (Henke, 1966). In this fashion, the computer program calculated the three lowest formant frequencies of the vocal tract filter system which specify the acoustic properties of a vowel (Chiba and Kajiyama, 1958; Fant, 1960).

In Figure 11, the first and second formant frequencies of the vowels of American English are plotted for a sample of seventy-six adult men, adult women, and children (Peterson and Barney, 1952). The labeled, closed loops indicate the data points that accounted for 90 percent of the samples in each vowel category. The points plotted in Figure 12 represent the formant frequencies that corresponded to our simulated Neanderthal vocal tract. We have duplicated the vowel "loops" of Figure 11 in Figure 12. Note that the Neanderthal vocal tract cannot produce the range of sounds plotted for the human speakers in Figure 11. We have compared the formant frequencies of the simulated Neanderthal vocal tract with this comparatively large sample of human speakers, since it shows that the speech deficiencies of the Neanderthal vocal tract are different in kind from the differences that characterize human speakers, even when the sample includes adult men, adult women, and children. The acoustic vowel space of American English would not appear to be anomalously large compared to other languages, although exhaustive acoustic data is lacking for many languages (Chiba and Kajiyama, 1958; Fant, 1960). It is not necessary to attempt to simulate the sounds of all languages with the computer-implemented Neanderthal vocal tract since the main point that we are trying to establish is whether

AREA FUNCTIONS OF THE SUPRALARYNGEAL VOCAL TRACT OF NEANDERTHAL RECONSTRUCTION MODELED ON COMPUTER



. 10

Note: The area function from 0 to 2 cm is derived from) 60) and represents the distance from the vocal folds to the opening o arynx into the pharynx. Curve 1 is the unperturbed tract. Cur represent functions directed toward a "best match" to the human vowe urves 5-8 are functions directed toward a "best match" to /a/. 9-13 are directed toward /u/.

FORMANT FREQUENCIES OF AMERICAN-ENGLISH VOWELS

FOR A SAMPLE OF 76 ADULT MEN, ADULT WOMEN, AND CHILDREN

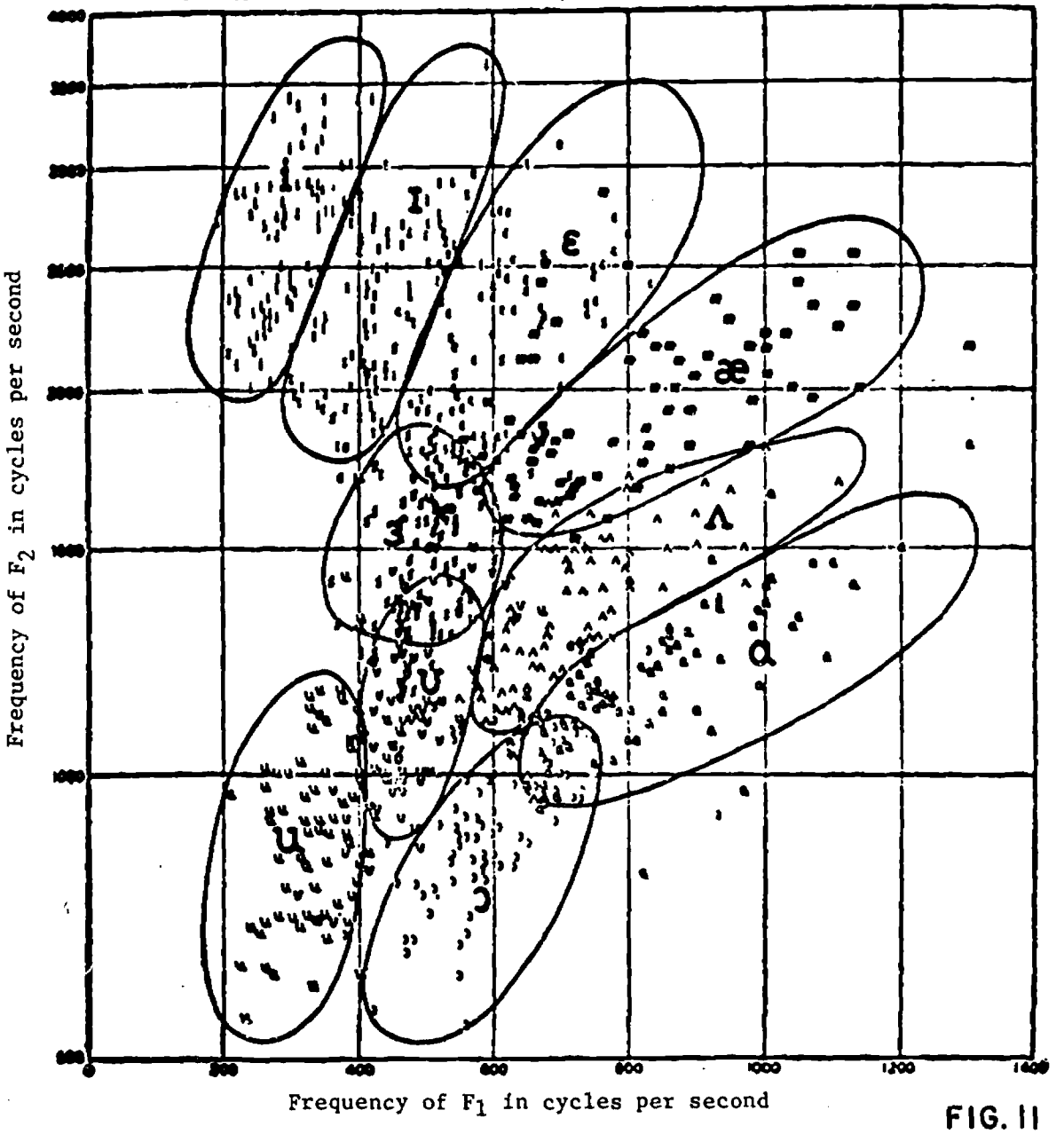
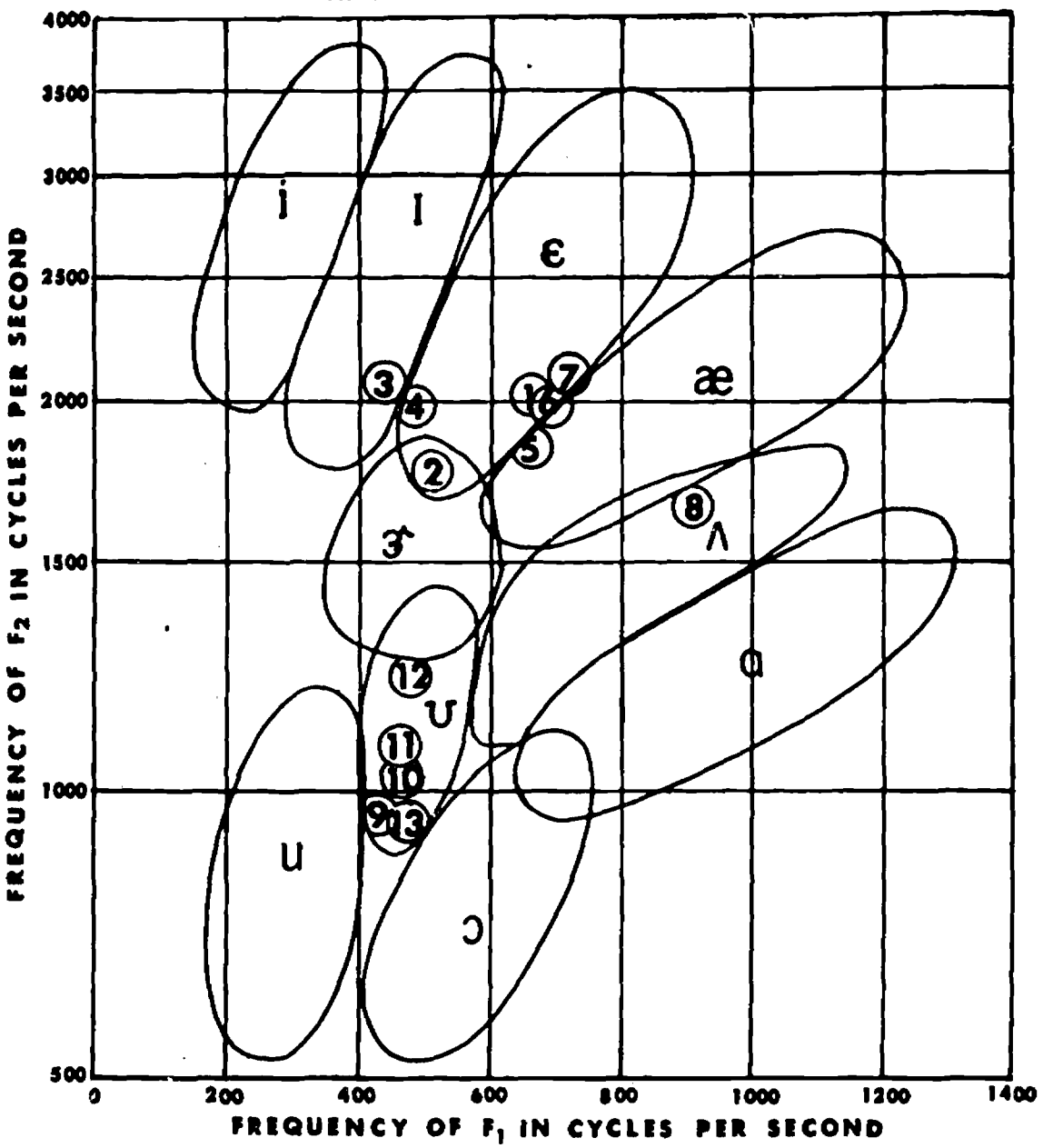


FIG. 11

Note: Closed loops inclose 90 percent of the data points in each vowel category, after Peterson and Barney (1952).

FORMANT FREQUENCIES CALCULATED BY COMPUTER PROGRAM

FOR NEANDERTHAL RECONSTRUCTION



Note: Numbers refer to area functions in Figure 10. The vowel loops of Figure 11 are repeated.

FIG. 12

Neanderthal man could produce the full range of human speech. Figures 11 and 12 show that the Neanderthal vocal tract cannot produce the full range of American English vowels. Note the absence of data points in the vowel loops for /u/, /i/, /a/, and /ɔ/ in Figure 12. Since all human speakers can inherently produce all the vowels of American English, we have established that the Neanderthal phonetic repertoire is inherently limited. In some instances, we generated area functions that would be appropriately human-like, even though we felt that we were forcing the articulatory limits of the reconstructed Neanderthal vocal tract, e.g., functions 3, 9, and 13 in Figure 10. However, even with these articulatory gymnastics, the Neanderthal vocal tract could not produce the vowel range of American English. The computer simulation was also used to generate consonantal vocal tract functions. It indicated that the Neanderthal vocal tract was limited to labial and dental consonants like /b/ and /d/.

The Neanderthal vocal tract also might lack the ability to produce contrasts between nasal and nonnasal sounds. In human speech the nasal cavity acts as a parallel resonator when the velum of the soft palate is lowered, e.g., in the initial consonant of the word mat. The parallel resonator introduces energy minima into the acoustic spectrum and widens the bandwidths of formants (Fant, 1960). In the Neanderthal vocal tract, the posterior pharyngeal cavity, which leads to the esophagus, will act as a parallel resonator whether or not the nasal cavity is coupled to the rest of the vocal tract. The energy minima associated with the parallel pharyngeal resonator, however, occur at rather high frequencies, and it is not clear whether they will have a perceptual effect. Our computer simulation did not allow us to introduce parallel resonators, so we could not investigate this phenomenon quantitatively. It is possible that all Neanderthal vocalizations had a "nasal" or "seminasal" quality.

We modeled the Newborn vocal tract in the same manner as the Neanderthal vocal tract. The computer output of the Newborn vocal tract was in accord with instrumental analyses of Newborn cry and perceptual transcriptions of Newborn vocalizations (Lieberman et al., 1968). The modeling of the Newborn vocal tract thus served as a control on the way in which we estimated the range of supralaryngeal area functions and the synthesis procedure. If we had not been able to synthesize the full range of Newborn vocalizations, we would have known that we were underestimating the range of supralaryngeal vocal tract variation in Neanderthal. But since we followed the same procedures for the Neanderthal and Newborn vocal tracts, and indeed "forced" the Neanderthal vocal tract to

its limits, it is reasonable to conclude that we have not underestimated the phonetic range of the reconstructed Neanderthal vocal tract.

Our computer simulation thus shows that the supralaryngeal vocal tract of Neanderthal man was inherently incapable of producing the range of sounds that is necessary for the full range of human speech. Neanderthal man could not produce vowels like /a/, /i/, /u/, or /ɔ/ (the vowel in the word brought), nor could he produce consonants like /g/ or /k/. All of these sounds involve the use of a variable pharyngeal region like Man's, where the dorsal part of the tongue can effect abrupt and extreme changes in the cross-sectional area of the pharyngeal region independent of activity in the oral region.[2] The area functions in Figure 13 are typical of the human vowels /a/, /u/, and /i/.

The Neanderthal vocal tract, however, has more "speech" ability than that of nonhuman primates. The large, cross-sectional area function variations that can be made in the Neanderthal oral region make this possible, since the Neanderthal mandible has no trace of a simian shelf (Boule, 1911-13) and the tongue is comparatively thick. It can produce vowels like /I/, /e/, /U/, and /æ/ (the vowels in the words bit, bet, put, and bat) in addition to the reduced schwa vowel (the first vowel in about). Dental and labial consonants like /d/, /b/, /s/, /z/, /v/, and /f/ are also possible, although nasal versus nonnasal contrasts may not have been possible. If Neanderthal man were able to execute the rapid, controlled articulatory maneuvers that are necessary to produce these consonants and had he the neural mechanisms that are necessary to perceive rapid formant transitions [special neural mechanisms appear to be involved in Man (Whitfield, 1969; Liberman et al., 1967)], he would have been able to communicate by means of sound. Of course, we do not know whether Neanderthal man had these neural skills, but even if he were able to make optimum use of his speech-producing apparatus, the constraints of his supralaryngeal vocal tract would have made it impossible for him to produce "articulate" human speech, i.e., the full range of phonetic contrasts employed by modern Man.

---

[2]Several studies (Negus, 1949; DeBrul, 1958; Coon, 1966) have suggested that the evolution of the human pharyngeal region played a part in making "articulate" speech possible. Negus (1949) indeed presents a series of sketches based on reconstructions by Arthur Keith where he shows a high laryngeal position for Neanderthal man.

On the Evolutionary Status of Neanderthal Man: Speech Apparatus, Brain, and Language

Of all the living primates, only Man has an extensive supralaryngeal-pharyngeal region that allows all of the intrinsic and extrinsic pharyngeal musculature to function at a maximum for speech production by changing the shape of the supralaryngeal vocal tract (Negus, 1949). It appears that the ontological development of the vocal apparatus in Man is a recapitulation of his evolutionary phylogeny.[3] If so, Neanderthal was an early offshoot from the mainstream of hominids that evolved into modern Man, just as Boule (1911-13) recognized. It is unlikely that Neanderthal man can represent a specialized form of modern Man (Coon, 1966) or an extremely specialized species that evolved from Homo sapiens (Leakey and Goodall, 1969).

---

[3]Apart from the absence of brow ridges and certain other specializations, the total form of the Newborn and Neanderthal skulls makes them members of the same class as differentiated from modern adult Man. The various anatomical features that we have discussed indicate this similarity, but the total similarity of the complex form is most evident to the human pattern recognizer. Human observers are still the best "pattern recognition systems" that exist. Modern statistical and computer techniques, while they are often helpful, have yet to achieve the success of human observers whether music, speech, or "simple" visual forms, like cloud patterns, form the input. Both the Neanderthal and the Newborn skulls have a "flattened out" base where there is space for the larynx to assume a high position with respect to the palate. The anatomical similarities between the Newborn and the Neanderthal skulls are also evident in the La Ferrassie I and Monte Circeo skulls, as well as the La Quina child's skull (estimated age, 8 years).

The La Quina skull, which lacks the massive brow ridges of the adult Neanderthal skulls, retains the anatomical features that result in a flattened out base. These similarities, of course, recall Haeckel's "Law of Recapitulation" (1907). Neanderthal man and modern Man probably had a common ancestor who had a flattened out skull base and a high laryngeal position, but who lacked massive brow ridges. The skulls of Newborn modern man and the La Quina Neanderthal child both point to this common ancestor insofar as they lack massive brow ridges, although they retain the aforementioned similarities. Classic Neanderthal man and the ancestors of modern Man diverged. The massive brow ridges of adult Neanderthal man reflect this divergence. They are a specialization of Neanderthal man. We do not find any trace of brow ridges in Newborn modern man since classic Neanderthal man is not a direct ancestor of modern man. He perhaps is a "cousin." The evidence which many scholars have interpreted as a general and complete refutation of Haeckel's theory should be reconsidered. The process of mutation and natural selection of necessity results in many variations. It is not surprising to find the presence of what appear to be many fossil species that are not in the direct line of human evolution. There is no reason to assume that all of the evolutionary hominid "experiments" are direct ancestors of modern man or that all fossil species of elephants are direct ancestors of modern elephants, etc. Many discussions of Haeckel's theory implicitly make this erroneous assumption when they review ontogenetic and phylogenetic data. Ontogenetic evidence can provide valuable insights into the evolution of living species.

Natural selection would act for the retention of mutations that developed a pharyngeal region like Man's because these developments increase the number of "stable" acoustic signals that can be used for communication. The sounds used in human language tend to be acoustically "stable." They are the result of supralaryngeal vocal tract configurations where deviations from the "ideal" shape result in signals that do not differ greatly from the acoustic signals that the ideal shape produces (Stevens, in press). Errors in articulation thus have minimal effect on the acoustic character of the signal. The vowels /a/, /i/, and /u/ are the most stable vowels. The Neanderthal supralaryngeal vocal tract cannot produce these vowels which involve a variable pharyngeal region and the associated musculature (Figs. 7, 9, and 13). The descent of the larynx to its lower position in adult Man thus would follow from the advantages this confers in communication. The adult human laryngeal position is not advantageous for either swallowing or respiration. The shift of the larynx from its position in Newborn and Neanderthal is advantageous for acquiring articulate speech but has the disadvantage of greatly increasing the chances of choking to death when a swallowed object gets lodged in the pharynx. In this respect, nonhuman primates also have an anatomical advantage (Negus, 1949). The only function for which the adult human vocal tract is better suited is speech.

In our synthesis procedure, we made maximum use of the reconstructed Neanderthal vocal tract. This perhaps yielded a wider range of sounds than Neanderthal man actually produced. It is possible, however, that Neanderthal man, who had a large brain, also made maximum use of his essentially nonhuman vocal tract to establish vocal communication. This would provide the basis for mutations that lowered the larynx and expanded the range of vocal communication in modern Man's ancestral forms.

Whether or not he did possess this mental ability may never be known. A fairly good intracranial cast was made from the La Chapelle-aux-Saints fossil (Boule and Vallois, 1957). Although Neanderthal has a cranial capacity equal to that of modern Man, this cannot be regarded as a reliable indicator of his mental ability. Cranial capacity varies greatly in modern Man and cannot be correlated with individual mental ability. There are indications that Neanderthal may not have had a sufficiently developed brain for articulate speech since his brain, although large, had relatively small frontal lobes (Fig. 14). From the developmental and phylogenetic viewpoints, it is the differences in the frontal lobes that most distinguish the human from the

SCHEMATIZED AREA FUNCTIONS FOR THE HUMAN VOWELS /a/, /u/, AND /i/

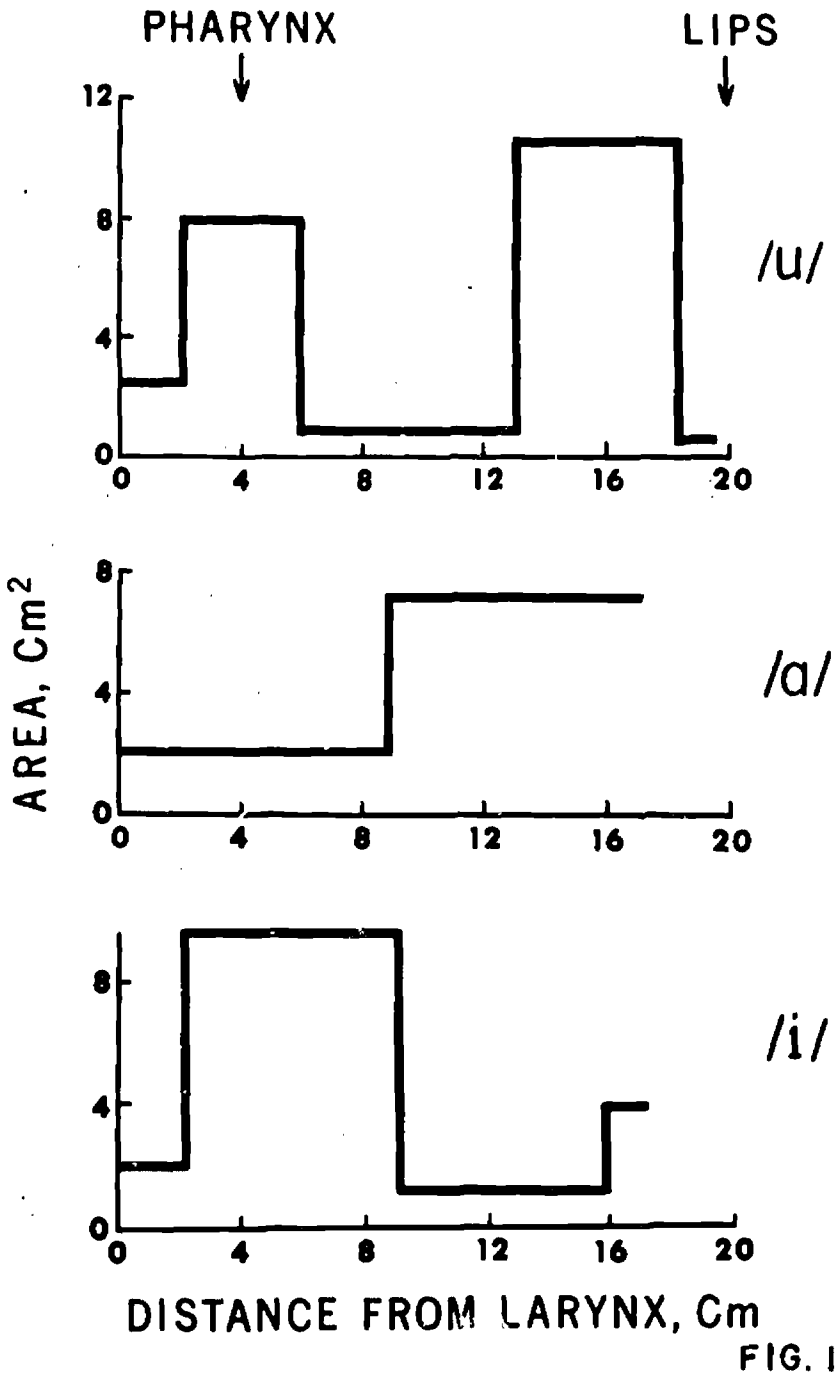
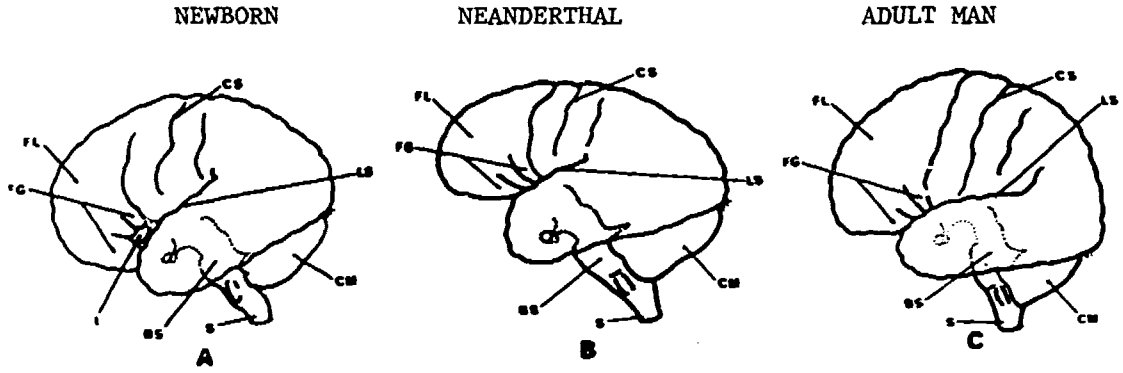


FIG. 13

Note: Area function changes abruptly, and the area of the pharyngeal region is independent of the area of the front part of the supralaryngeal vocal tract, after Lieberman et al. (1969).

LATERAL VIEW OF BRAIN



FL - Frontal Lobe
FG - Inferior Frontal Gyrus
CS - Central Sulcus
LS - Lateral Sulcus
BS - Brain Stem
CM - Cerebellum
S - Spinal Medulla (cord)
I - Insula

Note: The Neanderthal view is based on the introcranial cast of Boule and Vallois (1957).

FIG. 14

subhuman brain (Crosby et al., 1962). Although the frontal lobes of the Newborn are well developed, the brain has some grossly primitive features (Crelin, 1969).

The incline of the basilar part of the occipital bone of the Newborn skull results in a corresponding incline of the adjacent brain stem away from the vertical plane to form a marked angle where it passes vertically out of the foramen magnum to become the spinal medulla (cord). In adult Man the vertically oriented brain stem follows from the inclination of the adjacent basilar part of the occipital bone (Fig. 9). Since the base of the Neanderthal skull is so similar to that of the Newborn, we may assume that the brain stem was similarly inclined (Fig. 14). Boule and Vallois (1957) noted that on the Neanderthal intracranial cast the lateral sulcus of the brain gaped anteriorly. They interpreted this as an exposure of the insula. If this is true, it is another similarity between the Neanderthal and the Newborn brains. During brain development in modern Man, the insula gradually becomes completely covered by the enlarging inferior frontal gyrus. At birth, the insula is still exposed (Crelin, 1969; see Fig. 14). Since the insula also becomes completely covered by the inferior frontal gyrus in apes, it would be illogical to suppose that it would not do so in Neanderthal as well (Connolly, 1950). Therefore, that interpretation of the exposure of the insula in the Neanderthal brain is disputed.

Note that we are not claiming that neural developments played no role in the evolution of speech and language. We are simply stating that the anatomical mechanism for speech production is also necessary. The two factors together produce the conditions sufficient for the development of language. There is, indeed, some evidence that shows that the speech output mechanism and neural perceptual mechanisms may interact in a positive way. In recent years, a "motor" theory of speech perception has been developed (Liberman et al., 1967). This theory shows that speech is "decoded" by Man in terms of the articulatory maneuvers that are involved in its production. Signals that are quite different acoustically are identified as being the same by means of neural processing that is structured in terms of the anatomical constraints of Man's speech production apparatus. Signals that are acoustically similar may, in different contexts, be identified as being dissimilar by the same process. Animals like bullfrogs also "decode" their meaningful sounds by means of detectors that are structured in terms of the anatomical constraints of their sound-producing systems (Capranica, 1965). These neural processes are species-specific, and they obviously can evolve only as, or after, the

species develops the ability to produce specific sounds. The brain and the anatomical structures associated with signaling thus evolve together. Enhanced signaling, i.e., phonetic ability, correlates with general linguistic ability in the living primates, where modern man and the nonhuman primates are the extremes (Lieberman, 1968; Lieberman et al., 1969).

The articulatory maneuvers that underlie human speech constrain the entire neural embodiment of the grammar of language. The range of sounds and phonetic contrasts of speech form "natural" dimensions that structure the phonologic, syntactic, and lexical properties of all human languages. (Jakobson et al., 1963; Postal, 1968; Lieberman, 1970). The hypothetical language that Neanderthal man could have employed would have been more "primitive" in a meaningful sense than any human language. Fewer phonetic contrasts would have been available for the linguistic code.

Fully developed "articulate" human speech and language appear to have been comparatively recent developments in Man's evolution. They may be the primary factors in the accelerated pace of cultural change. Our conclusions regarding Neanderthal man's linguistic ability, which are based on anatomical and acoustic factors, are consistent with the inferences that have been drawn from the rapid development of culture in the last 30,000 years in contrast to the slow rate of change before that period (Dart, 1959).

Conclusion

Neanderthal man did not have the anatomical prerequisites for producing the full range of human speech.[4] He probably also lacked some of the neural detectors that are involved in the perception of human speech. He was not as well equipped for language as modern man. His phonetic ability was, however, more advanced than those of present day nonhuman primates, and his brain may have been sufficiently well developed for him to have established a language based on the speech signals at his command. The general level of Neanderthal

[4] Debetz (1961), in connection with attempts to explain directly the causes for the appearance of certain characteristics belonging to Homo sapiens, notes that, "...the pecularities of the skull, whose importance in the evolution of man is not in any case less important than the peculiarities in the structure of the hand and of the entire body, remain inexplicable." We have shown that some of the differences between the skull structure of classic Neanderthal man and Homo sapiens are relevant to the production of the full range of human speech. Earlier unsuccessful attempts at deducing the presence of speech from skeletal structures, which are discussed by Vallois (1961), were hampered by the absence of both a quantitative acoustic theory of speech production and suitable anatomical comparisons with living primates that lack the physical basis for articulate human speech.

culture was such that this limited phonetic ability was probably utilized and that some form of language existed. Neanderthal man thus represents an intermediate stage in the evolution of language. This indicates that the evolution of language was gradual, that it was not an abrupt phenomenon. The reason that human linguistic ability appears to be so distinct and unique is that the intermediate stages in its evolution are represented by extinct species.

Neanderthal culture developed at a slow rate. We may speculate on the disappearance of Neanderthan man, and we can note that his successors, for example, Cro Magnon man, who inhabited some of the old Neanderthal sites in the Dordogne (Boule and Vallois, 1957), had the skeletal structure that is typical of Man's speech mechanism. Neanderthal man's disappearance may have been a consequence of his linguistic--hence intellectual--deficiencies with respect to his spaiens competitors. In short, we can conclude that Man is human because he can say so.

References

Boule, M. (1911-13) L'Homme fossile de La Chapelle-aux-Saints. Annales de Paleontologie 6, 109; 7, 21, 85; 8, 1.

Boule, M., and H.V. Vallois. (1957) Fossil Men (Dryden Press, New York).

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

Chiba, T., and M. Kajiyama. (1958) The Vowel, Its Nature and Structure (Phonetic Society of Japan, Tokyo).

Chomsky, N. (1966) Cartesian Linguistics (Harper and Row, New York).

Connolly, C.J. (1950) External Morphology of the Primate Brain (C.C. Thomas, Springfield, Ill.).

Coon, C.S. (1966) The Origin of Races (Knopf, New York).

Crelin, E.S. (1969) Anatomy of the Newborn: An Atlas (Lea and Febiger, Philadelphia).

Crosby, E.C., T. Humphrey, and E.W. Laver. (1962) Correlative Anatomy of the Nervous System (Macmillan Co., New York).

Dart, R.A. (1959) On the evolution of language and articulate speech. HOMO 10, 154-165.

Debetz, G.F. (1961) Soviet anthropological theory. In Social Life of Early Man, S.L. Washburn, Ed. (Aldine, Chicago).

DuBrul, E.L. (1958) Evolution of the Speech Apparatus (C.C. Thomas, Springfield, Ill.).

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

Henke, W.L. (1966) Dynamic Articulatory Model of Spe oduction Using Computer Simulation. Thesis, Mass. Instit. of . Cambridge, Appendix B.

Howells, W.W. (1968) Mount Carmel Man: morphologica tionships. In Proceedings, VIIIth Int'l Cong. Anthro. and Ethn nces, Vol. I, Anthropology (Tokyo).

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

Keith, A. (1925) The Antiquity of Man (Williams an gate, London).

Leakey, L.S.B., and V.M. Goodall. (1969) Unveiling igins (Schenkman, Cambridge, Mass.).

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

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

Lieberman, P. (1970) Towards a unified phonetic theory. Linguistic Inquiry 1, 307-322.

Lieberman, P., K.S. Harris, P. Wolff, and L.H. Russell. (1968) Newborn infant cry and nonhuman primate vocalizations. Status Report 17/18 (Haskins Laboratories, New York City). Scheduled J. Speech and Hearing Res.

Lieberman, P., D.H. Klatt, and W.W. Wilson. (1969) Vocal tract limitations on the vowel repertories of rhesus monkey and other nonhuman primates. Science 164, 1185-1187.

Negus, V.E. (1949) The Comparative Anatomy and Physiology of the Larynx (Hafner, New York).

Perkell, J.S. (1969) Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study (MIT Press, Cambridge, Mass.)

Peterson, G.E. and H.L. Barney. (1952) Control methods used in a study of the vowels. J. Acoust. Soc. Am. 24, 175-184.

Postal, P.M. (1968) Aspects of Phonological Theory (Harper and Row, New York).

Stevens, K.N. (forthcoming) Quantal nature of speech. In Human Communication: A Unified View, E.E. David and P.B. Denes, Eds., (McGraw Hill, New York).

Straus, W.L., Jr., and A.J.E. Cave. (1957) Pathology and posture of Neanderthal Man. Quart. Rev. Biol. 32, 348-363.

Truby, H.M., J.F. Bosma, and J. Lind. (1965) Newborn Infant Cry (Almqvist and Wiksells, Uppsala, Sweden).

Vallois, H.V. (1961) The evidence of skeletons. In Social Life of Early Man, S.L. Washburn, Ed. (Aldine, Chicago).

Whitfield, I.C. (1969) Response of the auditory nervous system to simple time-dependent acoustic stimuli. Annals of N.Y. Acad. Sci. 156, 671-677.

SLIDE I

Glottal Adjustments for English Obstruents*

Masayuki Sawashima+
Haskins Laboratories, New Haven

Observation of the larynx for articulation of English consonants in running speech were made by using a coherent fiberoptics bundle. The procedures were as follows:

The fiberoptics bundle was inserted through the nose and positioned in the hypopharynx so as to obtain a good view of the glottis. A 16mm cinecamera was attached to the external end of the optics. The cinecamera was driven by a synchronous motor at sixty frames per second. Simultaneously with the filming, speech signals were recorded on tape together with synchronization time marks.

A list of sentences consisting of from three to fifteen syllables each and containing voiced and voiceless consonants were read aloud by three native American English talkers. Slide 1 shows selected frames of the motion picture for the sentence "Rub Billy's head with this towel." Each frame is correlated with the proper point in the sound spectrogram. A narrow-band trace is displayed above the wide-band pattern to show the voicing during speech. The symbols at the bottom are those of a broad phonetic transcription of the utterance.

In the leftmost frame, we see the larynx in inspiratory position with wide-open glottis before the utterance. The next frame shows the situation immediately before voice onset. The larynx is in phonatory position and the arytenoids are closed, while a narrow spindle-shaped opening is seen along the membranous portion of the glottis. The next frame shows almost the same position of the larynx, in which the blurred edges of the vocal folds indicate vibratory motion. The next frame is for the [b]-closure of "Rub Billy's." Here also, the larynx is in phonatory position with vibrating vocal folds. Its appearance is almost the same as in the next frame for the following vowel.

---

*Paper presented at the meeting of New York Speech and Hearing Association, May 1970.

+On leave from the University of Tokyo.

The sixth frame is for the transition from [z] to [h] of "Billy's head." The glottis is open with separated arytenoids. A sharp definition of the vocal fold edges indicates the cessation of vibration. The last frame shows the glottis just before the release of [t] of "towel." The opening of the glottis is as large as, or a little larger than, during the transition from [z] to [h].

On the films taken for the three subjects, we made a frame-by-frame analysis of the laryngeal state during the articulation of various consonants. In the frame analysis, the following features were examined:

1) opening and closing timing for the arytenoid cartilages
2) interruption and resumption of vocal fold vibration
3) maximum width of glottal aperture
4) width of glottal aperture at the time of oral release of the stop closure.

The corresponding spectrograms were used to fix the times of supraglottal articulatory gestures, as well as those of interruption and resumption of glottal pulses.

Our data revealed that, in voiceless aspirated stops and voiceless fricatives, there was a wide opening of the glottis with separation of the arytenoids, as well as interruption of glottal vibration. On the other hand, findings for the voiceless unaspirated stops and voiced consonants were somewhat complicated.

In Slide 2, the voiced and voiceless unaspirated stops, /b,g/ and /p,k/, are classified in two ways, depending on whether or not the vocal folds ceased to vibrate and whether or not the arytenoids were separated. In the lower right quadrant are the pooled data for three subjects.

In general, the sets /b,g/ and /p,k/ can be described as follows: most /b,g/ tokens show no arytenoid separation and no interruption of glottal vibration. Most cases of /p,k/ show both separation of arytenoids and interruption of vibration. At the same time, we should note that a few cases showed separation of the arytenoids and that some had an interruption of glottal vibration. There are, moreover, a large number (fifteen cases) of /p,k/ tokens in which no separation of the arytenoids was observable, while a few showed no interruption of glottal vibration.

Looking at the behavior of individual subjects, we can recognize certain differences between them, although the number of observations is perhaps too small to draw firm conclusions. For example, subject C has a considerable number of [p]'s without separation of the arytenoids, while subject A has all the [b]'s with separation of the arytenoids and a fair number of the [p]'s without arytenoid separation.

In distinguishing between voiced and voiceless categories, subjects C and L have no difficulty. In the case of subject A, there seems to be some overlap

Arytenoid Separation and Interruption of Glottal Vibration for English Voiced and Voiceless Unaspirated Stops

subject C

		b	p	g	k
Aryt. Sep.	+	0	11	0	15
	−	9	6	3	2
I.G.V.	+	0	16	0	17
	−	9	1	3	0

subject A

		b	p	g	k
Aryt. Sep.	+	4	8	0	13
	−	10	4	3	1
I.G.V.	+	5	9	0	14
	−	9	3	3	0

subject L

		b	p	g	k
Aryt. Sep.	+	0	9	0	15
	−	8	1	5	1
I.G.V.	+	0	10	0	15
	−	8	0	5	1

Pooled Data

		b	p	g	k
A.S.	+	4	28	0	43
	−	27	11	11	4
I.G.V.	+	5	35	0	46
	−	26	4	11	1

SLIDE 2

Arytenoid Separation and Interruption of Glottal Vibration for English Voiced Fricatives and Affricates

subject C

		Z	ð	V	j
Aryt. Sep.	+	8	0	0	2
	–	3	10	6	1
I.G.V.	+	0	0	0	0
	–	11	10	6	3

subject A

		Z	ð	V	j
Aryt. Sep.	+	10	--	0	4
	–	1	--	5	0
I.G.V.	+	3	--	0	3
	–	8	--	5	1

subject L

		Z	ð	V	j
Aryt. Sep.	+	1	0	0	0
	–	8	6	3	2
I.G.V.	+	0	0	0	0
	–	9	6	3	2

Pooled Data

		Z	ð	V	j
A.S.	+	19	0	0	6
	–	12	16	14	3
I.G.V.	+	3	0	0	3
	–	28	16	14	6

SLIDE 3

between /p/ and /b/, but the overlap disappears if the items are separated according to context.

Slide 3 shows a similar display for voiced fricatives and affricates. In the pooled data, we see that all the /ð/ and /v/ tokens were produced with closed arytenoids and continuation of glottal vibration, while the situation is complicated in /z/ and /ʒ/. Looking at data for individual subjects, we again see certain differences among them. In subject L, almost all the /z/ and /ʒ/ tokens show neither separation of the arytenoids nor interruption of glottal vibration. For subjects C and A, most of the /z/ and /ʒ/ tokens were produced with arytenoid separation. Furthermore, subject A has all the tokens which showed interruption of vibration.

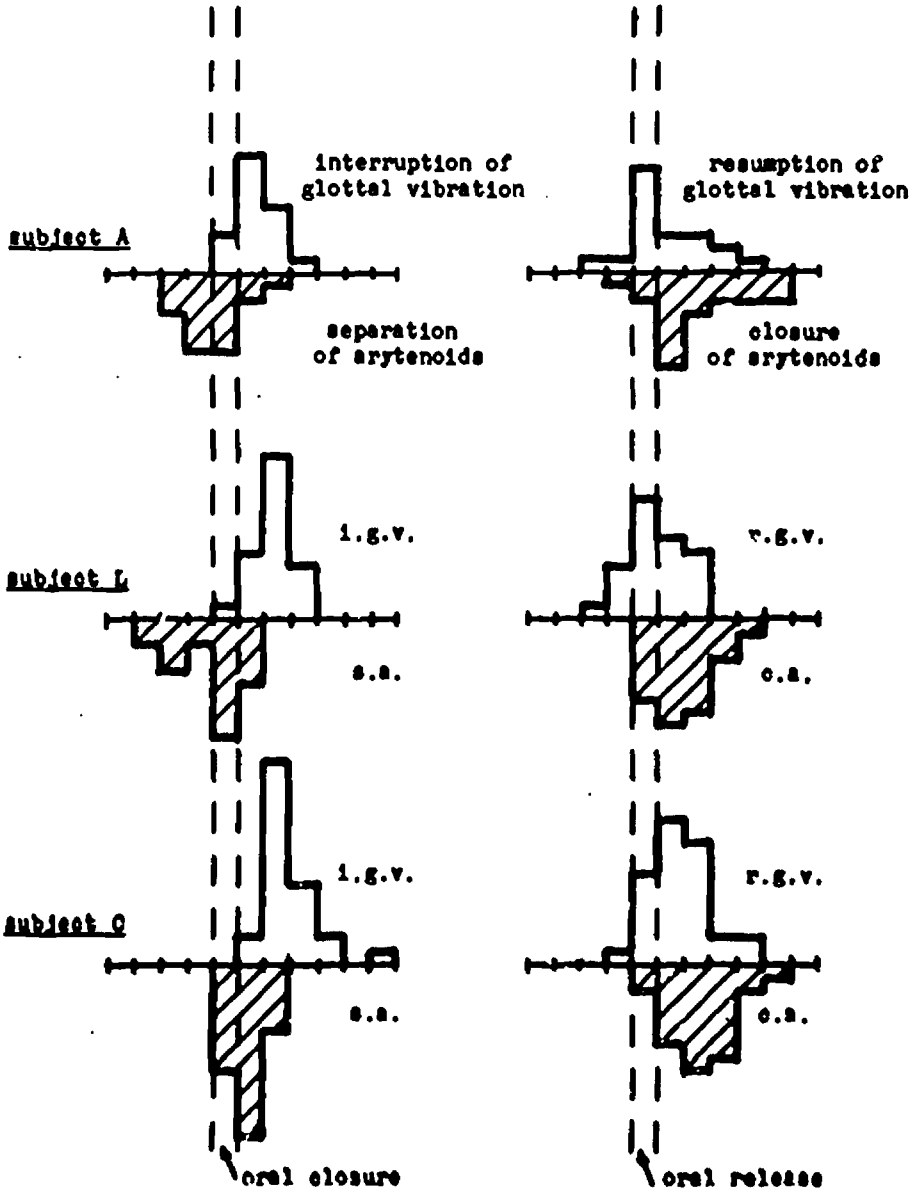
Now let us focus attention on the time relations between laryngeal and supraglottal articulatory gestures for voiceless consonants.

Slide 4 shows such time relations for the voiceless unaspirated stops. In the left column, three graphs indicate when interruption of glottal vibration and separation of the arytenoids occurred relative to the stop occlusion. The abscissa is marked off in time intervals representing the film frames in sequence. The ordinate indicates the frequency of occurrence along the abscissa. Blank graphs above the abscissae are distribution patterns for the interruption of glottal vibration, and shaded graphs below the abscissae are for the separation of the arytenoids. In the right column, a similar display is shown for the timing of resumption of glottal vibration and the closure of the arytenoids, relative to the stop release.

In the left graphs, we see that, in most cases, interruption of glottal vibration occurs one or two frames after the beginning of the closure. Separation of the arytenoids shows a relative timing that varies considerably, occurring both before and after oral closure, with some intersubject difference. There is a clear tendency for arytenoid separation to begin earlier than interruption of vibration, although there is some overlap in distribution patterns. Examination of the time relation for each token showed that, in sixty-three tokens out of sixty-six, interruption of vibration took place after arytenoid separation and that there was only one case in which the time relation was reversed.

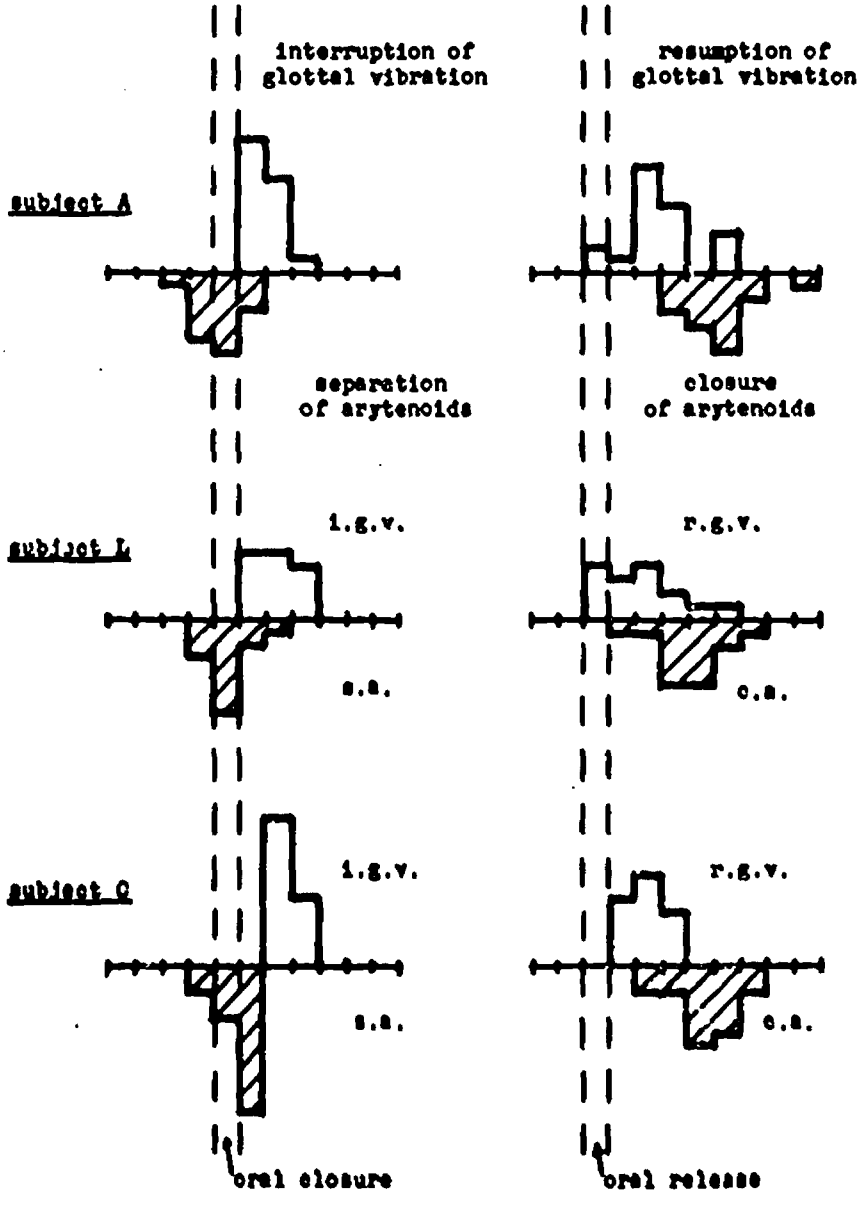
In the graphs on the right, we see that the resumption of vibration takes place, in most cases, just at or immediately after stop release, while arytenoid closure is achieved following the release. There seems to be a tendency for arytenoid closure to be completed shortly after resumption of glottal vibration.

Voiceless Unaspirated Stops



SLIDE 4

Voiceless Aspirated Stops



SLIDE 5

Examination of each token revealed that, in most cases, resumption of glottal vibration preceded the arytenoid closure, although there was a considerable number of tokens, particularly for subjects L and C, in which the two occurred at the same time.

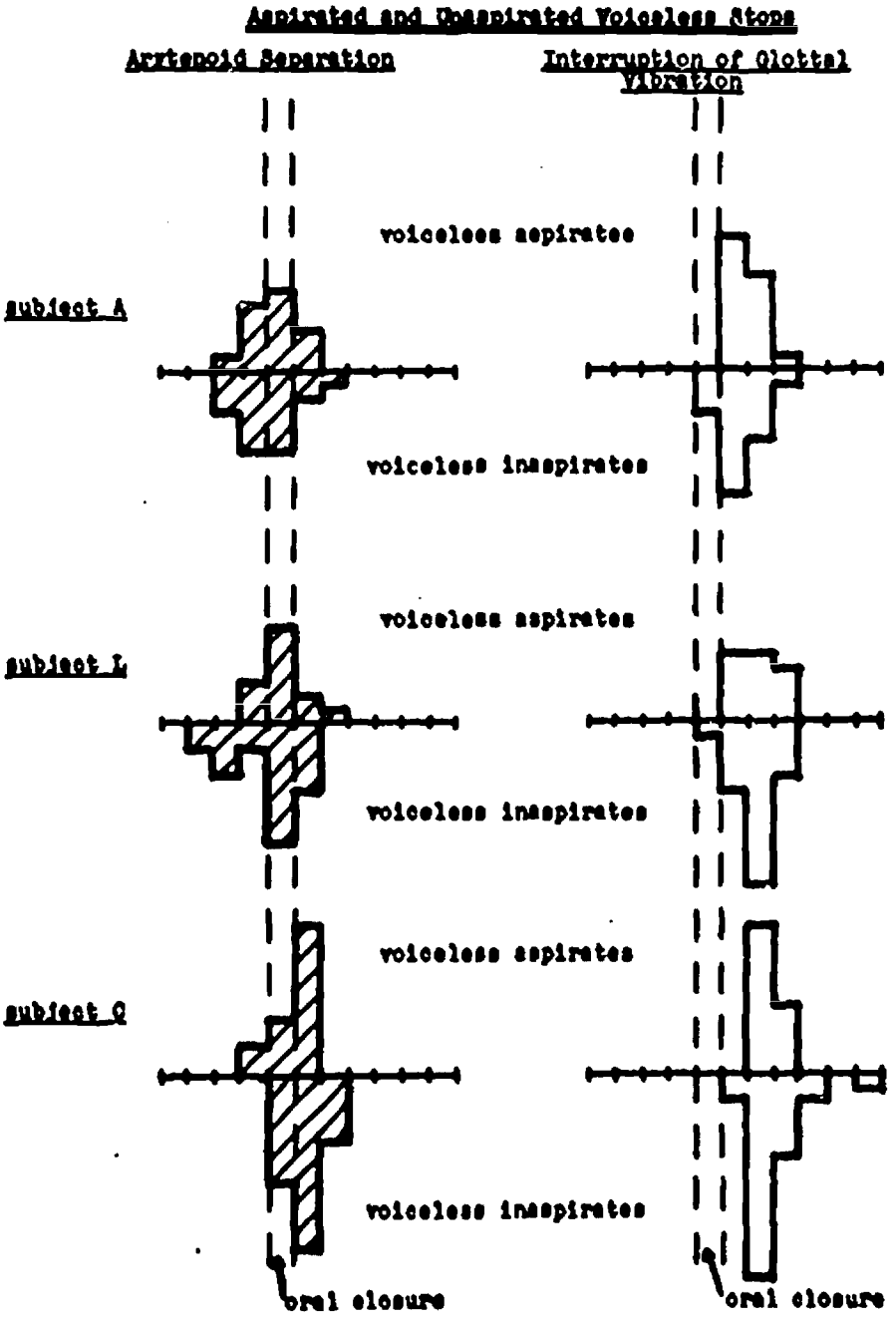
Slide 5 shows the same display for voiceless aspirated stops. The timing of the interruption of vibration and arytenoid separation relative to the stop closure, shown on the left, is quite similar to the situation for the voiceless inaspirates. Here also, separation of the arytenoids regularly precedes interruption of vibration for every token. On the right, we see that the arytenoid closure is achieved long after the stop release. Examination of the relative timing between arytenoid closure and resumption of vibration revealed that, in almost all tokens, the arytenoids were closed after resumption of vibration.

Slide 6 shows the timing of the laryngeal gesture at the beginning of voiceless aspirated stops in comparison with that of voiceless inaspirates. Graphs below the abscissae are those for the inaspirates. Shaded graphs on the left are for arytenoid separation, and blank ones on the right are for interruption of glottal vibration. Distribution patterns for the aspirates are well matched to those of the inaspirates. The display indicates that there is no difference in timing of laryngeal gestures between aspirates and inaspirates at the beginning of stop closure.

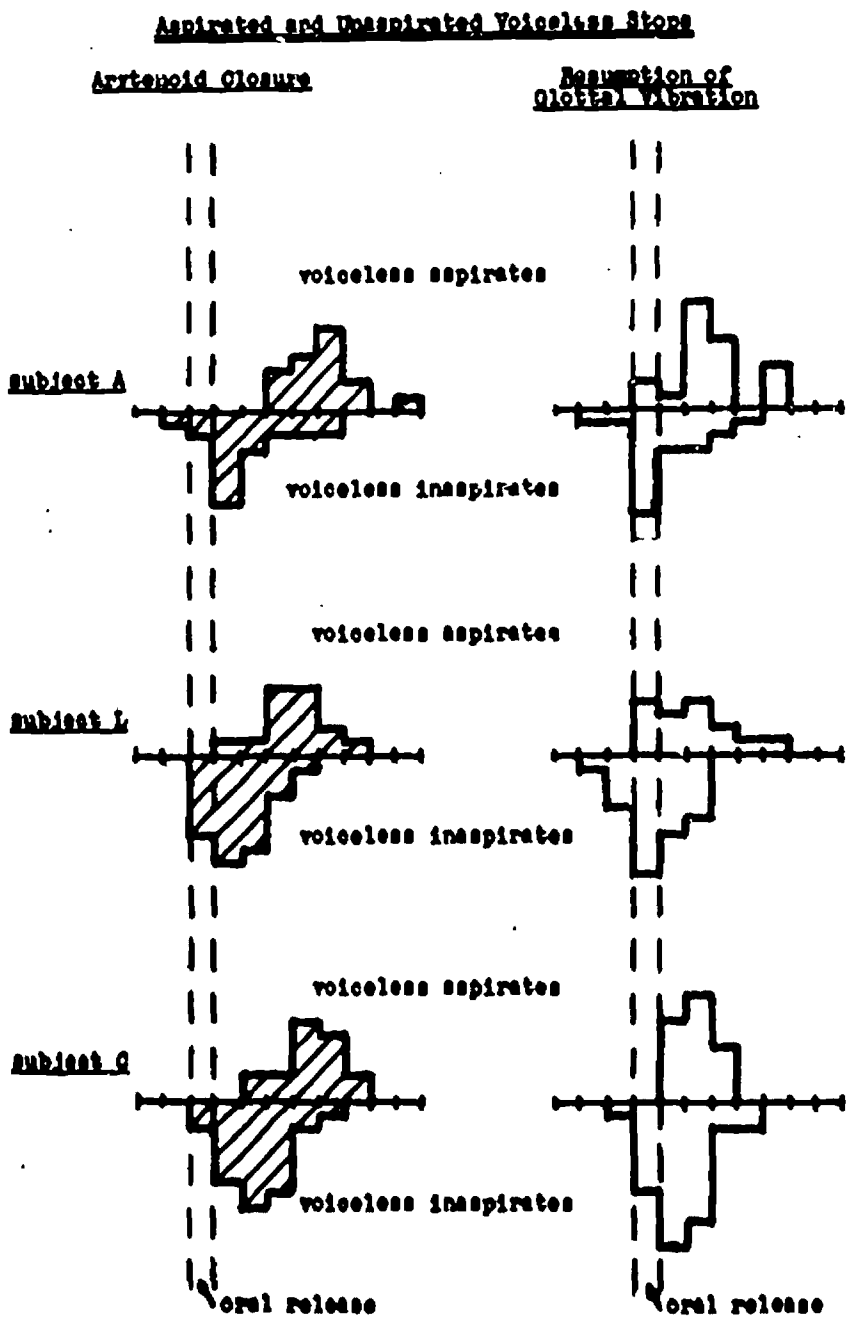
On the other hand, the difference in the timing of the laryngeal gesture for release of stop closure is clearly seen in Slide 7. On the left, we see that the arytenoid closure is achieved later in the aspirates than in the inaspirates. A similar tendency is observable for resumption of vibration, as shown on the right half of the slide, although there is more overlapping in the distribution patterns.

Slide 8 shows the time relation between laryngeal and upper articulatory gestures for voiceless fricatives. The graphs show patterns similar to those for voiceless unaspirated stops. For every token, the arytenoids begin to separate before interruption of glottal vibration.

For estimating width of glottal opening, we measured the distance between the vocal fold edges on magnified traces of the films. Slide 9 shows the maximum opening of the glottis during the articulation of voiceless consonants. We classified the width of the opening in 5mm steps as indicated on the abscissa. It should be noted that the values on the abscissa do not indicate the absolute values of the actual glottal opening. The ordinate indicates numbers of cases along the abscissa. The glottal aperture is smaller in inaspirates than in

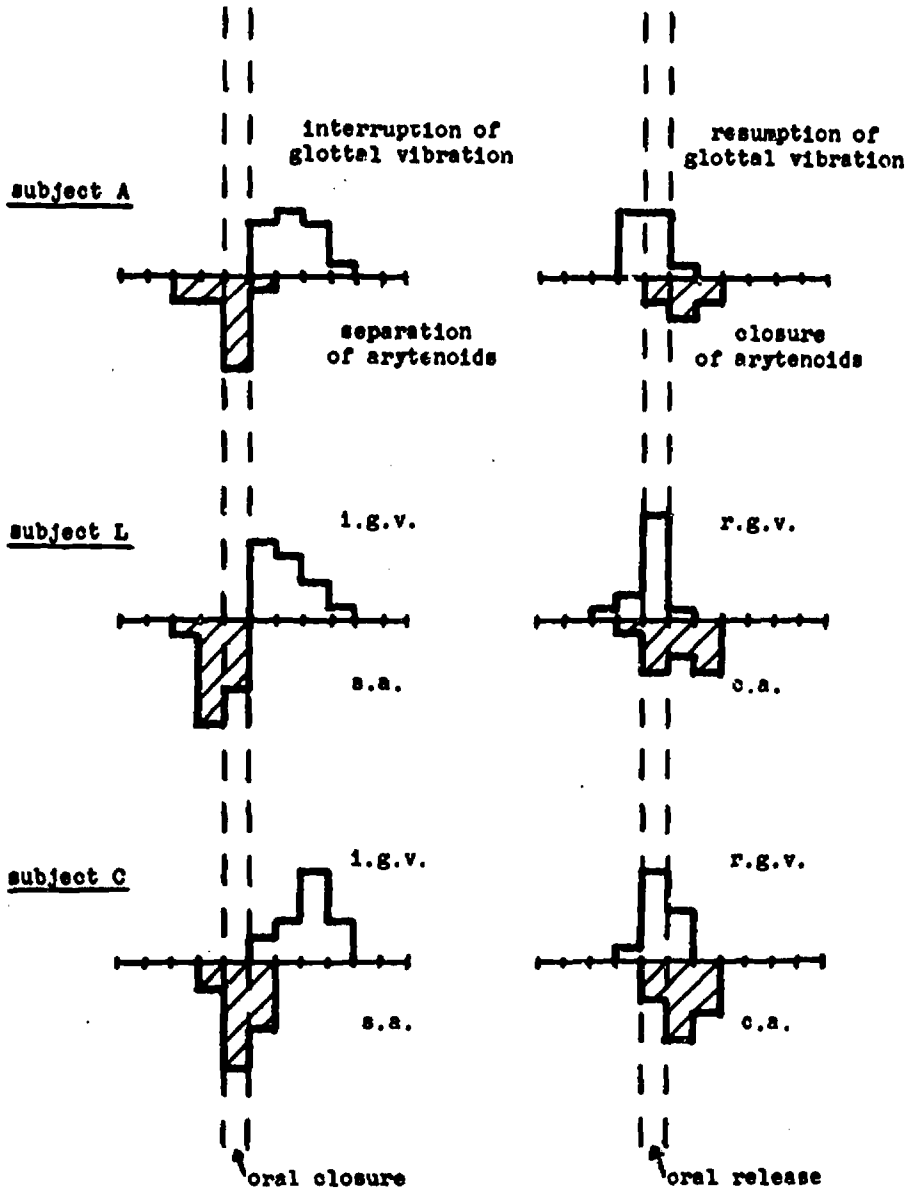

Aspirated and Unaspirated Voiceless Stops
Arytenoid Separation
Interruption of Glottal Vibration
voiceless aspirates
subject A
voiceless inaspirates
voiceless aspirates
subject L
voiceless inaspirates
voiceless aspirates
subject Q
voiceless inaspirates
oral closure
oral closure

SLIDE 6


Aspirated and Unaspirated Voiceless Stops
Arytenoid Closure
Resumption of Glottal Vibration
voiceless aspirates
Subject A
voiceless inaspirates
voiceless aspirates
Subject L
voiceless inaspirates
voiceless aspirates
Subject C
voiceless inaspirates
oral release
oral release

SLIDE 7

Voiceless Fricatives



SLIDE 8

Maximum Glottal Aperture for Voiceless Consonants
(in millimeters)

SLIDE 9

voiceless inaspirates
voiceless aspirates
subject A
subject L
subject C
5 10 15 20 25 30 35
Glottal Aperture at Stop Release
(in millimeters)

SLIDE 10

aspirates and fricatives. There seems to be no difference between the aspirates and fricatives.

Slide 10 shows the glottal opening at stop release. The opening in voiceless unaspirated stops is definitely smaller than that in voiceless aspirated stops. The data are consistent with those for the difference in timing of laryngeal gestures at stop release.

Findings presented here concern some basic features of laryngeal gestures, mainly for intervocalic consonants. In further studies, we plan to examine variations in these basic features for various consonant clusters and to extend these studies to include cross-language observations.

Cinegraphic Observations of the Larynx During Voiced and Voiceless Stops*

Leigh Lisker,[+] Masayuki Sawashima,[++] Arthur S. Abramson,[+++] and Franklin S. Cooper
Haskins Laboratories, New Haven

At the last meeting of the Society (J. Acoust. Soc. Am. 47, 105 (A), 1970), we reported certain observations of laryngeal activity associated with the production of English stop and fricative consonants in running speech. The method involved introducing a coherent fiberoptics bundle into the pharynx via the nose and coupling its external end to a cinecamera set to operate at 60 frames per second. From data on a single talker, it appeared that certain classes of sounds may be distinguished by whether or not the arytenoid cartilages move apart during their production. Thus, the voiceless fricatives /s,ʃ,f/ regularly show separation of the arytenoids, while the voiced stops do not. But some consonant classes show a degree of variability in this respect, in particular those variants of the voiceless stops described as unaspirated, which are found before unstressed vowels. Tokens of the set /b,g/ are sometimes produced without voicing during buccal closure, and of these, some are produced with separation of the arytenoids. Because these two consonant classes, the set /b,g/ with frequent lack of voicing during articulatory closure and the unaspirated set /p,k/, seemed to offer the most difficulty to the view that English stops can be neatly partitioned on the basis of whether or not the arytenoids execute an opening gesture, we chose to pay them special attention.

The present findings are derived from recordings of three native Americans, who read a list of sentences consisting of from three to fifteen syllables each. The sentences were designed to include a good selection of stops and fricatives in a variety of contexts. In conjunction with the filming, use was made of both a conventional and a throat microphone. Timing pulses enabled us to synchronize the photographic and acoustic recordings. An illustration is pro-

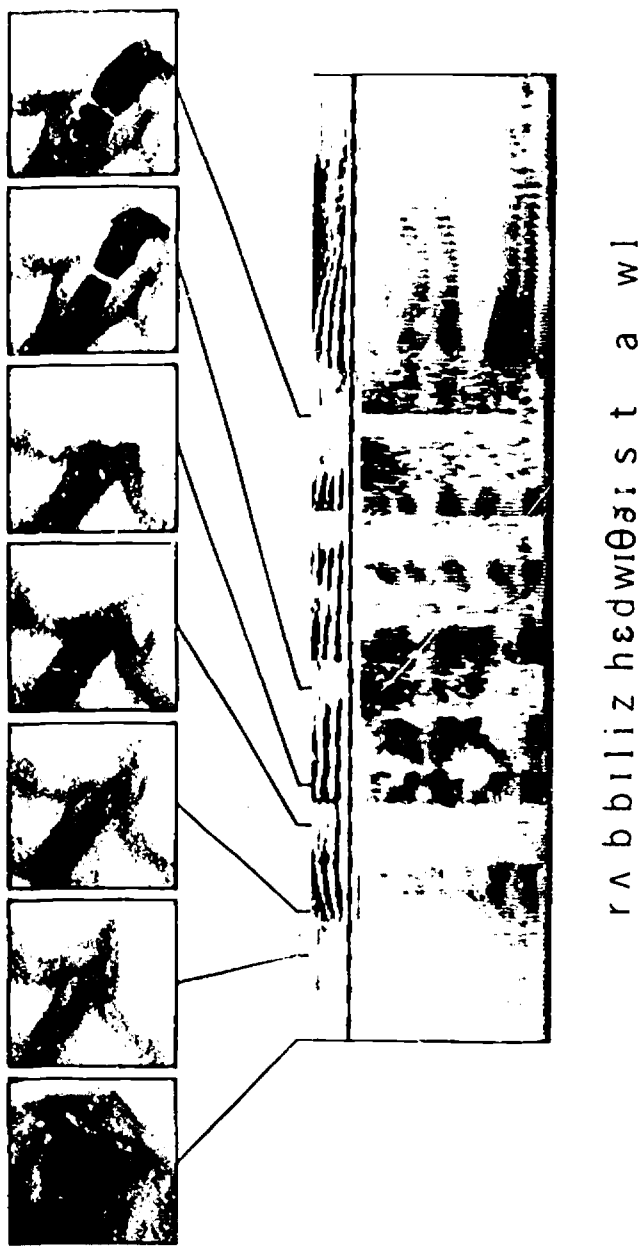
---

*Contributed paper given at the 79th Meeting of the Acoustical Society of America, Atlantic City, N.J., 21-24 April 1970.

+Also, University of Pennsylvania, Philadelphia.

++On leave from the University of Tokyo.

+++Also, University of Connecticut, Storrs.



SLIDE I

vided by Slide 1, in which selected frames are matched with locations in the spectrogram of the utterance Rub Billy's head with this towel. The seven frames are a sample of the range of glottal states observed in our study. Appropriate frame-sequences for the stop consonants in the utterances recorded were examined for the following features:

1. opening and closing movements of the arytenoid cartilages
2. interruption and resumption of vocal-fold vibration
3. maximum width of glottal aperture
4. width of the glottal aperture at the time of oral release of the stop.

The corresponding spectrograms were used primarily to fix the times of the stop-closure and release.

In Slide 2, the voiced and voiceless unaspirated stops are classified in two ways, depending on whether or not the arytenoids were seen to separate and on whether or not the vocal folds ceased to vibrate. In the lower right quadrant are the pooled data for the three subjects.

In general, the sets /b,g/ and /p,k/ can be described as follows: most /b,g/ tokens show no arytenoid separation and no interruption of glottal vibration, while most instances of /p,k/ have both separation of the cartilages and interruption of glottal vibration. At the same time, we should note that a few cases of [b] showed separation of the arytenoids and that some had an interruption of vibration. There are, in addition, some fifteen cases of /p,k/ tokens in which no separation of the arytenoids was detected, while a few, moreover, showed no interruption of glottal vibration.

Certain differences were observed among individual subjects, but the number of observations is perhaps too small for us to draw very firm conclusions. Subject C, for example, contributed most of the [p]'s without arytenoid separation, while subject A contributed all the [b]'s with arytenoid separation, all the [b]'s with interruption of vibration, and a fair number of the [p]'s without arytenoid separation. In distinguishing between voiced and voiceless categories, subjects C and L offer no difficulty. In the case of subject A, there seems to be some overlap between /b/ and /p/, but even there, this largely disappears if items are separated according to context.

Turning away from the question of distinguishing the two linguistic categories, we can learn something of the time relations between laryngeal and supraglottal articulatory gestures from our data. Slide 3 shows such time relations for the voiceless unaspirated stops. The three plots on the left indicate when interruption of glottal vibration occurred relative to the stop

Arytenoid Separation and Interruption of Glottal Vibration for English Voiced and Voiceless Unaspirated Stops

subject C

		b	p	g	k
Aryt. Sep.	+	0	11	0	15
	−	9	6	3	2
I.G.V.	+	0	16	0	17
	−	9	1	3	0

subject A

		b	p	g	k
Aryt. Sep.	+	4	8	0	13
	−	10	4	3	1
I.G.V.	+	5	9	0	14
	−	9	3	3	0

subject L

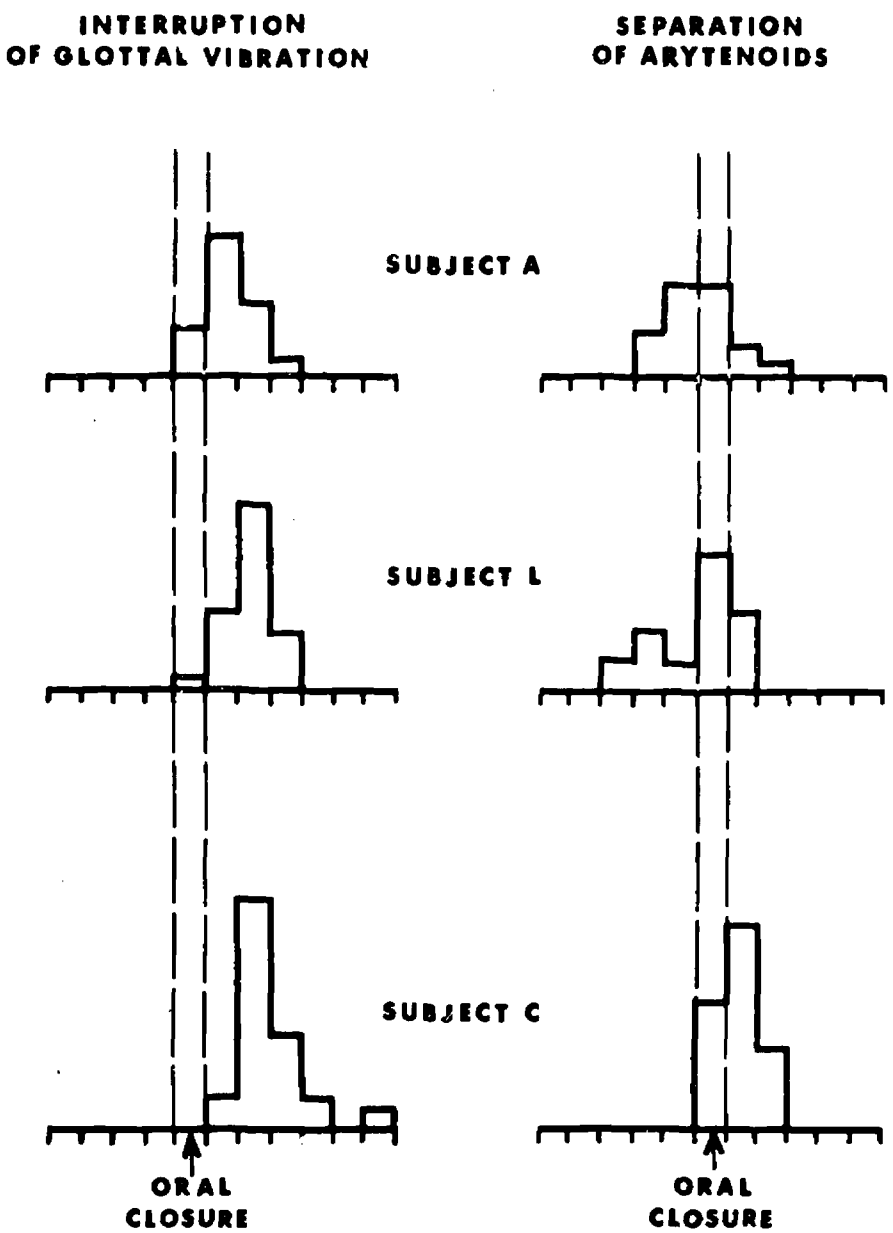
		b	p	g	k
Aryt. Sep.	+	0	9	0	15
	−	8	1	5	1
I.G.V.	+	0	10	0	15
	−	8	0	5	1

Pooled Data

		b	p	g	k
A.S.	+	4	28	0	43
	−	27	11	11	4
I.G.V.	+	5	35	0	46
	−	26	4	11	1

SLIDE 2

VOICELESS UNASPIRATED STOPS



SLIDE 3

occlusion. The abscissa is marked off in intervals representing the film frames in sequence. The ordinate indicates the distribution of values along the abscissa. On the right, the timing relation between the beginning of arytenoid separation and stop occlusion is shown in the same way. We see here that the interruption of glottal vibration usually occurs one or two frames after the beginning of the occlusion. The separation of the arytenoids shows a relative timing that varies considerably, occurring both before and after oral closure with some intersubject differences. Although not apparent from this display, separation of the arytenoids never begins after the interruption of vibration.

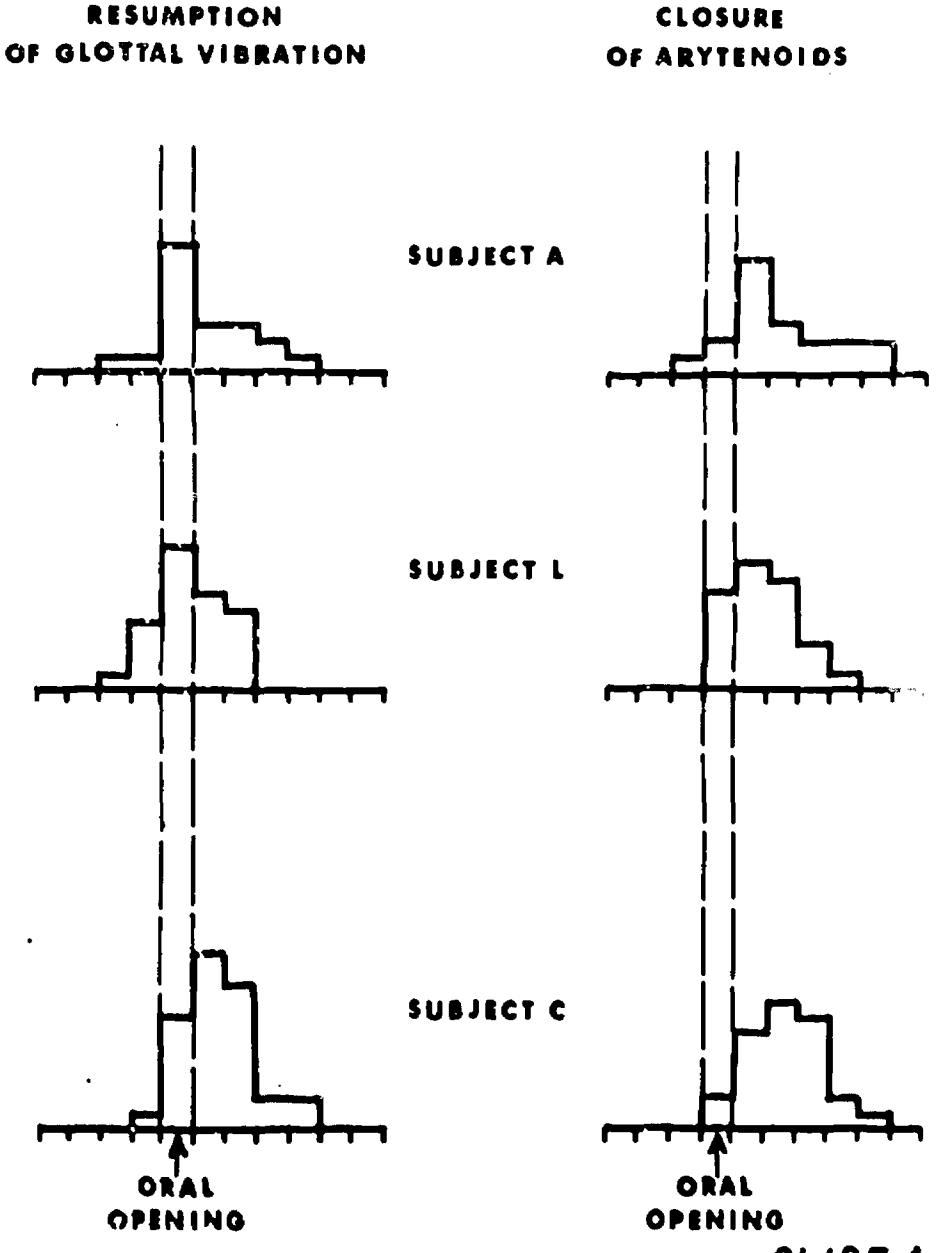
Slide 4 presents similar displays for the resumption of glottal vibration and the return of the arytenoids to closed position. In most cases, our films show resumption of vibration just at or immediately following stop release, while a closed state of the arytenoids is achieved, in most cases, just after release. There seems to be a tendency for arytenoid closure to be completed shortly after resumption of glottal vibration.

Because there were in our sample only five tokens of /b/ for which an interruption of glottal vibration was observed and four for which the arytenoids separated, we cannot say much about timing differences between voiced and voiceless unaspirated categories. The five /b/'s with interruption of vibration showed persistence of vibration for several frames into the interval of stop occlusion. Moreover, since for those stops vibration resumed directly upon release, the interval over which the vocal folds appeared to be still was very brief, usually a single frame. For the four /b/'s with arytenoid separation, this took place just at the beginning of oral occlusion, and the arytenoids were back together by the end of the occlusion.

The timing relations observed for the unaspirated stops may be compared with those for the voiceless aspirates, which are shown in Slide 5. The movement of the vocal folds is brought to a halt only after oral closure has been established, particularly in the case of subject C, who showed a similar tendency in his productions of voiceless inaspirates. Arytenoid separation occurs in close synchrony with oral closure. Here too, subject C lags behind. Slide 5 does not show that the magnitude of separation is decidedly greater for these stops than for both classes of inaspirates.

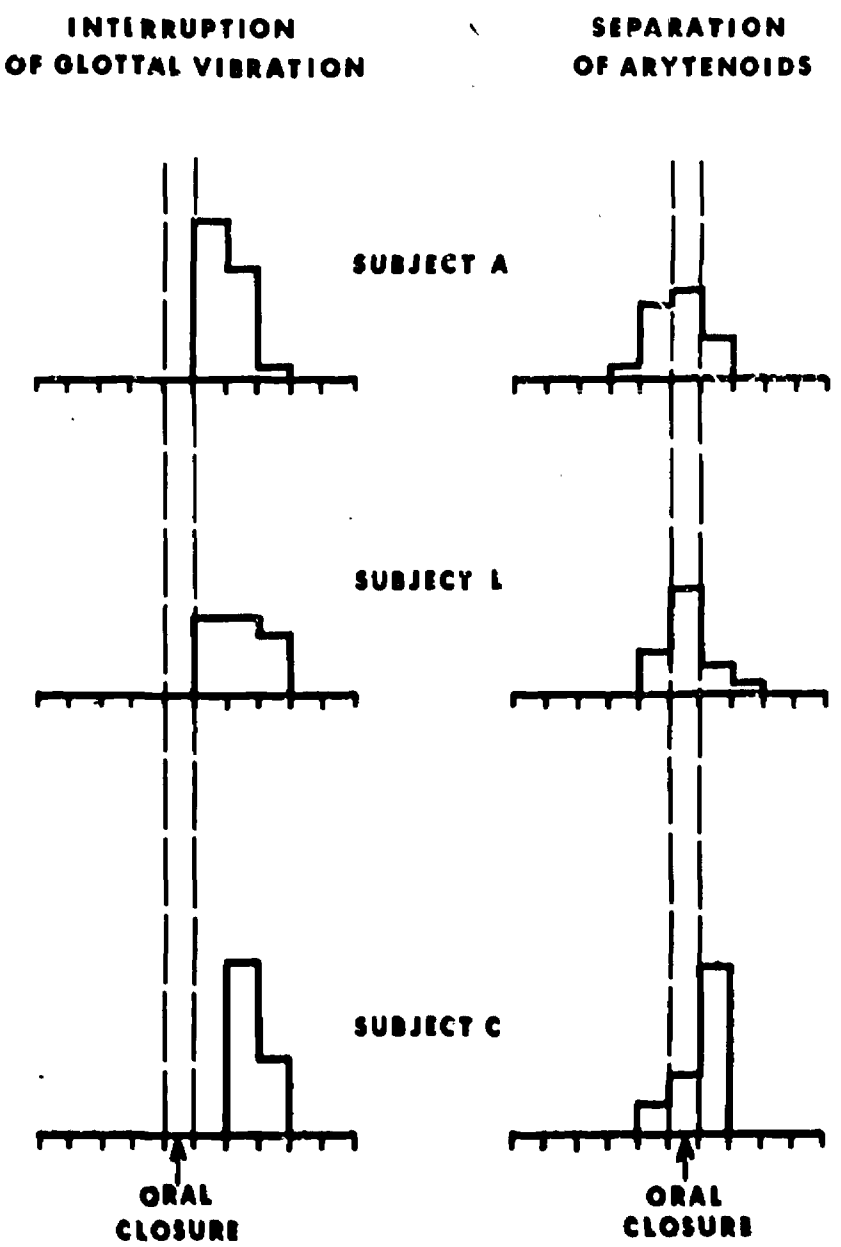
Slide 6 represents timing relations at the termination of occlusion for the voiceless aspirates. The resumption of vibration is somewhat later here than for the voiceless inaspirates, as we might expect. At the same time, we

VOICELESS UNASPIRATED STOPS



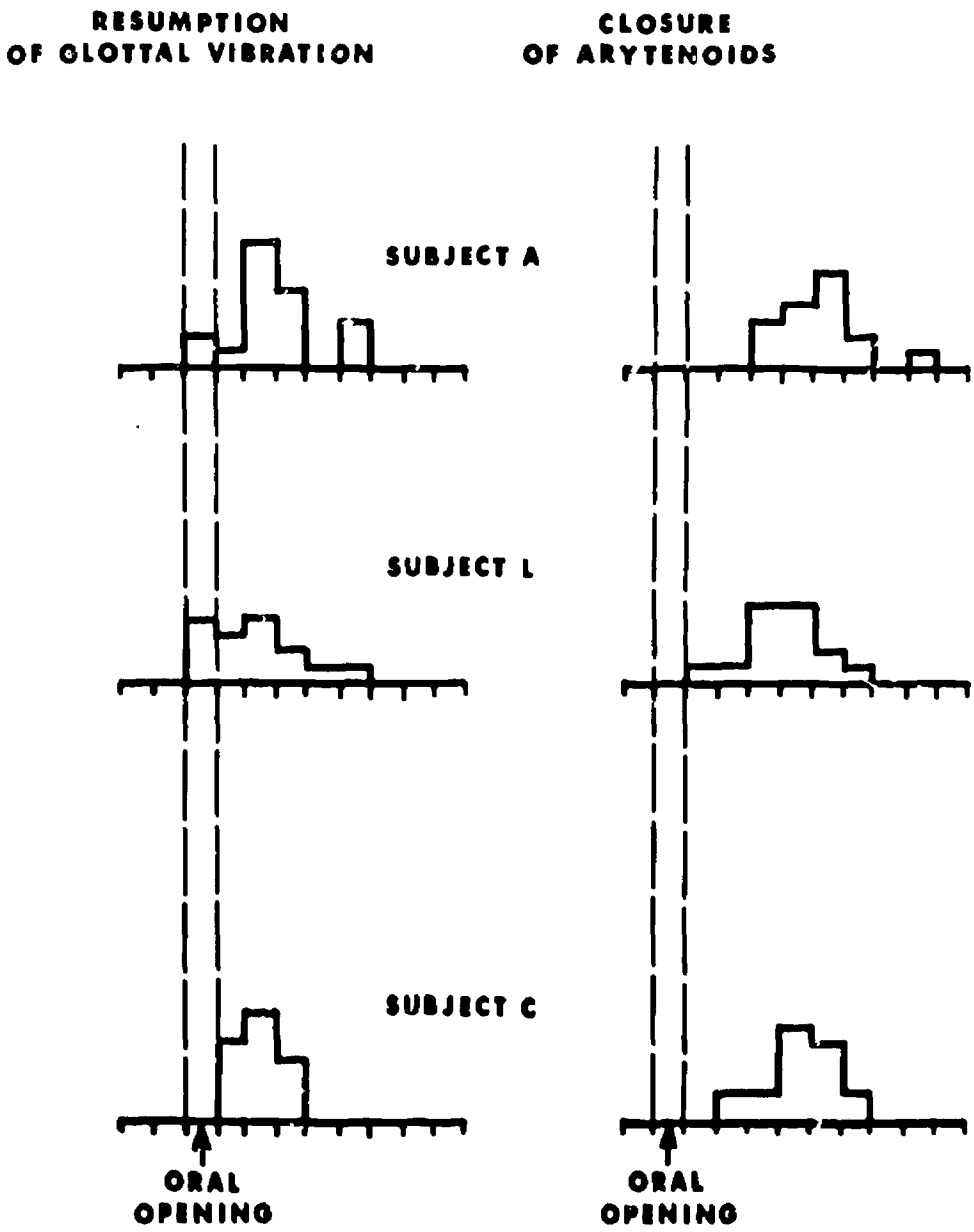
SLIDE 4

VOICELESS ASPIRATED STOPS



SLIDE 5

VOICELESS ASPIRATED STOPS



SLIDE 6

should note that, particularly for subject L, a number of items show resumption of vibration co-occurring with oral release. The arytenoids resume a closed position well after oral release, on the average after the onset of vibration.

Allowing for a certain amount of noise in our observations, which we will not go into here, it appears that the classes of phonetic events we have been considering are produced with rather different laryngeal gestures, in respect both to magnitude of opening and to timing relative to supraglottal events.

MANUSCRIPTS FOR PUBLICATION, REPORTS, AND ORAL PAPERS*

Manuscripts and Publications

The Voicing Dimension: Some Experiments in Comparative Phonetics. Leigh Lisker and Arthur S. Abramson. Proc. Sixth Intl. Cong. Phon. Sci. (Prague: Academia, 1970) 563-567.

Discriminability Along the Voicing Continuum: Cross-Language Tests. Arthur S. Abramson and Leigh Lisker. Proc. Sixth Intl. Cong. Phon. Sci. (Prague: Academia, 1970) 569-573.

Hemispheric Specialization for Speech Perception. Michael Studdert-Kennedy and Donald Shankweiler. J. Acoust. Soc. Amer. 48 (August 1970) 579-594.

Opposed Effects of a Delayed Channel on Perception of Dichotically and Monotically Presented CV Syllables. Michael Studdert-Kennedy, Donald Shankweiler, and S. Schulman. J. Acoust. Soc. Amer. 48 (August 1970) 599-602.

Effects of Filtering and Vowel Environment on Consonantal Perception. Thomas Gay. J. Acoust. Soc. Amer. 48, 4 (Part 2) (October 1970) 993-998.

Supraglottal Air Pressure in the Production of English Stops. Leigh Lisker. To be published in Lang. and Speech 13, Part 4 (December 1970)

Ear Differences in the Recall of Fricatives and Vowels. C.J. Darwin. To be published in Quart. J. Exptl. Psych. (1971).

On the Speech of Neanderthal Man. Philip Lieberman and Edmund S. Crelin. To be published in Linguistic Inquiry 2, 2 (March 1971).

Discrimination in Speech and Nonspeech Modes. Ignatius G. Mattingly, Alvin M. Liberman, Ann K. Syrdal, and Terry Halwes. This paper incorporates data, some of which have been reported earlier by the same authors in J. Acoust. Soc. Amer. 45 and 48 and by A.M. Liberman in "Some Characteristics of Perception in the Speech Mode," Proc. Association for Research in Nervous and Mental Diseases (in press).

A Direct Magnitude Scaling Method to Investigate Categorical Versus Continuous Modes of Speech Perception. Michael D. Vinegrad.

Texts of Oral Reports

Temporal Order Judgments in Speech: Are Individuals Language-Bound or Stimulus-Bound? Ruth S. Day. Paper presented at the Ninth Annual Meeting of the Psychonomic Society, St. Louis, Mo., November 1969.

---

*The contents of this Report, SR 21/22, are included in this listing.

MANUSCRIPTS FOR PUBLICATION, REPORTS, AND ORAL PAPERS (Cont.)

Cinegraphic Observations of the Larynx During Voiced and Voiceless Stops. Leigh Lisker, Masayuki Sawashima, Arthur S. Abramson, and Franklin S. Cooper. Paper presented at the Seventy-ninth Meeting of the Acoustical Society of America, Atlantic City, New Jersey, 21-24 April 1970.

Ear Differences in the Recall of Vowels Produced by Different Sized Vocal Tracts. C.J. Dawrin. Paper presented at the Seventy-ninth Meeting of the Acoustical Society of America, Atlantic City, New Jersey, 21-24 April 1970.

Selective Listening for Temporally Staggered Dichotic CV Syllables. Emily Kirstein. Paper presented at the Seventy-ninth Meeting of the Acoustical Society of America, Atlantic City, New Jersey, 21-24 April 1970.

Glottal Adjustments for English Obstruents. Masayuki Sawashima. Paper presented at Meeting of the New York Speech and Hearing Association, 4 May 1970.

Other Oral Reports

Colloquium. Donald Shankweiler. Stanford University Medical School, June 1969.

Dichotic Listening as a Method for Analysis of Speech Perception. Donald Shankweiler. Seventeenth International Congress of Psychology, University College, London, July 1969.

On the Nature of Cerebral Dominance. Donald Shankweiler. Forty-fifth Annual Convention of the American Speech and Hearing Association, Chicago, November 1969.

Cerebral Dominance in Speech Perception. Donald Shankweiler. Annual Convention of the American Association for the Advancement of Science, Boston, Mass., 30 December 1969.

Temporal Order Perception of a Reversible Phoneme Cluster. Ruth S. Day. Seventy-ninth Meeting of the Acoustical Society of America, Atlantic City, New Jersey, 21-24 April 1970.

Invited Address. Ruth S. Day. Connecticut State Speech and Hearing Association, May 1970.

The Nature of Hemispheric Specialization for Speech. Donald Shankweiler. Psychology Colloquium, University of Iowa, Iowa City, 15 May 1970.

Colloquium. Ruth S. Day. Department of Psychology, University of Pennsylvania, June 1970.

Invited Address. Ruth S. Day. Bell Telephone Laboratories, July 1970.