DOCUMENT RESUME

ED 318 055	CS 507 126
AUTHOR	Pisoni, David B.; And Others
TITLE	Research on Speech Perception. Progress Report No. 9, January 1983-December 1983.
INSTITUTION	Indiana Univ., Bloomington. Dept. of Psychology.
SPONS AGENCY	Air Force Systems Command, Washington, D.C.; National Institutes of Health (DHHS), Bethesda, Md.; National Inst. of Mental Health (DHHS), Rockville, MD.; National Science Foundation, Washington, D.C.
PUB DATE	83
CONTRACT	AF-F-33615-83-K-0501
GRANT	BNS-83-05387; MH-24027-07; NS-07134-05; NS-12179-08
NOTE	360p.; For other reports in this series, see CS 507 123-129.
PUB TYPE	Reports - Research/Technical (143) -- Collected Works - General (020) -- Information Analyses (070)
EDRS PRICE	MF01/PC15 Plus Postage.
DESCRIPTORS	*Acoustic Phonetics; Auditory Discrimination; *Auditory Perception; Communication Research; Computer Software Development; Infants; *Language Processing; Language Research; Linguistics; Speech; Speech Synthesizers
IDENTIFIERS	Indiana University Bloomington; *Speech Perception; Speech Research; Theory Development

ABSTRACT

Summarizing research activities from January 1983 to December 1983, this is the ninth annual report of research on speech perception, analysis and synthesis conducted in the Speech Research Laboratory of the Department of Psychology at Indiana University. The report includes extended manuscripts, short reports, and progress reports. The report contains the following 17 articles: "Contextual Effects on the Consonant/Vowel Ratio in Speech Production" (P. A. Luce and J. Charles-Luce); "Some Effects of Training on the Perception of Synthetic Speech" (E. C. Schwab and others); "Vowel Categorization by Three-Year-Old Children" (C. A. Kubaska and R. N. Aslin); "Effects of Speech Rate and Pitch Contour on the Perception of Synthetic Speech" (L. M. Slowiaczek and H. C. Nusbaum); "Recognition of Speech Spectrograms" (B. G. Greene and others); "Developmental Trends in the Classification and Perceived Similarity of Spoken Syllables" (A. C. Walley and others); "Speech Perception: Some New Directions in Research and Theory" (D. B. Pisoni); "Linguistic Experience and Infant Speech Perceptions: A Re-examination of Eilers, Gavin and Oller (1982)" (P. W. Jusczyk and others); "Contextual Variability and the Problem of Acoustic-Phonetic Invariance in Speech" (D. B. Pisoni); "Converging Approaches Towards Establishing Invariant Acoustic Correlates of Stop Consonants" (D. Kewley-Port); "Identification of Speech Spectrograms: Comparisons of Naive and Trained Observers" (B. G. Greene and others); "Perceptual Evaluation of Synthetic Speech: Some Constraints on the Use of Voice Response Systems" (H. C. Nusbaum and others); "Capacity-Demanding Encoding of Synthetic Speech in Serial-Ordered Recall" (P. A. Luce and D. B. Pisoni); "The Representation of Synthetic Speech in Precategorical Acoustic Storage" (P. A. Luce and D. B. Pisoni); "The Role of Fundamental Frequency and Duration in the Perception of Clause Boundaries: Evidence from a Speeded Verification Task" (P. A. Luce and J. Charles-Luce); "Perception of Synthetic Speech by Children" (B. G. Greene); and "Perceptual Evaluation of Synthetic Speech: Some Considerations of the User/System Interface" (D. B. Pisoni and others). Lists of publications and of laboratory staff, associated faculty and personnel conclude the report. (SR)

RESEARCH ON SPEECH PERCEPTION

Progress Report No. 9
January 1983 — December 1983

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana
47405

Supported by

Department of Health and Human Services
U.S. Public Health Service

National Institute of Mental Health
Research Grant No. MH-24027-07

National Institutes of Health
Research Grant No. NS-12179-08

National Institutes of Health
Training Grant No. NS-07134-05

National Science Foundation
Research Grant No. BNS-83-05387

and

Air Force Office of Scientific Research
Air Force Systems Command
Contract No. AF-F-33615-83-K-0501

RESEARCH ON SPEECH PERCEPTION

Progress Report No. 9

January 1983 - December 1983

David B. Pisoni, Ph.D.

Principal Investigator

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

Research Supported by:

Department of Health and Human Services
U.S. Public Health Service

National Institute of Mental Health
Research Grant No. MH-24027-07

National Institutes of Health
Research Grant No. NS-12179-08

National Institutes of Health
Training Grant No. NS-07134-05

National Science Foundation
Research Grant No. BNS 83-05387

and

Air Force Office of Scientific Research
Air Force Systems Command
Contract No. AF-F-33615-83-K-0501

Table of Contents

II. Short Reports and Work-in-Progress (Cont.)

INTRODUCTION

This is the ninth annual report summarizing the research activities on speech perception, analysis and synthesis conducted in the Speech Research Laboratory, Department of Psychology, Indiana University in Bloomington. As with previous reports, our main goal has been to summarize various research activities over the past year and make them readily available to granting agencies, sponsors and interested colleagues in the field. Some of the papers contained in this report are extended manuscripts that have been prepared for formal publication as journal articles or book chapters. Other papers are simply short reports of research presented at professional meetings during the past year or brief summaries of "on-going" research projects in the laboratory. We also have included new information on instrumentation developments and software support when we think this information would be of interest or help to others. We have found the sharing of this information to be very useful in facilitating our own research.

We are distributing reports of our research activities because of the ever increasing lag in journal publications and the resulting delay in the dissemination of new information and research findings in the field of speech processing. We are, of course, very interested in following the work of other colleagues who are carrying out research on speech perception, production, analysis and synthesis and, therefore, would be grateful if you would send us copies of your own recent reprints, preprints and progress reports as they become available so that we can keep up with your latest findings. Please address all correspondence to:

Professor David B. Pisoni
Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405
U.S.A.

Copies of this report are being sent primarily to libraries and specific research institutions rather than individual scientists. Because of the rising costs of publication and printing and the continued decline in funding for research it is not possible to provide multiple copies of this report or issue copies to individuals. We are eager to enter into exchange agreements with other institutions for their reports and publications. Please write to the above address.

The information contained in the report is freely available to the public and is not restricted in any way. The views expressed in these research reports are those of the individual authors and do not reflect the opinions of the granting agencies or sponsors of the specific research.

I. EXTENDED MANUSCRIPTS

Contextual effects on the consonant/vowel ratio in speech production*

Paul A. Luce

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

and

Jan Charles-Luce

Speech Research Laboratory
Department of Psychology
and
Department of Linguistics
Indiana University
Bloomington, Indiana 47405

*We are extremely grateful to David Pisoni for his enthusiastic support and encouragement of this project and for his many helpful comments on earlier versions of this paper. We would also like to thank Howard Nusbaum, Robert Port, Diane Kewley-Port, and Amanda Walley for their comments and Eileen Schwab for her statistical advice. This work was supported by a grant from NIH (NS-12179).

Abstract

An experiment was conducted to determine the extent to which the consonant/vowel ratio is a context-independent acoustic correlate of phonological voicing of syllable-final stop consonants in speech production. We measured vowel and closure durations of test words containing vowels of three different inherent durations produced in two different sentence positions and two different local phonetic environments. Six subjects read 144 CVC test words that occurred either in non-phrase-final or phrase-final position within sentence frames. The test words contained one of three vowels (/I/, /i/, and /a/) and ended in either a voiced or voiceless velar, bilabial, or dental stop. Each test word was also produced in one of two local phonetic environments: The word immediately following the test word began with the reduced vowel /ə/ or the voiceless dental stop /t/. Our results showed that the C/V ratio is a highly context-dependent correlate of voicing. In addition, we found that for test words ending in velar and bilabial stops, vowel duration was a more reliable correlate of voicing than the C/V ratio across changes in intrinsic vowel duration and local and sentential contexts. These results argue against the claim that the durational correlate of voicing of word-final stop is best viewed as a combination of vowel and post-vocalic closure durations.

Contextual effects on the consonant/vowel ratio in speech production

Two primary temporal attributes of phonological voicing of syllable-final stop consonants are well-known in the literature: First, vowels preceding voiced stops are longer than those preceding vo'celess stops (Crystal and House, 1982; Denes, 1955; House, 1961; House and Fairbanks, 1953; Klatt, 1973; Lisker, 1978; Malecot, 1970; Peterson and Lehiste, 1960) and second, closure durations for voiced stops are shortcr than those for voicaless stops (Lisker, 1957; Port, 1979, 1981a). A number of studies have shown that either vowel duration or closure duration can control perception of voicing (Denes, 1955; Lisker, 1978; Raphael, 1972; Raphael and Dorman, 1980; Liberman, Harris, Eimas, Lisker, and Bastian, 1961; Port and Dalby, 1982). However, other studies have shown that perception of voicing on the basis of either of these temporal cues alone can be systematically shifted by manipulation of the speaking rate of the sentence in which the test stimulus is embedded (Fitch, 1981; Miller and Grosjean, 1981; Miller and Liberman, 1979; Port, 1977, 1978, 1979). Because the same temporal cues that serve to distinguish between two phonemic categories also cue speaking rate, several researchers (eg., "latt, 1976; Fitch, 1981; Port and Dalby, 1982) have noted a confounding of the temporal cues used to perceive segmental phonetic contrasts with the temporal cues used to determine speaking rate: Because these durational cues are affected by both speaking rate and phonological voicing, the perceptual system must somehow determine speaking rate to make a phonetic judgment and, conversely, make a phonetic judgment in order to determine speaking rate (see also Miller, 1981).

Port (1981a, 1981b; Port and Dalby, 1982) has recently argued that the confound of temporal intervals simultaneously cueing speaking rate and voicing may be circumvented if vowel and closure duration are not viewed as independent cues to voicing. Instead, Port (see also Denes, 1955) proposed that the ratio of closure duration to vowel duration (the C/V ratio) may serve as an invariant relational cue to voicing of syllable-final stops in English.[1] In speech production studies, Port (1981a) has shown that the C/V ratio is relationally invariant across speaking rate, number of syllables in the test word, and vowel tensity--all three of which affect the absolute durations of both the vowel and closure intervals.[2] Therefore, although speakers produce vowel and closure intervals within a wide range of durations, the C/V ratio appears to remain relationally invariant in speech production due to temporal compensation between vowel and closure durations. According to Port, because the C/V ratio is a relational cue, and therefore not dependent on the absolute duration of a given segment, it is thus rate-independent and tends to avoid the confounding discussed above.

Port and Dalby (1982) and Fitch (1981) have also provided evidence supporting the perceptual invariance of the C/V ratio. Port and Dalby demonstrated that when other cues to voicing (eg. voicing into closure) are ambiguous, the C/V ratio is apparently the primary cue for perception of voicing of both bilabial and velar stop consonants.[3] Port and Dalby (1982) and Fitch (1981) furthermore have shown that perception of voicing based on C/V ratios remains relatively constant across changes in speaking rate. In both studies, the duration of the test words and the speaking rate of the sentence frames in which they were embedded were varied orthogonally. Port and Dalby found that the effects of speaking rate on perception of voicing were independent of the effects of syllable duration when the C/V ratio was taken to be the primary voicing cue.

Similarly, Fitch (1981) found small changes in the location of voicing boundaries in terms of the C/V ratio for test words embedded in sentence frames produced at fast, normal, and slow speaking rates. Thus, data from studies of both speech production and perception suggest that the temporal factors of vowel duration and closure duration taken in combination provide a relationally invariant cue to voicing, whereas either of these factors alone does not.

Recently, Massaro and Cohen (1983) have reinterpreted Port and Dalby's (1982) results in terms of a fuzzy-logical model in which vowel and closure duration are extracted from the speech signal as independent cues to voicing. This view may be contrasted to Port and Dalby's claim that the C/V ratio is perceived directly as an abstract relational cue to voicing. Although Massaro and Cohen convincingly argue that the Port and Dalby data are better described by their model, they do not adequately address the claim that the C/V ratio is a rate-independent cue to voicing. In particular, they do not address the Fitch (1981) study that showed that at slow, normal, and fast speaking rates, the locus of voicing boundaries in terms of C/V ratios remain almost constant in perception. Thus, although objections to the claim that the C/V ratio is a cue to voicing have been raised, the crucial aspect of the invariance claim for the C/V ratio has not been directly addressed; namely, that the C/V ratio avoids the problem of temporal intervals simultaneously providing cues to segmental phonetic contrasts and speaking rate.

To address this issue, we performed a production study in which we indirectly manipulated "local" rate (cf. Miller, 1981) by exploiting the phenomenon of phrase-final lengthening.[4] Numerous studies have shown that syllables preceding phrase boundaries are longer than syllables produced in non-phrase-final position (Cooper, 1975; Klatt, 1975, 1976; Oller, 1973). Klatt (1975, 1976), Oller (1973), and Umeda (1977) have shown that this lengthening is primarily due to an increase in vowel duration. However, no systematic study has been performed to determine precisely how vowel and closure durations are affected by phrase-final lengthening. Moreover, little if any work has specified how changes in intrinsic vowel duration or local phonetic environment influence the durations of vowels and closures in non-phrase-final compared to phrase-final sentence positions. We were therefore interested, in part, in determining to what extent the C/V ratio would remain invariant across sentence positions due to changes in vowel and closure durations.

However, our primary interest was not in demonstrating the presence or absence of the absolute invariance of the C/V ratio across sentence positions. Studies by Klatt (1975,1976), Oller (1973), and Umeda (1977) already suggest that the C/V ratio may change across sentence positions. Instead, our goal was to determine across which contexts the C/V ratio remains relationally invariant, or, in other words, how independent the C/V ratio is of context. Specifically, we were interested in determining the extent to which the C/V ratio serves to distinguish voicing across changes in intrinsic vowel duration, local phonetic environment, and sentence position independently of intrinsic vowel duration or local or sentential context. By manipulating local rate, we were able to obtain test words articulated at various rates without explicitly asking subjects to speed or slow their speaking rate unnaturally. In addition, we were able to examine in precise detail the effects that phrase-final lengthening has on word-final vowel and consonant closure durations.

1. Method

A. Subjects

Three male and three female volunteers recruited from the laboratory staff served as unpaid subjects. All subjects were native English speakers of a Midwestern dialect and reported no history of speech or hearing disorders. Subjects were naive to the purpose and design of the study.

B. Materials

Nine minimal pairs of consonant-vowel-consonant (CVC) test words were used. Words within a minimal pair differed only on the voicing of the final consonant. Three of the pairs ended in a bilabial stop, three ended in a dental stop, and three ended in a velar stop. Each word-final consonant was preceded by three vowels, /I/, /i/, and /a/. The vowels were chosen to examine the effects of intrinsic vowel duration on the C/V ratio. On the average, /a/ tends to be longer in duration than /i/, and /i/ longer than /I/ (Crystal and House, 1982; House, 1961; House and Fairbanks, 1953; Peterson and Lehiste, 1960). The test words are shown in Table 1.

Insert Table 1 about here

Previous research by Port (1981a) has shown that place of articulation of the initial stop consonant in a CVC syllable has little or no effect on the durations of the intervals relevant to the present study. The initial stop consonants were thus selected only to facilitate measurement of vowel duration. Furthermore, in order to construct such minimally contrastive pairs, it was necessary to include some nonsense words in the set of test words. All nonsense words, however, were phonologically permissible English words.

Each test word was embedded in one of four sentence frames, which are also given in Table 1. In sentence frames (1) and (3), the test word occurred phrase-finally, whereas in sentence frames (2) and (4) one word occurred between the test word and the phrase boundary. The sentential contexts of the test words were thus either non-phrase-final or phrase-final. In addition, in sentence frames (1) and (2), the test word was followed by the reduced vowel /ə/. In sentence frames (3) and (4) the test word was followed by the voiceless dental stop /t/. The local phonetic environments of the test words were thus either a stop or a vowel. (Local phonetic environment here refers only to the initial phoneme of the word following the test word.)

Table 1. Test words and sentence frames produced by the talkers.

Test words

Velars:	/pIg/-/pIk/	/pig/-/pik/	/kag-/kak/
Bilabials:	/dIb/-/dIp/	/dib/-/dip/	/kab/-/kap/
Dentals:	/pId/-/pIt/	/kid/-/kit/	/kad/-/kat/

Sentence Frames

(1) When Mark read ____, Elaine made a checkmark.

(2) When Mark read ____ aloud, Elaine made a checkmark.

(3) If Ted says ____, Tom will leave the room.

(4) If Ted says ____ today, Tom will leave the room.

C. Procedure

Two repetitions of each word in each sentence frame were read by the speakers. This resulted in 144 test sentences (18 test words X 4 sentence frames X 2 repetitions) for each speaker. Each subject read two blocks of 144 stimuli, only one block of which comprised the materials for the present experiment. The order of presentation of the blocks was balanced across subjects. Each subject received a different randomization of the test sentences.

At the beginning of a session, subjects were instructed to read each sentence in as natural a manner as possible and to avoid placing any undue stress on any of the words. Subjects then read a short practice list of sentences to familiarize them with the materials and to allow the experimenter to adjust the levels on the tape recorder. The utterances were recorded in a sound attenuated booth (IAC Model 401A) using an Electro-Voice D054 microphone and an Ampex AG-500 tape recorder.

D. Measurements

All 144 test sentences for each of the six speakers were low-pass filtered at 4.8 kHz and digitized via a twelve-bit analog-to-digital converter. Measurements were made from a visual waveform display using a digital waveform editor (see Luce and Carrell, 1981). For each test word, vowel duration and closure duration for the final stop consonant were measured. Vowel duration was measured from onset of periodic energy to a marked decrease in the periodic energy in the waveform. Closure duration was measured from this decrease in periodic energy to the onset of burst frication of the word-final stop.[5]

2. Results and Discussion

C/V ratios were computed for each test word produced by each speaker. The C/V ratios, vowel durations, and post-vocalic closure durations for each repetition were then averaged for each speaker. Analyses of variance were subsequently performed on these mean values. The results will be discussed separately for each place of articulation of the final consonant.

A. Velars

Insert Figure 1 about here

The mean C/V ratios for the velars are shown in Fig. 1. The bars representing the mean C/V ratios are shown for each vowel and are grouped according to sentence position (non-phrase-final and phrase-final). The top panel shows the mean ratios for /g/ and /k/ produced in the local vowel environment and the bottom panel shows the mean ratios for /g/ and /k/ produced in the local stop environment.

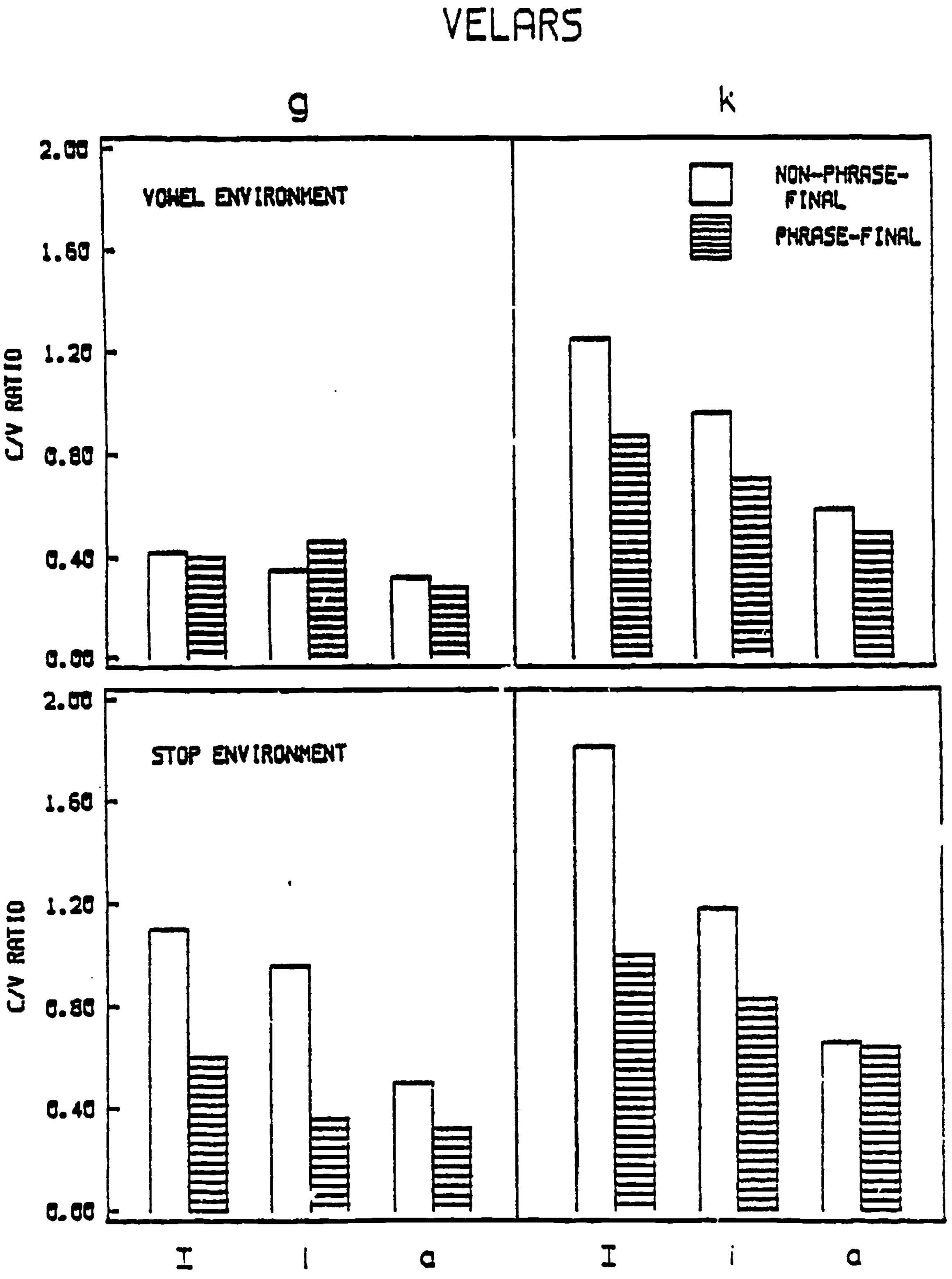


Figure 1. Mean C/V ratios for word-final velar stops. The upper panel shows the mean C/V ratios for test words ending in /g/ and /k/ produced in the vowel environment. The bottom panel shows the mean C/V ratios for test words ending in /g/ and /k/ produced in the stop environment. Mean C/V ratios are shown for each vowel (/I/, /i/, /a/). The open bars refer to the C/V ratios for test words produced in non-phrase-final position; the hatched bars refer to C/V ratios for test words produced in phrase-final position.

Insert Table 2 about here

A four-way (Sentence Position X Local Environment X Voicing X Vowel Duration) repeated measures analysis of variance was performed on the C/V ratios for the velar stops. The _F_ values and significance levels for the significant main effects and interactions are shown in Table 2.

Sentence position, local phonetic environment, voicing, and intrinsic vowel duration all significantly affected the values of the C/V ratios. Specifically, C/V ratios were, on the average, larger for test words produced in non-phrase-final than phrase-final position, as would be expected if vowel durations increased more than closure durations from non-phrase-final to phrase-final positions (see Klatt, 1975, 1976; Oller, 1973; Umeda, 1973). Local phonetic environment also affected the C/V ratios such that the ratios for test words produced in the stop environment were larger than the ratios for test words produced in the vowel environment. The significant main effect of voicing indicates that, on the average, the C/V ratio reliably distinguished voiced from voiceless velar stops, such that the ratios for voiceless stops were larger than those for voiced stops (with a number of exceptions; see below). Finally, the magnitude of the C/V ratios varie inversely and systematically with intrinsic vowel duration.

Four two-way interactions were also significant. The interaction of sentence position and local phonetic environment reveals that local environmental effects on the C/V ratios were stronger in non-phrase-final than in phrase-final position. This finding was not unexpected because any effects of local phonetic environment must operate regressively across a phrase boundary. Our results are therefore in agreement with previous work that has shown that such cross-boundary contextual conditioning is relatively weak or non-existent (cf. Cooper and Paccia-Cooper, 1980).

Sentence position also interacted significantly with vowel duration for the C/V ratios. The magnitude of increase of the C/V ratio from non-phrase-final to phrase-final positions decreased as intrinsic vowel duration increased. A similar effect of local phonetic environment was also obtained, resulting in a significant local phonetic environment by intrinsic vowel duration interaction. Specifically, these two interactions demonstrate that the effects of sentence position and local environment on the C/V ratio are greater for test words containing shorter intrinsic vowel durations.

Finally, the interaction of intrinsic vowel duration and voicing was significant for the C/V ratios. Inspection of Fig. 1 reveals that the differences between the ratios for /g/ and /k/ were smaller overall for test words containing /a/ than for test words containing /i/, and smaller for test words containing /i/ than for those containing /I/. In short, as vowel duration increased, the difference between the ratios for /g/ and /k/ decreased.

Table 2. Significant effects, *F* values, and significance levels for the C/V ratios for the velars.

Effect	*F* value[a]	Significance level
Position	$F(1,5)=18.19$	$p<0.009$
Environment	$F(1,5)=24.70$	$p<0.005$
Voicing	$F(2,10)=57.00$	$p<0.001$
Vowel	$F(2,10)=13.07$	$p<0.002$
Position X Environment	$F(1,5)=6.980$	$p<0.050$
Position X Vowel	$F(2,10)=12.06$	$p<0.003$
Environment X Vowel	$F(2,10)=12.06$	$p<0.006$
Voicing X Vowel	$F(2,10)=6.20$	$p<0.020$

[a] Degrees of freedom are given in parentheses.

Because this interaction suggests that the voicing contrast in terms of the C/V ratio may be absent statistically for test words containing longer vowel durations simple-effects post-hoc tests comparing the ratios for /g/ and /k/ were performed for each vowel. These tests revealed that the C/V ratios were significantly longer for /k/ than for /g/ only for words containing the shortest vowel, /I/, $F(1,6)=9.63$, $p<0.01$. Thus, although Fig. 1 shows longer ratios for test words ending in /k/ for all three vowels, statistically significant effects of voicing on the C/V ratio were obtained only for test words containing /I/.

Inspection of Fig. 1 also reveals that there is no strict separation of the C/V ratios for /g/ and /k/ when differences in sentence position, local phonetic environment, and intrinsic vowel duration are ignored. Instead, in a number of instances, the C/V ratios for test words ending in /g/ are longer than for those ending in /k/. (Compare the ratios for /g/ in the stop environment to those for /k/ in the vowel and stop environments.) Only by taking vowel duration, local phonetic environment, and sentence position into account simultaneously do the C/V ratios distinguish voicing, and then only significantly for words containing /I/.

Insert Figure 2 and Table 3 about here

To examine more closely the results obtained for the C/V ratios, four-way repeated measures analyses of variance were performed on the vowel and post-vocalic closure durations separately. The means for the vowel and closure durations are shown graphically in Fig. 2; the numerical values are given in Table 3. The organization of the bars for the vowel and closure durations in Fig. 2 parallels that in Fig. 1. Vowel durations are shown in the left-hand panel and closure durations are shown in the right-hand panel.

Insert Table 4 about here

Significant main effects for the vowel durations are shown in Table 4. No interactions were significant. Vowel durations were longer for test words produced in phrase-final position than in non-phrase-final position, thus replicating the already well-documented evidence that vowels lengthen phrase-finally (Cooper, 1975; Klatt, 1975, 1976; Oller, 1973). Vowel durations for test words produced in the local vowel environment were significantly longer than the durations for test words produced in the local stop environment. And vowel durations were longer before /g/ than before /k/, again in agreement with previous findings (Crystal and House, 1982; Denes, 1955; House, 1961; House and Fairbanks, 1953; Klatt, 1973; Lisker, 1978; Malecot, 1970; Peterson and Lehiste, 1960). In addition, /I/ was consistently shorter in duration than /i/, and /i/ was likewise consistently shorter than /a/.

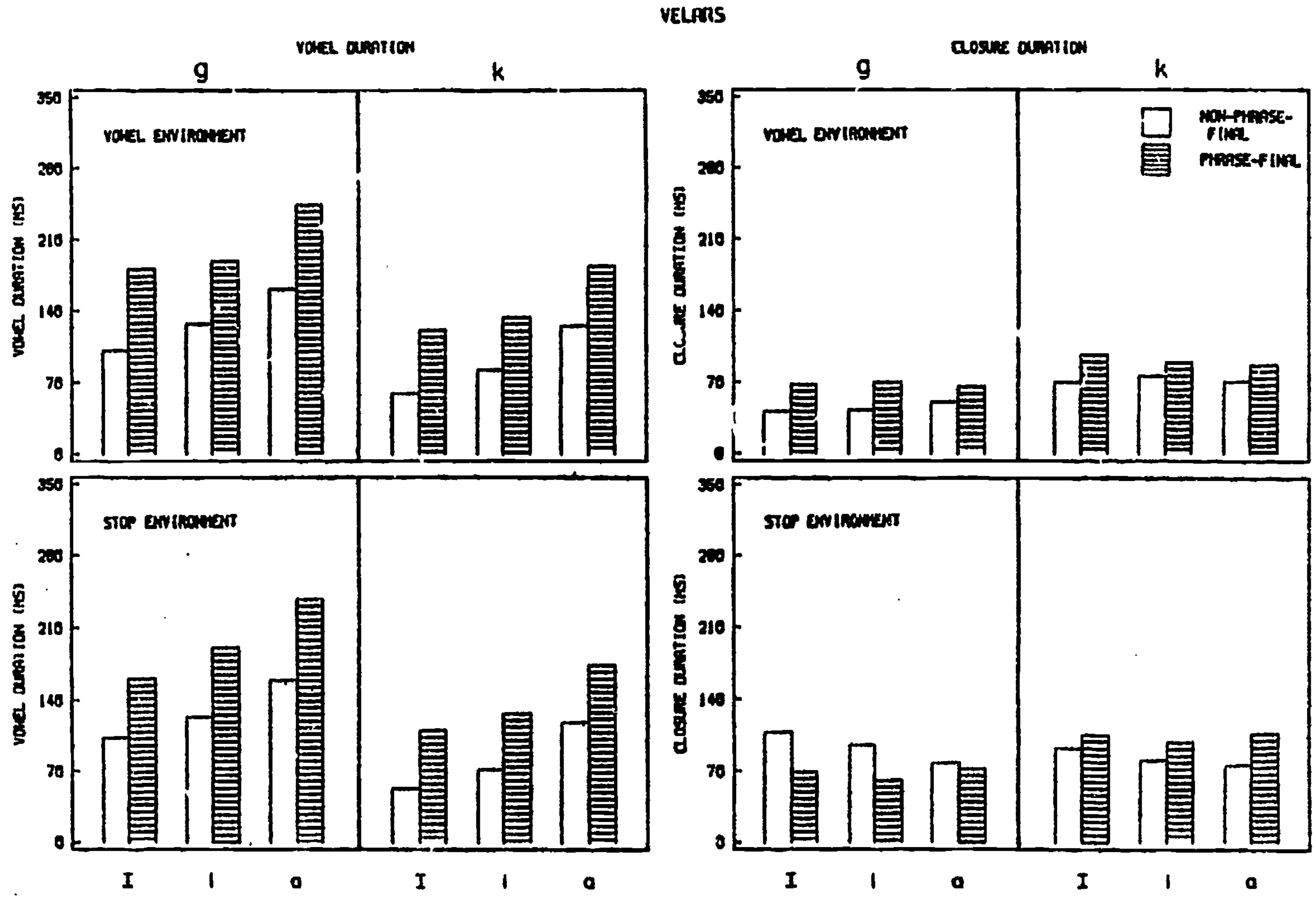


Figure 2. Mean vowel (left panel) and closure (right panel) durations for word-final velar stops. The upper panels show the mean durations for test words ending in /g/ and /k/ produced in the vowel environment. The bottom panels show the mean durations for test words ending in /g/ and /k/ produced in the stop environment. Mean durations are shown for each vowel (/I/, /i/, /a/). The open bars refer to the durations for test words produced in non-phrase-final position; the hatched bars refer to durations for test words produced in phrase-final position.

Table 3. Mean vowel and closure durations in ms for /g/ and /k/. Mean durations are shown for test words containing each vowel (/I/, /i/, /a/) produced in each local environment (stop and vowel) at each sentence position (non-phrase-final and phrase-final).

Vowel Durations:

		Vowel Environment	
		/g/	/k/
Non-phrase final	/I/	101.4	59.6
	/i/	126.9	82.4
	/a/	161.0	125.2
Phrase-final	/I/	180.8	120.9
	/i/	188.6	133.3
	/a/	244.5	183.6

		Stop Environment	
		/g/	/k/
Non-phrase final	/I/	103.2	52.8
	/i/	123.5	71.1
	/a/	159.0	117.4
Phrase-final	/I/	161.1	109.9
	/i/	190.8	126.4
	/a/	237.3	173.3

Closure Durations:

		Vowel Environment	
		/g/	/k/
Non-phrase final	/I/	41.4	70.2
	/i/	42.7	76.6
	/a/	50.9	70.7
Phrase-final	/I/	67.8	96.8
	/i/	70.6	89.9
	/a/	65.5	87.1

		Stop Environment	
		/g/	/k/
Non-phrase final	/I/	108.2	92.3
	/i/	95.8	81.0
	/a/	78.3	76.2
Phrase-final	/I/	69.4	105.3
	/i/	61.1	98.4
	/a/	72.2	106.7

Table 4. Significant effects, *F* values, and significance levels for vowel and closure durations for the velars.

Vowel Durations:

Effect	*F* value[a]	Significance level
Position	*F*(1,5)=17.49	*p*<0.009
Environment	*F*(1,5)=10.97	*p*<0.030
Voicing	*F*(1,5)=52.03	*p*<0.001
Vowel	*F*(2,10)=58.40	*p*<0.001

Closure Durations:

Effect	*F* value[a]	Significance level
Environment	*F*(1,5)=24.60	*p*<0.005
Voicing	*F*(1,5)=33.07	*p*<0.003
Position X Environment	*F*(1,5)=11.29	*p*<0.030
Position X Voicing	*F*(1,5)=12.73	*p*<0.020
Position X Environment X Voicing	*F*(1,5)=35.09	*p*<0.002
Environment X Vowel	*F*(2,10)=4.13	*p*<0.050

[a] Degrees of freedom are given in parentheses.

Clearly, vowel duration was affected by each of our four manipulations. Of special interest, however, is the finding that the effect of voicing on vowel duration did not interact with sentence position, local phonetic environment, or intrinsic vowel duration. Thus, vowel duration consistently distinguished /g/ and /k/ statistically, whereas the C/V ratio did not. Although it is indeed true that vowel duration, like the C/V ratio, distinguishes voicing only when all contextual variables are specified, the absence of any statistically significant interactions demonstrates the independence of the voicing effect for vowel duration from the other variables manipulated in this study.

Turning now to the data for the closure durations, we found significantly longer closure durations for test words produced in the stop environment than for test words produced in the vowel environment. Closure durations were also significantly longer for /k/ than for /g/. Although no significant main effect of sentence position was obtained, it did enter into three significant interactions (sentence position by local phonetic environment, sentence position by voicing, and sentence position by local phonetic environment by voicing). Briefly, the pattern of these interactions indicates that the effect of local phonetic environment was stronger in non-phrase-final position than phrase-final position and that the difference in closure durations between /g/ and /k/ was greater in phrase-final position than in non-phrase-final position. The significant three-way interaction was presumably due to a slightly longer mean closure duration for /g/ than for /k/ in the stop environment in non-phrase-final position. A possible explanation of this result is that word-final /g/ in non-phrase-final position becomes devoiced as a result of regressive voicing assimilation with the following word-initial /t/, thus producing closure durations equal to (or in this case, slightly longer than) closure durations for word-final /k/. This effect was not observed in phrase-final position, however, because regressive assimilation would not be expected to occur across a clause boundary (see also Flege and Brown, 1982; Hyman, 1975; Lisker, 1957).

Although closure durations were largely independent of intrinsic vowel duration (as demonstrated by the nonsignificant main effect of vowel duration), a vowel duration effect on closure durations was implicated in a significant two-way interaction between intrinsic vowel duration and local phonetic environment. This result was due to slightly longer mean closure durations for test words containing /I/ than for test words containing /i/ or /a/ in the stop environment.

Of particular interest for the closure durations, however, is the finding that the voicing effect varied as a function of both sentence position and local phonetic environment (as indicated by the two significant interactions involving voicing). Unlike the results for the vowel durations in which the voicing effect was independent of intrinsic vowel duration, sentence position, and local phonetic environment, the voicing effect in terms of the closure durations was clearly dependent on sentence position and local environment.

B. Bilabials

Insert Figure 3 and Table 5 about here

The mean C/V ratios for the bilabials are shown in Fig. 3. The significant main effects and interactions are given in Table 5.

As with the velars, significant main effects of sentence position, local phonetic environment, voicing, and intrinsic vowel duration were obtained for the C/V ratios. Of special interest, however, are the two significant interactions of sentence position and voicing and local phonetic environment and voicing. These interactions suggest that the voicing effect in terms of the C/V ratio was attenuated both in phrase-final position and in the stop environment. To determine the source of these interactions, simple effects tests were performed comparing the ratios for /b/ and /p/ at each sentence position and each local phonetic environment. These tests revealed that the ratios for /b/ and /p/ differed significantly only for test words produced in non-phrase-final position in the local vowel environment, $\underline{F}(1,5)=6.61$, $\underline{p}<0.05$.

The significant interaction of sentence position and local phonetic environment indicates that local phonetic environmental effects were again stronger non-phrase-finally than phrase-finally. The remaining two significant interactions of sentence position and intrinsic vowel duration and local phonetic environment and intrinsic vowel duration were due to greater effects of sentence position and local environment on the C/V ratios for words containing /I/ than for those containing /i/ and greater effects of sentence position and environment on words containing /i/ than those containing /a/. As with the velars, the effects of local and sentential context on the C/V ratios were influenced by the intrinsic duration of the vowel.

Again, we found that the C/V ratio failed to consistently distinguish voicing of syllable-final stops. However, unlike the results obtained for the velars in which intrinsic vowel duration interacted with voicing, for the bilabials, sentence position and local phonetic environment affected voicing in terms of the C/V ratio. Statistically significant differences in C/V ratios for test words ending in /b/ and /p/ were obtained only for test words produced in the vowel environment in non-phrase-final position.

Inspection of Fig. 3 also reveals that the C/V ratios for /b/ and /p/ did not fall into strictly separate classes. As with the velars, only when intrinsic vowel duration, local phonetic environment, and sentence position are simultaneously taken into account can the C/V ratio be said to differ consistently for /b/ and /p/. But, as before, many of these differences did not reach statistical significance.

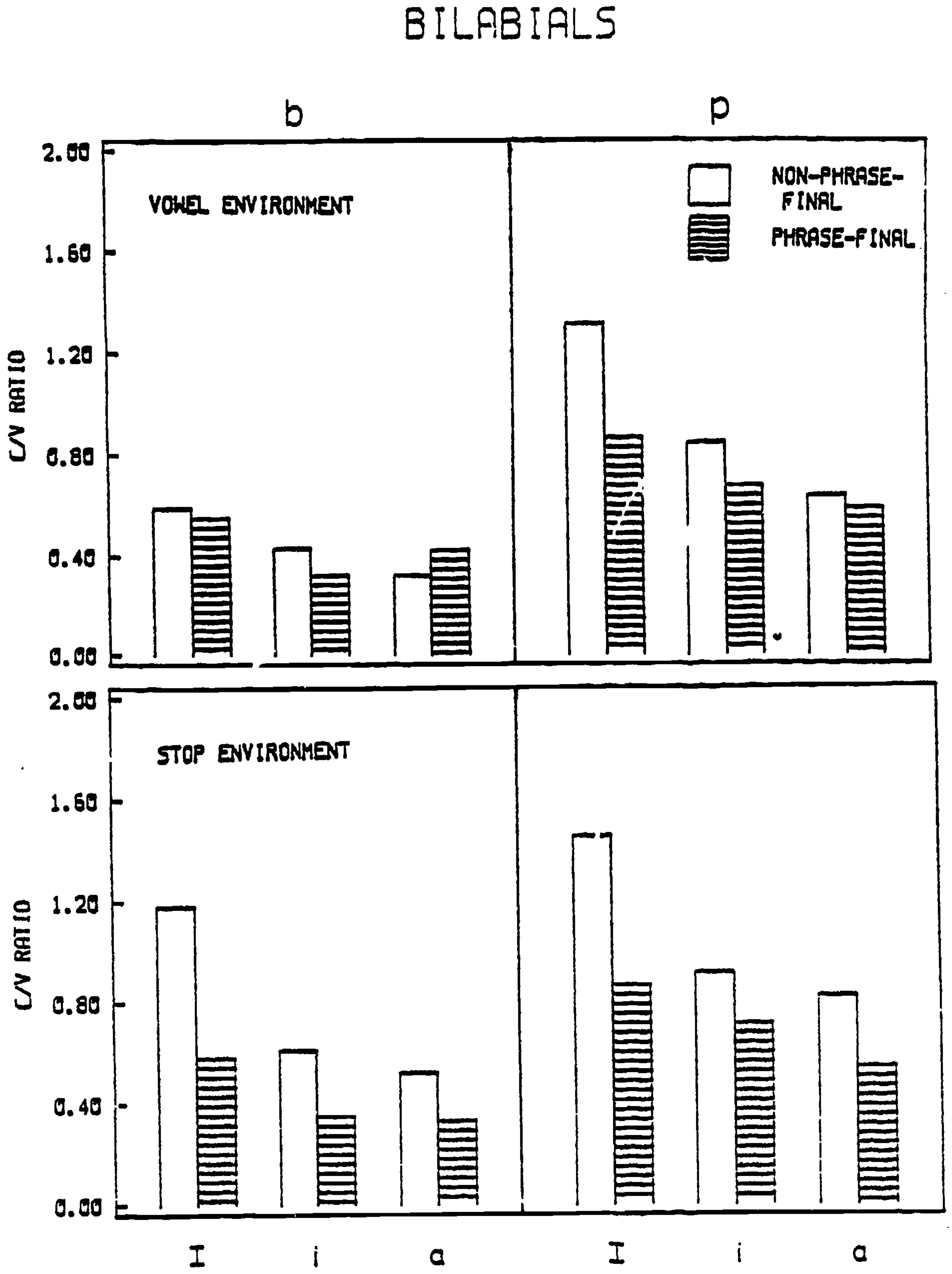


Figure 3. Mean C/V ratios for word-final bilabial stops. The upper panel shows the mean C/V ratios for test words ending in /b/ and /p/ produced in the vowel environment. The bottom panel shows the mean C/V ratios for test words ending in /b/ and /p/ produced in the stop environment. Mean C/V ratios are shown for each vowel (/I/, /i/, /a/). The open bars refer to the C/V ratios for test words produced in non-phrase-final position; the hatched bars refer to C/V ratios for test words produced in phrase-final position.

Table 5. Significant effects, F values, and significance levels for the C/V ratios for the bilabials.

Effect	F value[a]	Significance level
Position	$F(1,5)=31.51$	$p<0.003$
Environment	$F(1,5)=55.88$	$p<0.001$
Voicing	$F(1,5)=227.61$	$p<0.001$
Vowel	$F(2,10)=49.64$	$p<0.001$
Position X Environment	$F(1,5)=30.85$	$p<0.003$
Position X Voicing	$F(1,5)=12.51$	$p<0.020$
Environment X Voicing	$F(1,5)=12.06$	$p<0.020$
Position X Vowel	$F(2,10)=8.86$	$p<0.007$
Environment X Vowel	$F(2,10)=5.49$	$p<0.030$

[a] Degrees of freedom are given in parentheses.

Insert Figure 4 and Table 6 about here

Mean vowel and closure durations for the bilabials are graphically displayed in Fig. 4. The numerical values are given in Table 6.

Insert Table 7 about here

The significant effects for the vowel and closure durations are summarized in Table 7. As with the velars, the four main effects of sentence position, local phonetic environment, voicing, and intrinsic vowel duration were significant for the vowel durations. No significant interactions were obtained. The effect of voicing on vowel duration therefore again proved statistically reliable for vowel duration and was unaffected by intrinsic vowel duration or local or sentential context.

For the closure durations, significant main effects of sentence position and voicing were obtained. Again, no significant interactions were obtained. The finding that the effect of voicing was independent of intrinsic vowel duration and local and sentential context for both the vowel and closure durations reveals that, for the bilabials, the durations of the vowel and closure are more reliable correlates of voicing when considered separately than when combined in the C/V ratio.

C. Dentals

Insert Figure 5 and Table 8 about here

Mean C/V ratios for the dentals are shown in Fig. 5. Significant main effects and interactions are summarized in Table 8.

As with the velars and bilabials, the main effects of sentence position, local phonetic environment, voicing and intrinsic vowel duration for the dentals were all significant. However, unlike the results for the velars and bilabials, the effect of voicing did not interact with any of the other three independent variables. Thus, for the dentals, the C/V ratio reliably distinguished voicing across intrinsic vowel durations and local and sentential contexts.

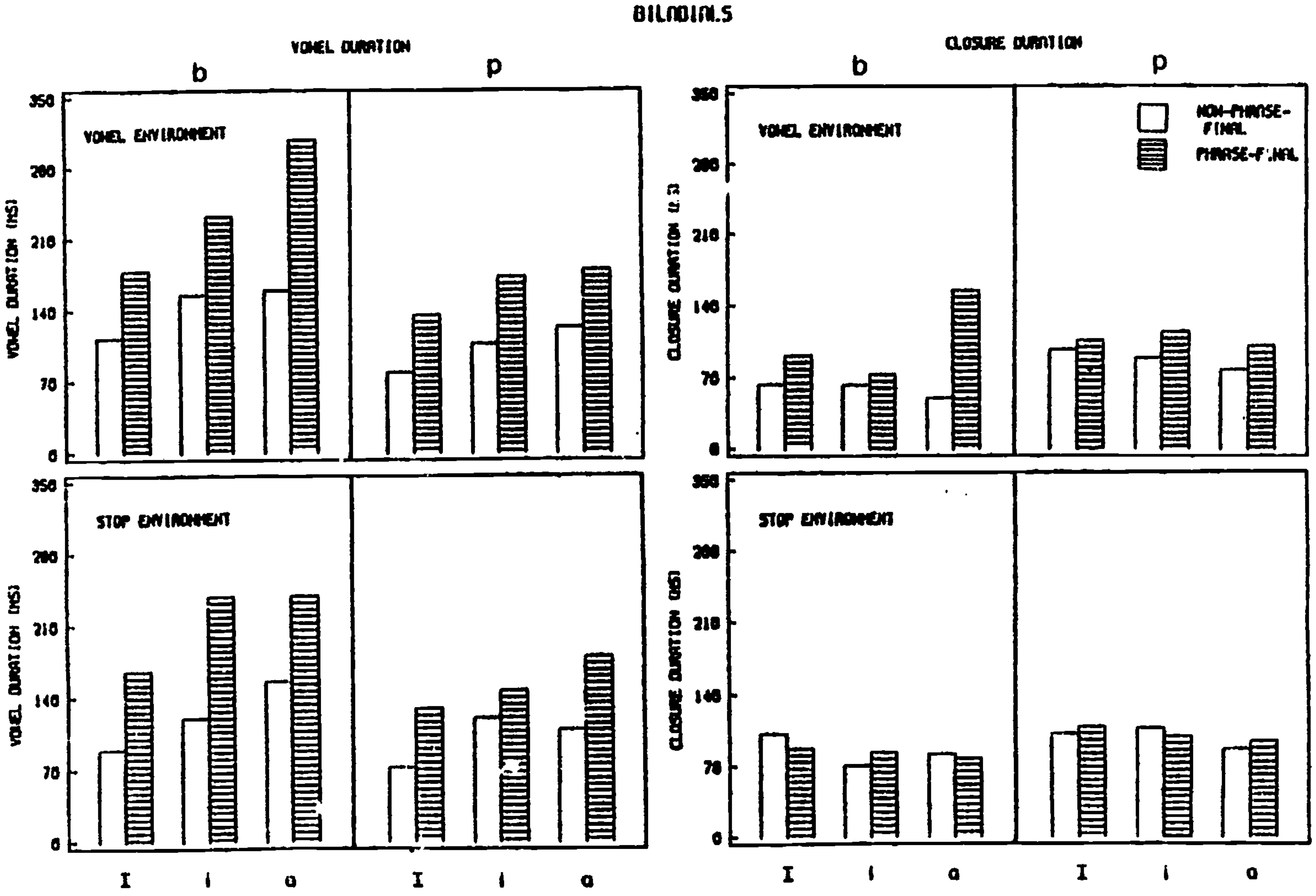


Figure 4. Mean vowel (left panel) and closure (right panel) durations for word-final bilabial stops. The upper panels show the mean durations for test words endi~g in /b/ and /p/ produced in the vowel environment. The bottom panels show the mean durations for test words ending in /b/ and /p/ produced in the stop environment. Mean durations are shown for each vowel (/I/, /i/, /a/). The open bars refer to the durations for test words produced in non-phrase-final position; the hatched bars refer to durations for test words produced in phrase-final position.

Table 6. Mean vowel and closure durations in ms for /b/ and /p/. Mean durations are shown for test words containing each vowel (/I/, /i/, /a/) produced in each local environment (stop and vowel) at each sentence position (non-phrase-final and phrase-final).

Vowel Durations:		Vowel Environment	
		/b/	/p/
Non-phrase final	/I/	112.8	78.5
	/i/	154.8	106.8
	/a/	159.8	123.5
Phrase-final	/I/	178.7	135.0
	/i/	232.4	172.6
	/a/	307.9	179.8
		Stop Environment	
		/b/	/p/
Non-phrase final	/I/	89.3	72.2
	/i/	120.0	120.3
	/a/	155.8	108.7
Phrase-final	/I/	165.7	129.6
	/i/	238.0	146.8
	/a/	238.8	180.4

Closure Durations:		Vowel Environment	
		/b/	/p/
Non-phrase final	/I/	63.5	96.6
	/i/	62.8	88.8
	/a/	50.3	77.3
Phrase-final	/I/	91.1	105.9
	/i/	72.8	114.3
	/a/	154.5	100.1
		Stop Environment	
		/b/	/p/
Non-phrase final	/I/	101.6	100.8
	/i/	70.4	106.5
	/a/	81.3	86.1
Phrase-final	/I/	87.0	107.9
	/i/	83.2	98.5
	/a/	77.3	93.5

Table 7. Significant effects, F values, and significance levels for vowel and closure durations for the bilabials.

Vowel Durations:		
Effect	F value[a]	Significance level
Position	$F(1,5)=30.92$	$p<0.003$
Environment	$F(1,5)=7.090$	$p<0.050$
Voicing	$F(1,5)=152.85$	$p<0.001$
Vowel	$F(2,10)=27.46$	$p<0.001$
Closure Durations:		
Effect	F value[a]	Significance level
Position	$F(1,5)=7.71$	$p<0.040$
Voicing	$F(1,5)=6.96$	$p<0.050$

[a] Degrees of freedom are given in parentheses.

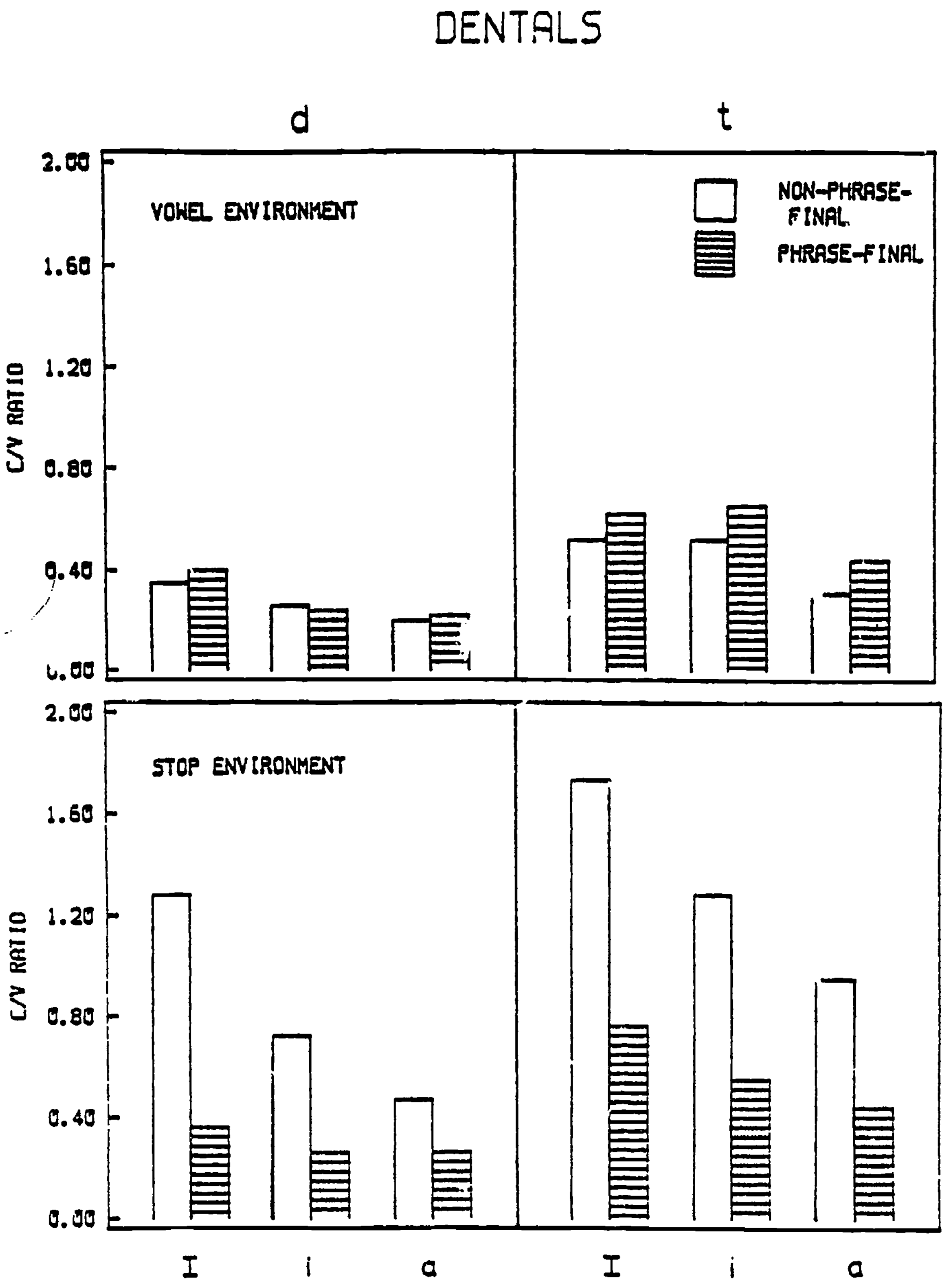


Figure 5. Mean C/V ratios for word-final dental stops. The upper panel shows the mean C/V ratios for test words ending in /d/ and /t/ produced in the vowel environment. The bottom panel shows the mean C/V ratios for test words ending in /d/ and /t/ produced in the stop environment. Mean C/V ratios are shown for each vowel (/I/, /i/, /a/). The open bars refer to the C/V ratios for test words produced in non-phrase-final position; the hatched bars refer to C/V ratios for test words produced in phrase-final position.

Table 8. Significant effects, *F* values, and significance levels for the C/V ratios for the dentals.

Effect	*F* value[a]	Significance level
Position	$F(1,5)=11.53$	$p<0.020$
Environment	$F(1,5)=39.97$	$p<0.002$
Voicing	$F(2,10)=112.9$	$p<0.001$
Vowel	$F(2,10)=17.74$	$p<0.001$
Position X Environment	$F(1,5)=33.86$	$p<0.003$
Position X Vowel	$F(2,10)=9.58$	$p<0.005$
Environment X Vowel	$F(2,10)=13.60$	$p<0.002$
Position X Environment X Vowel	$F(2,10)=14.98$	$p<0.001$

[a] Degrees of freedom are given in parentheses.

Also, as as in our previous analyses, different effects of local phonetic environment were observed for C/V ratios at non-phrase-final and phrase-final positions, resulting in a significant sentence position by local environment interaction: The effects of local environment were stronger non-phrase-finally than phrase-finally. And again, the shorter the duration of the vowel in the test word, the larger the effects of sentence position and local phonetic environment, as indicated by the significant interactions of local environment and intrinsic vowel duration and sentence position, local environment, and intrinsitic vowel duration.

Insert Figure 6 and Table 9 about here

Mean vowel and closure durations for the dentals are shown in Fig. 6. The numerical values are given in Table 9.

Insert Table 10 about here

The significant main effects and interactions for the vowel and closure durations are summarized in Table 10. For vowel duration, significant main effects of sentence position, voicing, and intrinsic vowel duration were obtained. Although the main effect of local phonetic environment failed to reach significance, an effect of local environment was implicated in a significant sentence position by local phonetic environment interaction. As with the velars and bilabials, local environmental effects on vowel duration were generally restricted to non-phrase-final position in which vowel durations for test words produced in the vowel environment were longer than vowel durations for test words produced in the stop environment.

The significant interactions of most interest, however, are those involving voicing. Unlike the velars and bilabials, the voicing effect on the vowel durations for the dentals varied with both sentence position and intrinsic vowel duration. To evaluate the source of these two interactions, two sets of simple effects post-hoc tests were performed. These tests revealed significant effects of voicing for vowel duration only in phrase-final position (_F_(1,6)=7.83, _p_<.04) and only for the vowel /i/ (_F_(1,5)=7.26, _p_<.05). Thus, for the dentals, vowel duration failed to consistently distinguish voicing for syllable-final /d/ and /t/ across intrinsic vowel duration or sentence position.

For the closure durations, significant effects of local phonetic environment and voicing were obtained. Local phonetic environment also interacted with sentence position such that larger effects of local environment occurred in non-phrase-final position than phrase-final position. Although the main effect

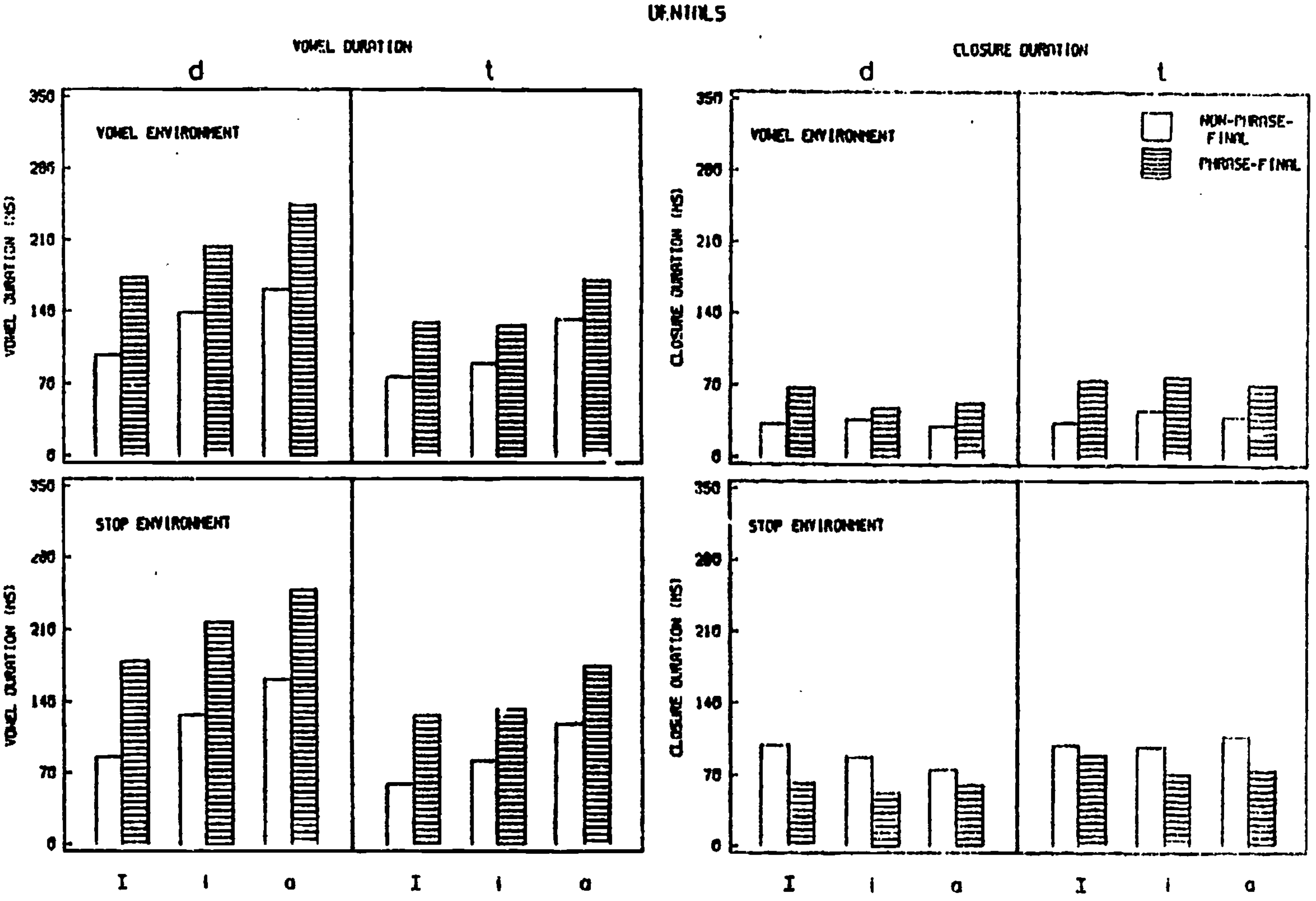


Figure 6. Mean vowel (left panel) and closure (right panel) durations for word-final dental stops. The upper panels show the mean durations for test words ending in /d/ and /t/ produced in the vowel environment. The bottom panel shows the mean durations for test words ending in /d/ and /t/ produced in the stop environment. Mean durations are shown for each vowel (/I/, /i/, /a/).

Table 9. Mean vowel and closure durations in ms for /d/ and /t/. Mean durations are shown for test words containing each vowel (/I/, /i/, /a/) produced in each local environment (stop and vowel) at each sentence position (non-phrase-final and phrase-final).

Vowel Durations:

		Vowel Environment	
		/d/	/t/
Non-phrase final	/I/	97.5	76.0
	/i/	138.7	89.8
	/a/	161.4	132.0
Phrase-final	/I/	172.9	128.7
	/i/	203.0	125.3
	/a/	244.1	170.4

		Stop Environment	
		/d/	/t/
Non-phrase final	/I/	85.5	59.1
	/i/	126.4	81.5
	/a/	161.2	117.9
Phrase-final	/I/	179.4	126.0
	/i/	216.7	131.6
	/a/	247.9	174.2

Closure Durations:

		Vowel Environment	
		/d/	/t/
Non-phrase final	/I/	32.3	34.3
	/i/	36.5	46.1
	/a/	30.4	40.3
Phrase-final	/I/	67.0	74.8
	/i/	47.4	78.7
	/a/	52.2	71.1

		Stop Environment	
		/d/	/t/
Non-phrase final	/I/	99.1	98.9
	/i/	87.5	97.6
	/a/	74.7	106.9
Phrase-final	/I/	62.5	88.9
	/i/	52.8	71.3
	/a/	59.9	74.2

Table 10. Significart effects, *F* values, and significance levels for vowel and closure durations for the dentals.

Vowel Durations:

Effect	*F* value[a]	Significance level
Position	$F(1,5)=23.18$	$p<0.005$
Voicing	$F(1,5)=99.15$	$p<0.001$
Vowel	$F(2,10)=213.3$	$p<0.001$
Position X Environment	$F(1,5)=10.62$	$p<0.030$
Position X Voicing	$F(1,5)=17.47$	$p<0.009$
Voicing X Vowel	$F(2,10)=18.50$	$p<0.001$

Closure Durations:

Effect	*F* value[a]	Significance level
Environment	$F(1,5)=57.19$	$p<0.001$
Voicing	$F(1,5)=23.53$	$p<0.005$
Position X Environment	$F(1,5)=39.17$	$p<0.002$

[a] Degrees of freedom are given in parentheses.

of voicing was significant and independent of intrinsic vowel duration or context, inspection of Fig. 6 reveals the effect of voicing to be relatively small and, in some specific cases, virtually nonexistent. Also of interest in Fig. 6 is the tendency for closure durations in the stop environment to be longer in non-phrase-final than phrase-final position. In non-phrase-final position, the final dental stop of the test words may have frequently been unreleased, probably due to assimilation of the word-final dental stop with the /t/ of the following word. Thus, longer closure durations due to assimilation in non-phrase-final position may have resulted in decreases in closure duration from non-phrase-final to phrase-final positions.

3. General Discussion

Our primary intent in this study was to investigate the extent to which the C/V ratio is a context-independent correlate of phonological voicing of word-final stops. We believe the results of this study clearly demonstrate that for velar, bilabial, and dental word-final stop consonants the C/V ratio varies with intrinsic vowel duration, sentence position, and local phonetic environment. These findings were not unexpected. Port (1981a) has shown similar effects of intrinsic vowel duration on the C/V ratio for the vowels /I/ and /i/. The effects of sentence position were likewise predicted by previous work on phrase-final lengthening of acoustic-phonetic segments (Klatt, 1975, 1976; Oller, 1973; Umeda, 1973). And, finally, the effects of local phonetic environment reflect well-known influences of phonetic context on closure durations (Umeda, 1977; see also Pickett and Decker, 1960). However, we have demonstrated that the combined influences of these variables makes the C/V ratio a highly context-dependent correlate of phonological voicing. When intrinsic vowel duration, local phonetic environment, and sentence position are not specified, the C/V ratio fails to uniquely distinguish voicing; only when all such variables are taken into account can the C/V ratio be assumed to be an invariant correlate of voicing.

On the basis of this finding it is clear that the C/V ratio does not circumvent the confound discussed by Klatt (1976) and Port and Dalby (1982) that segmental durations appear to simultaneously provide cues for both speaking rate and phonological voicing. Were the perceptual system to employ the C/V ratio, our results suggest that relatively smaller C/V ratios would cue clause boundaries and once a clause boundary has been identified, this information would then have to be taken into consideration in judging voicing on the basis of the C/V ratio. In addition, a perceptual mechanism employing the C/V ratio must also be sensitive to the intrinsic duration of the vowel and the local phonetic environment in adjusting perception of voicing at clause boundaries. In short, a perceptual system that evaluates the C/V ratio directly as a cue to segmental voicing must have simultaneous knowledge of sentence position, local phonetic environment, and intrinsic vowel duration. The C/V ratio thus succumbs to the very problems it was originally postulated to alleviate.

Of particular interest, however, is the finding that the effect of voicing in terms of the C/V ratio for velar and bilabial word-final stops interacted with intrinsic vowel duration (for the velars) and local phonetic environment and sentence position (for the bilabials). Upon closer analysis of these interactions, we found that for the velars and bilabials, the C/V ratios for voiced and voiceless stops failed to differ significantly in a number of instances, thus calling into question the claim that the C/V ratio is a reliable correlate of voicing in speech production. Moreover, we found that for the bilabials and velars, vowel duration alone consistently distinguished voicing. Although vowel duration, like the C/V ratio, proved to be a highly context-dependent correlate of voicing, it nonetheless reliably distinguished voicing for the velars and bilabials, whereas the C/V ratio did not. In the case of the dentals, however, the opposite pattern of results was obtained. The C/V ratio reliably distinguished voicing, although vowel duration alone did not.

Although the voicing effect for the closure durations for the dentals and bilabials was statistically significant, we found considerable variation in the reliability of this interval as a correlate of voicing. Even when the main effect of voicing for the closure durations was statistically independent of any of the other variables manipulated in this study (specifically for the bilabials), we observed a number of instances in which the magnitude of the differences between the closure durations for voiced and voiceless stops was so small as to be of questionable utility in perception of voicing. In addition, no main effect of voicing for the closure durations was obtained for the velars.

Taken together, our results suggest that for the velar and bilabial word-final stops, vowel duration alone (or independent of closure duration) may serve as the most reliable durational correlate of voicing across various intrinsic vowel durations and local and sentential environments. The present results therefore tentatively suggest that the perception of voicing, when based on durational cues, is probably best accomplished by evaluation of vowel duration alone when the listener is confronted with the contextual conditioning of segmental durations found in continuous speech. Further research is necessary, however, to determine whether the C/V ratio or the vowel duration alone is the more reliable cue in perception of the voicing feature under the conditions examined in the present study.

In addition to these findings, we replicated a number of effects of intrinsic vowel duration and local and sentential contexts on the durations of vowels and post-vocalic stop closures. We confirmed the previous findings that phrase-final lengthening is primarily due to changes in vowel duration, although some small but significant lengthening of closure durations was observed. We found that local phonetic environment (the vowel /ə/ compared to the stop /t/) exerts considerable influence on the duration of the preceding post-vocalic stop and, in some instances, on the duration of the preceding vowel. In addition to these effects of local phonetic environment and sentence position, we found that, in general, the magnitude of the effects of local and sentential environment on C/V ratios and vowel durations is greatest for vowels of shorter inherent duration.

In summary, the results of this study demonstrate that the C/V ratio as a correlate of word-final voicing is highly dependent on intrinsic vowel duration and local and sentential environment. Most important, however, we have shown that vowel duration may be a more reliable correlate of voicing than the C/V ratio for velar and bilabials stops. On the other hand, for the dental stops we found that the opposite was true: The C/V ratio, and not vowel duration, proved to be the most reliable correlate of voicing. Our results also demonstrate that the C/V ratio does not overcome the confounding of segmental durations simultaneously cueing speaking rate and phonological voicing. Finally, we believe the present results emphasize the need for researchers to investigate the effects of local and sentential contexts in their search for possible durational correlates of phonetic contrasts. Although limiting the sentential contexts in which specific test words are produced may make the problems of measurement and variability more tractable, we believe that important questions regarding the roles of temporal cues in speech production must be studied with full cognizance of the myriad of phonetic and extra-phonetic influences on segmental durations that exists in continuous speech.

Footnotes

1. By "relational invariance" we mean that the C/V ratios for voiced stops will always be smaller than those for voiceless stops, regardless of variation in the absolute values of the ratios.

2. Port (1981a, 1981b) has pointed out, however, that the C/V ratio will serve as an invariant cue to voicing only when other, nontemporal cues are ambiguous (see Barry, 1979; Hogan and Rozsypal, 1980; Lisker, 1981; Wardrip-Fruin, 1982).

3. Port (Port and Dalby, 1982) does not believe the C/V ratio is a likely cue to voicing for syllable-final dental stops because of the tendency for flapping in intervocalic positions.

4. We use the term "local rate" throughout this paper to refer to the effect of phrase-final lengthening in order to be consistent with current terminology (cf. Miller, 1981). However, we are not entirely certain that phrase-final lengthening is a rate-dependent effect per se or simply a syntactically determined adjustment of vowel duration at phrase boundaries.

5. Vowel duration and closure duration here are strictly defined in terms of the acoustic criteria established for measurement.

References

Barry, W. (1979). "Complex encoding in word-final voiced and voiceless stops," J. Phonetica 36, 361-372.

Cooper, W. E. (1975). "Syntactic control of speech timing," Ph. D. thesis (MIT) (unpublished).

Cooper, W. E., and Paccia-Cooper, J. (1980). Syntax and Speech (Harvard University Press, Cambridge, Mass.).

Crystal, T. H., and House, A. S. (1982). "Segmental durations in connected speech signals: Preliminary results," J. Acoust. Soc. Am. 72, 705-716.

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

Fitch, H. L. (1981). "Distinguishing temporal information for speaking rate from temporal information for intervocalic stop consonant voicing," Haskins Laboratories Status Report on Speech Research SR-65, 1-32.

Hogan, J. T. and Rozsypal, A. J. (1980). "Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant," J. Acoust. Soc. Am. 67, 1764-1771.

House, A. S. (1961). "On vowel duration," J. Acoust. Soc. Am. 33, 1174-1178.

House, A. S., and Fairbanks, G. (1953). "The influence of consonantal environment upon the secondary acoustical characteristics of vowels," J. Acoust. Soc. Am. 25, 105-113.

Hyman, L. (1975). Phonology, Theory and Analysis (Holt, Rinehart, and Winston, New York).

Klatt, D. K. (1973). "Interaction between two factors that influence vowel duration," J. Acoust. Soc. Am 54, 1102-1104.

Klatt, D. K. (1975). "Vowel lengthening is syntactically determined in a connected discourse," J. Phonetics 3, 129-140.

Klatt, D. K. (1976). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. 59, 1208-1221.

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

Lisker, L. (1957). "Closure duration and the intervocalic voiced-voiceless distinction in English," Language 33, 42-49.

Lisker, L. (1978). "Rapid vs. Rabid: A catalogue of acoustic features that may cue the distinction," Haskins Laboratories Status Report on Speech Research SR-54, 127-132.

Lisker, L. (1981). "On generalizing the rabid-rapid distinction based on silent gap duration," Haskins Status Report on Speech Research SR-65, 251-259.

Luce, P. A., and Carrell, T. D. (1981). "Creating and editing waveforms using WAVES," Research on Speech Perception: Progress Report No. 7, Indiana University, 287-297.

Malecot, A. (1970). "The lenis-fortis opposition: Its physiological parameters," J. Acoust. Soc. Am. 47, 1588-1592.

Massaro, D. W., and Cohen, M. M. (1983). "Consonant/vowel ratio: An improbable cue in speech," Percpt. Psychophys. 33, 501-505.

Miller, J. L. (1981). "Effects of speaking rate on segmental distinctions," in Perspectives on the Study of Speech, edited by P. D. Eimas and J. L. Miller (Erlbaum, Hillsdale, N. J.).

Miller, J. L., and Grosjean, F. (1981). "How the components of speaking rate influence perception of phonetic segments," Journal of Experimental Psychology: Human Perception and Performance 7, 208-215.

Oller, D. K. (1973). "The effect of position in utterance on speech segment duration in English," J. Acoust. Soc. Am. 54, 1235-1247.

Peterson, G., and Lehiste, I. (1960). "Duration of syllabic nuclei in English," J. Acoust. Soc. Am. 32, 693-703.

Pickett, J. M., and Decker, L. R. "Time factors in perception of a double consonant," Language and Speech 3, 11-17.

Port, R. F. (1977). The Influence of Speaking Tempo on the Duration of Stressed Vowel and Medial Stop in English Trochee Words (Indiana University, Linguistics Club, Bloomington).

Port, R. F. (1978). "Effects of word-internal versus word-external tempo on the voicing boundary for medial stop closure," Haskins Laboratories Status Report on Speech Research SR-55/56, 189-198.

Port, R. F. (1979). "Influence of tempo on stop closure duration as a cue for voicing and place," J. Phonetics 7, 45-56.

Port, R. F. (1981a). "Linguistic timing factors in combination," J. Acoust. Soc. Am. 69, 262-274.

Port, R. F. (1981b). "On the structure of the phonetic space with special reference to speech timing," Lingua 55, 181-219.

Port, R. F., and Dalby, J. (1982). "Consonant/vowel ratio as a cue for voicing in English," Percept. Psychophys. 32, 141-152.

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

Raphael, L. J., and Dorman, M. F. (1980). "Silence as a cue to the perception of syllable-initial and syllable-final stop consonants," J. Phonetics 8, 269-275.

Sharf, D. J. (1964). "Duration of post-stress intervocalic stops and preceding vowels," Language and Speech 5, 26-30.

Umeda, N. (1977). "Consonant duration in American English," J. Acoust. Soc. Am. 61, 846-858.

Wardrip-Fruin, C. (1982). "On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue to voicing in final stop consonants," J. Acoust. Soc. Am. 71, 187-195.

Some Effects of Training on the Perception of Synthetic Speech*

Eileen C. Schwab, Howard C. Nusbaum, and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

*This research was supported by NINCDS research grant NS- 12179, and contract F33615-83-K-0501 with the Air Force Office of Scientific Research, Air Force System Command. We thank Beth Greene, Mike Dedina, and Pat Butler for their help and assistance in various phases of this project.

Abstract

An experiment was conducted to determine the effects of training on the perception of synthetic speech. To separate the effects of task familiarity due to training from specific improvements in perception, three groups of subjects received different types of training experience. All three groups were tested on Day 1 (pre-test) and Day 10 (post-test) of the experiment with synthetic speech. Training was conducted on Days 2-9 of the experiment. One group was trained with the synthetic speech. A second group went through the identical training procedures but heard natural speech. The third group received no training. Stimulus materials consisted of isolated words, normal English sentences, syntactically correct but semantically anomalous sentences, and continuous prose passages. Pre-test performance was not significantly different for any of the three groups. However, the groups differed significantly on the post- test measures of performance with synthetic speech: the group trained with synthetic speech performed much better than the other two groups. A six-month follow-up indicated that the group trained with synthetic speech displayed long-term retention of the knowledge and experience gained with prior exposure to synthetic speech. The results have implications for the design and implementation of voice output systems using synthetic speech.

Some Effects of Training on the Perception of Synthetic Speech

In the next few years we can expect to see an increase in the use of voice response systems in a variety of consumer, military, and industrial applications. Many of these systems will use synthetically produced speech that is generated by rule from a text-to-speech system. Despite the potential applications for this technology, several recent studies have demonstrated that synthetic speech, even high-quality synthetic speech, is more difficult to perceive and understand than natural speech (Pisoni, 1982; Pisoni & Hunnicutt, 1980). The differences in perception between natural and synthetic speech could be a major obstacle to the widespread use of synthetic speech in voice- output systems, particularly under high-information-load conditions such as in the cockpit of an aircraft or a command- control application. To overcome this problem, it is necessary to improve the intelligibility of synthetic speech as much as possible, so that performance with synthetic speech approximates performance with natural speech as closely as possible. There are two possible ways to achieve this goal. One way to improve intelligibility is to increase the quality and naturalness of the synthetic speech, so that it contains more of the critical acoustic cues and redundancy of natural speech. A number of efforts are currently being devoted to improving synthesis via greater attention to the acoustic-phonetic detail and effects of context and prosody on perception (e.g., Allen, 1981). Unfortunately, the results of these efforts may not be incorporated into commercially available products for quite some time. The second possibility is to train the listeners to perceive synthetic speech more accurately. Since humans are very flexible and are capable of learning new ways of processing information, perceptual training and exposure could change the strategies humans use in perceiving synthetic speech. Thus, it might be possible to improve intelligibility, and hence the usability of this new technology through short-term selective exposure and training.

Indeed, several researchers have reported a rapid improvement in the recognition of synthetic speech during the course of their experiments. Carlson, Granstrom, and Larsson (1976) studied the intelligibility of synthetic speech (produced from text) for lists of sentences and found large improvements in performance throughout their experiment. Subjects increased their performance from approximately 55% correct word identification at the beginning, to approximately 90% correct word identification by the end of the experiment. In a different study, Pisoni and Hunnicutt (1980) also reported consistent improvements in the perception of synthetically produced words in sentence contexts and in the comprehension of synthetic prose passages. These results suggest that exposure may improve the intelligibility of synthetic speech.

However, it is also possible that the observed improvements in performance were due to an increased mastery of the experimental procedures and tasks rather than any changes in the intelligibility and recognition of the synthetic speech. For example, Pisoni (1981) used a lexical decision task with natural and synthetic speech and found that performance (as measured by response latencies) improved for both types of stimuli within a one-hour session. In another experiment, subjects performed in a lexical decision task for five days with natural and synthetic speech (Slowiaczek and Pisoni, 1982). Again, performance improved for both synthetic and natural speech. Thus, it is possible that previously reported improvements in perception of synthetic speech were actually due to an increased familiarity with the experimental procedures rather than a change in the perceptual processing of synthetic speech. This paper reports the results of a study designed to separate these two factors.

The basic design of the experiment is shown in Table 1. A pre-test was conducted on Day 1 of the experiment. This pre-test determined baseline performance for perception of sreech generated by the Votrax Type-'N-Talk text-to-speech system. The Votrax system was chosen primarily because of the poor quality of its segmental (i.e., consonant and vowel) synthesis. Thus, ceiling effects could not obscure any effects of training. On Day 1, all subjects listened to synthetic speech without feedback. On Day 10, a post-test was conducted to assess the improvements, if any, in recognition performance for the synthetic speech. As in the pre-test, all subjects again listened to the output of the Votrax Type-'N-Talk and received no feedback.

Of the three groups of subjects shown in Table 1, groups 2 and 3 were the control groups. The third group (Control Group) controlled for the effects of exposure to synthetic speech on the pre-test. This group had minimal exposure to the synthetic speech and the experimental procedures since they only participated in the pre-test and the post-test. Any improvements in performance on Day 10 would be due entirely to their exposure to the synthetic speech on Day 1. The second group (Natural Group) controlled for the effects of familiarity with the training procedures. This group also was only exposed to synthetic speech on the Day 1 pre-test and the Day 10 post-test. However, this group did receive training between the pre-test and the post-test. During training, the Natural Group listened to natural speech and performed the same tasks and procedures as the first group. Any improvements in performance on Day 10 that were greater than the improvements obtained for the Control Group should be due to the practice and familiarity with the experimental procedures alone, since both of these groups had the same amount of experience with synthetic speech. The first group of subjects (Votrax Group) listened to synthetic speech throughout the experiment (on testing days and on training days). If this group showed better levels of performance on Day 10 than the group trained with natural speech, we would have evidence for a selective improvement in the perceptual processing of synthetic speech. On the other hand, if the group trained with Votrax-generated speech showed the same level of performance on Day 10 as the group trained with natural speech, then any improvements in performance for this group would in fact be due only to mastery of the experimental procedures.

Insert Table 1 about here

Method

Subjects

Thirty-one subjects participated in this experiment. All of the subjects were students at Indiana University and were paid for their participation. All subjects were right-handed, native speakers of English with no reported history of any speech or hearing disorder. Subjects were recruited from a paid-subject pool maintained by the laboratory. A simple audiometric screening test (see Walker, 1982 for a detailed description) was administered to all subjects. The results indicated none of the subjects had any gross hearing loss.

TABLE 1

Type of Speech Presented to Each Group During Various Phases of the Training Experiment

	Day 1	Days 2-9	Day 10
	TESTING	TRAINING	TESTING
	(Pre-Test)		(Post-Test)
GROUP			
1	Votrax	Votrax	Votrax
2	Votrax	Natural	Votrax
3	Votrax	------	Votrax

Materials

All speech stimuli were recorded on audio-tape. The natural speech stimuli were produced by a male talker. The synthetic speech stimuli were produced by the Votrax Type-'N-Talk system. All stimulus materials were sampled at 10 kHz, low-pass filtered at 4.8 kHz, digitized through a 12-bit A/D converter, and stored in digital form on disk on a PDP-11/34 computer. The stimuli were played out in real time at 10 kHz through a 12-bit D/A converter and low-pass filtered at 4.8 kHz. Five sets of stimulus materials were used in this experiment.

PB Lists. The first set of stimuli consisted of 12 lists of 50 monosyllabic words. These lists were a subset of 20 lists originally designed for testing natural speech intelligibility (American National Standards, 1960; Egan, 1948). The stimuli are "phonetically balanced" (PB) because the relative frequency of English phonemes is maintained in each list of 50 words. The items in each list consist of familiar words and were designed to have the same average (and same range of) difficulty (see Egan, 1948).

MRT Lists. The second set of stimuli consisted of four sets of 50 monosyllabic words taken from the Modified Rhyme Test (House, Williams, Hecker, & Kryter, 1965). This test was designed primarily to assess the intelligibility of consonants in natural speech. In the MRT, a closed-response format is used. Subjects are presented with one stimulus -- a spoken word -- and six response alternatives. The six alternatives are CVC words that share a medial vowel and either an initial or final consonant. When the final consonant is shared by the six alternative response words, the initial consonants are all different, and the responses form six different English words. Conversely, when the initial consonants are the same, the responses are six words that differ only in their final consonant.

Harvard Sentences. The third set of materials consisted of 10 lists of 10 normal English sentences (Egan, 1948; "IEEE Recommended Practice", 1969). These sentences contained five key words, plus a variable number of function words, arranged in a variety of syntactic structures. Of the five key words, four were monosyllabic and the fifth was bisyllabic. An example of this type of stimulus is "The empty flask stood on the tin tray."

Haskins Sentences. The fourth set of materials consisted of 10 lists of 10 syntactically normal, but semantically anomalous sentences (Nye & Gaitenby, 1974). These sentences contained high frequency, monosyllabic words. Each stimulus had four key words which were presented in the following syntactic structure: The (adjective) (noun) (verb, past tense) the (noun). The sentences were semantically anomalous in order to reduce word identification cues based on the meaning of the utterance. An example of this type of sentence is "The round blood grew the wind."

Prose Passages. Finally, the fifth set of materials consisted of 38 texts taken from a variety of sources including popular magazines and reading comprehension tests. Eight of the passages were used as training stories and were taken from the Nelson-Denny Reading Comprehension Test (1930). Another nine passages were taken from other adult reading tests (Brown, 1973; Derrick, Harris,

& Walker, 1960; Farr, 1972; Gardener, Callis, Merwin, & Madden, 1972). The remaining 21 passages were modified versions of texts which appeared in magazines such as Time and Scientific American. For the eight training stories, the original multiple-choice comprehension questions taken from the comprehension test were used. For the remaining 30 stories, yes/no comprehension questions were constrıcted. The questions were designed to examine information at different levels of text comprehension. These levels ranged from a relatively low level where subjects decided if a particular word occurred in a text, to a high level where subjects evaluated the theme of the text. In addition, several questions were designed to examine inferences that could be drawn from the passages but were not explicitly stated in the passage. Questions with correct "yes" and "no" responses were presented equally often.

Insert Table 2 about here

Procedure

All stimuli were presented to subjects in real time under computer control through matched and calibrated TDH-39 headphones at 76 dB SPL measured using a VTVM. (Signal amplitude was calibrated for an isolated word produced by the Votrax system.) Different stimuli were presented on each day of the experiment; that is, on each day subjects heard different isolated words, different sentences, and different prose passages. Thus, the subjects were continually being exposed to novel utterances. This manipulation was designed to prevent subjects from learning any peculiarities of specific tokens produced by the Votrax system. Therefore, any improvements in performance observed as a result of training with the synthetic speech would have to be due to learning the general characteristics of the acoustic-phonetic structure of the speech generated by the system -- that is, the rules themselves since the stimulus materials varied from day to day.

The Votrax and Natural Groups participated in 10 experimental sessions (a pre-test, a post-test, and eight training sessions) and the Control Group participated only in the pre-test and post-test sessions. Each experimental session lasted approximately one hour. Since the experiment was run during the business week, a two day break intervened between Days 5 and 6 of the experiment.

Testing. The entire experiment was conducted during a two- week period. All subjects were tested on the first (Day 1) and last (Day 10) day of the experiment using the same procedures and order of tasks (see Table 2). However, the stimulus materials used each day were different to avoid stimulus-specific learning. On testing days, no feedback was provided to the subjects about their performance. For each session, the first test always consisted of identifying isolated words. Subjects were presented with two fifty-item PB Lists, one word at a time. After hearing a word, subjects were required to transcribe it onto a response sheet using a pen. Subjects were asked to write down an English word on every trial and to guess if they were not sure. After writing their response,

TABLE 2

Tasks Administered During the Pre-Test (Day 1) and the Post-Test (Day 10)

	STIMULI	RESPONSE TYPE	TOTAL RESPONSE	DEPENDENT MEASURES
I.	2 PB Lists	Open Class	100	Exact Correct Word
II.	2 MRT Lists	6 AFC	100	Accuracy and RT
III.	1 Harvard List (10 Sentences)	Open Class	50 Content Words	Exact Correct Word
IV.	1 Haskins List (10 Sentences)	Open Class	40 Content Words	Exact Correct Word
V.	3 Prose Passages	Y/N Questions	30	Accuracy and RT

the subjects pushed a button on a computer-controlled response box to indicate they were ready for the next trial. When all subjects had responded, the next word was presented.

For the second task, subjects listened to isolated words from the Modified Rhyme Test (MRT Lists). After hearing a stimulus, subjects were presented with six alternative words centered on a CRT screen. The subjects were required to choose one of the six alternatives as the word they had heard by pushing one of six buttons on a response box in front of them. The subjects were instructed to respond as quickly and as accurately as possible. Response latency and accuracy were recorded by the computer and stored on disk for later analysis.

For the third test, subjects listened to and then transcribed each sentence in a list of 10 Harvard Sentences. After each sentence was presented, subjects wrote down what they heard. Subjects were encouraged to guess if they were not sure of the words in the utterance. After transcribing the sentence, subjects pushed a response button to indicate they were ready for the next sentence.

The fourth task consisted of listening to and transcribing each of the 10 Haskins Sentences. The subjects were informed that these sentences were "bizarre" or unusual because they did not make any sense. After each sentence was presented, subjects wrote down what they heard, guessing if necessary. After transcribing the sentence, subjects pushed a response button to indicate they were ready for the next sentence.

Finally, in the last test, subjects listened to several short prose passages. After hearing each passage, 10 comprehension questions were presented, one at a time. The subjects answered "yes" or "no" to each question by pressing the appropriately labeled button on a response box. The subjects were required to respond to each question as quickly and as accurately as possible. Both the accuracy and the latency of the responses were recorded by computer and stored on disk for later analysis.

As shown in Table 1, three groups of subjects participated in testing. The first group (the Votrax Group) listened to synthetic speech throughout the entire experiment. The second group (the Natural Group) received synthetic speech on Day 1 and Day 10, but listened to natural speech during training on Days 2- 9. The third group (the Control Group) participated only on Day 1 and Day 10 of the experiment. All groups listened to synthetic speech and received the same materials, procedures, and tasks on Day 1 and Day 10.

Insert Table 3 about here

Training. Although three groups of subjects participated in the experiment, only two of the groups actually received any training between the pre-test on Day 1 and the post-test on Day 10. The third group (the Control Group) was tested but received no training between the pre-test and the post-test. The two groups of subjects that received training went through the same identical procedures on

TABLE 3

Procedures Used During the Training Phase of the Experiment (Days 2-9)

I. PB WORDS (50 Items)

Stimulus	Response	Stimulus	Response
(auditory)	(written & button)	(visual & auditory)	(button) indicating accuracy

II. HARVARD AND HASKINS SENTENCES (10 Sentences)

Stimulus	Response	Stimulus	Response
(auditory)	(written & button)	(visual & auditory)	(button) indicating accuracy

III. PROSE PASSAGES (4 Passages)

Passage 1 (Training Passage)

Stimulus	Stimulus	Responses (4)
(visual, on paper)	(visual & auditory)	(5 alternative multiple choice, paper & pencil)

Passages 2-4 (Testing Passages)

Stimulus	Response
(auditory)	(button, accuracy & RT) (Yes-No Questions)

Day 2 through Day 9 of the experiment. Table 3 summarizes the procedures used to train subjects on each day. Even though the procedures were the same each day, the stimuli were different each day (as in testing) so subjects were constantly exposed to novel utterances.

During each session, subjects received 50 isolated PB words, 10 Harvard Sentences, 10 Haskins Sentences, and four prose passages. The basic training procedure was the same for the isolated words and the two types of sentences. On each trial, subjects were presented with an utterance (i.e., a word or sentence) and were required to transcribe it onto a response sheet. When they finished transcribing the stimulus, each subject pressed a button on a computer-controlled response box. When all the subjects in the group had responded, feedback was presented. This feedback consisted of a second auditory repetition of the stimulus item together with a visual (orthographic) presentation of the stimulus on a CRT display located in front of the subject. Associating the auditory stimulus with a visual presentation served two functions. First, it provided information about the accuracy of the subjects' responses. Second, it provided a basis for learning the peculiarities of the rule system used to generate the synthetic speech. After feedback, subjects scored their response by pressing one of two buttons (labeled "correct" or "incorrect") to indicate their accuracy in transcribing the stimulus.

Training with the prose passages used a slightly different procedure. Subjects were given a printed version of a passage that was read silently before and during the presentation of the spoken passage. After hearing the training passage, the printed text was removed and subjects answered multiple-choice comprehension questions. Each day, after this training passage, subjects were presented with three additional test passages. For these passages, subjects listened to the spoken text and then answered comprehension questions as they did during testing. However, they did not see a printed version of these stories nor did they receive feedback.

The only difference between the two groups of subjects during the training phase was the type of speech that was presented. For the Natural Group, the stimuli were naturally produced words, sentences, and passages. For the Votrax Group, the stimuli were produced by the Votrax Type-'N-Talk text-to- speech system. Thus, both groups heard the same words, sentences, and prose passages, and performed the same tasks. The only difference between the groups was whether they received natural or synthetic speech.

Results and Discussion

Five subjects did not complete the experimental sessions and their data were excluded from the statistical analyses. Of the remaining 26 subjects who did complete the experiment, ten were in the group that received no training, seven were in the group that was trained with natural speech (three dropouts), and nine were in the group that was trained with synthetic speech (two dropouts).

Insert Figure 1 about here

Pre-test and Post-test Data

PB Lists. Transcription accuracy for the isolated PB words was determined in the following way. To be scored as a correct response, subjects had to transcribe the exact word with no phonemes added or deleted. An assistant scored each transcription response. If the word was "flew" and the subject transcribed "few" or "flute," the response was scored as incorrect. If the subject wrote "flue", the response was scored as correct.

A three-way analysis of variance (Groups x Days x Lists) was conducted on the pre-test and post-test scores. A significant main effect of lists was obtained [$F(1,23) = 83.88$, $p < .0001$]. Performance on the second PB list on each day was higher than performance on the first PB list. A significant Groups x Days interaction was also obtained [$F(2,23) = 134.93$, $p < .0001$]. This two-way interaction can be seen in Figure 1 which shows the mean accuracy for the three groups of subjects (collapsed over the two PB Lists presented) for Day 1 and Day 10. A simple-effects analysis of this interaction showed that on the first day of the experiment, all three groups of subjects had equivalent levels of performance [$F(2,35) = .73$, $p > .45$] -- subjects correctly identified approximately 25% of the stimuli. On Day 10 the groups showed very different levels of performance [$F(1,35) = 122.11$, $p < .0001$]. A Newman-Keuls analysis revealed that the Votrax Group performed better than either the Control or Natural Group ($p < .05$). No difference was found in performance between the Control and Natural Groups on Day 10 ($p > .05$). Simple-effects analyses also indicated that the Votrax Group showed a significant improvement from Day 1 to Day 10 [$F(1,23) = 705.55$, $p < .0001$]. Performance improved from approximately 25% correct identification on Day 1 to approximately 70% correct identification. The Control and Natural Groups also had a significant improvement from approximately 25% correct identification on Day 1 to approximately 35% correct word identification on Day 10 [$F(1,23) = 27.99$, $p < .0001$ and $F(1,23) = 56.39$, $p < .0001$, respectively].

Insert Figure 2 about here

MRT Lists. The pre-test and post-test performance scores for the Modified Rhyme Test are shown in Figure 2. Day 1 accuracy on the MRT was much higher than the PB words (25% versus 61%), a result that is due primarily to the limited response set size in the MRT test (i.e., six alternatives versus an open response set). A three-way analysis of variance revealed a significant two-way Days x Groups interaction [$F(2,23) = 12.89$, $p < .0005$]. Examination of this interaction

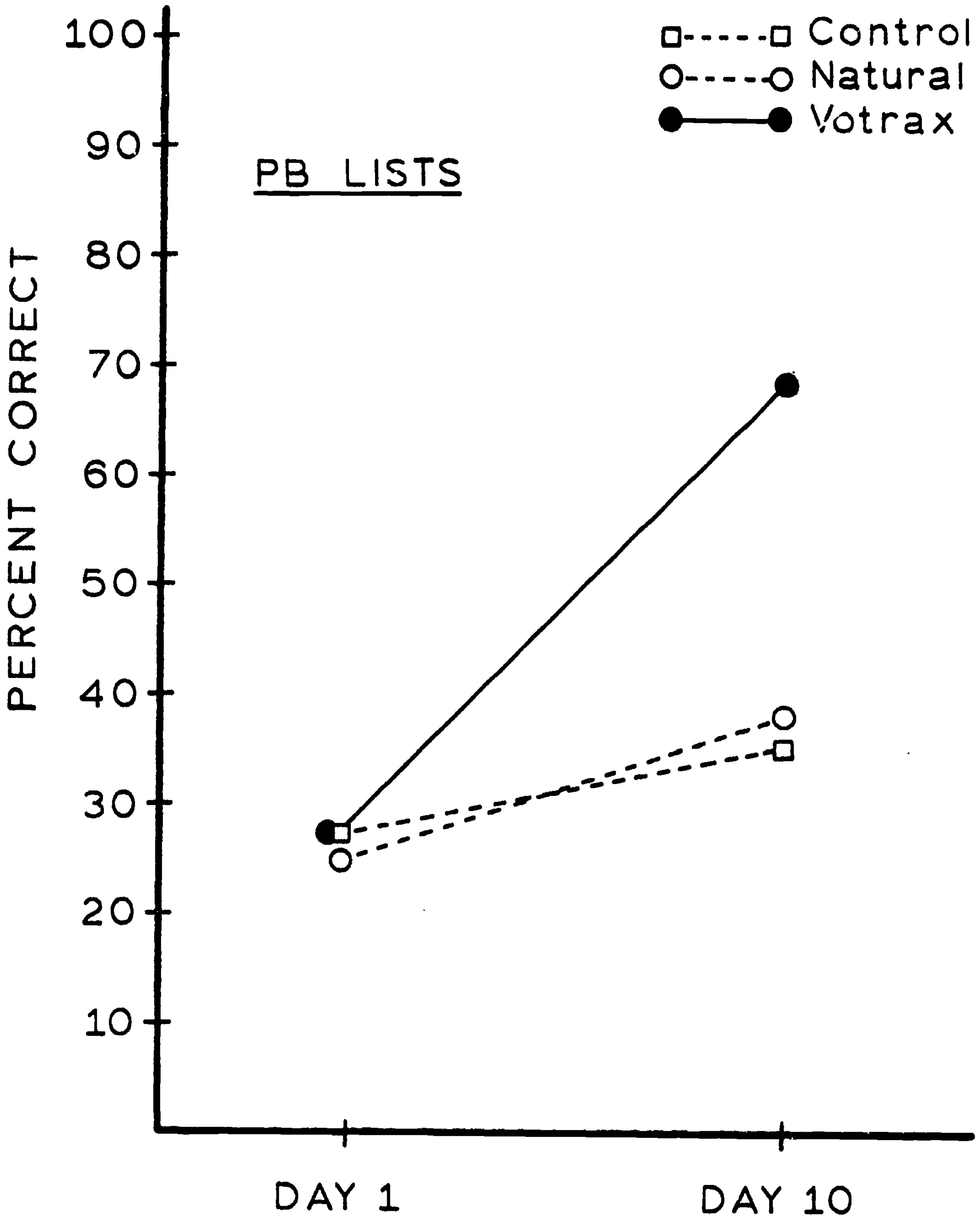


Figure 1. Mean accuracy for Day 1 (pre-test) and Day 10 (post- test) for transcribing isolated synthetic words (PB Lists).

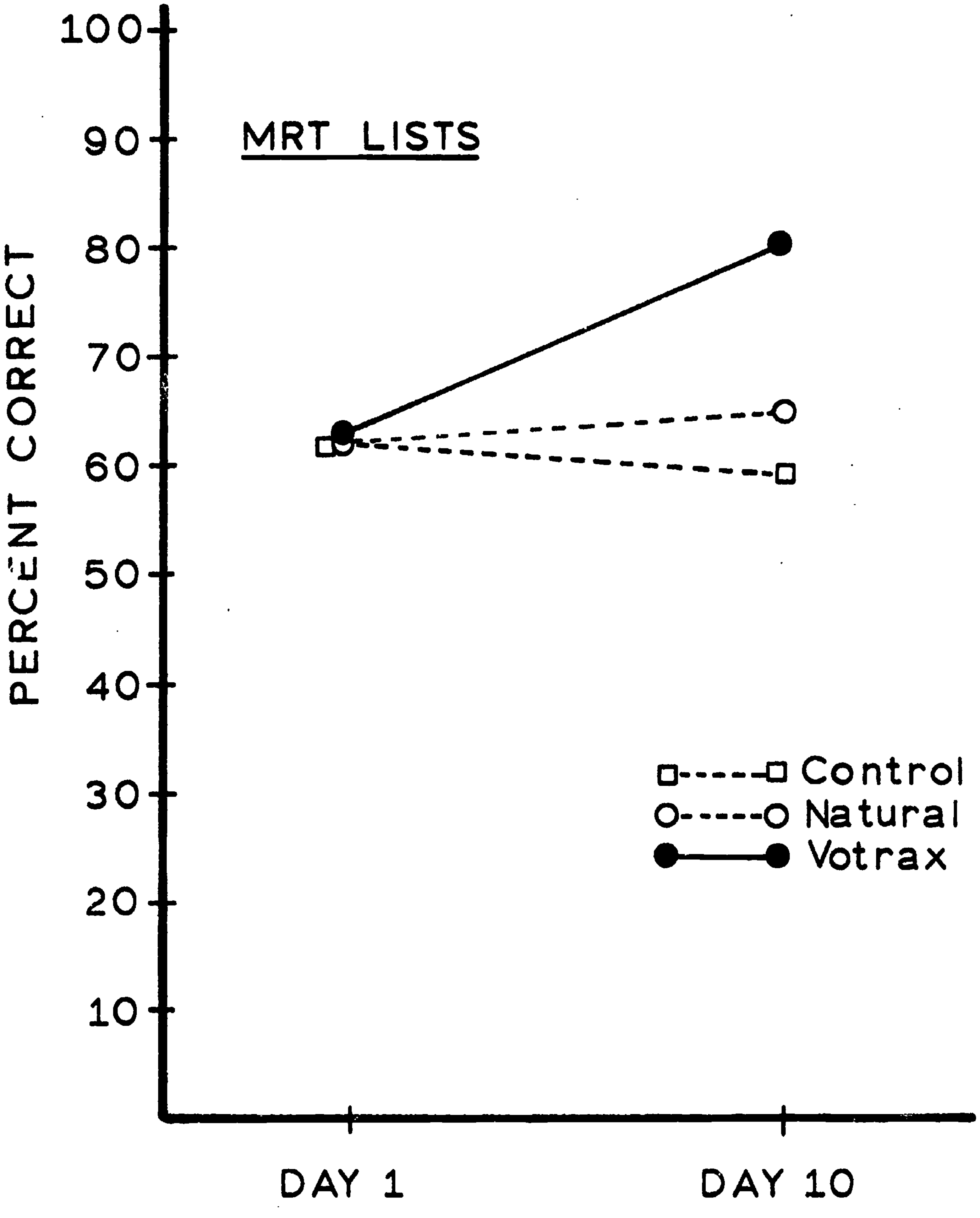


Figure 2. Mean accuracy for Day 1 (pre-test) and Day 10 (post- test) for the Modified Rhyme Task.

indicated that all three groups had equivalent levels of accuracy on Day 1, at approximately 60% correct word identification [$F(2,46) = .07$, $p > .9$]. However, on Day 10, the groups differed significantly [$F(2,46) = 28.9$, $p < .0001$]. A Newman-Keuls analysis indicated that the Votrax Group performed better ($p < .05$) than both the Control and Natural Groups which were not different from each other. A simple-effects analysis also indicated that, on Day 10, the Votrax Group had increased their performance from 63% to 80% correct word identification [$F(1,23) = 34.65$, $p < .0001$]. Performance for the Control and Natural Groups did not change from Day 1 to Day 10 [$F(1,23) = 1.05$, $p > .3$ and $F(1,23) = .70$, $p > .4$, respectively].

Insert Figure 3 about here

Figure 3 shows the mean response latencies for the three groups of subjects for the MRT stimuli on Days 1 and 10. The data in Figure 3 are collapsed over the two lists presented each day. A three-way analysis of variance on the response latencies from the MRT test revealed a significant Days x Groups interaction [$F(2,23) = 3.69$, $p < .05$]. Examination of this interaction indicated that the three groups of subjects did not have different latencies on Day 1 [$F(2,33) = .20$, $p > .8$]. However, on Day 10 the three groups showed significantly different mean latencies [$F(2,33) = 3.37$, $p < .05$]. Post-hoc analyses indicated that between Day 1 and Day 10, neither the Control Group nor the Natural Group showed any significant change in response latency [$F(1,23) = 1.14$, $p > .25$ and $F(1,23) = 3.4$, $p = .078$, respectively]. However, the Votrax Group showed a faster response latency on Day 10 relative to Day 1 [$F(1,23) = 22.03$, $p < .0002$]. Thus, by Day 10, the Votrax Group was not only more accurate in identifying the isolated words in the MRT, but they also responded faster than the other two groups.

Insert Figure 4 about here

Harvard Sentences. Each test sentence contained five key words that were scored for accuracy. In order for a word to be scored as correct, subjects had to transcribe the exact word with no additions or deletions. Figure 4 displays the mean percent correct word identification for words in the Harvard Sentences. A two-way analysis of variance (Groups x Days) revealed a significant two-way interaction [$F(2,23) = 39.47$, $p < .0001$] that can be seen in Figure 4. Post-hoc analysis of this interaction indicated that, on Day 1, all three groups had equivalent levels of performance, at approximately 40% correct word identification [$F(2,35) = 1.53$, $p > .2$]. By Day 10, however, a significant difference in performance was observed [$F(2,35) = 44.21$, $p < .0001$]. A Newman-Keuls analysis indicated that the Votrax Group performed better on Day 10 than either the Control ($p < .05$) or Natural ($p < .05$) Groups while the Control and Natural Groups did not differ from each other. Post-hoc analyses also revealed

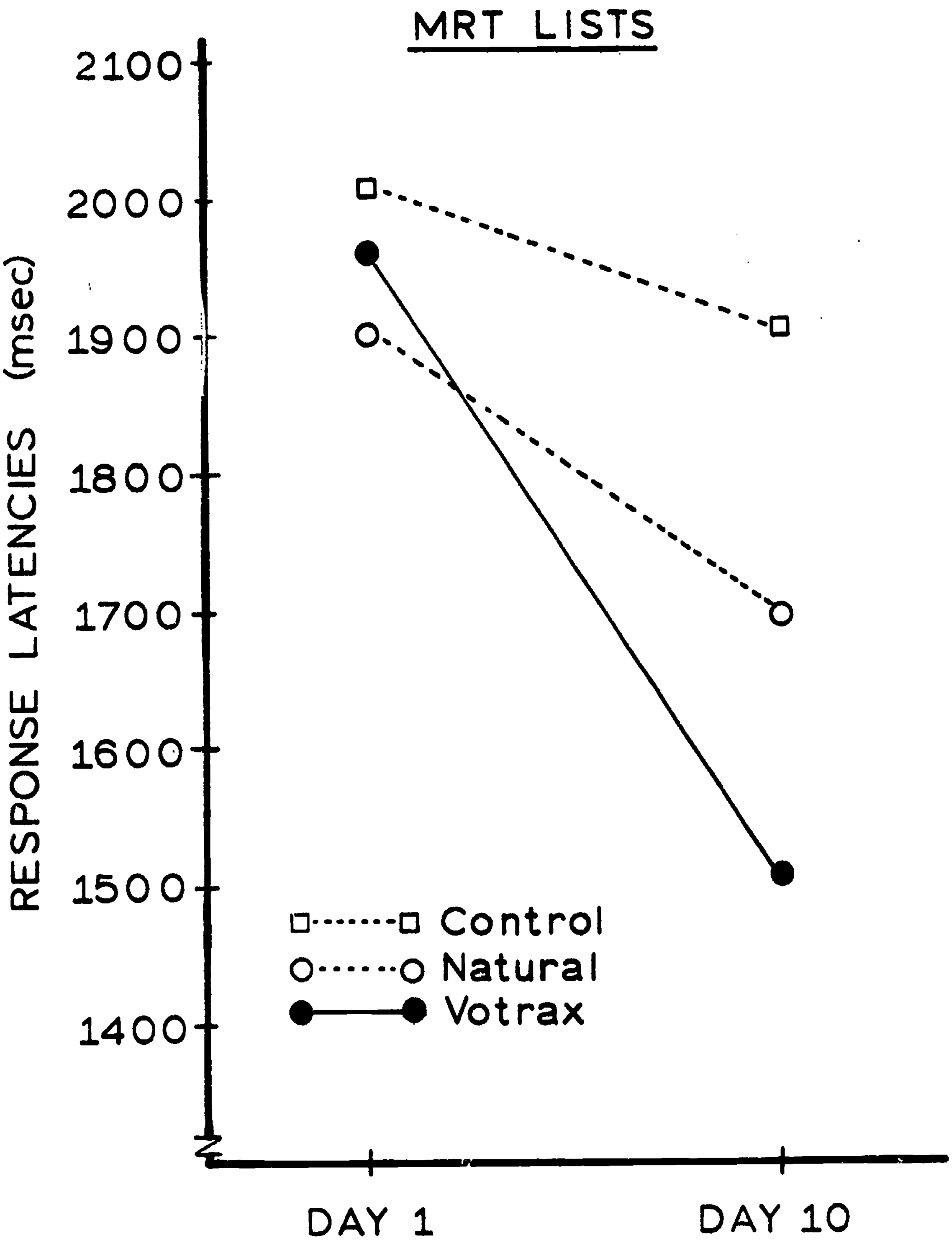


Figure 3. Mean response latencies for Day 1 (pre-test) and Day 10 (post-test) for identifying synthetic stimuli in the Modified Rhyme Task.

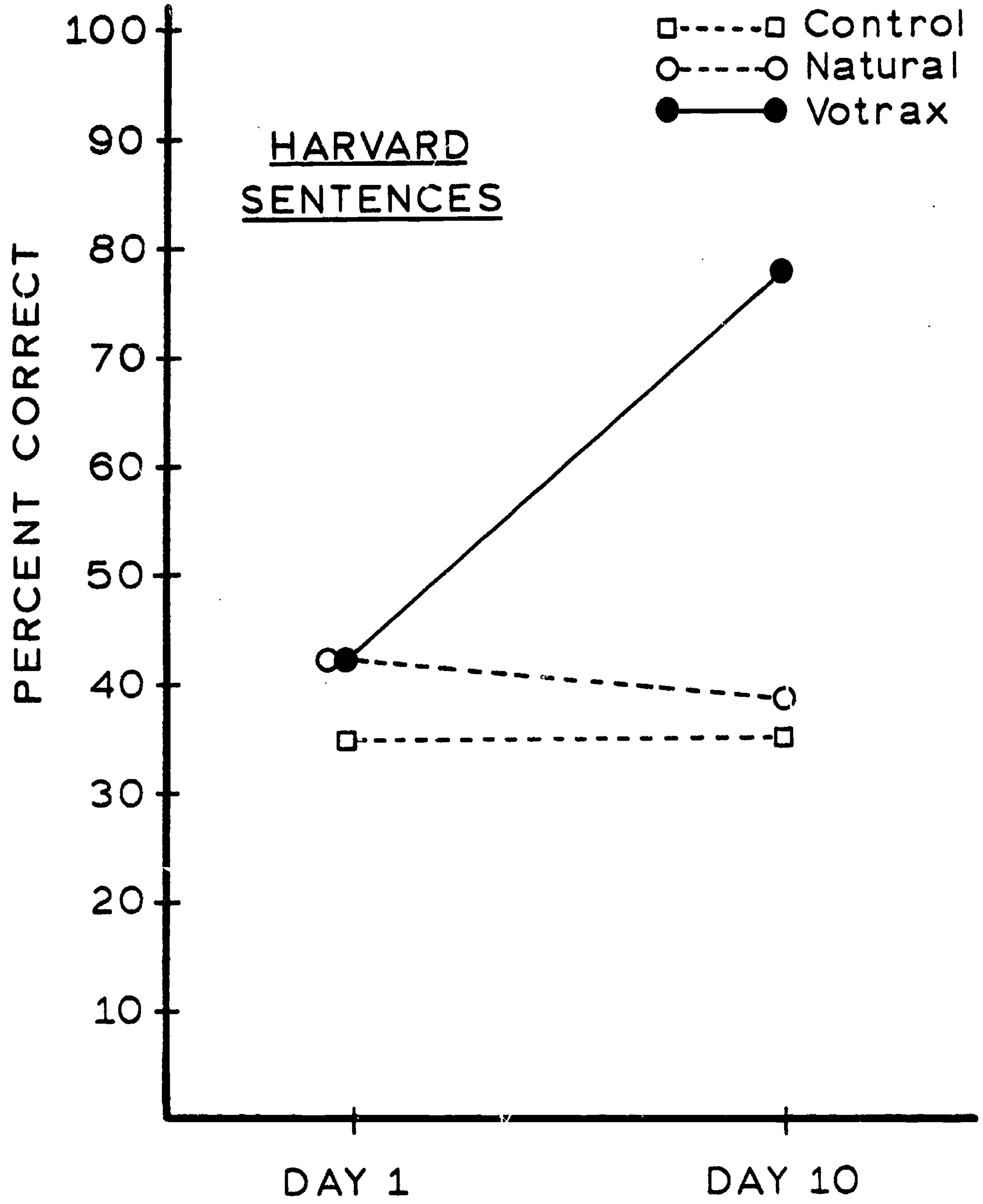


Figure 4. Mean correct synthetic word identification for Day 1 (pre-test) and Day 10 (post-test) for normal English sentences.

that the Votrax Group improved in performance from Day 1 to Day 10, from 42% correct to 78% correct word identification [*F*(1,23) = 55.54, *p* <.0001]. By contrast, the performance of the Control and Natural Groups did not change from Day 1 to Day 10 [*F*(1,23) < 1.00, *p* >.9 and *F*(1,23) = .40, *p* >.5, respectively].

Insert Figure 5 about here

Haskins Sentences. Each anomalous sentence contained four key words that were scored for accuracy. Figure 5 shows the mean percent correct word identification for the three groups of subjects. A two-way analysis of variance (Groups x Days) revealed a significant two-way interaction [*F*(2,23) = 14.32, *p* <.0005]. A post-hoc analysis of the Groups x Days interaction showed that on Day 1, all three groups obtained equivalent levels of performance at approximately 25% correct word identification [*F*(2,45) = 1.16, *p* >.3]. In contrast, a difference in accuracy for the three groups was observed on Day 10 [*F*(2,45) = 32.26, *p* <.0001]. A Newman-Keuls analysis indicated that the Votrax Group performed better than the Control (*p* <.05) and Natural (*p* <.05) Groups. As before, Day 10 performance for the Control and Natural Groups was not different. However, post-hoc analyses of the scores on Day 1 and Day 10 also indicated that on Day 10, all three groups showed improved word identification, relative to Day 1: Control [*F*(1,23) = 25.35, *p* <.0001], Natural [*F*(1,23) = 24.91, *p* <.0001], and Votrax [*F*(1,23) = 137.27, *p* <.0001]. Thus, all groups displayed some improvement as a result of exposure to synthetic speech on Day 1. However, this improvement was greatest for the subjects in the Votrax Group that received synthetic speech training.

Insert Figure 6 about here

Prose Passages. The mean accuracy for the comprehension questions is shown in Figure 6. A two-way analysis of variance was conducted on the mean percentage of correct responses to the comprehension questions. A significant effect of days was observed [*F*(1,23) = [illegible].65, *p* <.05]. All subjects were more accurate on Day 1 than on Day 10. No significant effect of groups was observed [*F*(2,23) = 1.60, *p* >.2] and the Groups x Days interaction was not significant [*F*(2,23) = .76, *p* >.4]. A two-way analysis of variance was also conducted on the response latencies and no effect of groups was observed [*F*(2,23) = 1.39, *p* >.25]. The Groups x Days interaction was only marginally significant [*F*(2,23) = 2.93, *p* = .073]. However, a significant effect of days was observed [*F*(1,23) = 4.53, *p* <.05], indicating subjects were faster to respond on Day 10 than on Day 1. An examination of the data demonstrated that the effect of days was due primarily to the Votrax Group. A simple-effects test showed a significant decrease in latencies for the Votrax Group [*F*(1,23) = 9.63, *p* <.001], but no significant change in latency for either the Control Group [*F*(1,23) = .04, *p* >.8] or the Natural Group [*F*(1,23) = .61, *p* >.4]. Thus, some improvement in performance was observed for the Votrax Group in responding to comprehension questions.

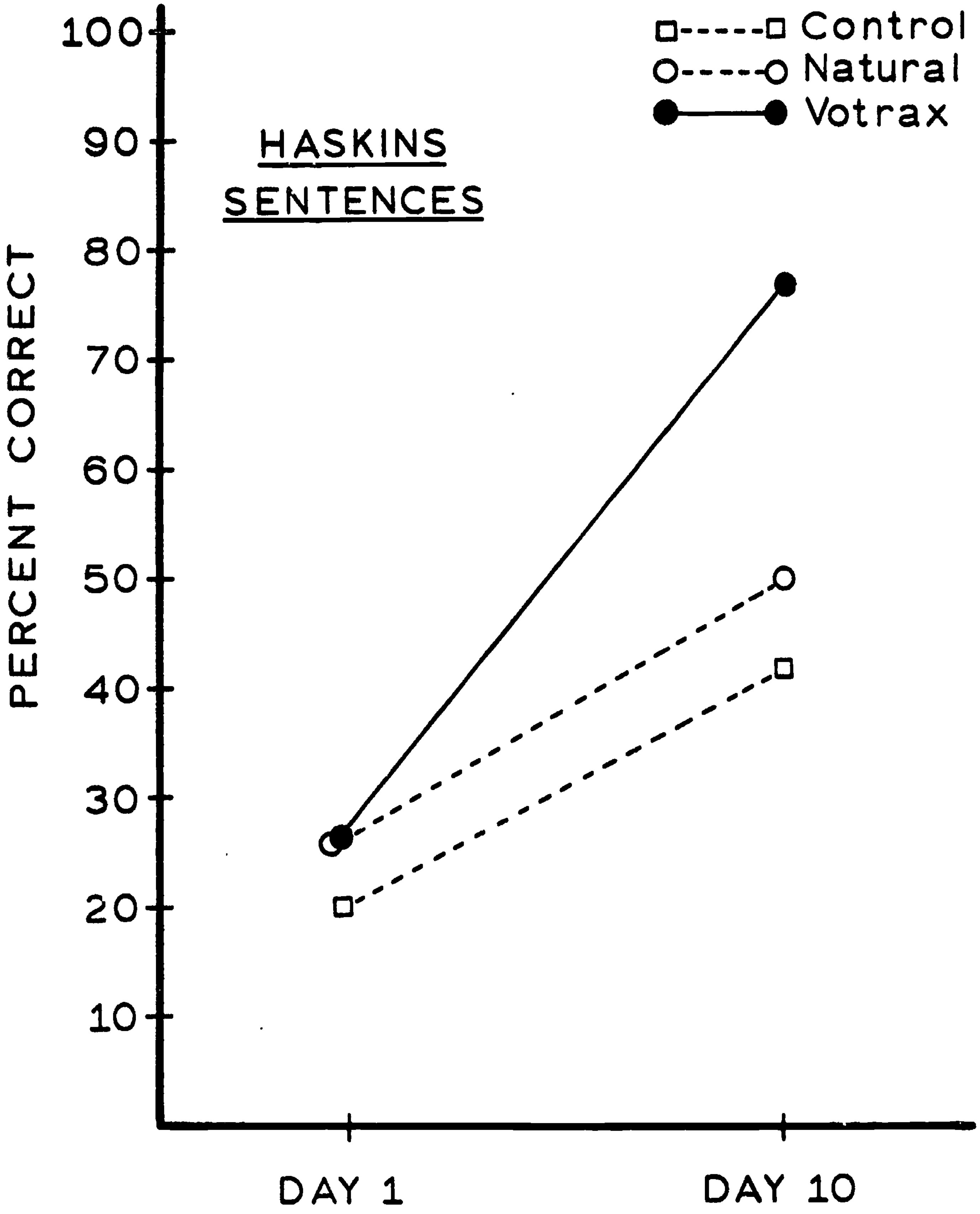


Figure 5. Mean correct synthetic word identification on Day 1 (pre-test) and Day 10 (post-test) for syntactically correct but semantically anomalous sentences.

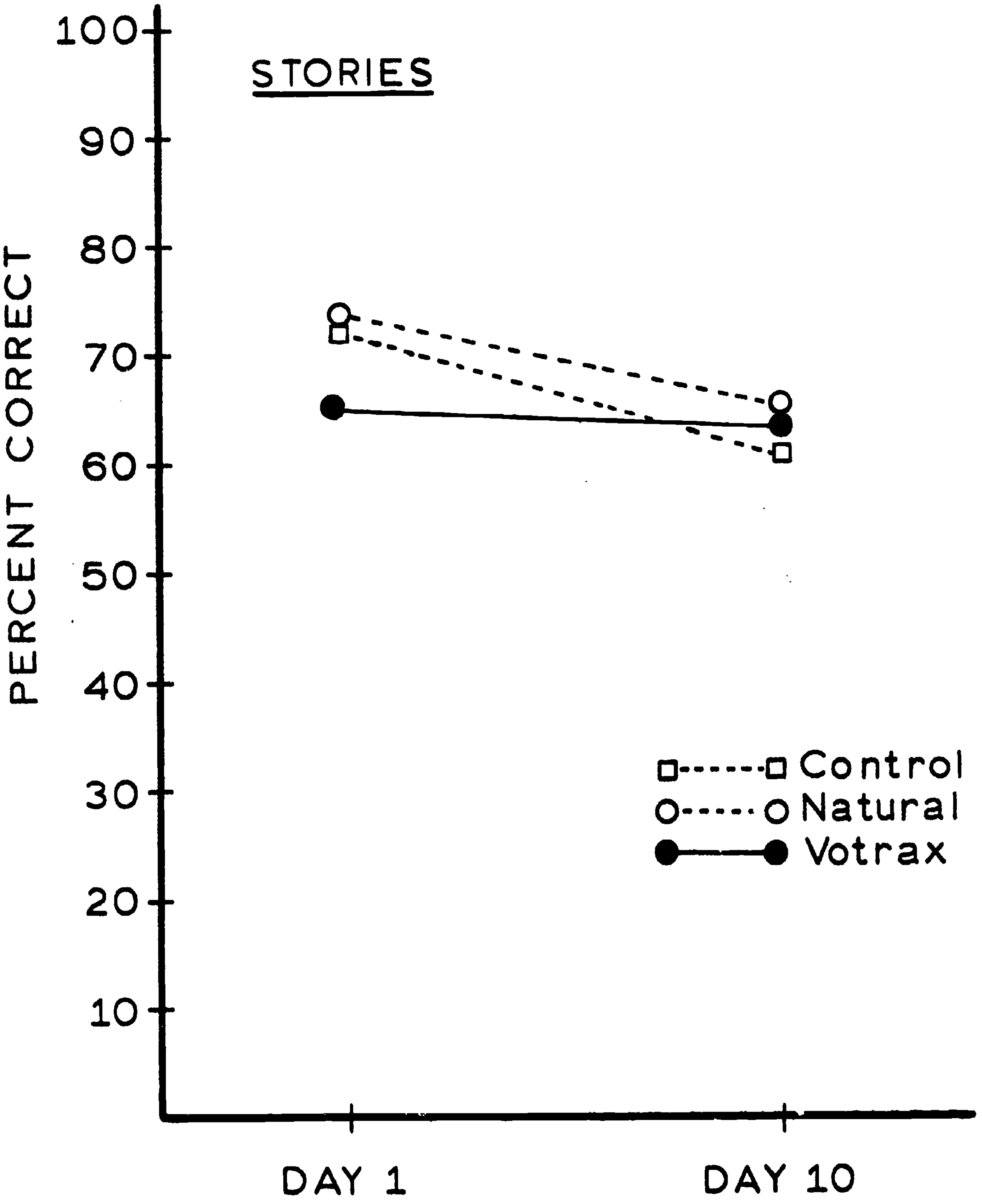


Figure 6. Mean correct responses for comprehension questions presented on Day 1 (pre-test) and Day 10 (post-test).

Summary. With the exception of the data for the prose passages, the results from the remaining tasks demonstrate that specific training with synthetic speech produces substantial improvements in performance. Moreover, the experimental design revealed that this improvement is primarily a function of specific training with and exposure to the synthetic speech rather than simple familiarity with the experimental procedures and tasks. After equivalent amounts of training to the Votrax Group, the Natural Group displayed performance levels that were comparable to the performance levels of the Control Group that had no training. Thus, experience with the experimental procedures did not affect performance. When performance was examined on Day 1 (i.e., the pre-test), no significant differences were obtained among the three groups of subjects. However, by Day 10, the Votrax Group performed better than either the Control or Natural Groups when identifying isolated words and words in sentences. The Votrax Group showed a significant improvement in performance from Day 1 to Day 10 as a function of training with synthetic speech stimuli. Their increase in performance was due to changes in the perceptual processing of the stimulus input.

These general results can be separated into two different patterns based upon task differences. One pattern of results was obtained with the MRT and Harvard Sentences and a slightly different pattern was obtained with the PB Lists and the Haskins Sentences. With the MRT and the Harvard Sentences, only the Votrax Group displayed any improvement in performance from Day 1 to Day 10. No change in performance for either the Control or Natural Groups was observed. This pattern of results may have been due to the fact that both the MRT and Harvard Sentences are highly constrained tasks. The MRT uses a closed response set, with only six alternatives. In addition, it has been shown that performance with this task is quite stable and generally does not change as a function of practice (House et al., 1965). The fact that the Votrax Group improved with this task indicates the subjects had changed the way they encoded the synthetic speech.

A similar pattern was observed with the Harvard Sentences which constrain word recognition to only a few plausible alternatives. Normal semantic and syntactic constraints substantially reduce the number of words that are possible at any point in the sentence (cf. Grosjean, 1980; Salasoo & Pisoni, 1982). For example, if a subject hears a word in a sentence such as "The funny clown made them laugh," that could not be identified, relatively few plausible substitutions for the word "laugh" exist. Thus, it is possible that subjects adopted a recognition strategy that exploited the inherent task constraints in the MRT and Harvard Sentences.

However, the PB Lists and the Haskins Sentences are much harder tasks due to the lack of constraints on either the response set or the set of plausible word candidates. Thus, a different pattern of results was obtained for these stimuli. With these stimuli, all groups improved their performance from Day 1 to Day 10. However, the Votrax Group performed better than either the Control or Natural Groups on Day 10. It is entirely possible that, given the reduced constraints on these tasks, these measures were more sensitive to the improvements in synthetic speech recognition. The improvements shown on these tasks by the Natural and Control Groups may have resulted from the single exposure to synthetic speech on the pre-test. Since the Natural Group received eight additional days of practice with the tasks, the Natural Group should have been better than the Control Group if the improvement were due to mastery of the procedures. Therefore, the lack of a difference between the Natural and Control Groups suggests that the slight improvement in performance shown by these groups may be attributable to some learning of the structure of the synthetic speech that occurred on the first day of the experiment -- the pre-test. It should be emphasized, however, that even though all groups improved on these two tasks, the Votrax Group showed the greatest improvement. Thus, the extra exposure and training with synthetic speech improved performance with synthetic speech.

The one set of stimuli for which subjects did not show a change in performance was the prose passages. For these materials, no increase in performance from the pre-test to the post-test was obtained for any group of subjects. All three groups had equivalent levels of accuracy for Day 1 and Day 10. In fact, accuracy on Day 10 was slightly worse than on Day 1. Analysis of the data indicated that the variability in performance was more a function of the particular passages heard and the particular test questions than the nature of the signal. Because of the lack of sensitivity of this measure to differences between natural and synthetic speech, a more detailed analysis of the results was not conducted.

Training Data

All training data were averaged over two-day periods in order to reduce variability due to the small number of subjects in each group. Since the Control Group received no training, the results are presented only for the Votrax and Natural Groups. It is important to reemphasize that these groups received exactly the same training procedures with the same words, sentences, and passages. The only difference was that the Votrax Group listened to synthetic speech while the Natural Group listened to natural speech produced by a male talker. As a result, performance for the Natural Group should be consistently higher than performance for the Votrax Group throughout training.

Insert Figure 7 about here

PB Lists. The mean accuracy for the transcription of isolated PB words for the two groups of subjects that received training on Days 2-9 is shown in Figure 7. Not surprisingly, the Natural Group showed ceiling levels of performance in recognizing isolated words throughout the course of training. These data contrast sharply with the performance shown by the Votrax Group. Subjects in this group showed a consistent increase in their performance over the course of training. A two-way analysis of variance (Groups x Days) was conducted on the accuracy scores. The different patterns of performance for the two groups produced a significant Groups x Days interaction [$F(3,42) = 74.90$, $p < .0001$]. As expected, post-hoc analysis of this interaction showed no changes in performance for the Natural Group [$F(3,42) = 1.76$, $p > .10$] over the course of the training. However, significant changes in accuracy were found for the Votrax Group [$F(3,42) = 210.67$, $p < .0001$]. A Newman-Keuls analysis indicated that consistent improvements were obtained by the Votrax Group for each successive two-day period

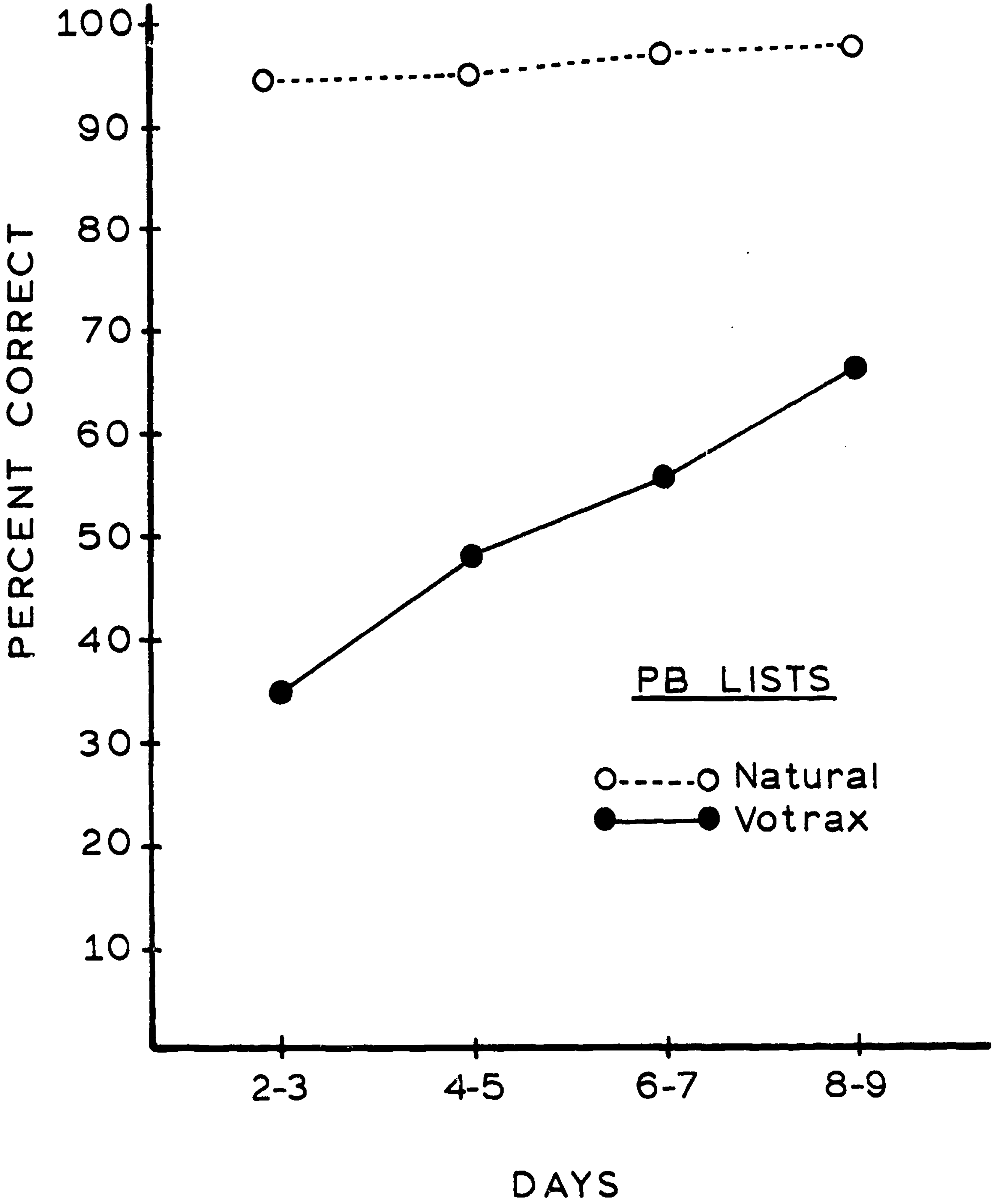


Figure 7. Mean accuracy for natural and synthetic speech on the training days of the experiment for the PB Lists.

(p <.05) over the course of training. Performance on the synthetic words improved from 35% correct at the beginning of training to 67% correct at the end of training.

Insert Figure 8 about here

Harvard Sentences. Figure 8 shows the mean percent correct identification of key words in meaningful sentences. Again, the Natural Group was at ceiling throughout training whereas the Votrax Group showed a large and consistent increase in performance during training. A two-way analysis of variance demonstrated that the pattern of data produced a significant Groups x Days interaction [$F(3,42) = 55.66$, $p < .0001$]. As expected, post-hoc analysis indicated no changes in performance for the Natural Group [$F(3,42) = .06$, $p > .9$]. However, a significant improvement in performance for the Votrax Group was obtained [$F(3,42) = 125.76$, $p < .0001$]. A Newman-Keuls analysis indicated that the improvement was significant for each two-day period during the course of training (p <.05). The subjects in the Votrax Group improved in their performance from 60% correct at the start of training to 87% correct at the end.

Insert Figure 9 about here

Haskins Sentences. Figure 9 displays the mean percent correct identification for the key words in the Haskins Sentences. As before, the Natural Group was at ceiling throughout training while the Votrax Group showed a large improvement in accuracy during training. Again, this pattern of results produced a significant Groups x Days interaction [$F(3,42) = 25.81$, $p < .0001$]. A post-hoc analysis indicated no change in the performance of the Natural Group during the course of training with natural speech [$F(3,42) = 0.1621$, $p > .9$]. However, an improvement in accuracy for the Votrax Group was observed [$F(3,42) = 64.56$, $p < .0001$]. A Newman-Keuls analysis indicated significant improvements in performance for each two-day period of training for the Votrax Group (p <.05). The Votrax subjects improved from 53% correct identification at the start of training to 81% correct at the end of training with synthetic sentences.

Insert Figure 10 about here

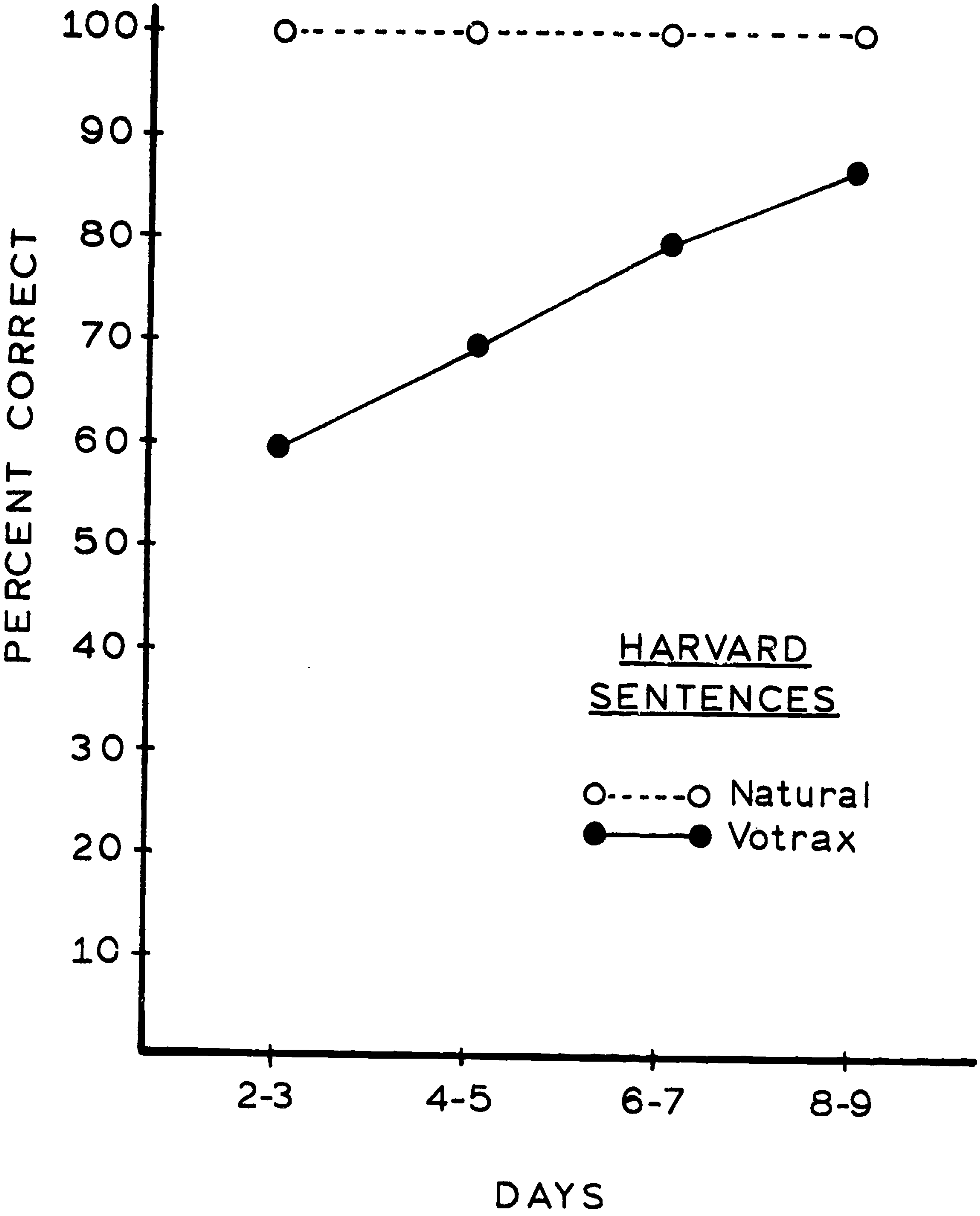


Figure 8. Mean correct word identification for natural and synthetic Harvard Sentences during the tiaining sessions.

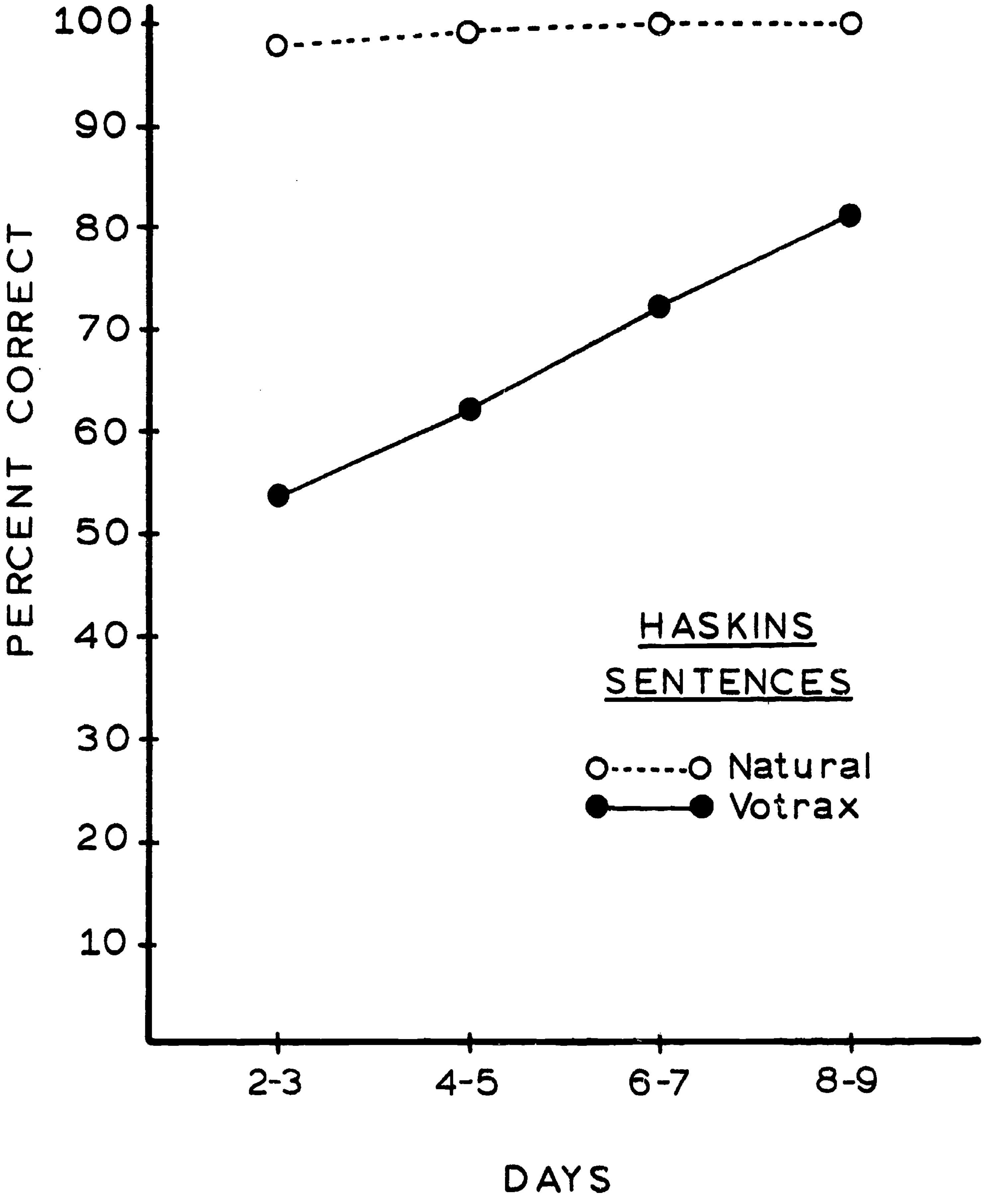


Figure 9. Mean correct word identification for natural and synthetic Haskins Sentences during the training sessions.

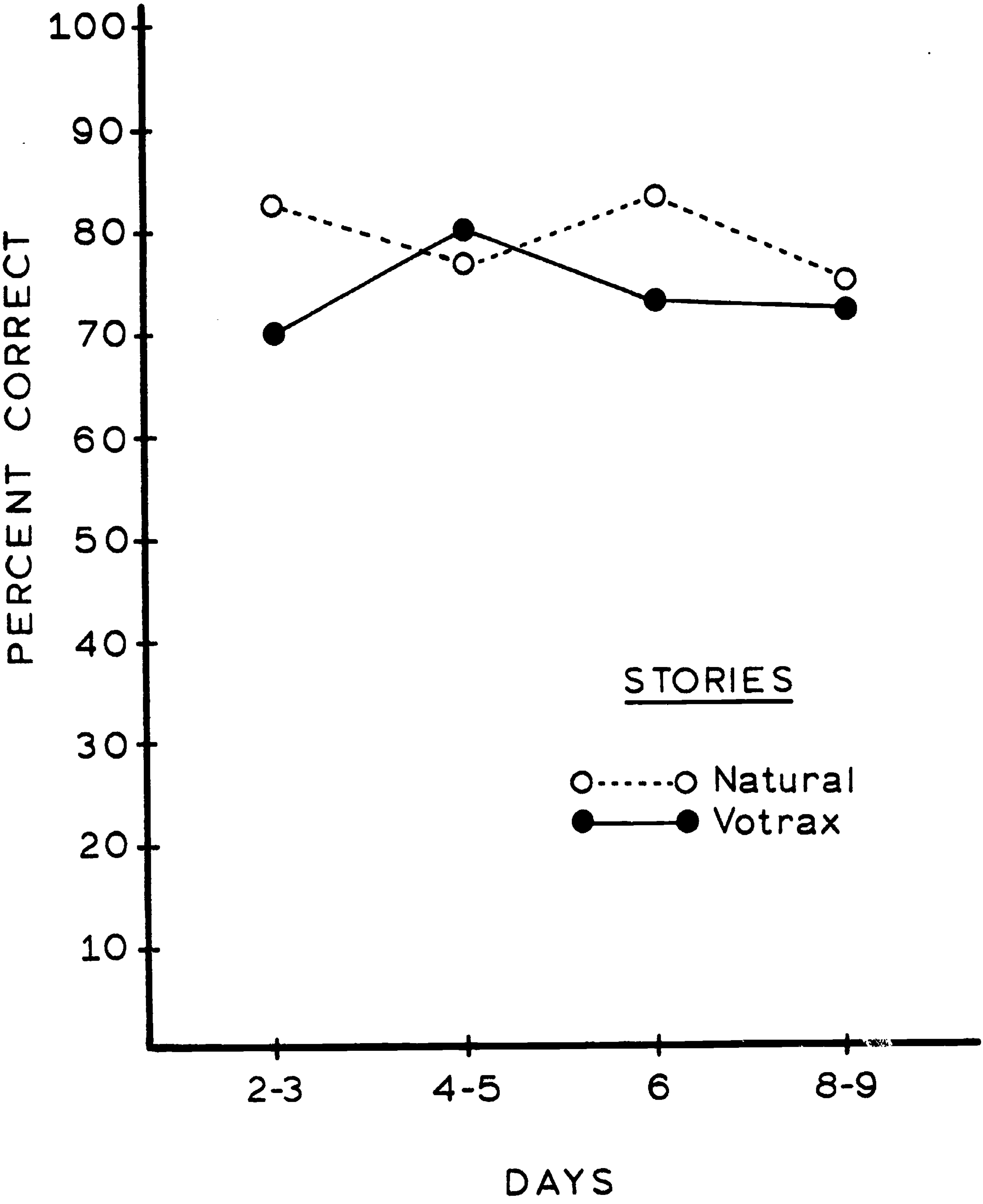


Figure 10. Mean correct responses to comprehension questions for natural speech and synthetic speech test stories presented during training sessions.

Prose Passages. Because of a computer malfunction, the prose data for the Votrax Group on Day 7 were accidently deleted and could not be recovered. To keep the design balanced, the comprehension data on Day 7 for the Natural Group were excluded from all subsequent analyses. Figure 10 displays the mean accuracy for comprehension questions for the three test passages presented on each training day. As shown here, a different pattern of results was obtained compared with the data from the previous tasks. No clear separation was observed in performance for the two groups. A Groups x Days analysis of variance was conducted and a significant two-way interaction was observed [*F*(3,42) = 5.37, *p* <.005]. A simple-effects analysis of this interaction revealed that the Natural Group performed better than the Votrax Group for Days 2-3 [*F*(1,39) = 9.07, *p* <.005] and Day 6 [*F*(1,39) = 7.53, *p* <.01] of the experiment. This result is not surprising since the Naturál Group listened to passages produced by a real talker. For the remaining days, there was no difference in accuracy between the two groups.

A Groups x Days analysis of variance was also conducted on the response latencies for the comprehension questions. The two- way interaction was not significant [*F*(3,42) = 1.11, *p* >.3]. However, a significant main effect for days was obtained [*F*(3,42) = 18.18, *p* <.0001]. A Newman-Keuls analysis indicated that the atencies for Days 2-3 were longer than the latencies for Days 4- 9 (*p* <.05). This difference may have been due to the specific passages and questions used on Days 2-3.

Summary. The day-by-day training data show a consistent pattern of results for the PB Lists, Harvard Sentences, and Haskins Sentences. As expected, subjects who listened to natural speech were at ceiling levels of performance throughout the course of training. By contrast, the Votrax subjects showed a very different pattern of results. These subjects showed a consistent and reliable improvement in their performance throughout the course of training. Moreover, based on a trend analysis, no evidence was obtained that subjects had reached asymptotic levels of performance even by the last day of training with the synthetic speech. Thus, it is entirely possible that further improvements could have been obtained if training had been carried out for a longer period of time.

The training data on comprehension of passages of text differ from the data collected with the other types of materials in several respects. First, the results indicated that on some days, the Natural Group showed higher levels of ·formance in answering comprehension questions than the Votrax Group. On other days, performance was equivalent for the two groups. No consistent overall pattern emerged from these results, perhaps due to the relative lack of sensitivity of our comprehension measures. Also, it appears that much of the variability in the data was produced by the specific stories and questions themselves, obscuring any possible effects of training.

In summary, the results of the testing and training sessions demonstrate several prominent effects of training on the perception of synthetic speech. First, training and exposure to synthetic speech improves the perception and recognition of isolated synthetic words and synthetic words in sentences. Second, training decreases response latencies to synthetic speech. Finally, and perhaps most importantly, the improvements in perception were not simply a function of mastery of the testing procedures or memorization of exemplars of the stimuli. Rather, improvements in performance were due to improved acoustic-phonetic encoding of the synthetic speech. By extension, this suggests that subjects have acquired tacit knowledge of the rules used by the Votrax Type-'N-Talk to produce synthetic speech.

The results demonstrate that it is possible to produce dramatic changes in the perception of synthetic words and sentences with only moderate amounts of training. Moreover, these improvements were obtained with synthetic speech produced by a low-cost commercially available text-to-speech system. Measuring performance for the output of a text-to-speech system with naive or unpracticed subjects may underestimate the performance characteristics of the system for certain applications. If a particular text-to-speech system is going to be used routinely on a daily basis by the same person, it may be difficult to predict daily performance with the synthetic speech by that individual from group recognition scores obtained with inexperienced subjects in a one-hour testing session in the laboratory.

Six-Month Follow-Up Evaluation

The overall results demonstrated that training was very effective in improving the perception of synthetic speech generated by a low-quality text-to-speech system. However, if a text-to-speech system will only be used occasionally, then the effectiveness of training as a means of improving intelligibility might be minimized. Thus, in addition to examining the effectiveness of training, it is important to examine the persistence of the training effects over some period of time. To investigate this problem, subjects were recalled from the Natural and Votrax Groups six months after training was completed. At that time, tests were conducted to assess performance and detail the possible changes that may have occurred.

Method

Subjects

Only five of the original subjects from each group were willing or able to participate in this follow-up evaluation. The subjects were paid for their participation as in the earlier study.

Materials and Procedure

In the six month follow-up, subjects were presented with Votrax-produced synthetic speech and were tested with the same procedures used for the pre-test and the post-test. The subjects transcribed words from two fifty-item PB Lists, two sets of 10 Harvard Sentences, and two sets of 10 Haskins Sentences. Subjects were also given two sets of 50 words from the Modified Rhyme Test. Prose passages were not used because of the lack of a clear pattern of performance in the earlier experiment. The procedures used were the same as those described for the testing days. No feedback was given for any of these tasks and all of the words and sentences were different tokens from the materials used six months earlier.

Results and Discussion

All responses were scored for accuracy in the same manner as described before. The Day 1 and Day 10 data presented in this section are on]y for the ten subjects who participated in the six-month f‹llow-up.

Insert Figure 11 about here

PB Lists. The mean accuracy for transcription of isolated words by the two groups of subjects for the three days of testing is shown in Figure 11. It can be seen that even after six months, subjects in the Votrax Group showed higher levels of performance in identifying these stimuli than subjects in the Natural Group. A two-way analysis of variance was conducted on the mean accuracy scores. This analysis revealed a significant Groups x Days interaction [$F(2,16)= 98.42$, $p < .0001$]. The post- hoc analyses indicated that, for the subjects who participated in the follow-up, both the Natural and the Votrax Groups showed significant changes in performance over the three test periods [$F(2,16) = 54.76$, $p < .0001$ and $F(2,16) = 454.16$, $p < .0001$, rspectively]. Newman-Keuls analyses indicated that both groups showed significant improvements in accuracy from Day 1 to Day 10 and that both groups showed a decrease in accuracy from Day 10 to 6 Months ($p < .05$). Simple-effects tests also showed that the Votrax Group was better in transcribing isolated words than the Natural Group on both Day 10 [$F(1,11) = 92.08$, $p < .0001$] and six months later [$F(1,11) = 61.18$, $p < .0001$]. This pattern of results suggests that although the effects of training decreased somewhat over six months, the Votrax Group retained a considerable amount of their experience with the synthetic output from the text-to-speech system.

Insert Figure 12 about here

MRT Lists. The mean percent correct word recognition for the Modified Rhyme Test is shown in Figure 12. A Groups x Days analysis of variance showed no significant interaction [$F(2,16) = 2.13$, $p > .15$]. Given the significance of the Groups x Days interaction in the original pre-test/post-test analysis ($p < .0005$), the absence of this interaction was surprising. However, this result might have been due to the small number of subjects participating in the follow-up evaluation. In order to determine if the Votrax Group performed better on Day 1 than the Natural Group, a simple-effects test was conducted. The results of this analysis revealed that on Day 1, the Votrax Group was not more accurate than the Natural Group [$F(1,6) = 5.2$, $p = .06$] in the MRT. However, on Day 10 and six months later, the Votrax Group was more accurate than the Natural Group [$F(1,6) = 40.83$, $p < .001$, and $F(1,6) = 26.13$, $p < .005$, respectively].

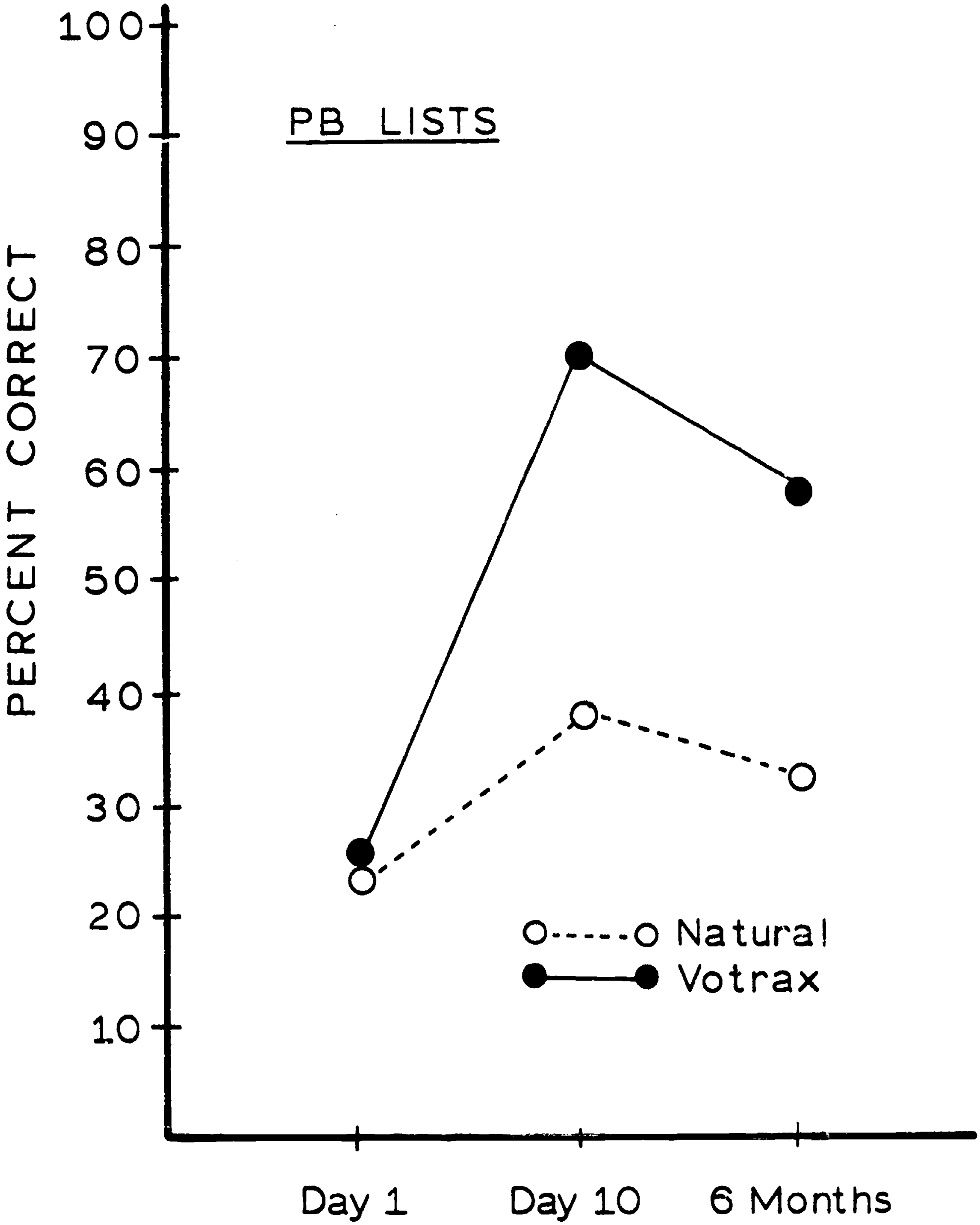


Figure 11. Mean correct responses for syntheti.: speech presented during test sessions on Day 1, Day 10 and 6 Months for the PB Lists.

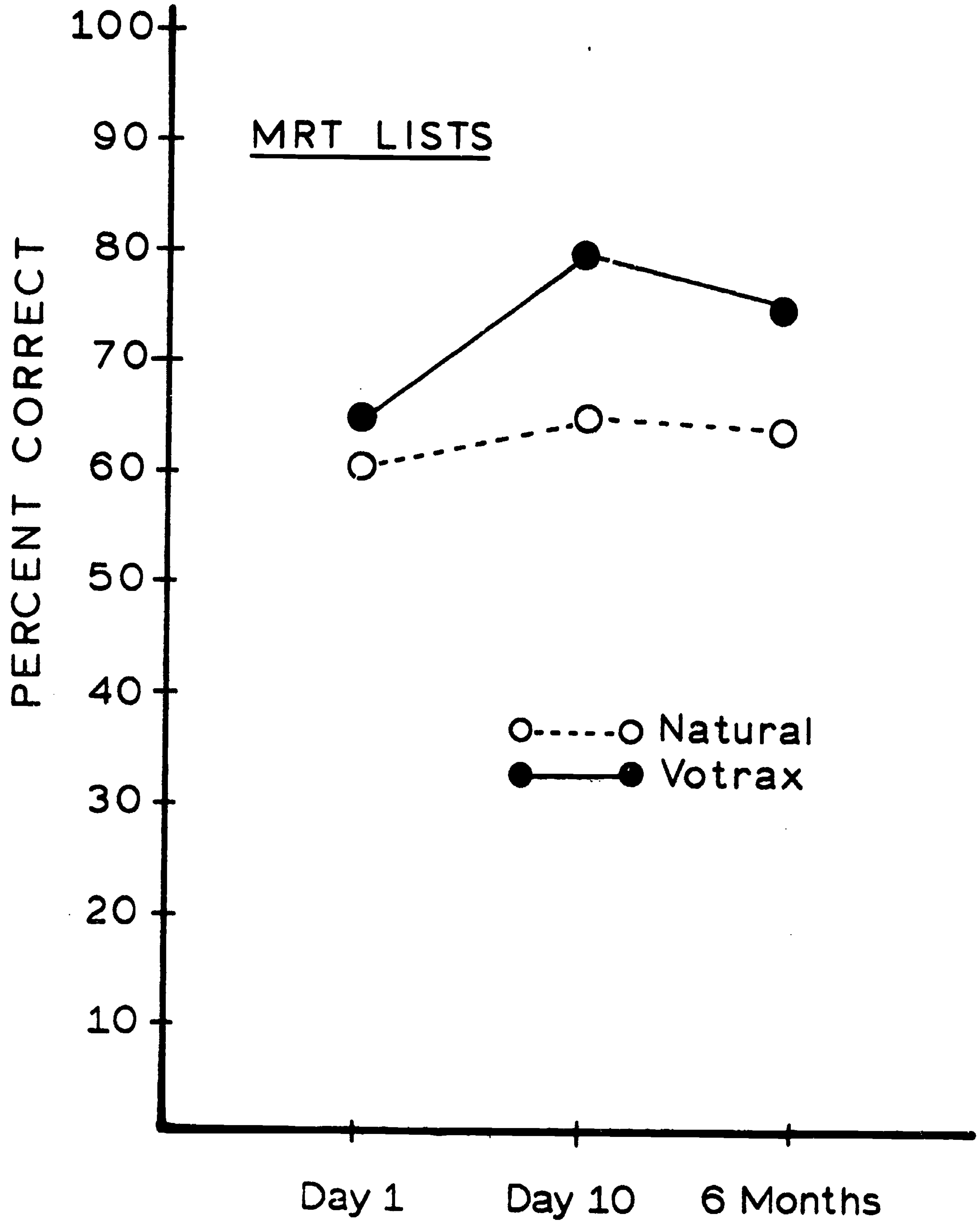


Figure 12. Mean correct responses for the synthetic speech MRT Lists presented on Day 1, Day 10, and 6 Months.

A two-way analysis of variance was also conducted on the latency data from the MRT. This analysis showed a significant Groups x Days interaction [F(2,16) = 4.11, p <.05]. A simple- effects analysis of this interaction indicated significant changes in latencies for both the Natural and Votrax Groups over the three test days [F(2,16) = 9.27, p <.005, and F(2,16) = 19.00, p <.0001, respectively]. Newman-Keuls analyses established that the Natural Group was faster to respond during the 6 Months test session than they were on Day 1 and Day 10 (p <.05). For the Natural Group, no difference in latencies between Days 1 and 10 was observed (p >.05). The Votrax Group responded faster on Day 10 and 6 Months than on Day 1 (p <.05). Moreover, no significant difference in latencies between Day 10 and 6 Months was observed for this group.

Insert Figure 13 about here

Harvard Sentences. The mean percent correct recognition of key words in the Harvard Sentences for each of the three test sessions is shown in Figure 13. A two-way analysis of variance showed a significant Groups x Days interaction [F(2,16) = 25.45, p =.0001]. A simple-effects analysis of this interaction indicated that word recognition for the two groups was different only on Day 10 of the experiment [F(1,11) = 25.24, p <.0005]. A Newman-Keuls analysis indicated the Votrax Group maintained their level of performance from Day 10 to 6 Months whereas the Natural Group improved their accuracy from Day 10 to 6 Months (p <.05). For the Natural Group, the improvement in performance six months after training is difficult to explain. Since the six-month follow-up was the third exposure to synthetic speech (for these subjects), it is possible that this improvement reflects the same type of learning effects found for the Votrax Group. Indeed, a comparison of the six-month follow-up data for the Natural Group with the training data for the Harvard Sentences on Days 2-3 for the Votrax Group (cf. Figure 13 and Figure 8) indicates comparable levels of performance. Thus, three exposures to synthetic speech seemed to produce roughly equivalent levels of performance for the Natural and Votrax Groups, even though the third exposure for the Natural Group came six months later. Similarly, comparable levels of performance were obtained for the third exposure to synthetic Haskins Sentences, for the Natural Group at six months and the Votrax Group on Days 2-3 (cf. Figure 14 and Figure 9). However, the increase in performance for the synthetic Haskins Sentences by the Natural group was not significant, possibly because of the small number of subjects in the follow-up.

Insert Figure 14 about here

Haskins Sentences. Figure 14 shows the mean percent correct word recognition for the Haskins sentences. A two-way analysis of variance showed a significant Groups x Days interaction [F(2,16) = 7.87, p <.005]. Post-hoc

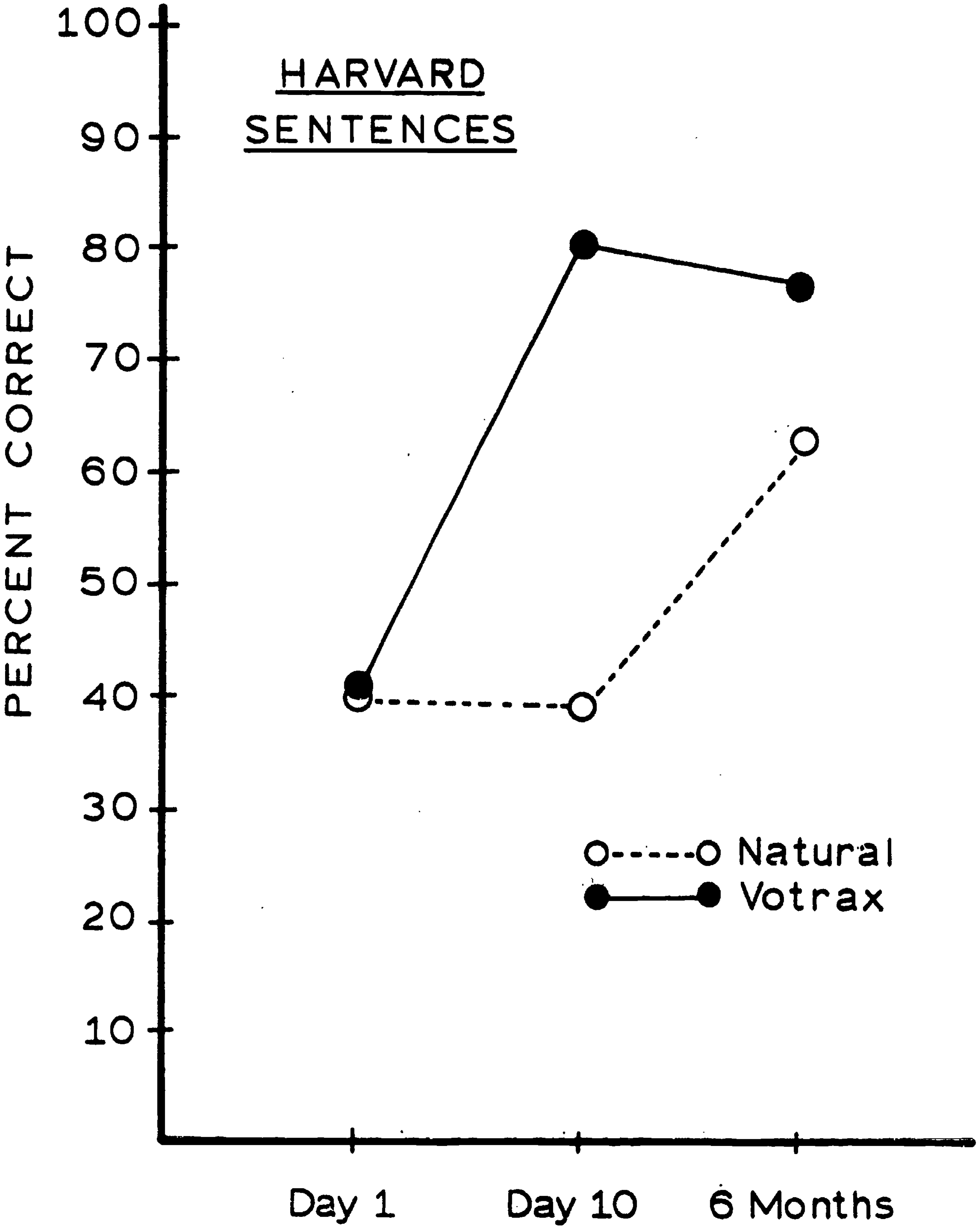


Figure 13. Mean correct synthetic word identification for Harvard Sentences presented on Day 1, Day 10, and 6 Months.

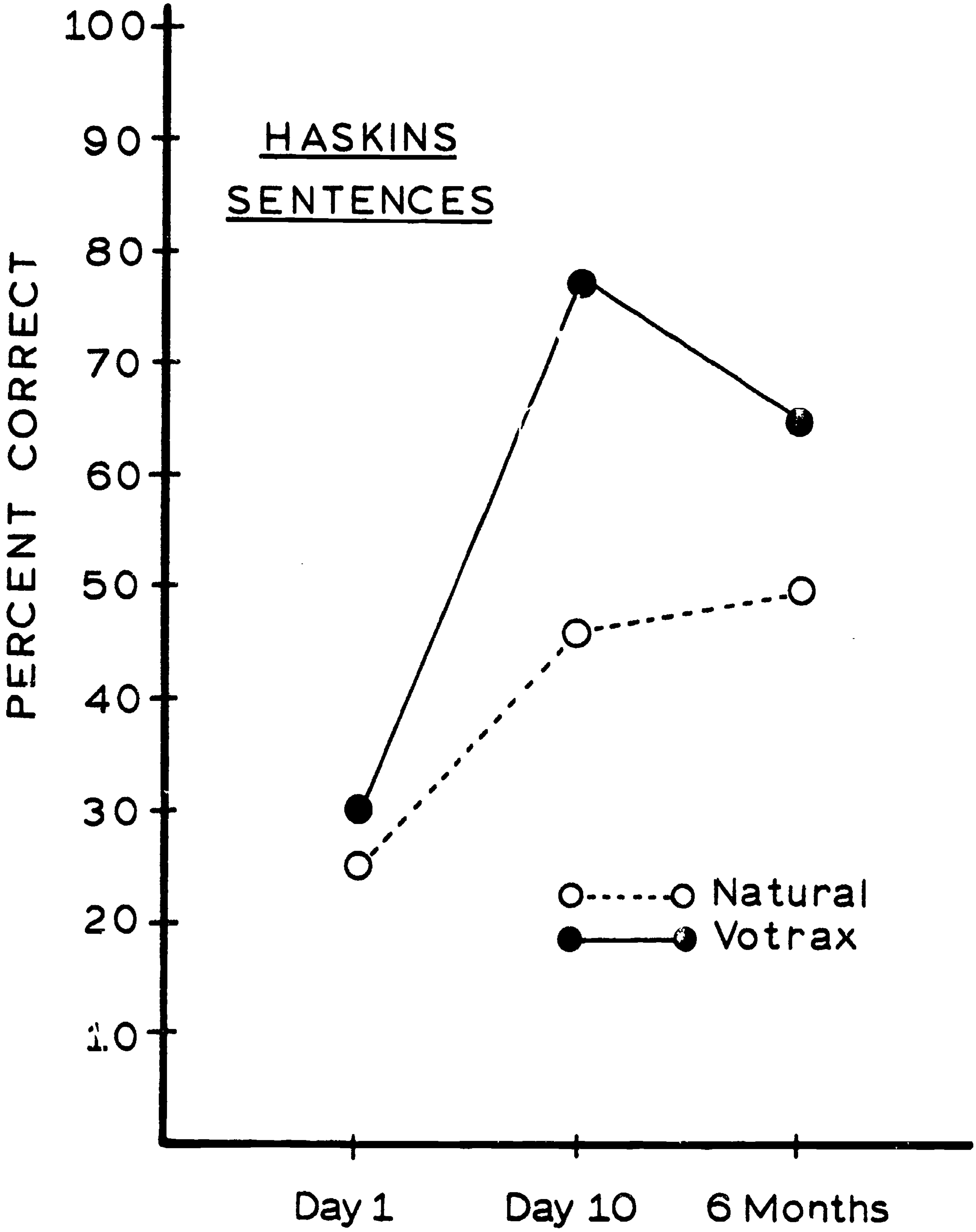


Figure 14. Mean correct synthetic word identification for Haskins Sentences pıesented on Day 1, Day 10, and 6 Months.

analyses indicated that the Votrax Group was more accurate than the Natural Group on Day 10 $[F(1,15) = 21.61, p < .0005]$ and 6 Months $[F(1,15) = 5.23, p < .05]$. In addition, Newman-Keuls analyses established that the Votrax Group decreased in accuracy from Day 10 to 6 Months ($p < .05$). Despite this decrease, performance for the Votrax Group was still better than the Natural Group.

Summary. The results of the six-month follow-up study showed that, in general, subjects maintained the gains in performance they had obtained as a result of training with the synthetic speech. The results with the PB Lists showed that, although performance for the Votrax Group decreased, performance was much higher six months after training than it was on Day 1 of the experiment. Furthermore, subjects in the Votrax Group retained some of the training even after six months, since their six-month follow-up performance was better than the performance of the Natural Group. This same pattern of results was observed for the Haskins Sentences. Although some drop in performance at six months was observed, subjects trained with synthetic speech performed much better than subjects trained with natural speech. Recognizing isolated PB words and words in the Haskins Sentences are the most difficult tests in the present study. In these tasks, the response set is extremely large and the subjects cannot use meaning as an aid to perception and understanding. The enhanced levels of performance six months after the completion of training indicate that subjects must have learned and retained fairly detailed information about the acoustic-phonetic properties of the Votrax text-to-speech system.

In the MRT, a task that is considerably easier because of the forced-choice format, no reduction in performance was observed for the Votrax Group even after six months. Moreover, for word recognition in the Harvard Sentences, where subjects can use linguistic knowledge and context to aid perception, we also found no reduction in performance for the Votrax Group. Thus, subjects trained on synthetic speech performed as well at the completion of training as they did six months later.

The results from the six-month follow-up demonstrate that the improvements in recognition of synthetic speech produced by training were maintained six months after training was completed. These results are even more interesting in our view considering that the subjects had no exposure to synthetic speech during the intervening period. Thus, no opportunity occurred for the subjects to actively rehearse the earlier effects of training by listening to synthetic speech during the intervening period.

Conclusions

The results of the present study suggest several important conclusions about the effects of training. First, the effect of training appears to be localized to a specific improvement in the encoding of synthetic speech produced by the Votrax Type-'N-Talk. Clearly, subjects did not simply learn to perform the various tasks better, since the subjects trained on natural speech showed little or no improvement in performance. Moreover, since different stimuli were presented on each day, subjects could not have memorized particular tokens from a fixed ensemble of stimuli. In addition, the training affected performance similarly with isolated words and words in sentences, and for closed and open response sets. This pattern of results indicates that subjects in the group trained on synthetic speech did not just learn special strategies for sophisticated "guessing"; that is, they did not simply learn to use linguistic knowledge or task constraints to improve recognition. Rather, subjects abstracted detailed information about the structural characteristics and rule system of this particular synthetic speech system. This conclusion is further supported by the design of our study. Improvements in performance were obtained on novel materials even though the subjects never heard the same words or sentences twice. In order to show improvements in performance, subjects were required to learn abstract information about the rules that are used to generate the detailed acoustic-phonetic properties of the synthetic speech produced by the system.

In addition, the follow-up study demonstrated that subjects retained much of the effects of training even after six months with no further contact with the synthetic speech. Thus, it appears that training produced a relatively stable and long-term, permanent change in the perceptual encoding processes used by subjects in responding to synthetic speech. Furthermore, it is likely that more extensive training would have produced even greater persistence of the training effects. If subjects had been trained to asymptotic levels of performance, the long-term effects of training might have been even more sizable than those observed here.

Thus, it appears that low-cost, poor-quality, commercially available, text-to-speech systems can be used to provide speech output in certain applications. If the cost of training is minimal compared to the cost of voice response systems, then for some well-defined applications a low-cost text-to-speech system may provide acceptable performance even when the system is not used on a regular basis. The extent to which performance can be modified is known to be affected by cognitive load, the degree of required processing and number of responses under different work load conditions. This study examined performance with synthetic speech in relatively easy tasks imposing little cognitive load. Obviously, future research will need to be directed towards the problems inherent in high-information load tasks where the observer may be required to carry out several attention-demanding activities simultaneously.

REFERENCES

Allen, J. Linguistic-based algorithms offer practical text-to- speech systems. Speech Technology, 1981, 1, 12-16.

Brown, J. I. Nelson-Denny Reading Test, Form D. Boston: Houghton- Mifflin, 1973.

Carlson, R., Granstrom, B., and Larsson, K. Evaluation of a text-to-speech system as a reading machine for the blind. Speech Transmission Laboratory, QPSR 2-3, 1976, 9-13.

Derrick, C., Harris, D. P., and Walker, B. Cooperative English Test: Reading Comprehension, Form 1B. Princeton N.J.: Educational Testing Service, 1960.

Egan, J. P. Articulation testing methods. Laryngoscope, 1948, 58, 955-991.

Farr, R. (Ed.) Iowa Silent Reading Tests, Level 3, Form E. New York: Harcourt Brace Jovanovich, 1972.

Gardener, E. F., Callis, R., Merwin, J. C., and Madden, R. Stanford Test of Academic Skills: Reading. College Level II- A. New York: Harcourt Brace Jovanovich, 1972.

Grosjean, F. Spoken word recognition processes and the gating paradigm. Perception & Psychophysics, 1980, 28, 267-283.

House, A. S., Williams, C. E., Hecker, M. H. L., and Kryter, K. D., Articulation-testing methods: Consonantal differentiation with a closed-response set. Journal of the Acoustical Society of America, 1965, 37, 158-166.

IEEE Recommended Practice for Speech Quality Measurements, IEEE No. 297, New York: IEEE, 1969.

Nelson, M. J., and Denny, E. C. Nelson-Denny Reading Test for Colleges and High Schools, Forms A and B. New York: Houghton-Mifflin, 1930.

Nye, P. W., and Gaitenby, J. The intelligibility of synthetic monosyllable words in short, syntactically normal sentences. Haskins Laboratories Status Report on Speech Research, 1974, 38, 169-190.

Pisoni, D. B. Speeded classification of natural and synthetic speech in a lexical decision task. Journal of the Acoustical Society of America, 1981, 70, S98.

Pisoni, D. B. Perception of speech: The human listener as a cognitive interface. Speech Technology, 1982, 1, 10-23.

Pisoni, D. B., and Hunnicutt, S. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1980, 572-575.

Salasoo, A., and Pisoni, D. B. Sources of knowledge in spoken word identification. Research on Speech Perception, Progress Report 8, 1982, 105-145.

Slowiaczek, L., and Pisoni, D. B. Effects of practice on speeded classification of natural and synthetic speech. Journal of the Acoustical Society of America, 1982, 71, S95.

Walker, L. A. Ears: a simple auditory screening test. Research on Speech Perception, Progress Report 8, 1982, 319-323.

Vowel Categorization by Three-Year-Old Children*

C. A. Kubaska and R. N. Aslin

Infant Perception Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

*The results reported in this paper were previously presented at the 105th and 106th meetings of the Acoustical Society of America. We wish to thank David Pisoni for providing the use of the Speech Research Laboratory for processing our speech stimuli. We also thank Jerry Forshee for software development, Mary Buuck for assistance in running subjects, and Howard Nusbaum and Sandra Shea for their helpful comments. This research was supported by NIH Training grant #NS 07134 and NICHHD Research Grant #HD11915 to Indiana University.

Abstract

A new two-alternative, forced-choice procedure was used in two experiments to test three-year-old children's categorization of edited, natural vowel tokens produced by several talkers. An appropriate pointing response (right or left) was visually reinforced with one of two televisions. In the first experiment, the stimuli were isolated tokens of /a/ and /i/ produced by a male adult, a female adult, a male child, and a female child. The stimuli in the second experiment were isolated tokens of /ae/ and /ʌ/ produced by the same talkers. In both experiments, three-year-olds spontaneously generalized their pointing response from the male adult vowel tokens to the corresponding vowels produced by the other talkers. Children reinforced for an arbitrary grouping of the two vowel categories persisted in categorizing on the basis of vowel quality. Results from both experiments demonstrate the usefulness of this new procedure for assessing speech categorization in preschoolers. They also demonstrate the presence of perceptual constancy for vowel tokens across talkers. In particular, the results from the second experiment provide evidence for normalization of isolated, quasi-steady-state vowel tokens because the formant values for tokens of /ʌ/ produced by the woman and the two children were closer to the formant frequencies of the male adult's /ae/ than the male adult's /ʌ/. Developmental trends in the use of dynamic versus steady-state information for vowel identification are discussed.

Vowel Categorization by Three-Year-Old Children

I. Introduction

It has been known for over thirty years that the first two formants of a vowel are important acoustic cues to its phonetic identity (Delattre, Liberman, Cooper and Gerstman, 1952). However, theories of vowel perception are complicated by the fact that the formant frequencies of a given vowel vary widely within and across talkers, due to differences in context, speaking rate, and vocal tract dimensions (Shankweiler, Strange, and Verbrugge, 1977). Despite this variability, listeners are able to identify accurately vowels that are produced by different talkers under a variety of experimental conditions, including vowels in consonantal contexts (Verbrugge, Strange, Shankweiler, and Edman, 1976; Strange, Verbrugge, Shankweiler, and Edman, 1976; Strange and Gottfried, 1980; Diehl, McCusker, and Chapman, 1981), vowels in isolation (Macchi, 1980; Diehl et al, 1981; Assmann, Nearey, and Hogan, 1982), gated vowel centers (Assmann et al, 1982), and syllables from which the steady state portion of the vowel has been deleted (Strange, Jenkins, and Johnson, 1983; Jenkins, Strange and Edman, 1983).

Theories of vowel perception have emphasized the importance of the information from either the steady-state formant frequencies or the dynamic cues distributed throughout the syllable. These two types of theories have been referred to as "target normalization" and "dynamic specification" theories by Strange et al (1983). "Target normalization" theories use the steady-state formant values of the vowel as input to a normalization process by which the vowel is categorized appropriately in the vowel space (Joos, 1948; Gerstman, 1968; Liebermar. 1973; Nearey, 1977, 1983). Other researchers (Shankweiler et al, 1977; Verbrugge et al, 1976; Strange et al, 1979; Strange et al, 1983) have stressed the role of dynamic spectral cues in vowel perception. According to this theory, steady-state formant information is considered a sufficient, though not necessary, cue for vowel identification (Strange and Gottfried, 1980). The listener uses the acoustic information distributed throughout the syllable, including the formant transitions and vowel duration, to identify the vowel. The results of studies which show lower error rates in identification for vowels in syllable context versus vowels in isolation (Verbrugge et al, 1976; Strange et al, 1976; Strange and Gottfried, 1980; Gottfried and Strange, 1980) are taken as evidence in support of dynamic specification theories. Additional evidence supporting this position comes from studies demonstrating accurate vowel identification for syllables without steady-state vowel segments (Strange et al, 1983; Jenkins et al, 1983).

An important question in developmental speech research is whether infants and young children perceive the similarity of phonetically equivalent but acoustically variable vowels spoken by different talkers. Using an operant headturning procedure, Kuhl (1979) obtained results which suggested that 6-month-old infants do have perceptual constancy for certain vowel categories. The stimuli were synthetic tokens of the vowels /a/ and /i/, with formant parameters set for a male adult, a female adult, and a child talker. Also, each token was synthesized with either a rising or a rising-falling pitch contour. The infants learned a headturning response to one of the male adult vowels with a rising-falling contour, and were able to generalize this response to presentations of the same vowel by the female adult and the child, as well as to the male adult token with a rising contour. In doing so, the infants ignored the variations in talker and intonation contour present in the background stimuli. Kuhl (1983) obtained similar results with synthetic tokens of /a/ and /ɔ/ corresponding to a male adult, female adult, and child talker.

In another study, Kuhl and Miller (1982) used a high amplitude sucking procedure to show that 1- to 4-month-old infants discriminated changes in vowel quality and pitch contour when only one of these components varied after a preshift habituation period. When both components varied, infants detected changes in vowel target, but not in intonation. Also, infants did not discriminate stimulus changes when vowel quality and pitch contour were recombined into novel targets after a preshift period in which the stimuli varied in both dimensions. Kuhl and Miller also measured the amount of time required by infants to habituate to the preshift stimuli. Generally, infants took longer to habituate during preshift periods with two stimuli, but when these periods contained stimuli with the same vowel quality, the time to habituation was not significantly different from preshift periods with only one stimulus. The results of these discrimination experiments by Kuhl and her colleagues show that young infants appear to recognize some similarity between members of a vowel category, despite differences in intonation and talker, and that vowel quality seems to be a particularly salient perceptual dimension for young infants.

In the present study, we tested three-year-old children's categorization of edited, natural vowel tokens produced by several talkers using a new two-alternative, forced choice procedure. In the first experiment, the stimuli were isolated tokens of /a/ and /i/ produced by a male adult, a female adult, a male child, and a female child. The stimuli in the second experiment were isolated tokens of /ae/ and /ʌ/ produced by the same talkers. This second contrast was chosen because the formant values for tokens of /ʌ/ produced by the woman and the two children were closer to the formant frequencies of the male adult's /ae/ than the male adult's /ʌ/. It might be expected that young children would have difficulty with this categorization task, particularly with the /ae/-/ʌ/ contrast, since the dynamic spectral cues from formant transitions and inherent vowel duration were absent. By eliminating these cues, vowel normalization of isolated, quasi-steady-state vowel tokens by three-year-olds could be assessed.

II. Method

Previous studies of speech perception in young children have used a two-alternative, forced choice procedure which typically involved a picture identification task (Simon and Fourcin, 1978; Greenlee, 1980; Strange and Broen, 1981; Krause, 1982). For example, Simon and Fourcin (1978) examined 2- to 14-year-old children's perception of a voiced-voiceless contrast when VOT and first formant onset were varied. The younger subjects in this experiment pointed to one of two pictures, each depicting an item from a minimal pair that contrasted initial stops. Similarly, Strange and Broen (1981) tested three-year-old children's perception of the /w/-/r/ and /r/-/l/ contrasts, using an almost identical procedure. In the present study, subjects also learned a two-alternative, forced choice task, but the procedure was adapted to assess categorization in young children without using any picture referents.

The basic layout of the testing apparatus is shown in Figure 1.

Insert Figure 1 about here

The child (C) was seated on his parent's lap (P) in one corner of a single-walled, sound attenuated booth (I.A.C. model #402). The test stimuli were presented over a loudspeaker (Radio Shack MC-1000) positioned in front of the child. Two color television sets (12" RCA XL-100) were located in the remaining corners of the booth. The child was trained to point to one television when a particular vowel was heard and to the second television upon hearing another vowel. If the correct pointing response was made, the television came on and displayed an 8 second segment of the Muppet Movie, or some other cartoon, via an RCA Selectavision video disc player (#SJT090). During the testing sessions, the parent listened to masking music over headphones to eliminate any bias. A camera was mounted outside the booth near a concealed window, and the experimenter (E) observed the child's responses on a TV monitor, scoring the directional responses on a button box. Stimulus presentation and reinforcement were controlled by a PDP 11/34 computer. The stimuli were stored in digital form on computer disk and presented to subjects on line via a 12 bit D-A converter, with a sampling rate of 10 kHz and a low pass filter at 4.8 kHz. During the testing sessions, the experimenter was unaware of the specific stimulus which was being presented; she served simply to initiate trials when the child was sitting still and to score the direction in which the child pointed.

This procedure is an improvement over previous two-alternative, picture choice tasks for the following reasons. First, most subjects were highly motivated. They enjoyed watching the cartoons and were thrilled by the seemingly magical quality by which they apparently turned on the TVs. There was no need to interrupt the testing session for rests or play breaks, or to cajole the child into finishing the task. Second, using this procedure, one can test children's perception of non-phonemic items or isolated sound segments because picture referents are not needed. This also eliminates the need for concern over the child's familiarity with picture items. This non-semantic aspect of the procedure enables the testing of children as young as two years of age. Finally, possible experimenter cueing was controlled. The computer determined the stimulus presentation order, and different randomizations were used for each session. The experimenter remained outside the booth and did not know which stimulus was being heard by the child.

The two experiments reported in this paper tested three-year-old children's categorization of vowel pairs produced by four talkers consisting of a male adult, a female adult, a male child, and a female child. The general procedure for both experiments included three phases: shaping, testing, and categorization. During shaping sessions, all subjects were trained to point to one TV upon hearing one male adult vowel token, and to the other TV upon hearing the second male adult vowel token. The pairing of TVs and vowels was counterbalanced across subjects. During the shaping sessions, parents did not wear headphones and encouraged their children to point at the appropriate TV when each of the two vowels was presented. The experimenter controlled the

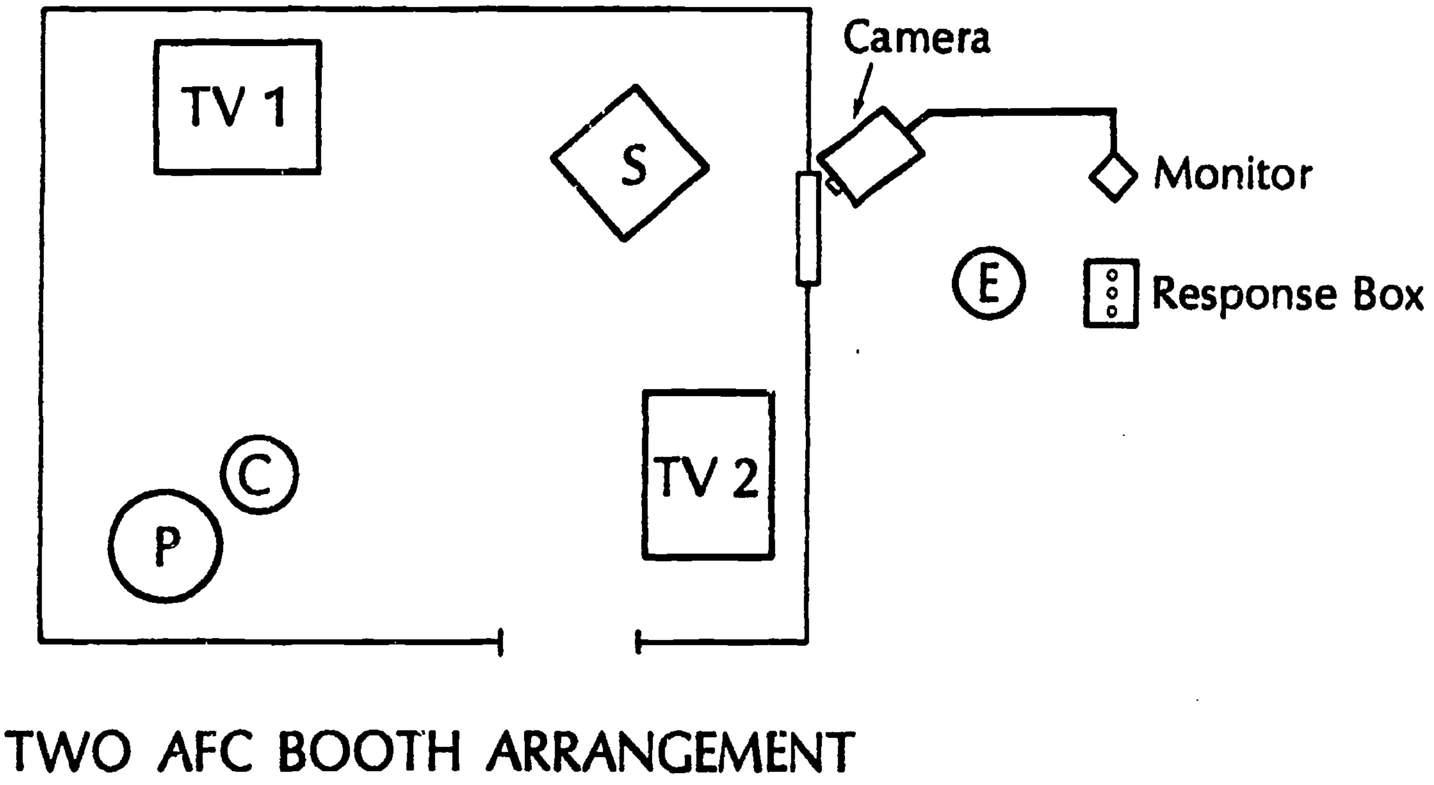


Figure 1: Testing apparatus used in both experiments.

presentation of one of the two stimuli on each trial, which consisted of the repetitive presentation of a given vowel with an interstimulus interval of 1.5 seconds. Each trial was terminated by the experimenter's button press, indicating a pointing response, or by a total of 20 repetitions of the vowel. The child was reinforced on every trial provided that the correct pointing response was made, even if it was preceded by an incorrect response.

A subject proceeded to the testing phase when the experimenter judged that the child's first pointing response was consistently correct, generally after one or two 10 minute sessions. During the testing sessions, parents wore headphones and listened to music while their children responded to the male adult's vowel tokens. The choice of the vowel stimulus and the presentation of reinforcement in this phase was now computer-controlled. In these testing sessions, the subject was required to point in the correct direction in order to view the cartoon reinforcer. If the initial pointing response was incorrect, the reinforcer was not presented and another trial was not initiated until the child had ceased pointing and was sitting still. An adjustment in the computer program was made for Experiment II, so that another trial could not be initiated for 8 seconds, which was equal to the amount of reinforcement time for a correct response. If the child responded correctly to 90% of 10 consecutive trials, the session immediately proceeded to the categorization phase, during which vowels produced by the other talkers were introduced, in addition to the male adult tokens. In this last stage, a session consisted of 5 trials of each vowel presented in random order.

III. Experiment 1

The first study was conducted to evaluate our new procedure for assessing 3-year-olds' categorization of the two acoustically distinct vowels /a/ and /i/.

A. Subjects

The subjects in this study were 20 three-year-old children from the Bloomington, Indiana area, ranging in age from 2;10.5 to 3;0.75 with a mean age of 2;11.75. Their names were obtained from birth announcements in the local newspaper, and their parents were solicited by mail and a follow-up telephone call. Subjects were tested twice a week for a total of 4 to 5 sessions, and were paid $3 per session. An additional 17 subjects were dropped after completing 2 or more sessions due to illness, scheduling conflicts, or inability to reach the testing criterion on the two training stimuli.

B. Stimuli

Isolated tokens of the vowels /a/ and /i/ spoken by a male adult (RJ), a female adult (MB), a male child (JS, age 12), and a female child (TS, age 10) were produced in a single-walled, sound-attenuated booth (I.A.C. model #401A), and recorded with an Ampex AG-500 tape recorder and an Electro-Voice D054 microphone. These stimuli were then digitized on a PDP 11/34 computer with a low pass filter at 4.8 kHz and a sampling rate of 10 kHz. A computer editing program (Luce and Carrell, 1981) was used to set the duration of all vowel tokens at approximately 300 msec with the constraint that the vowel endpoints were placed at the zero crossing at the end of a complete pitch period. The amplitude of these tokens was also controlled with a waveform modification program (Bernacki, 1981). All stimuli were presented at 63±2 db SPL, verified with a Triplett Sound Level Meter (model #370, C scale). Transients at the onset of the stimuli were eliminated by gradually increasing the amplitude of the initial 20 msec of a token with an RMS energy ramping function on the waveform modification program. Similarly, offset transients were eliminated by gradually decreasing the amplitude of the final 20 msec segment of a token. The formant frequencies were also measured at the midpoint of each stimulus with a spectral analysis program (Kewley-Port, 1978), and these values are listed in Table 1. In cases where accurate measurements could not be obtained with this program's formant tracking algorithm, a second program which allowed manual formant tracking (Carrell, forthcoming) was used to determine these values.

Insert Table 1 about here

C. Procedure

There were two experimental groups in this study. Ten subjects were assigned to the Constant Vowel group, in which all tokens of /a/ were reinforced by one television, and all tokens of /i/ by the other television. The pairing of TVs and vowels was counterbalanced across subjects; all tokens of /a/ were assigned to the left TV for 5 of these subjects, and to the right TV for the remaining 5 subjects. The two Constant Vowel stimulus sets are listed in Table 2.

Insert Table 2 about here

Another 10 subjects were assigned to the Mixed Vowel group, in which each subject received a different random combination of vowels, with the constraint that 2 /a/s and 2/i/s were paired with each TV. This second condition was a control for the possibility that children could simply memorize the TV-vowel associations. As in the Constant Vowel group, the pairing of TVs and vowels was counterbalanced across subjects; 5 subjects were taught to associate the male adult /a/ (one of the 2 training stimuli) with the left TV, and the remaining 5 subjects associated that token with the right TV. The Mixed Vowel stimulus sets used in this experiment are listed in Table 3.

Table 1

Formant Frequency Measurements - Experiment I

Speaker	/a/	/i/
Male Adult (RJ)	F1 = 653	F1 = 327
	F2 = 1058	F2 = 2104
	F3 = 2339	F3 = 3097
Female Adult (MB)	F1 = 1076	F1 = 384
	F2 = 1678	F2 = 2762
	F3 = 2871	F3 = 3465
Male Child (JS)	F1 = 1070	F1 = 315
	F2 = 1545	F2 = 2942
	F3 = 2927	F3 = 4022
Female Child (TS)	F1 = 1232	F1 = 429
	F2 = 1485	F2 = 3349
	F3 = 3696	F3 = 3955

Table 1: The first 3 formant measurements in Hz for isolated tokens of the vowels /a/ and /i/ produced by a male adult (RJ), female adult (MB), male child (JS, age 12), and female child (TS, age 10). Measurements were made at the midpoint of each stimulus.

Table 2

Constant Vowel Stimulus Sets

	Left TV	Right TV
A.	1. Male Adult /a/	5. Male Adult /i/
	2. Female Adult /a/	6. Female Adult /i/
	3. Male Child /a/	7. Male Child /i/
	4. Female Child /a/	8. Female Child /i/
B.	1. Male Adult /i/	5. Male Adult /a/
	2. Female Adult /i/	6. Female Adult /a/
	3. Male Child /i/	7. Male Child /a/
	4. Female Child /i/	8. Female Child /a/

Table 2: Constant Vowel stimulus sets used in Experiment I. The male adult vowel tokens, stimuli 1 and 5 in each set, were the training stimuli.

Insert Table 3 about here

Note that in every Mixed Vowel stimulus set, stimuli 2 and 6 match the vowel quality of the two training stimuli, numbers 1 and 5, while stimuli 3,4,7, and 8 do not. Hereafter, stimuli 1,2,5, and 6 will be referred to as set I stimuli, and numbers 3,4,7,and 8 will be referred to as set II stimuli.

After each subject met the testing phase criterion of 90% correct on 10 consecutive trials, data were collected in two or more sessions in the categorization phase. The results are based on the subjects' responses to 10 trials of each vowel token, or a total of 80 trials per subject.

D. Results and Discussion

Overall, subjects in the Constant Vowel group responded correctly on an average of 75% of the trials, while subjects in the Mixed Vowel group were correct on only 50% of the trials (see Figure 2).

Insert Figure 2 about here

The Constant Vowel group performed at the 75% correct level for both sets of stimuli. However, the response level for the Mixed Vowel group dropped from 75% correct for set I stimuli to 25% correct for set II stimuli (see Figure 3).

Insert Figure 3 about here

A two-factor, mixed design ANOVA with repeated measures was used to analyze these data. The results show a significant difference between the two subject groups (F(1,18) = 54.18, p <.001) and stimulus sets (F(1,18) = 24.20, p <.001), and a significant stimulus set x group interaction (F(1,18) = 26.11, p <.001). A posthoc F test for simple effects showed that the significant interaction was due to the Mixed Vowel group's responses to set II stimuli (F(1,18) = 64.87, p <.001). This interaction is reflected in the performance of the two groups for the two stimulus sets (see Figure 3). The Constant Vowel group performed at the 75% correct level for both set I and set II stimuli, since in this group, all tokens of one vowel were paired with one TV, and the remaining vowel tokens with the other TV. In contrast, the Mixed Vowel group responded correctly on 75% of the trials with set I stimuli, but this response rate dropped to 25% correct for set II stimuli. Recall that these set II stimuli did not match the original training stimuli's vowel quality.

Table 3

Mixed Vowel Stimulus Sets

	Left TV	Right TV
A.	1. Male Adult /a/	5. Male Adult /i/
	2. Female Adult /a/	6. Female Child /i/
	3. Female Adult /i/	7. Male Child /a/
	4. Male Child /i/	8. Female Child /a/
B.	1. Male Adult /a/	5. Male Adult /i/
	2. Male Child /a/	6. Female Child /i/
	3. Female Adult /i/	7. Female Adult /a/
	4. Male Child /i/	8. Female Child /a/
C.	1. Male Adult /a/	5. Male Adult /i/
	2. Female Child /a/	6. Female Child /i/
	3. Female Adult /i/	7. Female Adult /a/
	4. Male Child /i/	8. Male Child /a/
D.	1. Male Adult /a/	5. Male Adult /i/
	2. Female Child /a/	6. Male Child /i/
	3. Female Adult /i/	7. Female Adult /a/
	4. Female Child /i/	8. Male Child /a/
E.	1. Male Adult /a/	5. Male Adult /i/
	2. Female Child /a/	6. Female Adult /i/
	3. Male Child /i/	7. Female Adult /a/
	4. Female Child /i/	8. Male Child /a/
F.	1. Male Adult /i/	5. Male Adult /a/
	2. Male Child /i/	6. Female Adult /a/
	3. Male Child /a/	7. Female Adult /i/
	4. Female Child /a/	8. Female Child /i/
G.	1. Male Adult /i/	5. Male Adult /a/
	2. Female Adult /i/	6. Female Adult /a/
	3. Male Child /a/	7. Male Child /i/
	4. Female Child /a/	8. Female Child /i/

Table 3 continued

H.	1. Male Adult /i/	5. Male Adult /a/
	2. Male Child /i/	6. Male Child /a/
	3. Female Adult /a/	7. Female Adult /i/
	4. Female Child /a/	8. Female Child /i/
I.	1. Male Adult /i/	5. Male Adult /a/
	2. Female Adult /i/	6. Male Child /a/
	3. Female Adult /a/	7. Male Child /i/
	4. Female Child /a/	8. Female Child /i/
J.	1. Male Adult /i/	5. Male Adult /a/
	2. Female Adult /i/	6. Female Child /a/
	3. Female Adult /a/	7. Male Child /i/
	4. Male Child /a/	8. Female Child /i/

Table 3: Mixed Vowel stimulus sets used in Experiment I. The male adult vowel tokens, numbers 1 and 5 in each set, were the training stimuli.

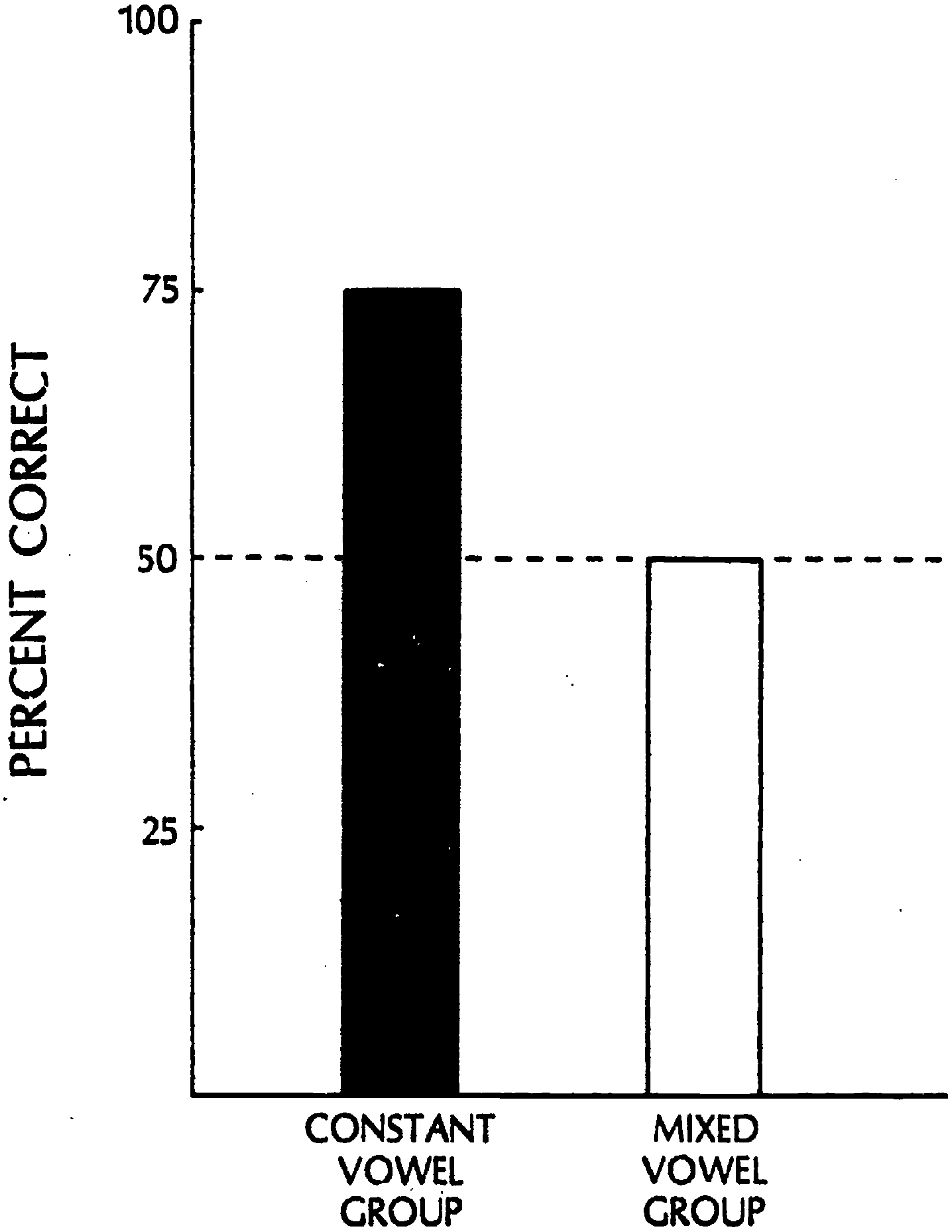


Figure 2: Mean percentage of correct responses for the Constant Vowel group and the Mixed Vowel group in Experiment I.

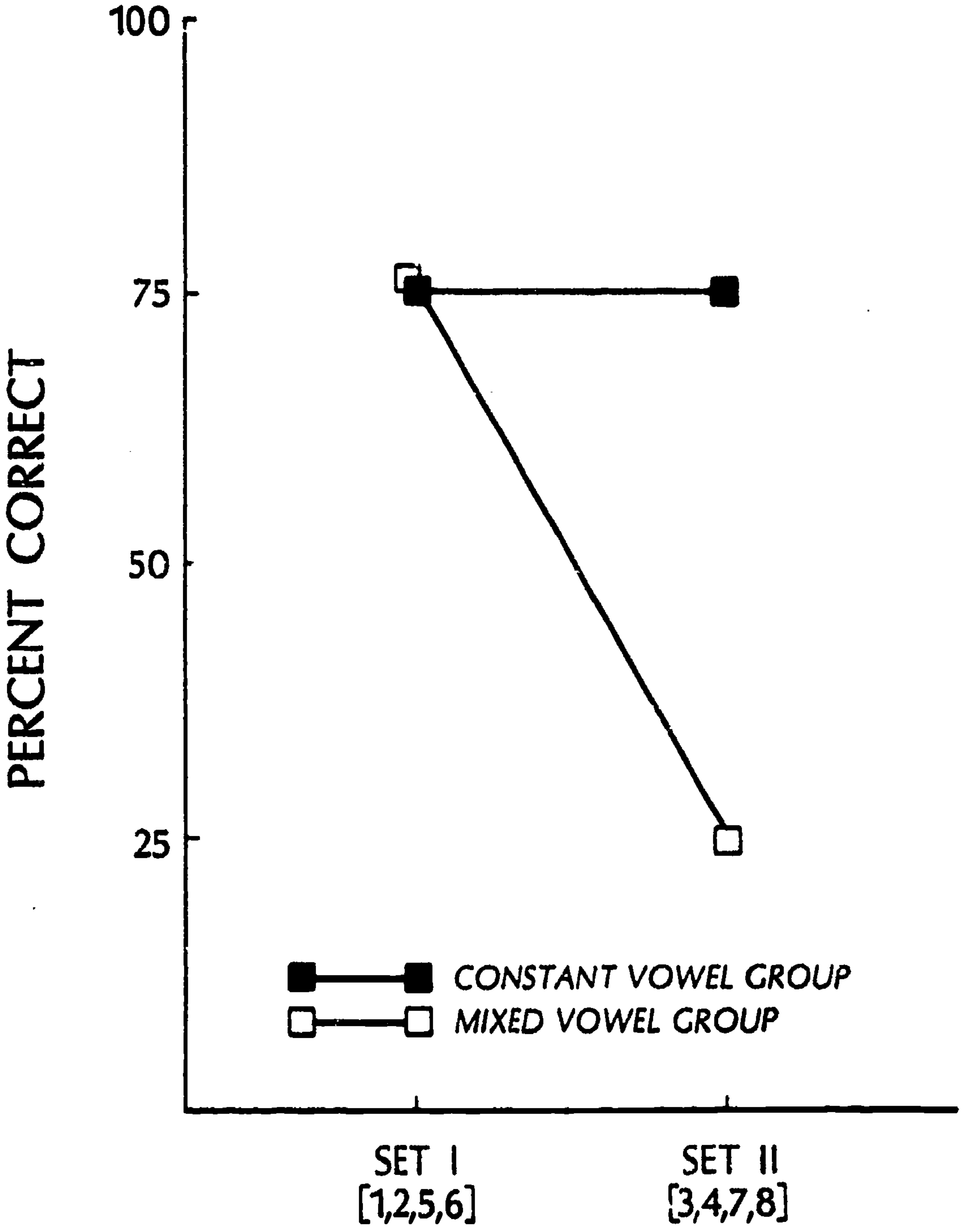


Figure 3: Mean percentage of correct responses to set I and set II stimuli for the Constant Vowel group and the Mixed group in Experiment I.

These results indicate that the Mixed Vowel group, as well as the Constant Vowel group, were generalizing their pointing response from the male adult vowel tokens to the similar vowel tokens from the other talkers. These results demonstrate three-year-old children's perceptual constancy for the vowels /a/ and /i/ across talkers. Moreover, they demonstrate that visual reinforcement for an arbitrary grouping of two vowel categories is not sufficient to overcome the natural tendency to categorize vowels by vowel quality.

IV. Experiment II

In Experiment I, three-year-old children demonstrated their ability to categorize the point vowels /a/ and /i/ across talkers. The purpose of Experiment II was to assess three-year-old children's vowel normalization of isolated, quasi-steady state tokens of /ae/ and /ʌ/ produced by the same talkers from Experiment I. These stimuli were chosen because the tokens of /ʌ/ by the speakers with smaller vocal tracts have formant values which are closer in frequency to the male adult's /ae/. Three-year-olds were tested to see if they were able to categorize these isolated tokens appropriately, despite the differences in absolute frequency, and with the absense of dynamic spectral or temporal cues, such as inherent duration or formant transitions.

A. Subjects

The subjects were 20 three-year-old children from the Bloomington, Indiana area; their ages ranged from 2;11.25 to 3;4.5 with a mean of 3;1. The children's names were obtained from birth announcements in the local newspaper, and their parents were solicited by mail and a follow-up telephone call. Subjects were tested twice a week for a total of 3-7 sessions and were paid $3 per session. None of the subjects from Experiment I participated in this experiment. An additional 40 subjects were dropped after completing 2 or more sessions, primarily due to inability to reach the testing criterion for this difficult contrast. Other reasons included illness and scheduling conflicts.

B. Stimuli

The stimuli were tokens of the vowels /ae/ and /ʌ/, taken from the words "had" and "hud", produced by the same talkers from Experiment I at a slow rate of speech. These words were recorded in a single-walled, sound-attenuated booth (I.A.C. model #401A) with an Ampex AG-500 tape recorder and an Electro-Voice D054 microphone. The stimuli were digitized on a PDP 11/34 computer with a low pass filter at 4.8 kHz and a sampling rate of 10 kHz. Vowels were isolated from the words using a waveform file editor (Luce and Carrell, 1981). These stimuli were also edited to control for duration (within one pitch period of 300 msec) and amplitude (63±2 db SPL), using the same procedures described in Experiment I.

Formant frequencies were measured at the midpoint of each stimulus using the same spectral analysis programs as in Experiment I (Kewley-Port, 1978; Carrell, forthcoming). The first three formant frequencies for each token are listed in Table 4 and the values for F1 and F2 are plotted in Figure 4.

Insert Table 4 about here

Insert Figure 4 about here

As this figure illustrates, the tokens of /ʌ/ by the female adult and the children have formant values which are closer in frequency to the male adult's production of /ae/.

C. Procedure

The procedure in this experiment parallels that of the first study. Ten subjects were assigned to the Constant Vowel group, in which all tokens of /ae/ were reinforced by one TV and all tokens of /ʌ/ by the other TV. Another ten subjects were assigned to the Mixed Vowel group, in which 2 /ae/s and 2 /ʌ/s were paired with each TV. The pairing of TVs and vowels was counterbalanced across subjects. The Constant Vowel stimulus sets and the Mixed Vowel stimulus sets are listed in Tables 5 and 6, respectively.

Insert Tables 5 and 6 about here

As in Experiment I, a criterion of 90% correct on 10 consecutive trials was required to proceed from the testing to the categorization phase. The results are based on the subjects' responses to 10 trials of each vowel token, or a total of 80 trials per subject.

D. Results and Discussion

As in the first experiment, the Constant Vowel group had a higher percentage of correct responses than the Mixed Vowel group. Subjects in the Constant Vowel group responded correctly on an average of 80% of the trials, while subjects in the Mixed Vowel group were correct on only 51% of the trials (see Figure 5).

Table 4

Formant Measurements - Experiment II

Speaker	/ae/	/ʌ/
Male Adult (RJ)	F1 = 704	F1 = 527
	F2 = 1602	F2 = 1230
	F3 = 2040	F3 = 2068
Female Adult (MB)	F1 = 773	F1 = 764
	F2 = 2167	F2 = 1636
	F3 = 2795	F3 = 3056
Male Child (JS)	F1 = 1039	F1 = 822
	F2 = 2118	F2 = 1614
	F3 = 2538	F3 = 3096
Female Child (TS)	F1 = 1088	F1 = 938
	F2 = 1957	F2 = 1686
	F3 = 2533	F3 = 3793

Table 4: The first 3 formant measurements in Hz for isolated tokens of the vowels /ae/ and /ʌ/ produced by a male adult (RJ), a female adult (MB), a male child (JS, age 12), and a female child (TS, age 10). Measurements were taken at the midpoint of each stimulus.

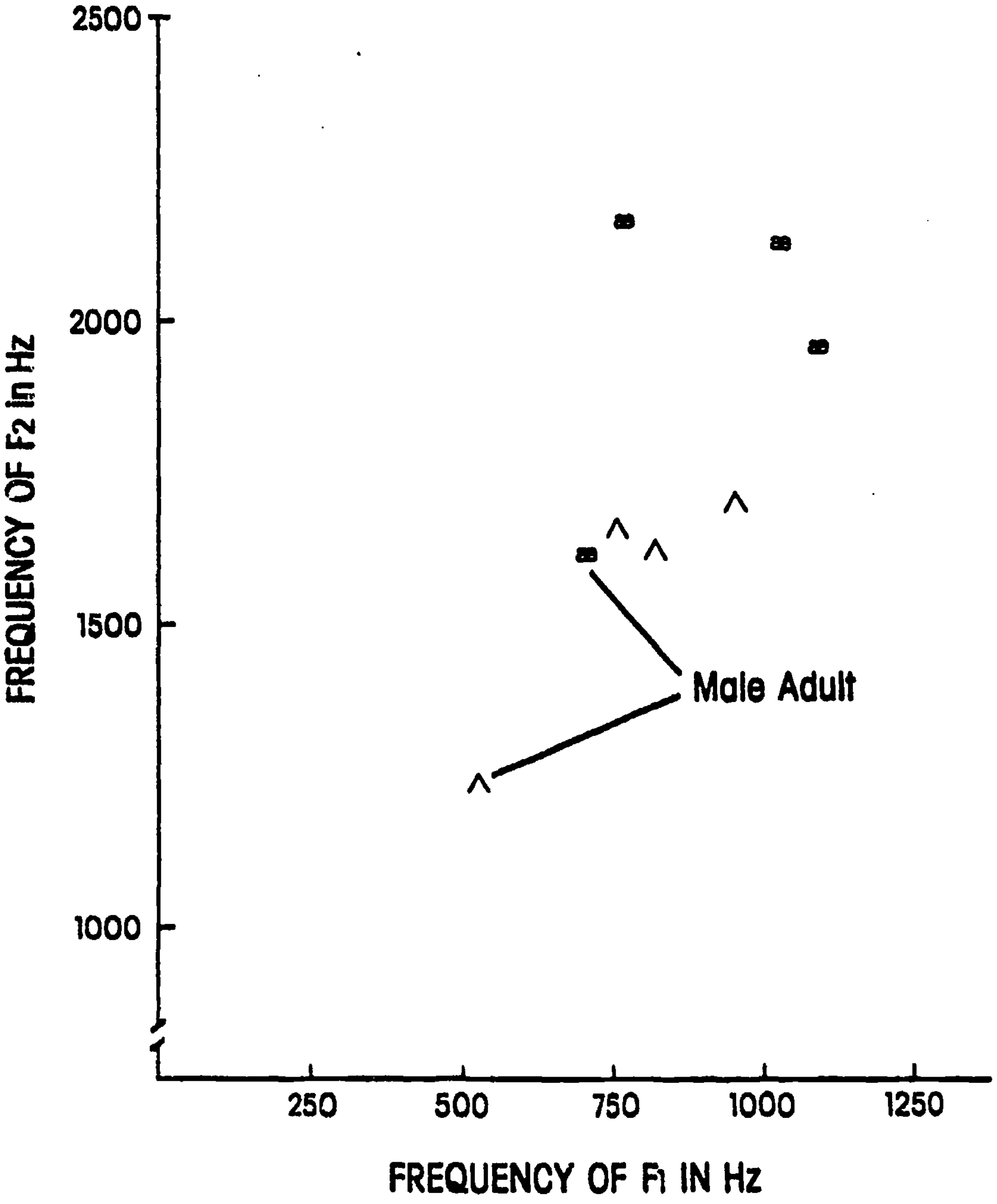


Figure 4: F1 and F2 measurements plotted for tokens of the vowels /ae/ and /ʌ/ produced by a male adult, a female adult, a male child, and a female child. The male adult's vowel tokens are labeled.

Table 5

Constant Vowel Stimulus Sets

	Left TV	Right TV
A.	1. Male Adult /ae/	5. Male Adult /ʌ/
	2. Female Adult /ae/	6. Female Adult /ʌ/
	3. Male Child /ae/	7. Male Child /ʌ/
	4. Female Child /ae/	8. Female Child /ʌ/
B.	1. Male Adult /ʌ/	5. Male Adult /ae/
	2. Female Adult /ʌ/	6. Female Adult /ae/
	3. Male Child /ʌ/	7. Male Child /ae/
	4. Female Child /ʌ/	8. Female Child /ae/

Table 5: Constant Vowel stimulus sets used in Experiment II. The male adult tokens, numbers 1 and 5 in each set, were the training stimuli.

Table 6

Mixed Vowel Stimulus Sets

	Left TV	Right TV
A.	1. Male Adult /ae/	5. Male Adult /ʌ/
	2. Female Adult /ae/	6. Female Child /ʌ/
	3. Female Adult /ʌ/	7. Male Child /ae/
	4. Male Child /ʌ/	8. Female Child /ae/
B.	1. Male Adult /ae/	5. Male Adult /ʌ/
	2. Male Child /ae/	6. Female Child /ʌ/
	3. Female Adult /ʌ/	7. Female Adult /ae/
	4. Male Child /ʌ/	8. Female Child /ae/
C.	1. Male Adult /ae/	5. Male Adult /ʌ/
	2. Female Child /ae/	6. Female Child /ʌ/
	3. Female Adult /ʌ/	7. Female Adult /ae/
	4. Male Child /ʌ/	8. Male Child /ae/
D.	1. Male Adult /ae/	5. Male Adult /ʌ/
	2. Female Child /ae/	6. Male Child /ʌ/
	3. Female Adult /ʌ/	7. Female Adult /ae/
	4. Female Child /ʌ/	8. Male Child /ae/
E.	1. Male Adult /ae/	5. Male Adult /ʌ/
	2. Female Child /ae/	6. Female Adult /ʌ/
	3. Male Child /ʌ/	7. Female Adult /ae/
	4. Female Child /ʌ/	8. Male Child /ae/
F.	1. Male Adult /ʌ/	5. Male Adult /ae/
	2. Male Child /ʌ/	6. Female Adult /ae/
	3. Male Child /ae/	7. Female Adult /ʌ/
	4. Female Child /ae/	8. Female Child /ʌ/
G.	1. Male Adult /ʌ/	5. Male Adult /ae/
	2. Female Adult /ʌ/	6. Female Adult /ae/
	3. Male Child /ae/	7. Male Child /ʌ/
	4. Female Child /ae/	8. Female Child /ʌ/

Table 6 continued

H.	1. Male Adult /ʌ/	5. Male Adult /ae/
	2. Male Child /ʌ/	6. Male Child /ae/
	3. Female Adult /ae/	7. Female Adult /ʌ/
	4. Female Child /ae/	8. Female Child /ʌ/
I.	1. Male Adult /ʌ/	5. Male Adult /ae/
	2. Female Adult /ʌ/	6. Male Child /ae/
	3. Female Adult /ae/	7. Male Child /ʌ/
	4. Female Child /ae/	8. Female Child /ʌ/
J.	1. Male Adult /ʌ/	5. Male Adult /ae/
	2. Female Adult /ʌ/	6. Female Child /ae/
	3. Female Adult /ae/	7. Male Child /ʌ/
	4. Male Child /ae/	8. Female Child /ʌ/

Table 6: Mixed Vowel stimulus sets for Experiment II. The male adult tokens, numbers 1 and 5 in each set, were the training stimuli.

Insert Figure 5 about here

As shown in Figure 6, the performance of Constant Vowel group averaged 81% correct on stimulus set I, and 80% correct on stimulus set II. In contrast, the Mixed Vowel group performed at a level of 66% correct on stimulus set I, and 36% correct on stimulus set II.

Insert Figure 6 about here

A two factor, mixed design ANOVA with repeated measures was used to analyze these data. There was a significant difference between the subject groups ($F(1,18) = 30.64$, $p < .001$) and stimulus sets ($F(1,18) = 10.91$, $p < .005$), and a significant stimulus set x group interaction ($F(1,18) = 9.57$, $p < .01$). Posthoc F tests for simple effects showed that the difference between the Constant Vowel group and the Mixed Vowel group for stimulus set I was not significant ($F(1,18) = 3.7$, $p > .05$), and that the interaction was due to the difference between the Constant Vowel and Mixed Vowel groups on stimulus set II ($F(1,18) = 45.00$, $p < .001$).

These results show that both groups were generalizing their pointing responses from the training stimuli to the similar vowel tokens from the other talkers, despite the formant frequency overlap between vowel categories. Thus, these data provide evidence of vowel normalization for isolated, quasi-steady state vowel tokens by very young children.

V. General Discussion

In both experiments, three-year-olds were able to spontaneously generalize a directional response from the male adult vowel tokens to the appropriate vowels produced by a female adult and two children. In the first experiment, subjects readily categorized the vowels /a/ and /i/ across talkers, thus demonstrating their perceptual constancy for these isolated tokens. Stimuli within vowel categories shared a similar spectral pattern, although the absolute frequencies differed. These vowels were categorized by vowel quality even by subjects in the Mixed Vowel group, who were reinforced for responding to the stimuli in a different way. This suggests that vowel quality was a particularly salient dimension despite the variation in other aspects of the stimuli, such as voice quality and fundamental frequency.

Similar results were obtained in the second experiment for the isolated, quasi-steady-state tokens of the vowels /ae/ and /ʌ/. These vowels were chosen specifically because of the formant frequency overlap between the male adult's /ae/ and the other talkers' productions of /ʌ/. Despite this overlap, subjects

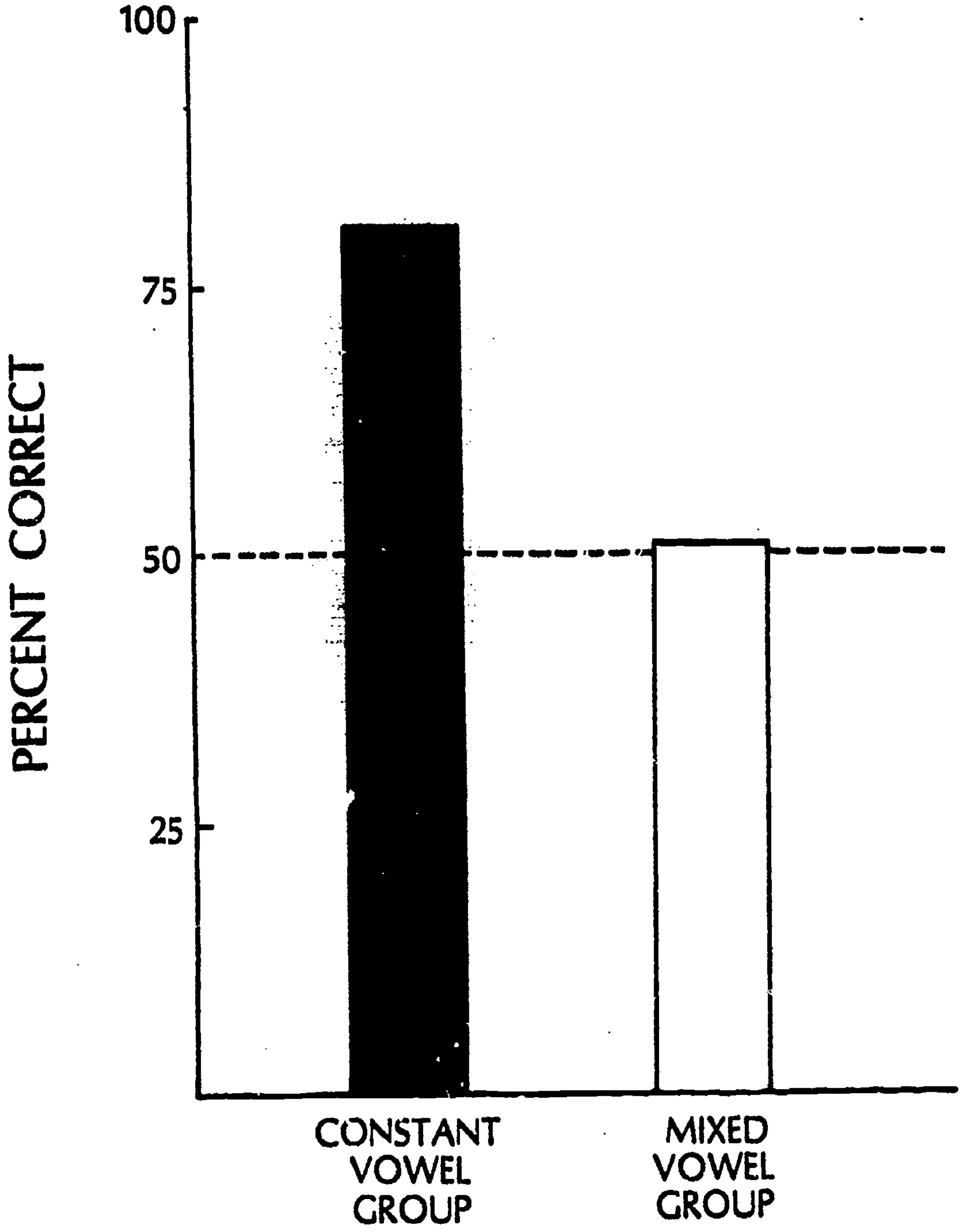


Figure 5: Mean percentage of correct responses for the Constant Vowel group and the Mixed Vowel group in Experiment II.

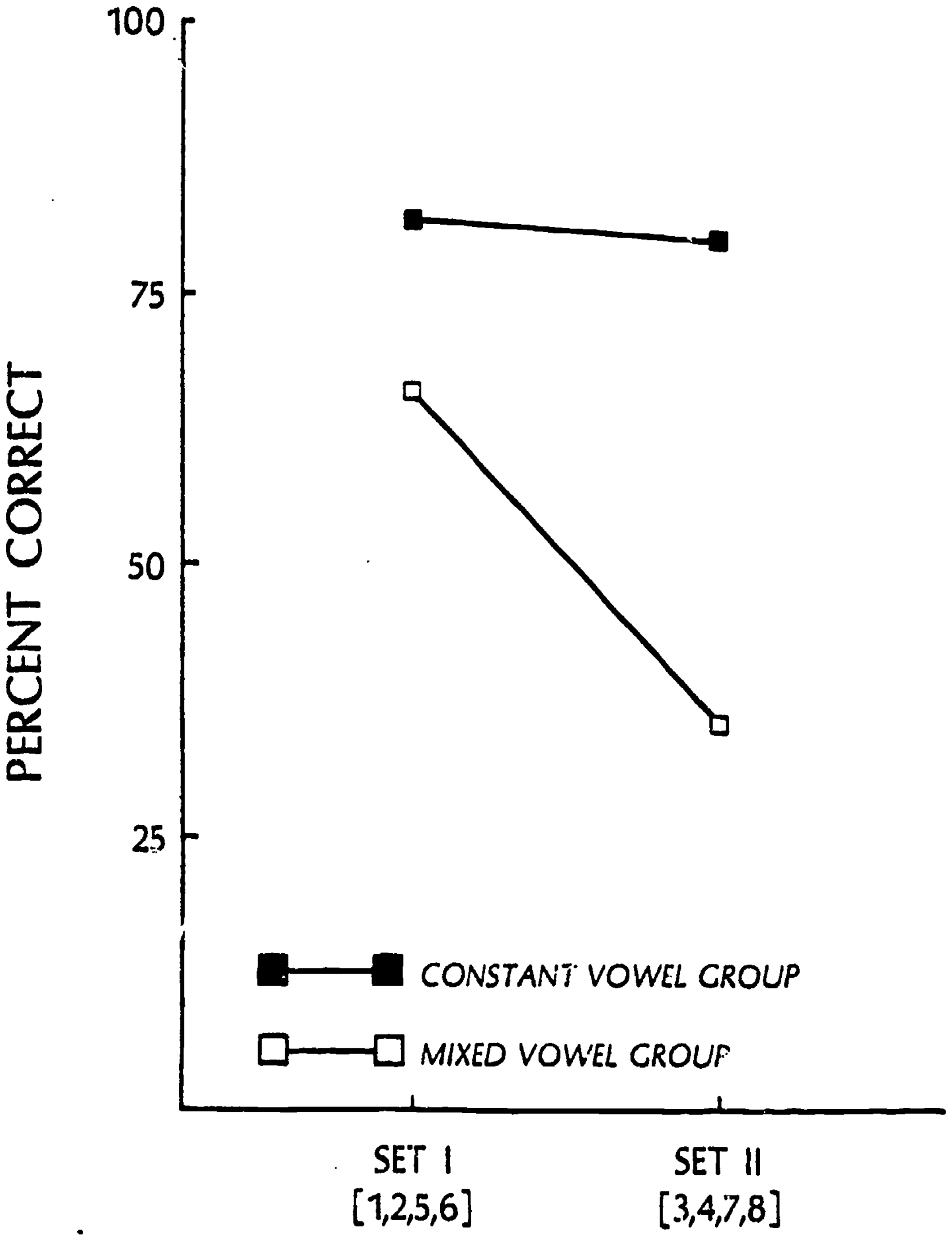


Figure 6: Mean percentage of correct responses to set I and set II stimuli for the Constant Vowel group and the Mixed Vowel group in Experiment Ii.

categorized the female adult and the two children's tokens of /ʌ/ with the male adult /ʌ/, rather than with /ae/, which contained more similar formant values. Of particular importance is that this pattern of results held for the Mixed Vowel group, who were not reinforced for categorizing the vowels by vowel quality. Taken together, these results demonstrate that steady-state formant information was sufficient for vowel normalization by three-year-olds.

Although the results of Experiment II are the first empirical demonstration of vowel normalization in pre-schoolers, one could argue from the relatively sophisticated speech production skills of 3-year-olds that vowel normalization must have been operative at an earlier age. However, it is important to point out that our subjects were confronted with a task in which both dynamic spectral information and semantic context were absent. It is quite likely that very young children rely heavily on prosody and non-phonemic information, such as situational context and gestures, to extract the meaning of words, particularly since identification of minimal pairs is rarely required of the listener in the natural environment.

Several researchers have emphasized the role of dynamic spectral and temporal information in vowel normalization by adults, but it is not clear whether this information is more beneficial to young children's vowel normalization than the quasi-steady-state information contained in the isolated tokens used in our experiments. In a study investigating children's use of VOT and first formant transition as cues to the voiced/voiceless distinction, Simon and Fourcin (1978) found that English-speaking children did not make use of the first formant transition cue until 4-5 years of age, and that French-speaking children did not use it at all. Greenlee (1980) found that three-year-olds did not use vowel duration, another to determine the voicing quality of a final consonant and Krause (1982) found that three-year-olds required significantly longer durations than six-year-olds or adults in order to make this judgment. Also, Strange, Jenkins, and Johnson (1983) found that listeners of an unfamilar dialect made more errors on a vowel identification task when steady state formant information was absent. The situation of an adult trying to identify vowels of a strange dialect might be considered analogous to that of a young child with relatively less language experience identifying vowels in his own dialect. In addition, steady-state information may be available more often to a child listener than to an adult since a prominent characteristic of "motherese" is a slower rate of speech compared to adult-adult conversation (Fraser and Roberts, 1975; Garnica, 1977).

Finally results from these experiments indicate that our new two-alternative, forced choice procedure is a useful one for assessing sound categorization in young children. This procedure can also be adapted to test young children's identification of speech sound continua by changing the reinforcement contingencies during the categorization phase of testing. During an initial testing stage, children would hear the TV pairings for the two endpoint stimuli from the continuum. During the labeling phase, intermediate stimuli from the continuum would be introduced. To eliminate learning effects when the child labeled intermediate tokens from the continuum, all pointing responses to the intermediate stimuli would be reinforced. However, if the endpoint stimuli were not categorized correctly during the labeling phase, the child would be automatically returned to the testing phase to relearn the correct associations. With this modified procedure, we intend to explore children's identification cf various synthetic continua, as well as their perception of cue trading relationships.

References

Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. Journal of the Acoustical Society of America, 68, 975-989.

Bernacki, B. (1981). WAVMOD: A program to modify digital waveforms. Research on Speech Perception: Progress Report No. 7, Indiana University, 275-286.

Carrell, T. D. (forthcoming). The effects of fundamental frequency, formant spacing, and glottal waveform on the perception of talker identity. Doctoral dissertation, Indiana University.

Diehl, R. L., McCusker, S. B., & Chapman, L. S. (1981). Perceiving vowels in isolation and in consonantal context. Journal of the Acoustical Society of America, 69, 239-248.

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

Fraser, C. & Roberts, N. (1975). Mothers' speech to children of four different ages. Journal of Psycholinguistic Research, 4, 9-16.

Garnica, O. (1977). Some prosodic and paralinguistic features of speech to young children. In C. Snow & C. Ferguson (Eds.) Talking to children: Language input and acquisition (pp.63-88), Cambridge: Cambridge University Press.

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

Gottfried, T. and Strange, W. (1980). Identification of coarticulated vowels. Journal of the Acoustical Society of America, 68, 1626-1635.

Greenlee, M. (1980). Learning the phonetic cues to the voiced-voiceless distinction: A comparison of child and adult speech perception. Journal of Child Language, 7, 459-468.

Jenkins, J. J., Strange, W. & Edman, T. R. (1983). Identification of vowels in "vowelless" syllables. Perception and Psychophysics, 34, 441-450.

Joos, M. A. (1948). Acoustic phonetics. Language Supplement 2, 24, 1-136.

Kewley-Port, D. (1978). SPECTRUM: A program for analyzing the spectral properties of speech. Research on Speech Perception: Progress Report No. 5, Indiana University, 475-492.

Krause, S. E. (1982). Vowel duration as a perceptual cue to postvocalic consonant voicing in young children and adults. Journal of the Acoustical Society of America, 71, 990-995.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America*, *66*, 1668-1679.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, *6*, 263-285.

Kuhl, P. K. & Miller, J. D. (1982). Discrimination of auditory target dimensions in the presence or absense of variation in a second dimension by infants. *Perception and Psychophysics*, *31*, 279-292.

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

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

Luce, P. A. & Carrell, T. D. (1981). Creating and editing waveforms using WAVES. *Research on Speech Perception: Progress Report No. 7*, Indiana University, 287-298.

Macchi, M. (1980). Identification of vowels spoken in isolation versus vowels spoken in consonantal context. *Journal of the Acoustical Society of America*, *68*, 1636-1642.

Nearey, T. M. (1977). Phonetic feature systems for vowels. Doctoral dissertation, University of Connecticut, 1977 (reproduced by Indiana University Linguistics Club, 1978).

Nearey, T. M. (1983). Vowel-space normalization procedures and phone-preserving transformation of synthetic vowels. *Journal of the Acoustical Society of America Supplement 1*, *74*, S17.

Simon, C. & Fourcin, A. J. (1978). Cross-language study of speech-pattern learning. *Journal of the Acoustical Society of America*, *63*, 925-935.

Shankweiler, D. P., Strange, W., & Verbrugge, R. R. (1977). Speech and the problem of perceptual constancy. In R.E. Shaw & J. Bransford (Eds.) *Perceiving, acting and knowing: Toward an ecological psychology*, (pp.315-345), Hillsdale, NJ: Erlbaum.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. (1976). Consonant context specifies vowel identity. *Journal of the Acoustical Society of America*, *60*, 213-224.

Strange, W., Edman, T. R. & Jenkins, J. J. (1979). Acoustic and phonological factors in vowel identification. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 643-656.

Strange, W. & Gottfried, T. (1980). Task variables in the study of vowel perception. *Journal of the Acoustical Society of America*, *68*, 1622-1625.

Strange, W. & Broen, P. A. (1981). The relationship between perception and production of /w/, /r/, and /l/ by three-year-old children. _Journal of Experimental Child Psychology_, _31_, 81-102.

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. _Journal of the Acoustical Society of America_, _74_, 695-705.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? _Journal of the Acoustical Society of America_, _60_, 198-212.

Effects of speech rate and pitch contour on the perception of synthetic speech*

Louisa M. Slowiaczek

and

Howard C. Nusbaum

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

*This research was supported by NIH grant NS-12179 and NIH training grant NS-07134 to Indiana University and a contract with the Air Force Office of Scientific Research. We would like to thank David B. Pisoni for valuable comments and discussion.

Abstract

Research on the perception of fluent speech has become increasingly concerned with the role prosodic information plays in the perceptual analysis of the acoustic signal. Two experiments were conducted to examine the effects of pitch contour and speech rate on the perception of synthetic speech. In Experiment 1, subjects transcribed sentences that were either syntactically correct and meaningful or syntactically correct but semantically anomalous. In Experiment 2, subjects transcribed sentences that varied in length and syntactic structure. In both experiments a text-to-speech system generated synthetic speech at either 150 or 250 words per minute. Half of the test sentences were generated with a flat pitch (monotone) and half were generated with "normally inflected" clausal intonation. The results indicate that the identification of words in fluent synthetic speech is influenced by speaking rate, meaning, length, and to a lesser degree, pitch contour. Decreased performance appears to be primarily a result of a degradation of the acoustic signal. The comparison of these results with research using natural speech stimuli suggests a difference in perceptual performance between natural and synthetic speech as speaking rate is increased. However, for pitch contour, no difference in perceptual performance was found between synthetic speech stimuli and natural speech stimuli used in previous investigations.

Effects of speech rate and pitch contour on the perception of synthetic speech

The production and perception of fluent speech requires the interaction of several different sources of linguistic knowledge. In addition to phonetic, lexical, syntactic, and semantic information, the speech waveform contains important suprasegmental properties such as rhythm, stress, and intonation. In the past, most of the research on the perception of speech has concentrated on investigations of segmental perception to determine which acoustic cues are most important to a listener (see Darwin, 1976; Pisoni, 1978; for reviews). Recently, however, researchers have begun to study more closely the role of suprasegmental factors in speech perception (Cutler, 1976; Nooteboom, Brokx, & De Rooij, 1978; Pierrehumbert, 1979; Wingfield, 1975). This shift in research interest has resulted in the examination of how prosodic factors interact with the processing of segmental information in the perception of fluent speech.

Researchers have studied the role of prosodic information in both production and perception, investigating such specific prosodic factors as intonation contour (Cooper & Sorensen, 1977; Dooling, 1974; Lehiste, 1970; Pierrehumbert, 1979; Wingfield, 1975), segmental duration (Klatt, 1976; Cooper & Paccia-Cooper, 1980; Lehiste, 1970; Lehiste, Olive & Streeter, 1976), and stress placement (Cutler, 1976; Cutler & Darwin, 1981; Lehiste, 1970). Other research has been concerned with the effect of these prosodic factors on various perceptual tasks including the parsing of sentences (Collier & t'Hart, 1975; Nakatani & Schaffer, 1978), phoneme monitoring (Cutler, 1976), disambiguating sentences (Lehiste, Olive & Streeter, 1976), and segmental identification (Martin, 1979; Summerfield, 1975). One general conclusion that can be drawn from this research is that prosodic factors do play an important role in the production and perception of fluent speech. However, the exact nature of that role is not yet clear. Prosody may guide the perception of phonemes, aid in syntactic parsing and comprehension of sentences, and/or facilitate chunking of sentences in short-term memory. However, it is not clear whether suprasegmentals are necessary for acoustic analysis and subsequent linguistic processing or only supplemental to such processing. To the extent that prosodic factors are necessary, are certain variables more critical than others? These are only a few of the issues that must be considered by researchers studying the role of prosody in language processing.

One suprasegmental variable that has been studied extensively is speaking rate. Recent technological advances have made it possible to produce speech at rates beyond the abilities of humans to understand it (deHaan, 1977). Accelerating speech increases the word rate (number of words per minute) and decreases the message duration. Speech which is accelerated allows listeners to perceive spoken text at rates similar to silent reading. This has several practical applications, including the the development of reading machines for the blind and machines that proofread texts for copy editors.

In the past, three methods have been used to accelerate speech (for a detailed review see Foulke & Sticht, 1969). The simplest method relies on the fact that word rate can be consciously controlled by a speaker. Although it is relatively easy for a human talker to modify his or her speaking rate, this method is constrained by the rate at which speech sounds can be articulated. In addition, when speaking faster, other factors such as vocal inflection, intensity, and the relative durations of consonants and vowels are affected as well.

Another method for increasing speaking rate involves recording a message and then reproducing it at a tape speed which is faster than that used in the original recording. Accelerated speech produced using this method is called "speeded speech" (Foulke & Sticht, 1969). Accelerating speech in this way, however, shifts the formant and fundamental frequencies of the signal and thus the perceived pitch of the speech. Thus, if the speech rate is doubled, the component frequencies will be doubled, and vocal pitch will increase by one octave (Foulke & Sticht, 1969).

A third method takes advantage of the redundancy of the speech signal. Miller and Licklider (1950) removed sections of a speech signal at regular intervals and tested subjects' perception of the modified (interrupted) speech. They found that when the speech signal was interrupted in this way during its reproduction, intelligibility did not drop below 90% until over 50% of the original signal had been removed. This finding prompted Garvey (1953) to develop the sampling method of speech acceleration in which small sections of the speech signal are chopped out of the waveform and the remaining sections of the signal spliced together. By removing information at regular intervals in the waveform, Garvey was able to accelerate the speech rate without affecting other factors such as the formant and fundamental frequencies of the signal. Because this method of increasing speech rate involves shortening the duration of the speech waveform by eliminating redundancy and compressing the remaining information in time, this type of accelerated speech is referred to as "time-compressed."

More recently, two additional methods of accelerating speech have been developed. Linear predictive coding (LPC) of speech involves mathematically regenerating speech waveforms (Atal & Hanauer, 1971). When the speech is synthesized, a change in the frame rate produces a change in the rate of speech. Lastly, text-to-speech synthesizers with the capability to change speaking rate have been used to accelerate speech. These synthesizers often increase speaking rate in ways similar to humans ---vowel durations are decreased to a greater extent than consonant durations.

In order for accelerated speech to be used in practical applications (e.g., reading machines for the blind), however, the perception of such speech must be evaluated. Garvey (1953) evaluated the intelligibility of speech produced by time compression (sampling) and found that speakers could understand speech accelerated as much as 2.5 times the original speech rate with less than a 10% loss in intelligibility. More recently, deHaan (1977, deHaan & Schjelderup, 1978) has described a method for obtaining thresholds of intelligibility for time-compressed speech. Studies conducted by deHaan and his associates suggest that speech compressed with a sampling method (time-compressed speech) is more intelligible than speeded speech that contains pitch distortions. In other words, time-compressed speech is more intelligible at higher speech rates than speeded speech which contains pitch distortions.

A distinction is often made in perception between evaluating the intelligibility of accelerated speech and evaluating the comprehension of accelerated speech (Foulke & Sticht, 1969). An index of intelligibility of time-compressed speech is the ability to repeat a word, phrase, or short sentence accurately. On the other hand, the evaluation of the effects of comprehension involves a subject listening to a passage of speech which has been time-compressed and then testing the listener's comprehension of that selection with objective tests (Foulke & Sticht, 1969). As with intelligibility measures, it has often been found that comprehension declines as speech rate increases. However, unlike intelligibility measures, at speech rates above 250 to 275 words per minute, Foulke and Sticht (1969) reported a rapid decline in comprehension. These researchers suggested that the increase in rate at which comprehension declines above 275 words per minute may be due to factors in addition to signal degradation. Specifically, at speech rates greater than 275 words per minute, words may not be processed as fast as they are received resulting in some loss of the linguistic information in the acoustic signal.

Fundamental frequency (FO) contour is another suprasegmental variable that has been studied by a number of researchers. Production studies (Cooper & Sorensen, 1977) provide evidence that fundamental frequency contour is influenced by a speaker's internal representation of the syntactic structure of an utterance. This suggests that fundamental frequency may be important in organizing the syntactic processing of an utterance. In addition, several studies have also been conducted to examine the contribution of fundamental frequency (FO) contour to speech perception (Cutler, 1976; Larkey & Danly, Note 1; Wingfield, 1975; Wingfield & Klein, 1971).

Wingfield and Klein (1971) were interested in testing the hypothesis that intonation pattern serves to reduce uncertainty about sentence structure at or around clausal boundaries. Half of the sentences in their study were produced with normal intonation such that the prosody and syntactic structure were in agreement. The other half were cross-spliced so that the intonation boundary and the major syntactic boundary in the sentences did not coincide (i.e., anomalous conflicting intonation). Subjects heard the sentences over headphones and were required to indicate when the sentences were switched from one ear to the other ear during the presentation of each sentence. In the normal intonation condition, no difference was found in localization accuracy between sentences which were switched at the syntactically defined boundary and sentences which were switched before or after the syntactically defined boundary. However, in the anomalous intonation sentences, localization was significantly more accurate when the switch occurred at the intonationally marked boundary than when it occurred before or after that point. In short, subjects were not able to accurately localize a switch that occurred at a syntactically marked boundary when the intonation and syntax were not in agreement. Wingfield and Klein argued that subjects were relying on the intonation pattern to reduce uncertainty about the syntactic structure of the sentences. Intonation appeared to supply redundant information to aid the perception process. When the intonation boundary was in conflict with the syntactic boundary, subjects were unable to accurately localize a switch in ear of presentation if the switch occurred before or after the intonation boundary. In addition, Wingfield and Klein found a slight recall advantage for sentences with normal intonation over sentences with anomalous intonation.

Wingfield and Klein reasoned that if intonation provides redundancy to the sentence recognition process, then its effects should be magnified if the sentences were presented under conditions of increased information load. To test this hypothesis, Wingfield (1975) time-compressed the sentences used in the original Wingfield and Klein study. Subjects in this second study listened carefully to each sentence and reported as accurately as possible what was heard. Two findings were obtained. First, prosodic cues aid in the assignment of syntactic structure. Specifically, the errors made by subjects indicated that perception of the structure of sentences was determined as much by prosodic features as by the formal (syntactic) structure of the sentences. Second, word recognition accuracy was reduced with increases in speaking rate. However, prosodic information in the speech apparently facilitated the recognition of words when the prosody was appropriate for the syntactic structure, despite the fact that segmental intelligibility was impaired by increased speaking rate.

More recently, Larkey and Danly (Note 1) were also interested in assessing the contribution of fundamental frequency contour to the perception of speech. In their experiment, subjects performed a sentence verification task (true/false judgment) on declarative sentences. Reaction times to verify the truth of the sentences with normal FO contour were 48 msec faster than reaction times to verify the same sentences with monotone intonation contours. Although this result was small and based on a fairly gross post-perceptual measure of comprehension performance, it was statistically significant. Larkey and Danly argued that their findings were meaningful since their experiment was an extremely conservative test of the role of prosody in speech perception. The fact that they found an effect of FO contour, even though the sentences in their experiment were syntactically simple, suggests that the FO contour may be an important factor in understanding sentences, as well as perceiving them. Moreover, their data suggest that prosody may play an even greater role with more complex linguistic materials.

The research discussed thus far has been concerned with the effects of speech rate and pitch contour on the perception of natural speech. One general conclusion that can be drawn from these experiments is that, under the conditions studied, speech rate and pitch contour appear to influence speech perception in a number of ways. Specifically, an increase in speech rate results in a decrease in intelligibility of natural speech. This decrease in intelligibility may be related to a general degradation in the acoustic-phonetic specification of the speech signal caused by the shortening of phonetic segments as speech rate is increased (Miller, 1981). Cutler and Darwin (1981) suggested that fundamental frequency contour helps predict upcoming phonetic information when they found that subjects were faster to identify phonemes in target words which were stressed than in target words which were unstressed. Furthermore, fundamental frequency contour has been shown to aid in syntactic processing as well (Wingfield, 1975). Such assistance may be accomplished in two ways. First, the FO contour may provide redundant information which would allow listeners to predict the syntactic structure of the sentence and thus focus on the segmental and lexical information in the sentence that is most relevant to comprehension. Second, FO contour may enable the listener to organize the incoming signal to facilitate subsequent processing.

The study of prosodic factors need not be restricted to the domain of natural speech, however. A great deal of research has been conducted in recent years on the perception of synthetic speech stimuli. The general conclusion from research comparing the perception of natural speech with the perception of synthetic speech across a variety of experimental procedures is that natural speech is consistently perceived or recognized better than synthetic speech, even when the synthetic speech is of high quality (Pisoni, 1982, Note 2; Slowiaczek & Pisoni, Note 3). The decrease in performance of synthetic speech appears to be due to a variety of factors including the suprasegmental variables that have been shown to play a role in the perception of natural speech. That is, some of the difficulty encountered during the processing of synthetic speech may be caused by the absence of appropriate prosodic information or the presence of inappropriate prosodic information in the acoustic signal (Allen, 1981; Nickerson, 1975). Huggins (1978) measured the intelligibility of sentences synthesized by rule and found that sentences with an incorrect fundamental frequency contour showed a decrease to about two-thirds in intelligibility compared to sentences with correct fundamental frequency contour. An even greater effect was found when sentences had incorrect timing.

On the other hand, Bernstein (1977) found no effect of suprasegmental information when he modified high quality synthetic speech by modelling the segmental errors and the suprasegmental timing distortions of a particular deaf child's speech. The results of four indices of intelligibility showed that consonant distortions had a greater effect on intelligibility than suprasegmental factors. When the segmental information was good, no amount of suprasegmental distortion had an effect on intelligibility. However, even with perfectly natural timing, consonantal distortions in the speech reduced intelligibility.

The present experiments were designed to further investigate perception of synthetic speech and how it may be related to the absence of certain prosodic cues. In particular, we were interested in the effects of speaking rate and pitch contour on the perception of fluent synthetic speech. When speaking rate is increased, the phonetic segments encoded in the speech signal are modified in several ways which include shortening, reducing, and in some cases deleting acoustic-phonetic information. The shortening of phonetic segments causes intelligibility to be reduced. One consequence of this change in segmental information in the signal may be to overload the cognitive capacity of the listener (Dallett, 1964; Foulke & Sticht, 1969; Miller, 1956; Wingfield, 1975). If this is true, subjects should have a more difficult time processing the stimulus information when the speaking rate of the synthetic speech is increased, resulting in a decrease in intelligibility.

Pitch contour, on the other hand, may predict upcoming information and hence aid in phonetic and syntactic processing. In the absence of appropriate fundamental frequency and durational cues, we would expect phonetic perception and syntactic processing to be impaired. If these factors do influence perception as suggested, we would expect to find a trade-off betweent the use of segmental and suprasegmental cues as speech rate and pitch contour are manipulated. That is, perceptual performance should decrease as increases in speech rate affect the segmental information in the signal, and as the absence of useful pitch information affects the ability to predict subsequent phonetic and syntactic information.

Experiment 1

Experiment 1 was designed to investigate how pitch contour, speaking rate, and meaning affect the perception of synthetic speech. As observed with natural speech, an increa.e in speaking rate consistently resulted in a decrease in perceptual performance. If increased speaking rate causes an impairment in perception by degrading the speech signal, we would expect a greater dependence on the information provided by pitch contour. Similarly, the listener may be forced to rely more heavily on pitch information if certain linguistic information (e.g., semantic or syntactic information) is not present in the signal. To test this prediction, syntactically correct but semantically anomalous sentences and syntactically correct and meaningful sentences were used in the first experiment. If the absence of semantic information increases the processing difficulty for a listener, we would expect a greater reliance on the available pitch information for the anomalous sentences than for the meaningful sentences. Given these variables, Experiment 1 could provide some initial support for the hypothesis that there is a tradeoff of segmental and suprasegmental cues in speech perception.

Method

Subjects. Forty-eight undergraduate students from Indiana University participated in the experiment in partial fulfillment of an introductory psychology course requirement. All subjects were native English speakers with no reported history of hearing loss or speech disorders.

Materials. Twenty-five syntactically correct and meaningful sentences and twenty-five syntactically correct but semantically anomalous sentences were used as stimulus materials in this experiment. The meaningful sentences contained five content words. The anomalous sentences contained four content words (Egan, 1948; Nye & Gaitenby, 1974). Table 1 shows some examples of these sentences.

Insert Table 1 about here

Each of the fifty test sentences was typed into computer files which served as input to a Speech Plus Prose-2000 text-to-speech system. The sentences were produced at two speech rates: 150 words per minute (slow) and 250 words per minute (fast). The speech rate for each condition was determined by setting a rate parameter cn the Prose-2000 (Telesensory Speech Systems, 1982). The text-to-speech system increased speaking rate by an algorithm which made reference to the syntactic structure and stress pattern of the sentence being

Table 1

Example Sentences for Experiment 1

Sentence Type	Example
Meaningful	The hogs were fed chopped corn and garbage.
	Rice is often served in round bowls.
Anomalous	The old corn cost the blood.
	The end home held the press.

synthesized, as well as to the prosodic parameters selected prior to synthesis. In addition to the speech rate manipulation, one version of each of the sentences was generated with a flat (monotone) pitch and one version was generated with a "normally" inflected pitch as determined by the pitch algorithm of the Prose 2000 system. The factorial combination of these two variables produced four experimental conditions: (1) slow rate with inflected pitch, (2) slow rate with monotone pitch, (3) fast rate with inflected pitch, and (4) fast rate with monotone pitch. The sentences were recorded on audio tape, passed through a 4.8 kHz low-pass filter, and digitized with an A-D converter at 10 kHz. All experimental stimuli were subsequently stored as digital stimulus files on a computer disk for later presentation to subjects in the experiment.

Procedure. Twelve subjects were tested in each of the four experimental conditions for a total of 48 subjects. Subjects were run in small groups in an experimental testing room. The presentation of all stimuli was controlled on-line by a PDP 11/34 computer. The stimuli were presented at 80 dB SPL over a pair of TDH-39 headphones. Each subject heard all 50 sentences in one of the four experimental conditions. The sentences were presented in two blocks so that one block consisted of the 25 meaningful sentences and the other block consisted of the semantically anomalous sentences. The order of presentation of the blocks was counterbalanced and the order of the sentences in each block was randomized for each session. Subjects were instructed to listen carefully to each sentence and to write down every word they heard. Subjects were encouraged to write something on every trial, even if they had to guess. Responses were recorded in prepared answer booklets.

A typical trial sequence proceeded as follows: First a cue light was presented for one second at the top of a subject's response box to indicate the beginning of the trial. A sentence was presented over the headphones 500 msec after the cue light went out. Subjects responded by writing the sentence next to the appropriate trial number on their answer sheets. Subjects indicated that they were done responding by pushing a button on the respon.e box in front of them. After all of the subjects in a group responded or when the maximum response time of 30 seconds had elapsed, the computer program automatically initiated the next trial.

Results and Discussion

The percentage of words correctly identified was determined for each of the two types of sentences (meaningful and semantically anomalous) for each of the experimental conditions. Only the content words in each of the sentences were scored. Words were scored as correct if the response matched the entire target word. Omission or addition of affixes were scored as incorrect. Words with alternate spellings that phonetically matched the target word were scored as correct (e.g., gram, graham). An analysis of variance was performed on the percentages. Figure 1 shows percent correct identification as a function of speaking rate and pitch contour. The squares in the figure present identification performance for the meaningful sentences; the triangles show performance for the anomalous sentences. The dashed lines and open symbols show performance for sentences with monotone pitch; the solid lines and closed symbols show performance for sentences with inflected pitch.

Insert Figure 1 about here

As shown clearly in Figure 1, a significant main effect of rate ($\underline{F}(1,44) = 330.95$, $\underline{p} < 0.0001$) and meaning ($\underline{F}(1,44) = 189.94$, $\underline{p} < 0.0001$) was obtained. A significant interaction between the effects of speaking rate and meaning on the word identification performance ($\underline{F}(1,44) = 20.47$, $\underline{p} < 0.0001$) was also observed. This interaction indicates that the fast speaking rate had a more detrimental effect on the perception of the meaningful sentences than on the perception of the semantically anomalous sentences. At the slow rate, performance on the meaningful sentences was better than performance on the anomalous sentences. Unlike the meaningful sentences, performance with the anomalous sentences was so low at the slow rate that it could not get much lower when the rate was increased. Thus, the absence of a drop in performance for the anomalous sentences may have been due to a floor effect in performance.

A surprising result was the failure to find an effect of pitch contour at either speaking rate for the meaningful or the semantically anomalous sentences ($\underline{F}(1,44) < 1.0$). One explanation for the absence of a pitch contour effect here is that the sentences used in this experiment were much too simple (although cf. Larkey & Danly, Note 1). The meaningful sentences were all active sentences of one length and the syntactic frame of the anomalous sentences was constant (i.e., Determiner - Adjective - Noun - Verb - Determiner - Noun).

Therefore, contrary to our predictions, the anomalous sentences did not result in a greater reliance on pitch contour. However, meaningfulness and speech rate had an effect on the perception of these sentences. Clearly then, meaningfulness and speaking rate are more important to the perception of simple synthetic sentences than pitch inflection.

Experiment 2

In order to determine whether an effect of pitch contour could be obtained using more linguistically complex stimuli, a second experiment was conducted. In this experiment, we manipulated the syntactic structure and the length of the sentences used. If pitch contour aids in predicting syntactic structure of sentences or helps to organize the sentences in short-term memory, we would expect a greater impairment of processing for more complex sentences when pitch is not informative (monotone condition). Also, if cognitive load is increased when sentences become more complex, we might expect a somewhat greater reliance on pitch contour with longer sentences. Thus, in this experiment, we manipulated speech rate and pitch contour in more complex linguistic materials.

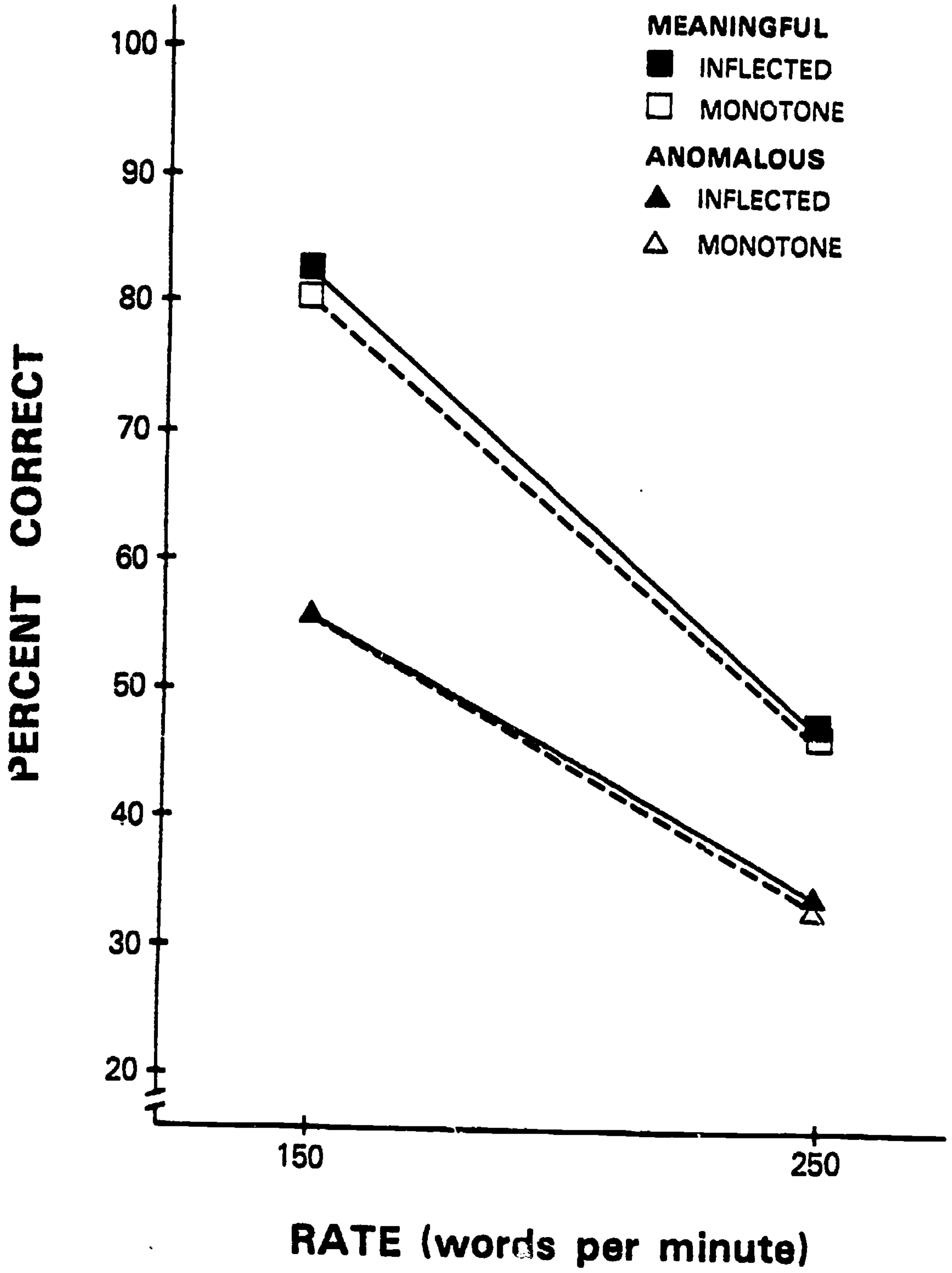


Figure 1. Percent correct identification for meaningful sentences and anomalous sentences as a function of rate of speech.

Method

Subjects. The subjects were sixty undergraduate students from Indiana University drawn from the same pool as in the first experiment. None of the subjects in this experiment had participated in Experiment 1.

Materials. Twenty active, 20 passive, and 20 center-embedded sentences were constructed for use in Experiment 2. Half of the sentences were short and half were long. Short sentences contained four content words and long sentences contained eight content words. Examples of the short and long sentences used are given in Table 2.

Insert Table 2 about here

To minimize the differences between sentences of different lengths or different sentence types (other than length or syntactic structure), the words in the sentences were equated on two additional variables. First, words used across different sentence types and lengths were matched in mean frequency based on the Kucera and Francis word counts (1967). The average frequency of words used in the three sentence types and two lengths were: active, 75.8; passive, 71.3; center-embedded, 65.8; short, 69.2; long, 72.8. In addition to word frequency, words were matched across sentence types and lengths on number of syllables. The average number of syllables per word used in the three sentence types were: active, 2.00; passive, 2.00; center-embedded, 1.92. The mean number of syllables was the same for the two sentence lengths--1.98 syllables per word.

As in Experiment 1, the sentences were generated by a Speech Plus Prose-2000 text-to-speech system. The sentences were produced at two different speaking rates (150 and 250 words per minute) and with two different pitch contours (monotone and inflected) by setting the appropriate rate and pitch parameters on the Prose-2000. The combination of these two variables resulted in the same four experimental conditions described in Experiment 1.

Procedure. The procedure was identical to the one used in Experiment 1, with the following exceptions. Each subject heard 60 sentences in one of the four experimental conditions. Fifteen subjects were tested in each of the four conditions for a total of 60 subjects.

Table 2

Example Sentences for Experiment 2

Length	Syntactic Type	Example
Short	Active	The compulsive clerk organized the reports.
	Passive	The scientific article was approved by the editor.
	Center-Embedded	The monkey the animal bit bled.
Long	Active	The angry customer returned the damaged merchandise to the department store manager.
	Passive	The expensive chrome sculpture was sold to the art gallery by the contemporary artist.
	Center-Embedded	The apathetic student the concerned dean advised failed the English test.

Results and Discussion

The percentage of words correctly identified in each of the experimental conditions was determined for the three sentence types (active, passive, and center-embedded) and two sentence lengths (short and long). The scoring was conducted as in Experiment 1. An analysis of variance was performed on the percentages. Figure 2 displays the results of the second experiment. Percent correct identification of words in the inflected pitch conditions is plotted in the left panel; the analogous monotone conditions are plotted in the right panel. The filled symbols correspond to the short sentences and the open symbols correspond to the long sentences. The squares represent the sentences presented at the slow speech rate and the triangles represent sentences presented at the fast speech rate.

Insert Figure 2 about here

The analysis of variance revealed a significant main effect of sentence length ($\underline{F}(1,56) = 185.61$, $\underline{p} < 0.0001$). Words in short sentences were identified consistently better than words in long sentences. In short sentences 78.4 percent of the words were correctly identified, while only 67.1 percent of the words in the long sentneces was correctly identified. Similarly, a significant effect of rate was observed ($\underline{F}(1,56) = 227.85$, $\underline{p} < 0.0001$). Words in slow sentences were consistently identified better than words in fast sentences. For slow sentences, 86.7 percent of the words were correctly identified and 58.9 percent of the words were correctly identified for fast sentences. In addition to the main effects for length and rate, a significant main effect was observed for syntax ($\underline{F}(2,112) = 39.47$, $\underline{p} < 0.0001$). The mean percent correct identification for words in active sentences was 75.2, for passive sentences, 74.2, and for center-embedded sentences, 68.8. Thus, correct word identification decreased with increasing syntactic complexity.

Finally, the analysis of variance revealed a main effect of pitch contour ($\underline{F}(1,56) = 6.47$, $\underline{p} < 0.01$). For pitch contour, inflected pitch produced 75.1 percent correct word identification and monotone pitch produced 70.4 percent correct identification. Thus, correct word identification was greater for the inflected pitch contour than for the monotone pitch contour. Clearly then, inflected pitch aided in word identification.

Insert Figure 3 about here

The only significant interactions among the variables were those involving syntax. The left panel of Figure 3 illustrates the syntax-by-length interaction ($\underline{F}(2,112) = 16.99$, $\underline{p} < 0.0001$). The solid line in this figure represents identification performance for words in short sentences. The dashed line

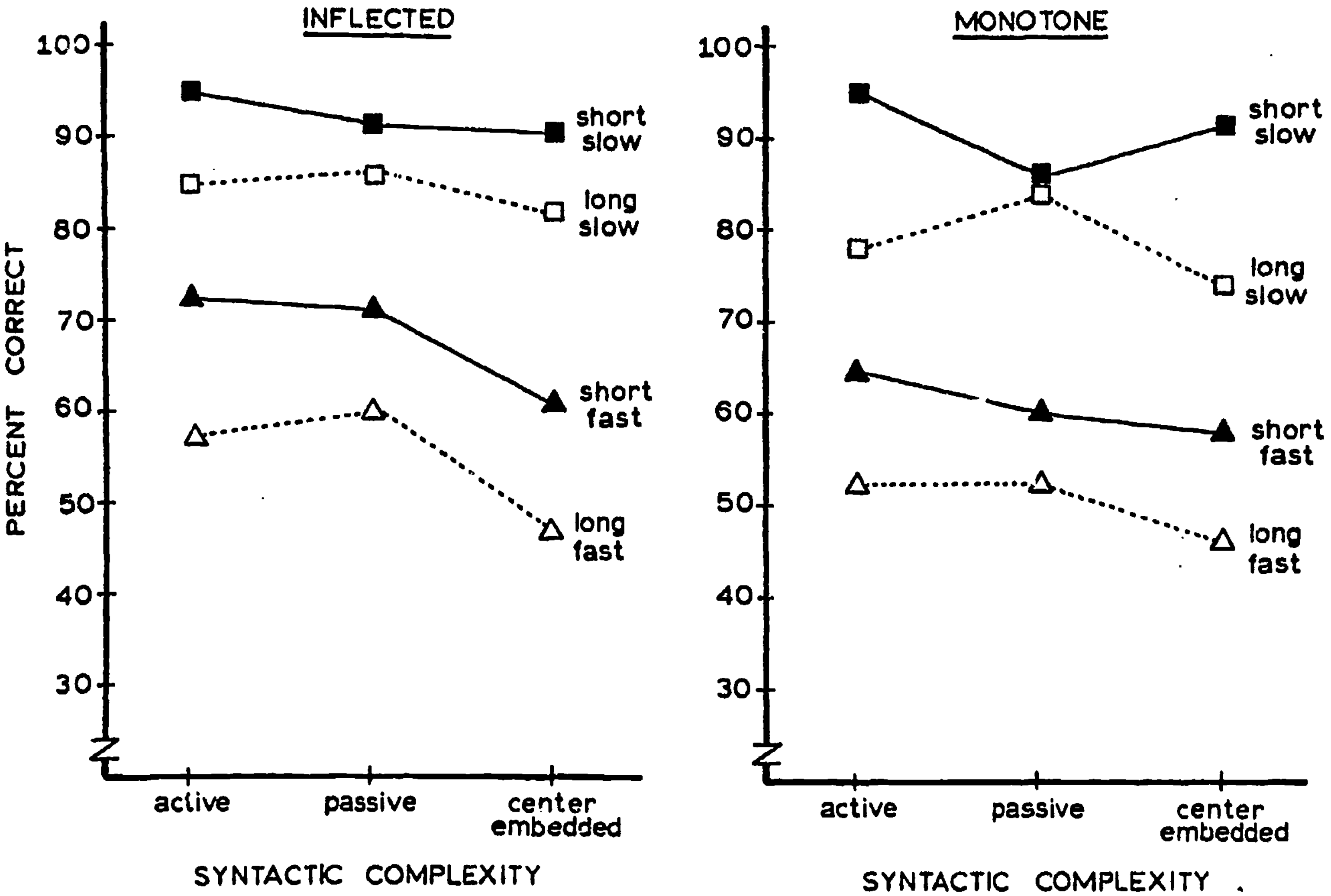


Figure 2. Percent correct identification for inflected conditions and monotone coditions as a function of syntactic comrlexity.

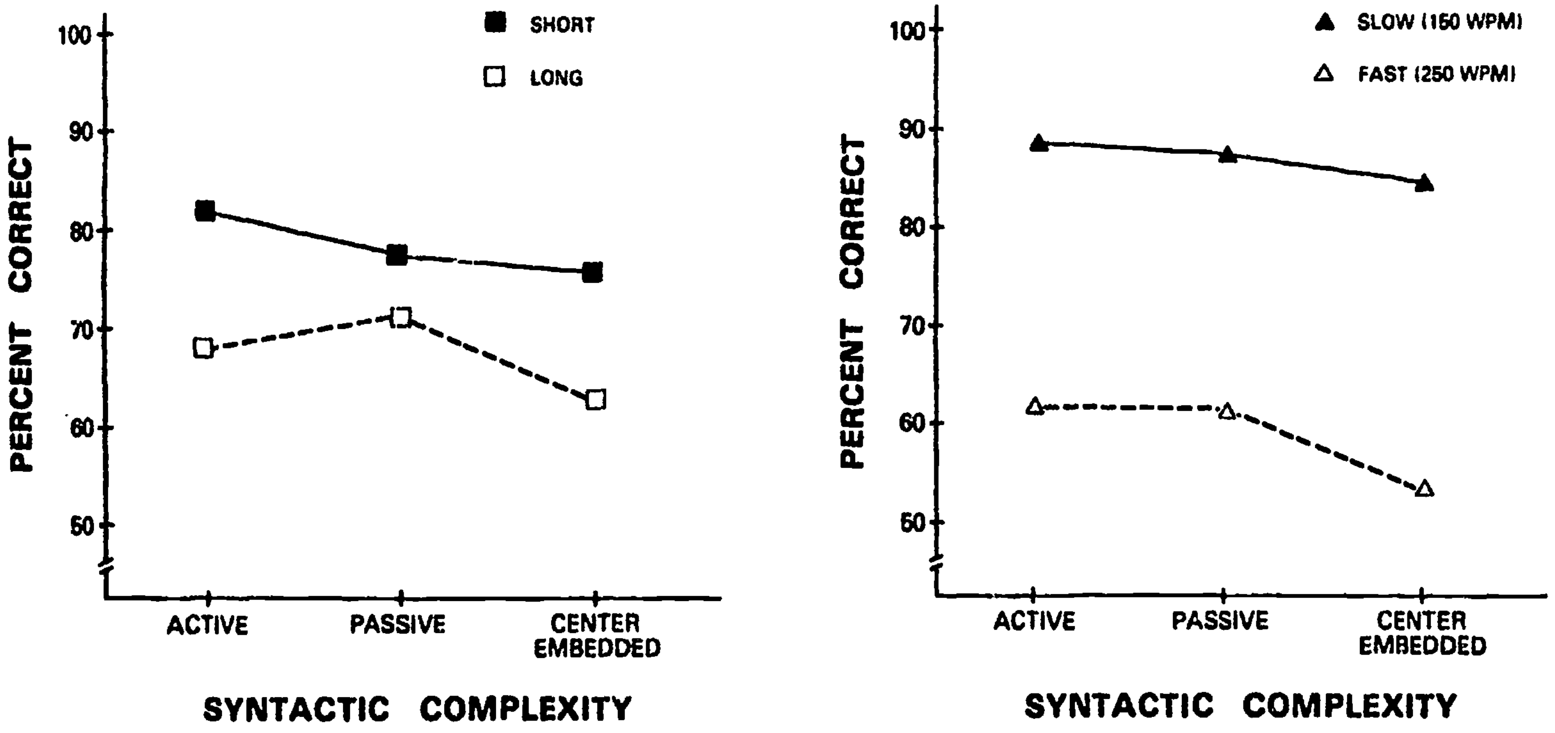


Figure 3. Percent correct identification at different levels of syntactic complexity for short and long sentences (left panel) and for slow and fast speaking rate (right panel).

represents identification of words in long sentences. Note that for the more difficult sentence condition (long), performance varied more across syntactic complexity than for the easy sentence condition (short). That is, for long sentences word identification performance was worst for center-embedded sentences than for the other two types of structures.

The syntax-by-rate interaction is shown in the right hand panel of Figure 3 ($\underline{F}(2,112) = 7.98$, $\underline{p} < 0.001$). The solid line in the figure corresponds to performance on sentences produced at the slow speaking rate (150 words per minute), the dashed line corresponds to performance on the sentences produced at the fast rate (250 words per minute). As observed earlier with the syntax-by-length interaction, performance on the more difficult condition (i.e., fast speaking rate) shows more variation across syntactic complexity than for the less difficult condition (slow speaking rate). This interaction is clearly illustrated in the right hand panel of Figure 3 by the decrease in performance at the fast speaking rate for center-embedded sentences.

In short, both interactions are due to a drop in performance for the center-embedded sentences for the more difficult length and rate conditions. These results demonstrate that the language processing mechanism can be overloaded by increased processing requirements of synthetic speech under these conditions.

General Discussion

The results of these two experiments demonstrate the effects of two prosodic variables on the perception of synthetic speech produced by the Prose-2000 text-to-speech system. In Experiment 1, using simple sentence materials, we found large effects of meaning and speaking rate on the perception of words in sentences, although no effect of pitch contour was observed. On the other hand, in Experiment 2, using more complex linguistic materials, we found a small but significant effect of pitch contour. Large effects of length, rate, and syntax on identification performance were also observed.

Taken together the results of these experiments suggest that increasing linguistic complexity will result in a decrease in perceptual performance. Specifically, with synthetic speech, removing semantic information from the stimulus, increasing the syntactic complexity of the sentences, accelerating the speaking rate, increasing the length, and, to a lesser degree, removing normal pitch cues, all decrease performance in identifying words in sentences.

Several of the results obtained in Experiment 1 were expected based on earlier work dealing with the effects of speaking rate. The effects of speech rate on synthetic speech replicate the results obtained by deHaan (1977) and Foulke and Sticht (1969) with natural speech. As noted in the introduction, Foulke and Sticht (1969) drew a distinction between intelligibility measurements and comprehension measurements of accelerated speech. According to these researchers, decrease in intelligibility of time-compressed speech is due to the degradation of the signal as the speech is accelerated. On the other hand, a decrease in comprehension may be attributed to exceeding the cognitive capacity limitations of the listener when processing the incoming signal, as well as signal degradation. The results obtained in Experiment 1 were based on a task measuring the identification of words in sentences and are similar to the decrease in intelligibility found by these researchers for time-compressed speech.

Several researchers have previously found that recognition is better for meaningful stimulus materials than it is for anomalous stimulus materials (Miller, Heise, & Lichten, 1951; Pisoni & Hunnicutt, Note 4). The main effect of meaning obtained in Experiment 1 provides further evidence that subjects rely on semantic information as well as acoustic-phonetic, lexical, and syntactic information when processing a sentence. Of greater interest, however, is the interaction in identification we obtained between speaking rate and sentence meaning. The results showed a smaller difference in performance between meaningful and anomalous sentences at the fast speech rate than at the slow speech rate. This interaction suggests an important difference in processing between the meaningful and semantically anomalous sentences. For both types of sentences, word identification was poor at the fast rate because the acoustic signal was impoverished by the increased rate of speech. For the slow speech rate performance on the anomalous sentences improved because the acoustic information in the signal was less impoverished. On the other hand, for the meaningful sentences, the amount of improvement for the slower speech rate was due to both the availability of semantic information in the signal and improved acoustic information. Thus two factors--improved acoustic information and available semantic information--influenced the improvement in perceptual performance for the meaningful sentences. For the anomalous sentences, only one factor--improved acoustic information--influenced the increase in performance at the slow speaking rate.

The results of Experiment 2 supported the major findings of Experiment 1 and provided additional evidence for the contribution of pitch contour in the perception of speech. As in Experiment 1, a main effect of speaking rate was observed: As speaking rate increased, intelligibility decreased. Once again, this result suggests that the acoustic signal becomes more impoverished as speech rate is increased resulting in decreased perceptual performance.

In addition to this effect, Experiment 2 revealed main effects of length and syntax. The length effect can be ascribed to memory limitations in language processing. The longer the stimulus item that the subject needs to identify, the harder it is to maintain it in short term memory for processing (Church, 1982; Frazier & Fodor, 1978). Therefore, identification of words should be impaired, not because the words could not be perceived, but rather because they could not be processed and maintained in short-term memory because of cognitive limitations (cf. Shiffrin, 1976).

The main effect of syntactic structure was also expected. Indeed, we selected the three syntactic structures so that the center-embedded type would be the most difficult to process and the passive type would be more difficult to process than the active sentence type (Forster & Olbrei, 1973). This variable was manipulated in order to assess the influence of stimulus complexity on the other variables. Specifically, as syntactic complexity increased, we expected decreased performance for the faster speech rate and the longer sentence length. With this reasoning in mind, the interactions of syntax-by-length and syntax-by-rate were expected. For both of these interactions, performance dropped most for the center-embedded sentences in the more difficult experimental conditions. This result indicates that the listener has a limited capacity that can be overloaded by both complex stimuli (center-embedded sentences) or increased processing requirements (increased speech rate and increased sentence length). Furthermore, the interaction between syntactic complexity and rate suggests an important interaction between top-down and bottom-up sources of knowledge in language processing tasks (Marslen-Wilson & Welsh, 1978; Marslen-Wilson & Tyler, 1980). Specifically, performance on the speech rate variable is indicative of phonetic processing. Differential performance on sentence types is indicative of syntactic processing and higher level constraints. Therefore, the interaction between these two variables suggests a trade-off in processing allocation between bottom-up and top-down knowledge sources. The fact that we did not find a three-way interaction of syntax, length and rate suggests that the effects of syntax and length were not due to the subjects inability to process the speech quickly enough.

The second experiment also revealed a main effect of pitch contour. Because pitch contour did not interact with any of the other variables, we can interpret this effect independently of the other variables considered. That is, pitch contour had an effect on processing regardless of the syntactic structure of the sentences, length of the sentences and rate of speech. This finding suggests that subjects are clearly using prosodic information provided by the fundamental frequency to identify words in sentences. This result therefore supports the conclusions made by other researchers that prosodic information apparently aids in phonetic processing (Cutler, 1976) and organizing the syntactic information in a sentence (Wingfield, 1978). We failed to find an effect of pitch contour in the first experiment primarily because the stimulus materials were much too simple. By increasing the stimulus complexity in Experiment 2, we found a pitch contour effect that did not interact with the other variables. It appears that when sentence structure was held constant in Experiment 1, subjects did not use pitch contour to process the stimuli. However, when sentence type varied, subjects relied on pitch contour to aid in the identification of the words in the sentences.

To summarize, the present experiments demonstrated effects of meaning, speaking rate, pitch contour, sentence length and sentence type on identification of words in sentences generated by a text-to-speech system. One conclusion which can be drawn from these results is that decreased performance appears to be primaily a result of a degradation of the acoustic signal. Specifically, the extent to which speaking rate affected perceptual performance may be due to the fact that the stimulus items were synthetically produced. Previous research conducted on the intelligibility of accelerated natural speech has indicated that, in general, intelligibility is extremely good at speaking rates up to 275 words per minute (deHaan, 1977; Foulke & Sticht, 1969; Garvey, 1953). The decrease in performance for the fast speaking rate may therefore be the result of an already degraded synthetic speech signal becoming even more impoverished by increasing speaking rate. Therefore, although important to our understanding of synthetic speech, the effects found with synthetic speech cannot be directly generalized to naturally produced speech (see Pisoni, 1982).

With regard to the pitch variable, the present experiments showed a decrease in word identification when the normally inflected pitch contour was removed from the acoustic signal. Studies of pitch contour using natural speech stimuli (Lehiste, 1970; Luce & Charles-Luce, Note 5) have demonstrated the importance of fundamental frequency contour to sentence perception. Although the precise nature of the role of pitch contour still must be specified, it does appear to be important in the perception of both natural and synthetic speech.

The conclusions to be drawn from this study on comparisons of natural and synthetic speech stimuli, therefore, are different depending on the variable under consideration. For speaking rate, the present investigation found a detrimental effect on the perception of synthetic speech as the speaking rate of the synthetic talker was increased. Previous research on speech rate using natural speech did not reveal such a dramatic decrement in performance at the speech rates we investigated. This difference in perceptual performance between natural and synthetic speech as speaking rate is increased may therefore reflect differences in the manner in which speaking rate is controlled by humans and by text-to-speech systems rather than inherent differences in the perceptual processing of the speech by human listeners. Further research obviously needs to be conducted to determine where the locus of this difficulty lies.

For the pitch variable, on the other hand, the results obtained using synthetic speech in the present investigation are consistent with previous results reported using natural speech. This consistent finding suggests that the pitch contour of an utterance is important to the processing of spoken language. However, the relative importance of such a finding in designing or redesigning text-to-speech systems may be minimal given the contribution of other variables to perceptual performance. Considering the relatively minor influence of pitch contour in these studies as compared to the large effect of speaking rate, research into the development of highly intelligible voice output systems might well concentrate more on improving the acoustic-phonetic quality of segmental information as a means of producing high-quality speech that can be easily understood, rather than improving pitch quality.

Reference Notes

1. Larkey, L.S., & Danly, M. Fundamental frequency and sentence comprehension. M.I.T., Working papers, Vol. II, 1983.

2. Pisoni, D.B. Speeded classification of natural and synthetic speech in a lexical decision task. Paper presented at the 102nd meeting of the Acoustical Society of America, Miami Beach, Florida, December, 1981.

3. Slowiaczek, L.M., & Pisoni, D.B. Effects of practice on speeded classification of natural and synthetic speech. Paper presented at the 103rd meeting of the Acoustical Society of America, Chicago, IL, April, 1982.

4. Pisoni,D.B., & Hunnicutt, S. Perceptual evaluation of MITalk: The M.I.T. unrestricted text-to-speech system. Paper presented at the International Conference on Acoustics and Signal Processing, 1980.

5. Luce, P.A., & Charles-Luce, J. The role of fundamental frequency and duration in the perception of clause boundaries: Evidence from a speeded verification task. Paper presented at the 105th meeting of the Acoustical Society of America, Cincinnati, Ohio, May 1983.

References

Allen, J. Linguistic-based algorithms offer practical text-to-speech systems. Speech Technology, 1981, 1, 12-16.

Atal, B.S., & Hanauer, S.L. Speech analysis and synthesis by linear prediction of the speech waveform. Journal of the Acoustical Society of America, 1971, 50, 637-655.

Bernstein, J. Intelligibility and simulated deaf-like segmental and timing errors. IEEE conference record on Acoustics, Speech, and Signal Processing Conference, 1977.

Church, K.W. On memory limitations in natural language processing. Bloomington, IN: IU Linguistics Club, 1982.

Collier, R., & t'Hart, J. The role of intonation in speech perception. In A. Cohen & S.G. Nooteboom (Eds.), Structure and Process in Speech Perception. Heidelberg, Germany: Springer-Verlag, 1975.

Cooper, W.E., & Paccia-Cooper, J. Syntax and Speech. Cambridge, Mass.: Harvard University Press, 1980.

Cooper, W.E., & Sorensen, J. Fundamental frequency contours at syntactic boundaries. Journal of the Acoustical Society of America, 1977, 62, 683-692.

Cutler, A. Phoneme monitoring reaction time as a function of preceding intonation contour. Perception & Psychophysics, 1976, 20, 55-60.

Cutler, A., & Darwin, C.J. Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency. Perception & Psychophysics, 1981, 29, 217-224.

Dallett, K.M. Intelligibility and short-term memory in the repetition of digit strings. Journal of Speech and Hearing Research, 1964, 7, 362-368.

Darwin, C.J. The perception of speech. In E.C. Carterette, & M.P. Friedman (Eds.), Handbook of Perception. New York: Academic Press, 1976.

deHaan, H.J. A speech-rate intelligibility threshold for speeded and time-compressed connected speech. Perception and Psychophysics, 1977, 22, 366-372.

deHaan, H.J. & Schjelderup, J.R. Threshold of intelligibility/comprehensibility of rapid connected speech: Method and instrumentation. Behavior Research Methods & Instrumentation, 1978, 10, 841-844.

Dooling, D.J. Rhythm and syntax in sentence perception. Journal of Verbal Learning and Verbal Behavior, 1974, 13, 255-264.

Egan, J.P. Articulation testing methods. Laryngoscope, 1948, 58, 955-991.

Forster, K.I., & Olbrei, I. Semantic heuristics and syntactic analysis. Cognition, 1973, 2, 319-347.

Foulke, E., & Sticht, T.G. Review of research on the intelligibility and comprehension of accelerated speech. Psychological Bulletin, 1969, 77, 50-62.

Frazier, L., & Fodor, J.D. The sausage machine: A new two-stage parsing model. Cognition, 1978, 6, 291-325.

Garvey, W.D. The intelligibility of speeded speech. Journal of Experimental Psychology, 1953, 45, 102-108.

Huggins, A.W.F. Speech timing and intelligibility. In J. Requin (Ed.), Attention and Performance VII. Hillsdale, N.J.: Erlbaum, 1978.

Klatt, D. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. Journal of the Acoustical Society of America, 1976, 59, 1208-1221.

Kucera, H., & Francis, W.N. Computational Analysis of Present-Day American English. Rhode Island: Brown University Press, 1967.

Lehiste, I. Suprasegmentals. Cambridge, MA: M.I.T. Press, 1970.

Lehiste, I., Olive, J., & Streeter, L. Role of duration in disambiguating syntactically ambiguous sentences. Journal of the Acoustical Society of America, 1976, 60, 1199-1202.

Marslen-Wilson, W.D., & Tyler, L.K. The temporal structure of spoken language understanding. Cognition, 1980, 8, 1-71.

Marslen-Wilson, W.D., & Welsh, A. Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology, 1978, 10, 29-63.

Martin, J. Rhythmic and segmental perceptions are not independent. Journal of the Acoustical Society of America, 1979, 65, 1286-1297.

Miller, G.A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 1956, 63, 81-97.

Miller, G.A., Heise, G. A., & Lichten, W. The intelligibility of speech as a function of the context of the test materials. Journal of Experimental Psychology, 1951, 41, 329-335.

Miller, G.A., & Licklider, J.C.R. The intelligibility of interrupted speech. Journal of the Acoustical Society of America, 1950, 22, 167-173.

Miller, J.L. Effects of speaking rate on segmental distinctions. In P.D. Eimas, & J.L. Miller (Eds.), Perspectives on the Study of Speech. Hillsdale, N.J.: Erlbaum, 1981.

Nakatani, L., & Schaffer, J. Hearing "words" without words: Prosodic cues for word perception. Journal of the Acoutical Society of America, 1978, 63, 234-245.

Nickerson, R.S. Characteristics of the speech of deaf persons. The Volta Review, 1975, 77, 342-367.

Nooteboom, S.G., Brokx, J.P.L., & DeRooij, J.J. Contributions of prosody to speech perception. In W.J.M. Levelt & G.B. Flores D'Arcais (Eds.), Studies in the perception of language. New York: Wiley, 1978.

Nye, P.N., & Gaitenby, J.H. The intelligibility of synthetic monosyllable words in short syntactically normal sentences. Haskins Laboratories Status Report on Speech Research, SR--37/38, 1974, 169-190.

Pierrehumbert, J. The perception of fundamental frequency declination. Journal of the Acoustical Society of America, 1979, 66, 363-369.

Pisoni, D.B. Speech perception. In W.K. Estes (Ed.), Handbook of Learning and Cognitive Processes (Vol. 6). Hillsdale, N.J.: Erlbaum Associates, 1978.

Pisoni, D.B. Perception of speech: The human listener as a cognitive interface. Speech Technology, 1982, 1, 10-23.

Shiffrin, R.M. Capacity limitations in information processing, attention, and memory. In W.K. Estes (Ed.), Handbook of Learning and Cognitive Processes (Vol. 4). Hillsdale, N.J.: Erlbaum Associates, 1976.

Summerfield, A.Q. How a full account of segmental perception depends on prosody, and vice versa. In A. Cohen & S.G. Nooteboom (Eds.), Structure and Process in Speech Perception. Heidelberg, Germany: Springer-Verlag, 1975.

Telesensory Speech Systems Prose 2000 Text-to-Speech Converter User's Manual. Palo Alto, California: Telesensory Systems, 1982.

Wingfield, A. The intonation-syntax interaction: Prosodic features in perceptual processing of sentences. In A. Cohen & S.G. Nooteboom (Eds.), Structure and Process in Speech Perception. Heidelberg, Germany: Springer-Verlag, 1975.

Wingfield, A. & Klein, J.F. Syntactic structure and acoustic pattern in speech perception. Perception & Psychophysics, 1971, 9, 23-25.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 9 (1983) Indiana University]

Recognition of speech spectrograms*

Beth G. Greene, David B. Pisoni and Thomas D. Carrell

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

*The research reported here was supported in part by Bio-Medical Science Research grant BSRG S07 RR07031 and National Science Foundation grant NSF BNS 80-23099 to Indiana University. We thank Paul Luce and Aita Salasoo for their help in stimulus preparation and Barry Green and Howard Nusbaum for their comments on an earlier version of this manuscript.

Abstract

The performance of eight naive observers in learning to identify speech spectrograms was studied over a two month period. Single tokens from a 50 word phonetically balanced (PB) list were recorded by several talkers and displayed on a Spectraphonics Speech Spectrographic Display system. Identification testing occurred immediately after daily training sessions. After approximately 20 hours of training, naive subjects correctly identified the 50 PB words from a single talker over 95% of the time. Generalization tests with the same words were then carried out with different tokens from the original talker, new tokens from another male talker, a female talker and finally a synthetic talker. The generalization results for these talkers showed recognition performance at 91%, 76%, 76% and 48%, respectively. Finally, generalization tests with a novel set of PB words produced by the original talker were also carried out to examine in detail the perceptual strategies and visual features that subjects abstracted from the training set. Our results demonstrate that even without formal training in phonetics or acoustics naive observers can learn to identify visual displays of speech at very high levels of accuracy. Analysis of subjects' performance in a verbal protocol task demonstrated that they rely on salient visual correlates of many phonetic features in speech.

Recognition of speech spectrograms

The speech spectrogram has been used for over thirty years as one of the major tools for speech research. By carefully examining spectrograms, researchers have identified many of the major acoustic correlates of the phonetic distinctions used in spoken language. For spectrograms to serve as an alternative display for deaf and severely hearing impaired individuals, rules relating acoustic-phonetic characteristics of speech to the visually salient features will also need to be established. It has been assumed that such a set of rules would enable an individual to "read" spectrographic displays and, thus, perceive speech through the visual modality. Similarly, such acoustic-phonetic rules should be extremely useful to researchers working on the problem of speech recognition by machines (Oshika, Zue, Weeks, Nue & Aurbach, 1975; Chen & Zue, 1983; Cole, Stern & Lasry, 1983; Lamel, 1983).

The early work of Potter, Kopp and Green (1947) on learning to recognize visible speech used a device known as the Visible Speech Translator which displayed spectrographic patterns of an utterance on a phosphor-coated moving belt in real time. Subjects learned to interpret the visible displays and, to some extent, converse via the visible speech translator. Potter et al. taught their subjects specific acoustic-phonetic rules for recognizing and interpreting the visual patterns and provided a great deal of training and feedback. Despite the scope of the Potter et al. study, very little data were provided in their book about the time course of perceptual learning or about the specific perceptual and cognitive strategies used by subjects in learning to recognize visible speech.

Several other studies have been carried out over the years to assess the feasibility of using one or another version of the Visual Speech Translator (House, Goldstein & Hughes, 1968; Stark, Cullen and Chase, 1968). House et al. (1968) presented subjects with eight CV syllables consisting of the consonant /t/ followed by each of eight English vowels. The experimental task required subjects to learn to press one of eight response buttons for each stimulus. Basically, subjects engaged in a traditional paired-associate learning task: each visual display was paired with a specific response button. The learning curves showed that subjects improved over the course of each daily session and over the four experimental sessions. Their performance reached 55-60% correct.

Stark et al. (1968) added a storage oscilloscope to the output of the Visible Speech Translator so they could display a modified real-time speech spectrogram and amplitude contour. Using this device school-age deaf children showed a variety of improvements in their speech productions such as lowering of pitch, decreased duration of utterances, and more accurate productions of /ba/, /ma/ and /wa/. Stark et al. concluded that the use of visual displays of speech could be a valuable aid for speech training with hearing impaired children and adults. At the present time there is such a device, the Spectraphonics Speech Spectrographic Display (SSD), in use in speech and hearing clinics, hospitals and schools for the deaf (Stewart, Houde and Larkin, 1976). Although the device does not produce the ideal display, clinicians report that the device is extremely useful in getting talkers to modify their vocalizations in selective ways (Houde, 1980; Maki, 1980; Lippman, 1982).

A number of other studies have been reported in which experimenters attempted to read spectrograms as a preliminary step to developing a set of rules that could serve as a front end for an automatic speech recognition device (Klatt and Stevens, 1973; Lindblom and Svennson, 1973; Kuhn and McGuire, 1974; Shockey and Reddy, 1974). In the Klatt and Stevens (1973) study, the task consisted of moving a 300 ms window (a piece of cut-out cardboard) across spectrograms of sentences in a single pass. The observer's task was to make a broad phonetic transcription of each utterance. The authors correctly transcribed 33% of the segments; an additional 40% were partially transcribed, producing a combined result of 73% correct phonetic segments.

In the spectrogram reading study of Swedish sentences carried out by Lindblom and Svennson (1973), two subjects were given written instructions for interpreting the spectrograms. One was told to use segmental (spectral) acoustic feature information (e.g., formant transitions, formant steady-state values, noise bursts); the second subject was told to use the segmental information plus some acoustic correlates of prosodic properties of the utterance (e.g., variations in FO intensity, syllable duration). Each subject correctly identified about 50% of the segments in the Swedish "pseudosentences" used in this experiment. When they then read spectrograms of meaningful Swedish sentences, performance averaged 50% correct for the subject who used the segmental instructions only but increased for the subject who used the segmental plus prosody strategy to an average of 70% correct.

Kuhn and McGuire (1974) examined their own performance in reading 432 spectrograms of nonsense words (VCVs spoken with stress on the second syllable). The goals of their study were to identify the consonants, to examine improvement in performance over time, and to examine the pattern of identification errors. Overall, Kuhn and McGuire correctly identified 83% of the consonants in these VCV syllables.

Shockey and Reddy (1974) described a project in which trained phoneticians provided transcriptions of a set of utterances under three experimental conditions: listening to the material, reading spectrograms and reading oscillograms of the utterances. The materials used in this experiment consisted of 50 samples of fluent conversational speech selected from 11 available languages. Each subject transcribed utterances in native, non-native and unknown languages. Shockey and Reddy assumed that transcribing in unfamiliar languages provided only the basic acoustic-phonetic information and would not provide any higher level cues such as morphology, syntax and semantics. For the three subjects who read spectrograms, the average percentage of matches to the phoneme expected by the experimenters was 24%. However, when their performance was scored for appropriate class membership (i.e., front vowels, nasals, voiced stops, voiceless stops, etc.), they reached 65% correct. The four phoneticians who listened to the utterances scored 56% for exact matches and 78% for class matches. The results obtained in this experiment are important because the use of exotic foreign language materials eliminated most of the higher order cues. Shockey and Reddy concluded that phones are articulated clearly enough in fluent speech to be identified as such by trained phoneticians but, on the other hand, spectrograms should not be relied on in determining the presence or absence of acoustic-phonetic features in the speech signal.

Recently, interest in spectrogram reading and in interpreting visual displays of speech has been revived by the dramatic results of an expert spectrogram reader, Dr. Victor Zue (Cole, Rudnicky, Zue & Reddy, 1980). Zue taught himself to read spectrograms over a long period of time. According to Cole et al., he devoted one hour a day for several years to this task for a total of some two thousand hours. In a quite different approach than the Potter et al. study, Zue applied his knowledge of acoustic-phonetics, phonological rules and speech anatomy and physiology (i.e., articulatory gestures and vocal tract shape vis a vis speech production) to the interpretation of static representations of speech as displayed in spectrograms. With sustained practice and attention to fine details in the spectrograms, Zue has developed a set of explicit rules that not only allow him to analyze spectrographic displays with great accuracy, but also permit him to teach others so that they too can analyze spectrographic displays of speech.

In carrying out segmental analyses of unknown utterances, Zue labeled 97% of the segments from spectrograms (see Cole et al., 1980 for detailed descriptions of procedures and analyses). When Zue's labeling of the segments was compared to transcriptions of the spoken utterances done by three phoneticians, he correctly identified about 85% of the segments. Zue's performance in interpreting spectrographic displays of speech may be considered to be the result of very long term practice plus the use of sophisticated and detailed (i.e., expert) knowledge of the acoustics and phonetics of speech (Zue, 1981; Zue & Cole, 1979 a,b; Cole, Rudnicky & Zue, 19;9). In this sense, he may be considered an expert system.

Taken together, previous studies of spectrogram reading show that both naive and sophisticated subjects can interpret the phonetic content of visual displays of speech when given appropriate training and instruction. Naive subjects generally acquire the requisite skills through carefully devised systematic training procedures predetermined by the experimenter. These studies were generally directed towards developing procedures, techniques or devices for improving the speech of hearing impaired individuals. For the sophisticated subjects, the task demands and goals were quite different. These subjects were generally concerned with questions regarding the use of spectrographic displays in speech recognition with the eventual goal of improving the front-end performance of automatic speech recognition systems.

The present study examined the performance of a group of naive observers in learning to recognize visual displays of isolated words. We were interested in studying the time-course of perceptual learning and the perceptual and cognitive strategies that subjects developed to carry out this task. If naive observers can learn to reliably recognize visual displays of speech, the specific cues or features that they use may provide important information about the salient visual features in the displays available for identifying isolated words. Such information will be relevant to researchers working on recognition of speech by machine as well as to basic issues surrounding the use of spectrographic displays as aids for the deaf.

I. METHOD

A. Subjects

Eight undergraduate students at Indiana University in Bloomington were paid to serve as subjects. They were native speakers of American English with no history of speech, hearing or visual disorders as revealed by a pretest questionnaire. None had any training or experience in acoustics, phonetics or signal processing techniques.

B. Materials

Stimulus materials for this experiment were selected from the phonetically balanced lists (PB) developed by Egan (1948). The PB lists were originally designed as a standard method for measuring monosyllabic word intelligibility. Each PB list contains 50 monosyllabic words chosen to represent speech sounds approximately according to their frequency of occurrence in normal speech. One list (PB List No. 1) was used in the training phase of the experiment. This list is included as Appendix I. Each of the 50 words on the list was recorded by a male talker (TDC) in a sound-attenuated room (IAC Model 401-A) using a high-quality microphone (Electrovoice Model DO54 with windscreen Model 355A) and professional-quality tape recorder (Ampex AG-500). The words were spoken in citation format with approximately one second between items. The audio tape recording was then low-pass filtered at 4.8 kHz and processed through a 12-bit analog-to-digital (A-D) converter running at a 10 kHz sampling rate in preparation for subsequent editing. A digital wave-form editor, available on the Speech Research Laboratory PDP 11/34 computer system, was used to edit all the stimulus materials for this project (see Luce & Carrell, 1981).

In addition to recording the materials needed for the training phase of the experiment, the original talker also produced a second recording of the original PB list. A different male talker (PAL) and a female talker (BGG) also recorded the same list. From another project, we also had available a set of synthetic tokens of this PB list produced by the MITalk system (Allen, 1981). Finally, the original talker recorded a different list (PB List No. 9) to be used as the novel set of words in generalization testing. This list is included as Appendix II. Each of the recorded PB lists was subjected to the identical recording, processing and digital editing procedures. Audio test tapes were constructed from digital representations of the stimuli.

All of the experimental tapes used in this experiment were produced using a computer-controlled audio tape making program. The digital waveforms of the stimuli were output through a 12-bit digital-to-analog (D-A) converter and recorded on audio tape at 7 1/2 ips. The tape making program provided a means to incorporate cue tones and timing marks between stimulus items. All stimulus tapes were reproduced on an Ampex AG-500 tape recorder which was interfaced to a Spectraphonics Speech Spectrographic Display (SSD)system (Model SSD-II).

C. Apparatus

The SSD is a real-time spectrograph originally developed for clinical use as a training aid for the deaf (Stewart, Houde and Larkin, 1976). The device produces and stores a frequency-time-intensity display on a video display monitor while the speech is spoken. The visual display resembles a conventional broad-band spectrogram in time, frequency and grey scale resolution (the spectrograms shown in Figure 4 are quite similar to the SSD display). The SSD displays a frequency range from approximately 100-5000 Hz linearly. The full screen display has two time scales, 0.75 s and 1.5 s. The intensity is displayed using the grey scale of the video monitor and more than a 40 dB dynamic range can be represented. For the present experiment, a 300 Hz analyzing filter suitable for displaying the speech of a male talker was selected. The 1.5 s time scale display was used throughout testing. The output of the SSD video display monitor was daisy-chained to CRT monitors in six booths in an experimental room. Thus, whatever appeared on the SSD video monitor also appeared simultaneously on individual CRT video screens in each booth.

D. General Procedure & Stimulus Presentation

The eight subjects were divided into two groups. Each group met with one of the authors (BGG) for a one hour session every day for 7 weeks. Each subject sat in an individual subject booth in a quiet room that was equipped with a CRT monitor (GBC Model MV-10A) located at eye level. Brightness and contrast levels for each CRT monitor were preset by the experimenter to obtain maximum visual distinctiveness for the spectrograms. These levels were checked daily using a standard stimulus word as a calibration procedure.

The experimenter was seated in the experimental room in a control booth which contained: (1) a CRT monitor, (2) a remote control unit to operate the SSD, (3) a second remote control unit to operate the tape recorder on playback, and (4) a pair of Telephonics TDH-39 headphones. The experimenter controlled the presentation of all stimulus materials by manually operating the tape recorder and SSD. Presentation of stimuli to subjects during the initial training phase of the experiment proceeded in the following manner. As soon as the experimenter heard a display cue tone, she pressed the display button on the SSD remote controller. The spectrogram for the word on the tape was immediately displayed in the center of each CRT monitor appearing on the screen from left-right in real-time (only the experimenter could hear as well as see the stimulus since she was wearing headphones). The spectrogram remained on the screen for approximately six seconds. When the experimenter heard a second cue tone, she pressed the erase button on the SSD remote controller and the CRT screen went blank. Two seconds later a new display cue tone was presented and the next spectrogram was displayed in the same manner. The duration of each trial was approximately twelve seconds.

The durations of the stimulus items used in this study ranged from a minimum of 435 ms to a maximum of 872 ms. All spectrograms were displayed in the center of the screen. On those training trials designated as "study" or "practice" trials, verbal feedback was given directly by the experimenter to subjects when the erase cue tone was heard. In all cases, feedback consisted of the experimenter speaking the word aloud to subjects. Before feedback was provided, subjects were required to write down a response for every spectrogram displayed on prepared sheets. On trials designated as test trials, no feedback was provided. Subjects were, however, required to write down a response on every trial even if they had to guess. Daily sessions lasted from twenty to sixty minutes. The experiment consisted of 37 daily sessions distributed over a seven week period.

E. Daily Study-Test Training Procedures

The 50 PB words were initially divided into twelve blocks. There were ten blocks of four words and two blocks of five words. Each of these twelve blocks was presented on a different day beginning with block one on Day 1 and ending with block twelve on Day 22. The experimental procedure involv d a "study-test" format. During the study or training phase, each word was presented five times each in random order. Thus, the first ten training blocks consisted of 20 items (one block of four words, five times each); the last two blocks consisted of 25 items in each set (one block of five words, five times each). Each training set was followed by a block of test trials. The test set consisted of the new words just presented that day plus all words previously learned to date. The words used in the testing phase were presented randomly and without feedback.

As the experiment progressed, the test sets increased in length as each new block of words was learned. In order to keep the daily sessions under one hour, a practice day was introduced on Day 16. This additional day also provided an opportunity to give subjects more trials with immediate feedback, thus insuring high levels of performance as well as maintaining a high level of motivation and positive attitude on the part of the subjects.

F. Testing Procedures

1. Test of absolute identification (Final Test). After all 50 words had been presented and learned under the study-test procedure described above, subjects were tested several times on the total set of 50 words. These repeated tests were intended to ensure that each subject could recognize and identify each spectrographic display with no errors.

2. Generalization tests. Four different generalization tests were presented on consecutive days. The first four generalization tests consisted of: (1) a second token of each original training word spoken by the original talker (TDC#2); (2) tokens spoken by another male talker (PAL); (3) tokens spoken by a female talker (BGG); (4) synthetic tokens (SYN).

On each day of generalization testing, each of the original 50 words (TDC#1) was reviewed with feedback. As before, subjects were required to write down a single word response for each display. Feedback was provided at this time more as a self-checking device than as training since the subjects had already reached asymptotic levels of performance in identification. Subjects were then given one of the four generalization tests for that day without feedback. Each spectrogram was displayed for eight seconds. The duration of each trial was again twelve seconds. Two random orders of each generalization set were presented yielding two responses for each spectrogram for each subject for each generalization set.

Prior to the administration of a fifth generalization test on Day 32, the original 50 words (TDC#1) were reviewed once again with feedback to ensure that their performance had remained at ceiling level.

The fifth and final generalization test consisted of the presentation of a novel set of words, spectrograms of words the subjects had never seen before. Another PB list (PB List No. 9) was recorded by the original talker for use as the novel set. Each word was presented to the subjects on the CRT monitors as in the earlier parts of the experiment. Subjects were told that they would be presented with spectrograms of words that they had never seen before. They were also told that each display would be a real English word and that they should write down whatever word they thought the display corresponded to even if they had to guess. Each display remained on the screen for 30 s to give subjects additional time to make their judgments and record their responses.

3. Protocol analysis. In order to obtain data on the perceptual strategies that subjects used in learning to identify the spectrographic displays, we carried out a protocol analysis for each PB word (Nisbett & Wilson, 1977; Ericsson & Simon, 1980). Subjects were shown each spectrographic display from the original PB list for <u>two minutes</u> and were asked to write a detailed verbal description of the cues, features or visual attributes that they used to identify each word. Additional time was provided if requested by subjects. Protocol responses were obtained on Days 14, 15, 25 and 26 before any generalization testing occurred. Subjects were asked the following questions to guide their responses:

(1) What is in the display that helps you recognize the word?

(2) What is it in the display that helps you remember the display?

(3) What specific aspects of the display do you find helpful?

Subjects were encouraged to provide as much detailed information in their responses as they possibly could to explain how they performed the task. They were asked to do this in their own words. Our goal in carrying out the protocol analysis was two-fold. First, we wanted to know, in the absence of any explicit formal acoustic-phonetic knowledge, what terminology naive subjects would select for describing salient and reliable acoustic attributes in the displays and what visual aspects of the display controlled their responses. We also wanted to examine some of the more subtle cognitive, linguistic and analytic processes that subjects used in this task. Subjects were provided with prepared response sheets for protocol analysis on which they were required to write the target word and then their subjective descriptions of the features and properties they used to identify and remember the specific word. These protocol responses were then analyzed independently by two graduate students in our laboratory who had extensive training in acoustic phonetics.

4. The reacquisition procedure. After subjects completed the five generalization tests, they were presented with the original 50 PB words once again to ensure that they were performing at asymptotic levels. Each of the first four generalization sets was then presented to the subjects again with feedback as a reacquisition set. Under this procedure, each generalization set,

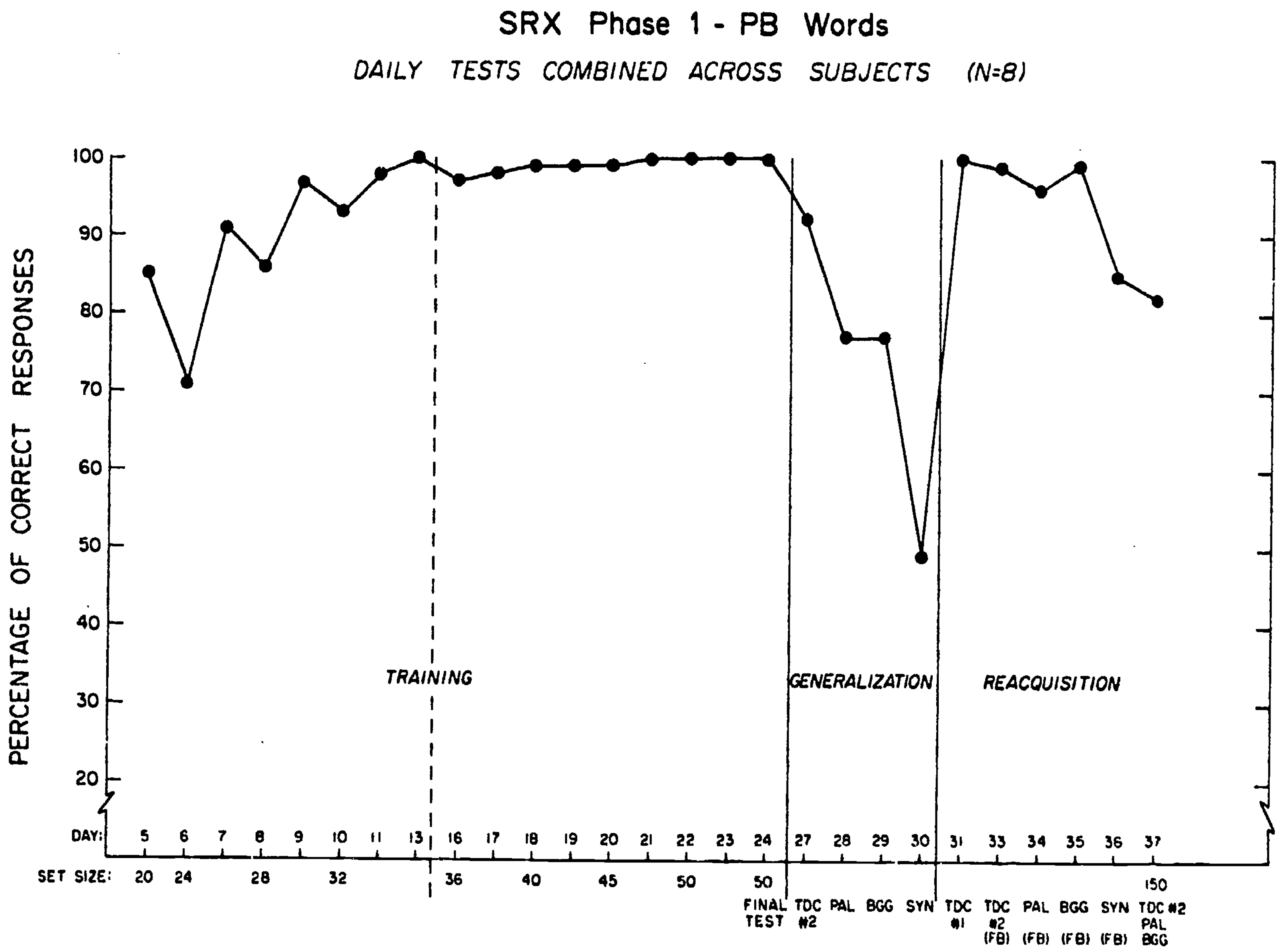


Figure 1. Daily test, generalization and reacquisition data graphed on a day-to-day basis summed across all subjects. Shown on the abscissa are: (1) the number of the daily session starting on Day 5 and ending on Day 37; (2) the set size, that is, the number of words subjects had learned at the time of each test. The left portion of the graph, labeled training, displays test results obtained during the training portion of the experiment. The portion of the graph labeled generalization displays the results of test using several talkers (see text). The right hand portion of the graph, labeled reacquisition, displays test results obtained during the reacquisition portion of the experiment (see text). Protocol responses were collected in Days 14, 15, 25 and 26. Responses with the novel set were collected on Day 32. No test data was obtained on those days and therefore those days are omitted from Figure 1.

that is, TDC#2, PAL, BGG and SYN, was presented with feedback as a one block training set of 50 items. The training set was immediately followed by two test sets of the same generalization tokens presented without feedback. Display times and durations were the same as during training and each reacquisition set was presented twice during testing.

II. RESULTS

A. Initial Training

The results of the training portion of the experiment are displayed in Figure 1. The left portion of this graph, labeled TRAINING shows the cumulative daily performance averaged over all 8 subjects. As can be seen in the figure, after 24 days of training in the experiment, all of our subjects had learned to identify the 50 words 100% of the time.

The cumulative learning curves combined over all eight subjects indicate that, on a day-by-day basis, subjects were performing close to ceiling levels. They learned four or five new items relatively easily and incorporated the most recent words into the total set with little difficulty. Most subjects consistently showed performance levels above 95% correct.

Insert Figure 1 about here

Ceiling level performance is particularly noticeable in the training portion of the learning curve after the first 32 items had been learned. The vertical dashed line at Day 16 indicates this point in the training phase. As described earlier, a practice day was introduced at this point to keep the daily session length under one hour.

As shown in Figure 1, after 24 sessions all eight subjects had learned to identify the fifty words 100% of the time. We allowed <u>three</u> days of testing, Days 22, 23 and 24, to be certain that all subjects had learned to identify each word at asymptotic levels. It can be seen from this figure that while there was some variability at early stages of training, for the most part, subjects were consistently performing at levels well beyond 95% correct for the bulk of the experimental sessions.

Figures 2 and 3 display the same data for the individual subjects in each group, respectively. As can be seen in these figures, the overall pattern of results is highly consistent across the two groups and across the eight individual subjects when examined separately.

Insert Figures 2 & 3 about here

B. Generalization

After the subjects had learned to identify the 50 PB words at asymptotic levels, generalization testing was undertaken. To review, subjects were tested for generalization with five sets of test materials: (1) a second token of the original talker, TDC#2; (2) a new male talker, PAL; (3) a female talker, BGG; (4) a synthetic talker, SYN and (5) a novel set of words (PB List No. 9). Illustrative spectrograms of these items are shown in Figure 4.

Insert Figure 4 about here

The generalization results for the first four tests are displayed in the second panel of Figure 1 (labeled GENERALIZATION). For the second token of the original talker (TDC #2), the overall percentage of correct responses was 91.3%; for the new male talker (PAL), performance was at 76% correct; for the female talker (BGG), performance was also at 76% correct; and for the synthetic talker (SYN), performance was at 48% correct. While the subjects' correct identification of the synthetic tokens was not high, their performance on the other three generalization sets was quite good, given that they had never seen these particular spectrograms before. Overall, in the absence of any specific training or feedback on these tokens, subjects were able to identify 73% of the spectrograms in the first four generalization tests. If the synthetic tokens are excluded, the overall percentage of correct responses increases to 81.3%.

Individual results are shown in Figures 2 and 3. Differences among individuals are more apparent for the generalization tests than for the daily training sets. However, the general pattern of performance on these four generalization tests shown in the group figures is consistent across all subjects.

C. Novel Set

Word identification performance on the fifth generalization test, the novel PB words, averaged only 6% correct. Subjects did not identify many of the words correctly. However, they sometimes correctly identified one or two phonemes in a novel word. Other times they provided the voiced counterpart of a voiceless consonant or substituted one place of articulation for another. For example, the word <u>fuse</u> was correctly identified by five subjects and the other responses were soothe, fizz and screws. The word <u>pact</u> led to eight different responses: put, putt, pen, pat, print, hat, dance, hint. For the word <u>weak</u>, the responses were weep, weed, week, weak, greet, meat, weep and white. While subjects identified

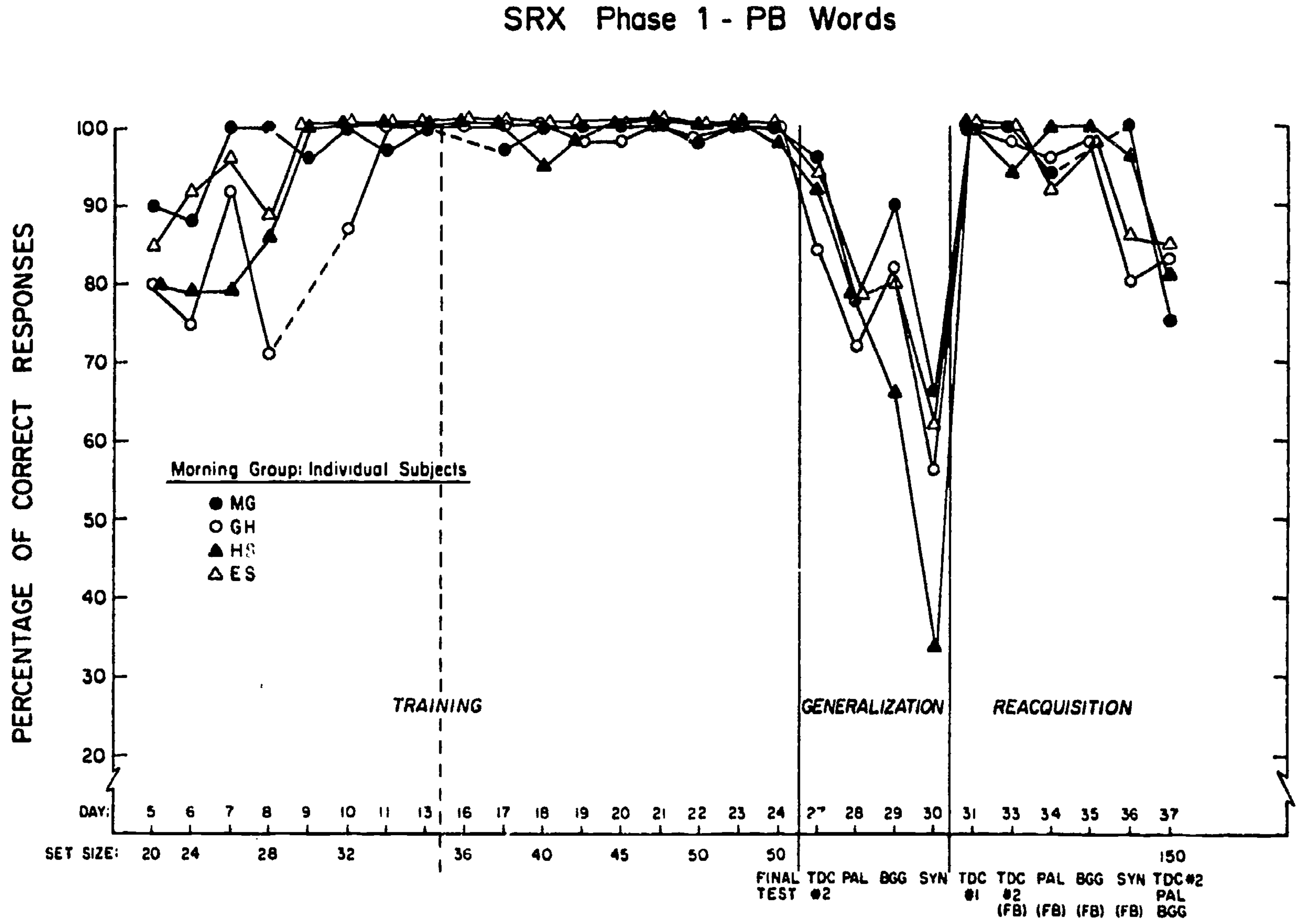
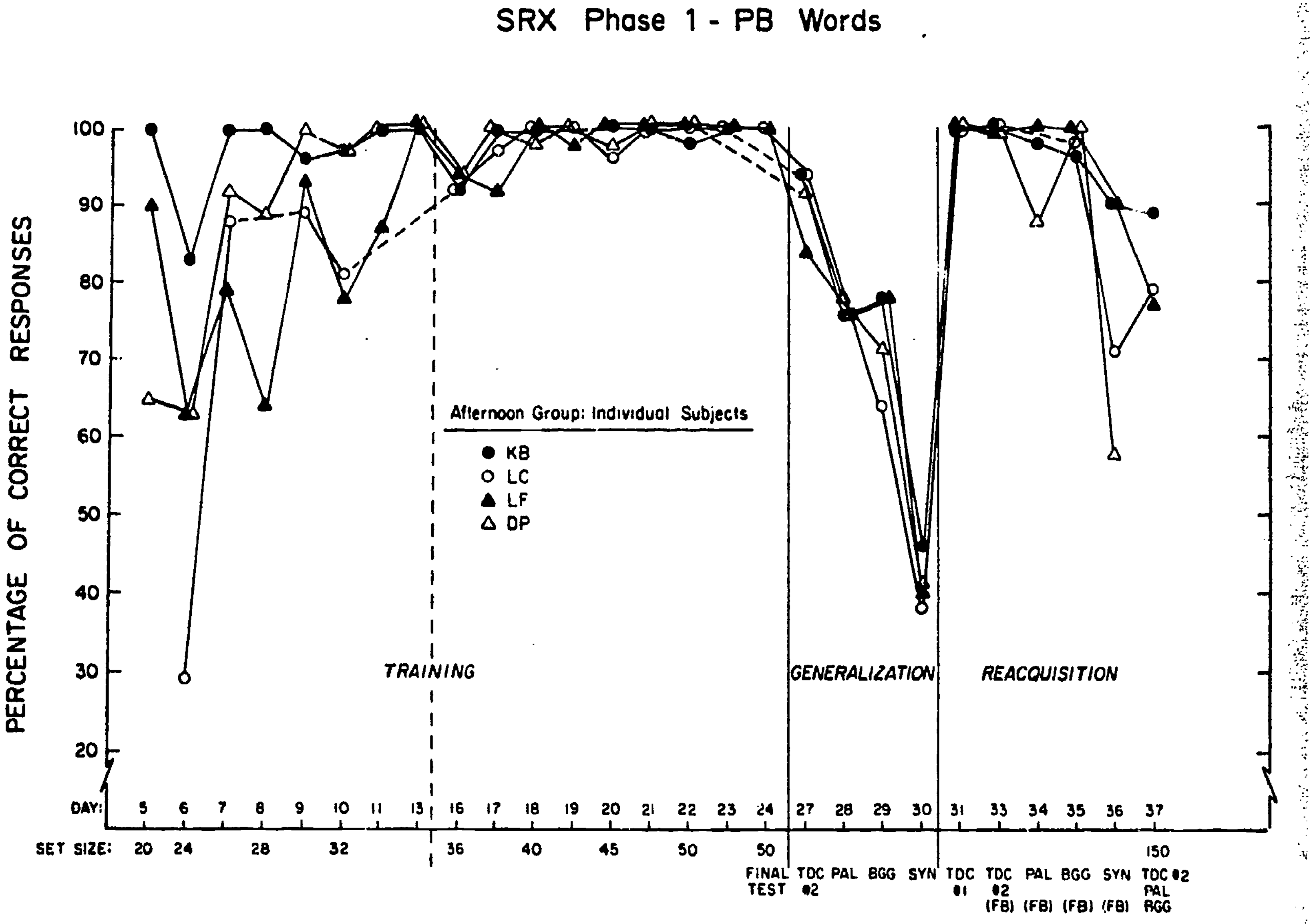


Figure 2. Daily test, generalization and reacquisition data graphed on a day-to-day basis for individual subjects.



Figures 3. Daily test, generalization and reacquisition data graphed on a day-to-day basis for individual subjects.

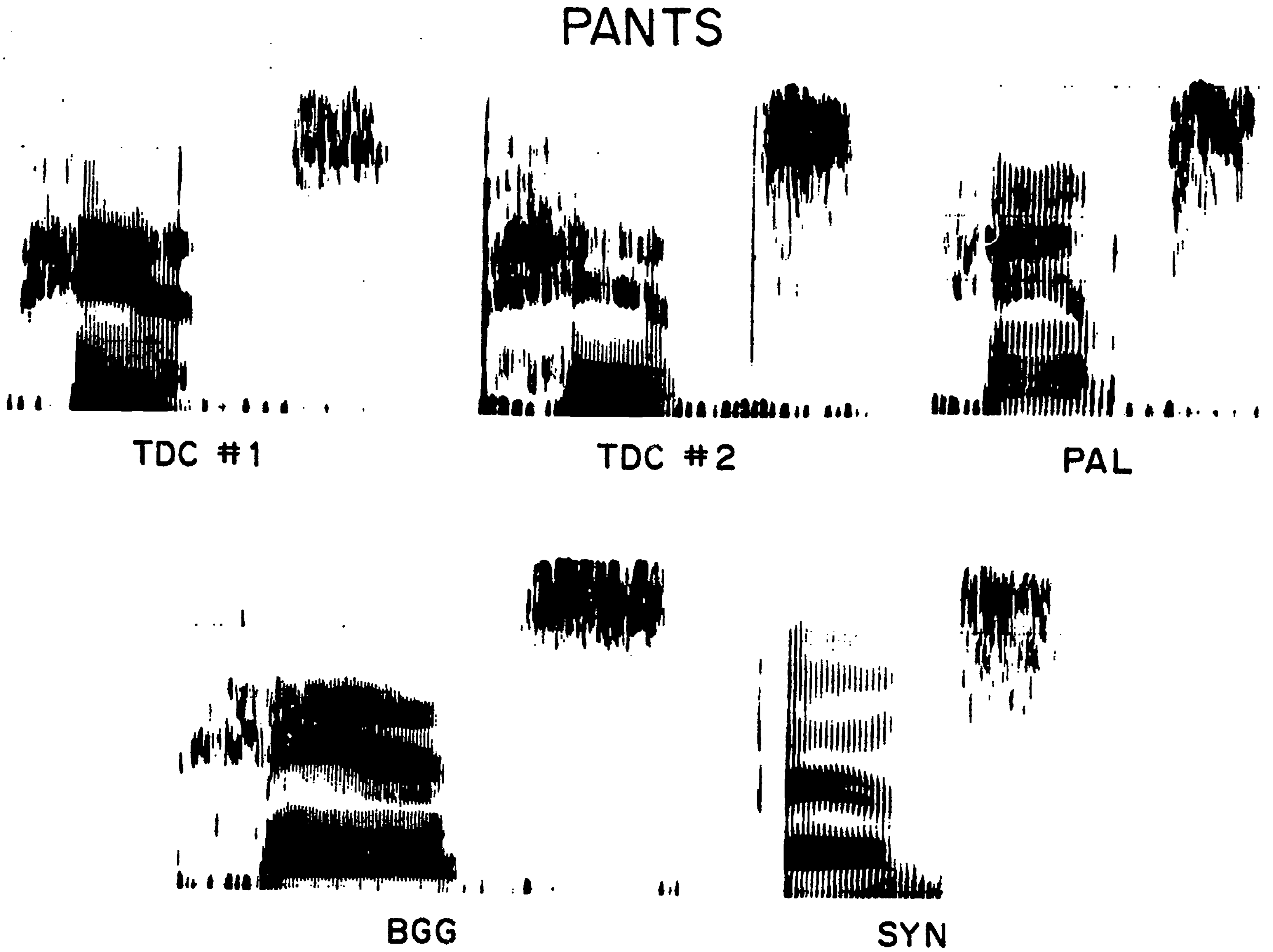


Figure 4. Conventional spectrograms of the word PANTS spoken by four talkers.

relatively few words correctly on an absolute basis, their responses certainly were not random and reflected important perceptual strategies. The examples noted above represent a wide range of responses. To quantify these observations, performance on the novel set of PB words was examined in terms of the exact consonant phoneme identified correctly, manner class identified correctly and vowel identified correctly. The novel set and all responses were transcribed using broad phonetic transcription. Each response for a specific phonetic segment in a specific position was compared to the correct phoneme to provide a score for exact phoneme identification. These data were also scored in terms of correct manner class identification whether exactly correct or not (i.e., if the correct phoneme was /t/, subjects were scored "manner class correct" if they indicated any stop /b,d,g,p,t,k/). Thus, for the word pact, five of eight responses were scored exactly correct for the initial phoneme /p/ while six of eight responses were scored manner class correct identification for the initial stop consonant. Recognition of initial and final consonant clusters was also examined in these analyses.

1. Exact phoneme identification. Subjects were able to identify the exact consonant phoneme 34% of the time. There was a total of 752 possible consonant responses, the majority of which were stop consonants or fricatives (328 stops and 264 fricatives). The remaining 160 consonants were liquids and glides (40), nasals (80), and semi-vowels (40).

Insert Table I about here

Table I displays the percentages of exact correct phoneme responses for each consonant class for initial and final position. Overall, subjects correctly identified 27% of the stops, 40% of the fricatives, 48% of the liquids and glides, 43% of nasals, and 23% of the semi-vowels. As shown in Table I, subjects did about equally well in identifying consonants in initial and final position with the exception of the fricatives. For the fricatives, identification was much better for initial position than for final position.

2. Manner class identification. Subjects were able to identify the correct manner class of a consonant 66% of the time. Table II displays the percentage of correct responses for each consonant manner class for both initial and final position broken down by manner of articulation. Overall, subjects correctly identified manner class 69% of the time for stops, 75% of the time for fricatives, 52% of the time for liquids and glides, 56% of the time for nasals and 23% of the time for semi-vowels. Correct manner class identification was roughly the same for initial and final position.

Insert Table II about here

Table I

Responses to Novel Set Scored for Exact Phoneme Identification

[8 subjects X 50 words (PB List 9)]

	Initial		Final		Total	
	Maximum # of correct responses	%	Maximum # of correct responses	%	Maximum # of correct responses	%
STOPS	144	26	184	28	328	27
FRICATIVES	120	50	144	31	264	40
LIQUIDS	32	47	8	50	40	48
NASALS	32	44	48	42	80	43
SEMI-VOWEL	40	23	-	-	40	23
VOWELS					400	23

Table II

Responses to Novel Set Scored for Manner
Correct Identification

[8 subjects X 50 words (PB List 9)]

	Initial		Final		Total	
	maximum # of correct responses	%	maximum # of correct responses	%	maximum # of correct responses	%
STOPS	144	71	184	68	328	69
FRICATIVES	120	79	144	71	264	75
LIQUIDS	32	53	8	50	40	52
NASALS	32	56	48	56	80	56
SEMI-VOWELS	40	23	-	-	40	23

3. Vowel identification. The overall vowel data are displayed in Table I. Subjects correctly identified vowels 23% of the time. The diphthong /aɪ/ produced the majority of the correct vowel identifications. It was correctly identified 71% of the time. The vowels /i/, /ɛ/, /æ/, /ɑ/, and /ɔ/ were each identified correctly at least 25% of the time.

4. Consonant clusters. Subjects were also able to recognize the presence of initial and final consonant clusters. While they did not always correctly identify both members of the cluster, they identified 40% of the consonant clusters that were present in the novel set. No difference was observed in identification of initial and final consonant clusters.

D. Protocol Analysis

The written responses provided by the subjects for the protocol analysis revealed that they were able to abstract consistent and reliable attributes or features from the visual displays. In many cases, subjects' responses could be mapped directly onto standard acoustic-phonetic categories.

1. Stops. The terminology used to describe consonantal closure included "gap," "space," and "blank space", each of which can be classified as referring to some property of articulatory closure. Descriptions of the release burst included "vertical line" and "abrupt beginning." References to the presence of the release burst portion of the display were three times more frequent than references to the closure interval. Other responses for stops were more idiosyncratic including vague references to the overall nature of the display or a simple identification ("When I see this I know it's a /p/").

2. Fricatives. In the case of fricatives, subjects generated responses that were primarily related to the intensity of the frication, (e.g., "dark patch," "light area"); the frequency location of the frication; or to the overall nature of the display, (e.g., "fuzzy" or "busy"). There were also several responses such as: "It has the /s/ at the beginning."

3. Other consonants. The consonants /n/ and /r/ were given fairly specific descriptions by the subjects. For example, /r/ was described by words such as "slope," "rises," "curves," or "inclines", corresponding to the prominence of the third formant. Descriptive labels for /n/ included reference to a concentrated "patch," "tail," "bar," or "dark bar", related to the presence of a nasal resonance.

4. Summary: Consonants. Of the 376 possible occurrences of stop consonant segments, subjects provided descriptive references to 140 of these segments or 37%. For the 304 fricative segments, 54% of them were given descriptions. Of the total possible 968 consonant segments, 361 were given some sort of verbal description for an overall response level of 37%. Thus, when providing written verbal descriptions of the subjective information used to identify a particular word, subjects described a consonant portion of the display on approximately 37% of the possible occurrences.

5. Vowels. In the case of vowels, most of the written descriptions referred to the location, shape and direction of movement of the formants. Of the 400 possible occurrences of vowels, 219 responses were made for a response level of 55%.

6. Other descriptors. A few references were made to the overall duration of the word. For example, four subjects simply responded, "This is the longest word." The two shortest words were also noted by subjects. Several subjects sometimes identified specific shapes, letters or figures in the visual pattern (e.g., "looks like an upside down smile", "letter "F" in pattern", "eye in pattern").

E. Reacquisition

Returning to the daily training sessions in Figure 1, the first data point in the right-hand portion (labeled REACQUISITION) shows the results obtained when subjects were tested on the original talker again, that is, retesting on the original tokens of TDC#1. Subjects were able to identify 100% of the original spectrograms. Over the next several days each of the first four generalization sets was presented one time with feedback. Subjects were then tested on the tokens from each set. This procedure was followed for the four generalization tests, TDC#2, PAL, BGG and SYN.

The data points in the REACQUISITION portion of the figure show that after only one block of trials with feedback, subjects identified spectrograms at considerably higher levels than during generalization testing. For the second token of the original talker (TDC#2), subjects performed at 91% correct during generalization testing and, after one trial with feedback during reacquisition, their performance increased to 99% correct. For the new male talker (PAL), performance went from 76% correct to 95% correct. For the female talker (BGG), performance went from 76% correct to 99% correct. And, for the synthetic talker (SYN), performance went from 48% correct to 85% correct. Finally, the last data point on the curve represents the results of a test that combined all three natural talkers (TDC#2, PAL, BGG) in a mixed list; performance reached 81% correct on this test. The corresponding percentage of correct responses for the individual talkers was: 89% for TDC#2, 75% for PAL, and 80% for BGG, respectively.

III. DISCUSSION

From the relatively few studies on speech spectrograms over the years, a number of general conclusions have been drawn about whether human observers can realistically interpret and use visual displays of speech. Prior to the reports of Zue's performance by Cole et al. (1979, 1980), the general consensus in the field was that spectrograms were extremely hard to read (Liberman et al., 1968), they were of questionable value as aids for the hearing impaired (Stark, 1974), and they are probably not the optimal means for deriving analysis parameters for automatic speech recognition (Shockey & Reddy, 1974). Although Potter et al. (1947) reported moderately successful results for their spectrogram reading study in 1947, only in the past five years has spectrogram reading emerged as an important area to pursue for research, development, teaching and clinical purposes. We suspect this is probably because of the recent developments in technology, the availability of a commercial real-time spectrograph and the renewed interest in automatic speech recognition. Moreover, the successful performance of Victor Zue with unknown spectrograms constitutes an "existence proof" that visual displays of speech can be identified reliably at very high levels of accuracy and that important properties of speech are represented in these displays.

The present study represents the first long-term investigation of spectrogram reading with unsophisticated subjects. In contrast to the Potter et al. (1947) stuuy, also a long-term project, subjects in the present study were provided little if any information about the specific content of spectrographic displays of speech. In the spectrogram reading studies of Klatt & Stevens (1973), Lindblom & Svensson (1973), and Kuhn & McGuire (1974), the observers had extensive knowledge of acoustic phonetics. In order to appreciate the significance of this investigation, we have broken the discussion section into several subsections, corresponding to the various parts of the project. First, we will discuss the initial training, then the generalization results. Finally, we will discuss the findings obtained from the protocol analysis and reacquisition results.

A. Initial Training

Subjects learned the original words in the training set relatively easily. They were able to integrate each new block of words into the set of words they had already learned with little difficulty. Examination of the ceiling and near ceiling levels of recognition performance demonstrate that subjects were able to learn to identify the spectrograms. While the individual learning curves (Figures 2 & 3) showed some variability, especially at the early stages of learning, subjects apparently could learn four new words at each session with little difficulty.

Despite the encouraging results, several aspects of the training procedure we used may have influenced the final performance levels. First, our procedure focused on holistic pattern learning. Subjects were shown a spectrogram and verbally told what the actual word was. They were never explicitly instructed to analyze the component parts of the spectrogram or to carry out a detailed segmental analysis of the internal structure of the words. Focusing attention on recognition of holistic patterns probably does not interfere with transfer (or generalization) to other tokens of the same words but such a strategy might affect the amount of generalization observed with novel words. The generalization results with the novel words will be discussed below.

Second, the experimental procedures focused attention on spectrograms of a single talker. The subjects' initial task was basically a 50-alternative forced-choice recognition task requiring absolute identification of an item from a pool of 50 candidates. The use of a holistic pattern learning strategy combined with presentation of tokens from only a single talker could have precluded generalization altogether since subjects may have simply memorized each display as a single visual image. As discussed below, this did not occur since subjects did show good generalization performance.

Several studies in the psychological literature have been concerned with recognition memory for pictures. Subjects are typically shown a large number of pictures presented as photographs or slides and, in most cases, display unusually good performance in a recognition memory task (Shepard, 1967, Standing, Conezio & Haber, 1970; Standing, 1973). Across a wide range of experimental procedures, retention intervals, exposure durations and set sizes, subjects' recognition memory for pictures exceeded 90% correct. Spectrograms are not the same as pictures, since pictures have numerous visual features that can be readily identified, labeled and integrated into a coherent whole. It is unlikely that our subjects treated the spectrograms in this way. Indeed, the results of the protocol analysis indicated that subjects did not usually ascribe a single holistic pictorial or visual representation to the spectrograms. Our subjects did not memorize the spectrographic displays in a simple paired-associates format.

Taken together, the results of the initial training portion of this experiment demonstrate that naive subjects can learn to identify visual displays of speech at very high levels of performance. The successful performance displayed by these subjects demonstrates that there are salient and reliable cues in spectrograms. We believe that the previous claim that spectrograms are too hard to read is probably unwarranted (cf. Liberman et. al., 1968).[2] This conclusion is not too surprising since to our knowledge there have been no previous laboratory studies of the perceptual and cognitive processes that naive subjects use in learning to identify speech spectrograms. The claims made about the difficulty of learning to interpret spectrograms of speech rest on little if any empirical data from observers in controlled laboratory situations.

B. Generalization

Subjects showed evidence of generalization in every test we conducted. The best performance was observed in the generalization to new tokens of the original words produced by the original talker. There are several ways to account for these successful generalization results. Subjects may be engaging in template or pattern matching. Each subject could absolutely identify the original 50 words 100% of the time. If the new spectrogram looked enough like the original, a simple pattern matching recognition scheme might be sufficient for identification. However, we believe this account is much to simple to deal with our subjects' performance.

In the present experiment, if a spectrogram in a generalization set differed markedly from the original, subjects would no doubt engage in some type of analytic process. They might perform partial pattern matching or they might segment the display and then do several partial pattern matches. It is also possible that subjects used an overall similarity metric; subjects may have identified a spectrogram in the generalization set as particular word by making a global overall similarity judgment.

We cannot say precisely at this time whether subjects used pattern matching, partial or multiple pattern matching, overall similarity judgments or feature extraction or some combination of all of these. When our generalization results for different talkers are taken together with the generalization tests for novel words and the protocol analysis, it does appear that subjects are doing more than just simple pattern matching and overall similarity comparisons. Subjects appear to be extracting visual correlates of criterial features -- some of which are the appropriate acoustic-phonetic cues and using these features in novel contexts. Thus, they appear to be abstracting out properties or features of the displays tacitly even though they were not explicitly instructed or trained to do so.

C. Novel Set

Naive subjects provided the correct phonetic label for about one-third of the possible phonetic segments in the novel set of 50 words. Since the subjects had never been explicitly told how to identify or recognize a spectrogram, we must assume that subjects developed individual identification or recognition strategies. Learning to recognize speech spectrograms with the study-test procedure we used may encourage implicit learning of the underlying dimensions (Reber, 1967; Reber & Lewis, 1977; Reber, Kassin, Lewis & Cantor, 1980).

We have examined in some detail the responses and the feature composition of the responses to the novel words.[3] Some manner classes were hard to recognize from spectrograms simply because the information was hard to see in the visual display. This was true for weak fricatives especially using the 5 kHz bandwidth. For example, in many cases a weak fricative may look more like a stop burst. The sophisticated observer may be able to see some weak energy in the display but the naive observer probably sees a gap with some faint lines in it. Within a particular manner class, the cues that distinguish one member of a class from another may be too subtle for the subjects to perceive. The direction and extent of the formant transitions for stop consonants may serve as a reliable cue for the sophisticated observer who knows what to look for in a display. For the unsophisticated observer, this cue may not be distinctive enough. Subjects may have known that the display contained a stop but they lacked an appropriate selection strategy to determine which particular stop it was. It may also be the case that the unsophisticated subjects are not attending to the appropriate cues in the displays when faced with partial or ambiguous information.

We undertook a further analysis that examined detection and identification of consonant clusters. Subjects detected the presence of 40% of the consonant clusters in the novel word set. In this case, we infer that they had some strategy that led them to conclude that more than one consonant was present in the display. Subjects might carry out a successful pattern match but then part of the display is left over that they have to account for. Another strategy could reflect durational cue information; that is, they might conclude that there was too much in the display to be just a single consonant so they decided to identify two or three consonants to fit the overall duration. On the other hand, subjects failed to detect 60% of the consonant clusters. Subjects may have identified the one consonant in the cluster that they knew and then simply ignored the other part. Our results show that subjects performed poorly on identification of non-initial and non-final consonants, that is, the second consonant in a consonant cluster. Subjects did not always have a reliable strategy to handle consonant clusters. Throughout training they were never exposed to the principle of coarticulation in speech nor its acoustic-phonetic consequences as revealed in spectrographic displays. In addition, their attention was never directed explicitly at noticing, for example, the difference between /s/ and /t/ as initial consonants compared to the /s/ portion and /t/

portion of the initial /st/ cluster. Obviously, a detailed knowledge of the contextual effects of one segment on another in speech and the concept of allophonic variation could improve performance on unknown displays substantially.

D. Protocol Analysis

Examination of the verbal reports made by individual subjects to the displays revealed that subjects do divide words into segments (i.e., segmentation) and then attempt to apply descriptive labels to each component (i.e., labeling). The abundance of stop consonants and fricatives in our set of test words may have served as obvious or natural points for such segmentation strategies. We feel confident in inferring from the protocol analysis that subjects used rudimentary phonetic or acoustic-phonetic recognition strategies to solve the task at hand.

We should note here that there are a number of differing opinions surrounding the use of protocol analysis as well as other subjective and introspective reports provided by subjects (Reber, 1967; Reber & Lewis, 1977; Brooks, 1978; Nisbett & Wilson, 1977; Ericsson & Simon, 1980). For our purposes in this study, the use of protocol analysis provided converging evidence that subjects were able to focus on the critical properties of these visual displays, that is, the precise properties that made the spectrographic displays distinctive. Subjects appeared to focus on these criterial properties implicitly without being told what to look for. There can be little doubt that the internal structure of words and phonotactic constraints of English strongly influence the way in which a subject described each spectrogram in these analyses.

E. Reacquisition

When subjects were undergoing generalization testing they received no feedback about their performance. At the end of each generalization set, spontaneous comments from subjects indicated that they found the task difficult, they felt they had performed poorly, and they had to guess on a large number of trials. These remarks are consistent with subjective and introspective reports made by subjects in other kinds of experiments including those in the field of implicit learning. In implicit learning studies (Reber, 1967; Reber et al., 1980), subjects learn an underlying set of rules or regularities unconsciously; the experimenter does not provide explicit instructions to look for specific rules or regularities in the stimulus patterns presented to them.

In the reacquisition phase, subjects showed very rapid learning of tokens of the new talkers with a minimum amount of feedback. After only one block of trials with feedback, performance approached ceiling levels. We consider the generalization data to new talkers to be a reflection of our subjects' ability to analyze the visual display into component attributes and features--i.e., subjects learned sets of distinctive cues that are present in other tokens produced by different talkers.

A number of limitations must be noted here. First, the training procedures used in this study differed from other previous studies in several respects. We used a "study-test" procedure in which tokens of a single talker were used to train the subjects. This procedure may be contrasted with the use of tokens from multiple talkers. Second, we presented spectrograms of isolated monosyllabic words in contrast to words in a carrier phrase, sentences, or samples of fluent conversational speech. With words presented in isolation rather than in context, important information about stress and intonation was unavailable in the displays. Third, subjects were explicitly required to identify whole words. Thus, subjects were not required to undertake a conscious detailed phonetic analysis of the visual displays. Finally, we required subjects to provide only a single response for every spectrographic display and we did not allow them to make multiple responses or entertain several hypotheses or word candidates. These points should be noted when drawing comparisons of our data with previous studies reported in the literature.

IV. SUMMARY AND CONCLUSIONS

In this experiment, eight naive subjects learned to identify spectrograms of 50 isolated monosyllabic words. After 22 one-hour sessions using a study-test procedure, our subjects correctly identified the original training words 100% of the time. When tested for generalization to other tokens of the words, they achieved 91.3% on different tokens from the original talker and 76.0% on tokens of new male and female talkers. Subsequent presentation of these new tokens using the study-test procedure resulted in rapid relearning of the new spectrograms. Subjects achieved scores close to ceiling levels after only a single presentation of each word with feedback. Protocol analysis revealed that subjects were able to extract features from the spectrograms that corresponded, in many cases, to well-known acoustic-phonetic features. Finally, the subjects correctly identified only 6% of the words from an unknown novel set. However, they correctly identified one third of the phonemes and correctly categorized two thirds of the manner class of the phonemes in these novel words.

APPENDIX I: PB List No. 1 (Egan, 1948): are, bad, bar, bask, box, cane, cleanse, clove, crash, creed, death, deed, dike, dish, end, feast, fern, folk, ford, fraud, fuss, grove, heap, hid, hive, hunt, is, mange, no, nook, not, pan, pants, pest, pile, plush, rag, rat, ride, rise, rub, slip, smile, strife, such, then, there, toe, use, wheat.

APPENDIX II: PB List No. 9 (Egan, 1948): arch, beef, birth, bit, boost, carve, chess, chest, clown, club, crowd, cud, ditch, flag, fluff, foe, fume, fuse, gate, give, grace, hoof, ice, itch, key, lit, mass, nerve, noose, nuts, odd, pact, phone, reed, root, rude, sip, smart, spud, ten, than, thank, throne, toad, troop, weak, wild, wipe, with, year.

FOOTNOTES

1. Some of these results were presented at the 103rd meeting of the Acoustical Society of America, Chicago, April, 1982.

2. For example, Liberman et al. (1968) state: "We believe that no amount of training will cause an appropriate speech decoder to develop for a visual input. The spee^h decoder is, we suspect, biologically linked to an auditory input and cannot be transferred or redeveloped for any modality. If that is so, then we are faced with an inherent and permanent limitation on our ability to read spectrograms fluently, and we cannot expect that the undecoded speech signal will be highly intelligible to the eye" (p. 131). Their assertion refers to reading spectrograms in much the same way as one reads connected text. However, the results reported by Cole et al. (1979, 1980) and Cole and Zue (1979a,b) demonstrate that it is, in principle and practice, possible to read spectrograms of fluent connected speech.

3. Comparisons of performance of unsophisticated and sophisticated observers on the novel set have been undertaken (see Pisoni, D. B., Greene, B. G. & Carrell, T. D. (1983), "Identification of visual displays of speech: Comparisons of naive and trained observers," J. Acoust. Soc. Am. Suppl. 1, 73, 102). This paper provides detailed phonetic analysis.

References

Allen, J. (1981). "Linguistic-based algorithms offer practical text-to-speech systems," Speech Technology, 1, 12-16.

Brooks, L. R. (1978). "Non-analytic concept formation and memory for instances," in *Cognition and categorization*, edited by E. Rosch & B. B. Lloyd, (Erlbaum, Hillsdale, NJ).

Chen, F. R. and Zue, V. W. (1983). "Exploring allophonic and lexical constraints in a continuous recognition system," J. Acoust. Soc. Am. Suppl. 1, 74, S15.

Cole, R. A., Rudnicky, A. I., and Zue, V. W. (1979). "Performance of an expert spectrogram reader", *Speech communication papers presented at the 97th meeting of the Acous. Soc. of Am.*, edited by J. J. Wolf and D. H. Klatt, (Acoust. Soc. Am., New York).

Cole, R. A., Rudnicky, A. I., Zue, V. W., and Reddy, D. R. (1980). "Speech as patterns on paper," edited by R. A. Cole, *Perception and production of fluent speech*. (Erlbaum, Hillsdale, NJ).

Cole, R. A., Stern, R. M. and Lasry, M. J. (1983). "Performing fine phonetic distinctions: Templates vs. features," paper presented at the Symposium on Variability and Invariance of Speech Processes, MIT, Cambridge, MA.

Egan, J. P. (1948). "Articulation testing methods," Laryngoscope, 58, 955-991.

Ericsson, K. A. and Simon, H. A. (1980). "Verbal reports as data," Psychol. Rev., 87, 215-251.

Houde, R. A. (1980). "Evaluation of drill with visual aids for speech training," in *Speech assessment and speech improvement for the hearing impaired*, edited by J. Subtelny, (A.G. Bell Society for the Deaf, Washington, DC).

House, A. S., Goldstein, D. P., and Hughes, G. W. (1968). "Perception of visual transforms of speech stimuli: Learning simple syllables," Am. Annals of the Deaf, 113, 2, 215-221.

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

Kuhn, G. M. and McGuire, R. McI. (1974). "Results of a VCV spectrogram reading experiment," Status Report on Speech Research SR 39/40, Haskins Laboratories, New Haven, CT.

Lamel, L. F. (1983). "The use of phonotactic constraints to determine word boundaries," J. Acoust. Soc. Am. Suppl. 1, 74, S15.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1968). "Why are speech spectrograms hard to read?" Am. Annals of the Deaf, 113, 2, 127-133.

Lindblom, B. E. F. and Svensson, S-G. (1973). "Interaction between segmental and nonsegmental factors in speech recognition," IEEE Trans. Audio and Electroacoust., AU-21, 6, 536-545.

Lippman, R. P. (1982). "A review of research on speech training aids for the deaf," in Speech and language: Advances in basic research and practice, Vol. 7, edited by N. J. Lass, (Academic Press, NY).

Luce, P. A. and Carrell, T. D. (1981). "Creating and editing waveforms using WAVES," in Research on Speech Perception. Progress Report No. 7, Speech Research Laboratory, Indiana University, Bloomington.

Maki, J. (1980). "Visual feedback as an aid to speech therapy," in Speech assessment and speech improvement for the hearing impaired, edited by J. Subtelny, (A.G. Bell Society for the Deaf, Washington, DC).

Nisbett, R. E. and Wilson, T. D. (1977). "Telling more than we can know: Verbal reports on mental processes," Psychol. Rev., 84, 231-259.

Oshika, B., Zue, V. W., Weeks, R. V., Nue, H., and Aurbach, J. (1975). "The role of phonological rules in speech understanding research," IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-23, 104-112.

Potter, R. K., Kopp, G. A., and Green, H. C. (1947). Visible Speech (Van Nostrand, New York; reprinted by Dover, 1966).

Reber, A. S. (1967). "Implicit learning of artificial grammars," J. of Verbal Learning and Verbal Behavior, 5, 855-863.

Reber, A. S. and Lewis, S. (1977). "Toward a theory of implicit learning: The analysis of the form and structure of a body of tacit knowledge," Cognition, 5, 333-361.

Reber, A. S., Kassin, S. M., Lewis, S. and Cantor, G. (1980). "On the relationship between implicit and explicit modes in the learning of a complex rule structure," J. of Exp. Psych.: Hum. Learning and Memory, 6, 492-502.

Shepard, R. N. (1967). "Recognition memory for words, sentences, and pictures," J. of Verbal Learning and Verbal Behavior, 6, 156-163.

Shockey, L. and Reddy, R. (1974). "Quantitative analysis of speech perception: Results from transcription of connected speech from unfamiliar languages." Paper presented at the Speech Communication Seminar, Stockholm, Sweden.

Standing, L. (1973). "Learning 10,000 pictures," Quart. J. of Exp. Psychol., 25, 207-222.

Standing, L., Conezio, J., and Haber, R. N. (1970). "Perception and memory for pictures: Single trial learning of 2560 visual stimuli," Psychonomic Science, 19, 73-74.

Stark, R. E. (1974). Sensory capabilities of hearing-impaired children (University Park Press, Baltimore, MD).

Stark, R. E., Cullen, J. K., and Chase, R. A. (1968). "Preliminary work with the new Bell Telephone Visible Speech Translator," Am. An. of the Deaf, 113, 2, 205-214.

Stewart, L. C., Houde, R. A., and Larkin, W. D. (1976). "A real time sound spectrograph with implications for speech training for the deaf," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, PA, pp. 590-593.

Zue, V. W. (1981). "Acoustic-phonetic knowledge representation: Implications from spectrogram reading experiments," paper presented at the 1981 NATO Advanced Summer Institute on Automatic Speech Analysis and Recognition, Bonas, France, pp. 203-222.

Zue, V. W. and Cole, R. A. (1979). "The use of context during spectrogram reading," Speech communication papers presented at the 97th meeting of the Acoust. Soc. of Am., edited by J. J. Wolf and D. H. Klatt, (Acoust. Soc. Am., New York). (a)

Zue, V. W. and Cole, R. A. (1979). "Experiments on spectrogram reading," IEEE International Conference on Acoustics, Speech, and Signal Processing, Washington, D. C., pp. 116-119. (b)

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 9 (1983) Indiana University]

Developmental trends in the classification and perceived similarity of spoken syllables*

Amanda C. Walley and Linda B. Smith

Department of Psychology
Indiana University,
Bloomington, Indiana 47401

and

Peter W. Jusczyk

Department of Psychology
University of Oregon
Eugene, Oregon

*We thank the children and teachers of Grandview Elementary School for their participation and co-operation in conducting this study. M. Intons-Peterson, D. B. Pisoni and R. Treiman provided valuable comments on an earlier version of this paper. Jackie Baker, Melanie Lockyear and Carol McCord assisted in data collection and manuscript preparation. The work reported here was supported by NSF grants BNS 78-13019 and 81-09888 to the second author, by N.S.E.R.C. (A-0282) and N.I.C.H.H.D. (HDNS-15795) grants to the third author, by NIMH (MH-24027) and NIH (NS-12179), and by a doctoral fellowship from the Social Sciences and Humanities Research Council of Canada awarded to the first author.

Abstract

In two experiments, we investigated the factors which influence the perceived similarity of speech sounds at two developmental levels. Kindergarteners and second graders were asked to classify nonsense words, which were related by syllable or phoneme correspondences. The results support the existence of a developmental trend from the wholistic to analytic perception of speech sounds. Moreover, one significant factor in determining perceived similarity even in young children appears to be the position of correspondences between speech sounds; attention to the beginnings of utterances may have developmental priority. An unexpected finding was that the linguistic status of the unit involved in a correspondence (whether it was a syllable or a phoneme) did not seem particularly important. Apparently, the factors which contribute to the perceived similarity of speech sounds in the classification task are not identical to those which underlie performance in explicit segmentation and manipulation tasks. Rather, the factors important for perceived similarity may tap a level of processing that is closer to the one entailed in word recognition and lexical access.

Developmental trends in the classification and perceived similarity of spoken syllables

Despite their rather sophisticated ability to prcduce and understand spoken language, young children of about four to five years of age experience considerable difficulty in tasks requiring explicit phonemic judgments and manipulations. For example, the young child is poor at judging the number of phonemic segments in spoken syllables, words and nonsense words (Elkonin, 1973; Liberman, Shankweiler, Fischer & Carter, 1974; Rozin, Bressman & Taft, 1974; Treiman & Baron, 1981), at making same-different judgments about phonemic segments (Calfee, Chapman & Venezky, 1972; Jusczyk, 1977; Savin, 1972) and at rearranging and deleting specified segments (Bruce, 1964; Rosner & Simon, 1971). In contrast to the phoneme, the syllable appears to be a more readily accessible unit. Young children are better at "counting" (tapping out or laying down poker chips corresponding to) the number of syllables than the number of phonemes in spoken words and nonsense words (Liberman et al., 1974; Treiman & Baron, 1981).

According to some investigators (e.g., Gleitman & Rozin, 1977; Liberman, Shankweiler, Liberman, Fowler & Fischer, 1977; Rozin & Gleitman, 1977), phonemic segments do have "psychological reality" for young children (i.e., constitute the basis for their internal representations of speech), but are not accessible for conscious inspection due to the context-dependent manner of their encoding in the speech waveform (see Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967); the ability to access these units may depend on, or at least be promoted by, experience in reading an alphabetic orthography, in which orthographic symbols map onto the spoken language at roughly the level of the phoneme (see Morais, Cary, Alegria & Bertelson, 1979). Syllables, like phonemes, exhibit contextual variability, but explicit syllable segmentation and manipulation by young children may be facilitated by the distinctive peak of acoustic energy associated with the vocalic nucleus contained in all syllables (Liberman *et al.*, 1977). The earlier accessibility of syllables in development provides some support for the notion that the syllable constitutes a perceptually more primary unit than does the phoneme (e.g., Aslin, Pisoni & Jusczyk, 1983; Bertoncini & Mehler, 1981; Jusczyk, 1982, 1983; Mehler, 1981· Savin & Bever, 1970; but see Foss & Swinney, 1973).

The greater difficulty of explicit phoneme judgments relative to syllable judgments for the young child may be one manifestation of a more general developmental trend. A number of findings in the perceptual-cognitive developmental literature suggests that the immature perceptual system tends to perceive (represent and/or process) stimulus arrays wholistically. For example, in free classification tasks, young children group visual stimuli that vary along (what are for adults) perceptually separable dimensions, such as color and form, by overall similarity. In such classifications, within-group similarity is maximized and between-group similarity is minimized. In contrast, older children and adults perceive such stimuli dimensionally and tend to make classifications based on shared values or identities on the separate dimensions (e.g., Smith & Kemler, 1977; 1978; Shepp, 1978). Recently, the notion of a developmental trend from wholistic to analytic perception has been extended to the realm of speech processing by Treiman and her associates (Treiman & Baron, 1981; Treiman & Breaux, 1982). The results of free and constrained classification tasks, as well as examinations of memory confusions, support the notion that young children are not completely insensitive to the sound structure of language. However, whereas older children and adults are sensitive to the common phoneme relations among spoken syllables, young children appear to be more sensitive to overall similarity relations. It may be that children can attend better to relations that are defined across the syllable than to those defined across the phoneme. Therefore, the child's poor performance in phonemic segmentation and manipulation tasks may not be a limitation which is uniquely tied to the contextual dependence of phonemic segments. Rather, it might be one instantiation of the fact that qualitatively different relations are primary in perception for the younger as opposed to the older child or adult.

In the present study, we sought further evidence bearing on the existence of a developmental trend from the wholistic to analytic perception of speech. What factors contribute to young children's documented tendency to classify speech sounds wholistically? In the stimulus sets that have been employed in Treiman's experiments, overall similarity relations could be computed by subjects in a number of ways: on the basis of featural overlap, the similarity of phonemic segments or overall goodness of fit (i.e., in template-matching fashion) at the level of the syllable. In the present experiments, the possible bases for wholistic perception were examined more closely by studying children's classifications of stimuli that were related to one another by component identities, the number, linguistic status and position of which were varied. One question of particular interest was whether any wholistic tendency in young children's classifications might manifest itself such that the syllable would appear to be a more important, salient unit than the phoneme early in perceptual development. In other words, the wholistic perception of speech by young children might be expressed not only in the primacy of overall similarity relations as defined by Treiman (i.e., based on the sum of adult similarity ratings for individual phonemes), but also in the primacy of a unit (the syllable) larger than the phoneme.

We were also interested, however, in assessing the extent to which the structure of the stimulus sets and the classification talk might promote classifications based on phonemic correspondences. Treiman's research has indicated that in a free classification task, young children prefer to make classifications on the basis of overall similarity relations. This work has also shown that even in situations where children are reinforced for making phonemic classifications, wholistic classifications predominate. However, in both types of tasks, subjects are typically presented, on any given trial, with triads of test items, in which pairs of items are related by a phonemic correspondence, overall similarity or not related at all. The availability of overall similarity relations might interfere with a subject's ability to note phonemic correspondences among stimuli. Another issue addressed in the present study then is what children's performance reflects about their basic sensitivity to such correspondences when conflicting stimulus relations are absent in a classification task.

Experiment 1

In this experiment, we were interested first in determining in what sense young children's perception of speech might be more wholistic than that of older children. In judging the similarity of speech patterns, do young children attend to the same type of information (e.g., phonemic segments) as older children, but "simply" require more of this information? Alternatively, different information might be important at different developmental levels; the syllable might, for example, have some special status for young perceivers. If this is the case, then young children might perform as well as older children in grouping together speech sounds which share a whole syllable and perform more poorly only when the relation is one of a single phonemic identity. Again, younger children's perceptions of speech might be more wholistic than older children's not in the sense that they are totally unable to decompose speech into constituent units, but rather in the sense that they decompose it into larger constituents than do older listeners.

A second goal of the present experiment was to determine the minimum correspondence (in terms of phoneme vs. syllable overlap) required by young children for them to perceive and judge speech patterns as similar. In other words, what amount and kind of overlap or correspondence is sufficient for perceived similarity and how does this compare with the information required by older children? In order to assess children's basic ability (vs. preferences) in making classifications, the stimulus arrays and classification task employed in the experiment were structured such that overall similarity relations and constituent identities (phonemes and syllables) did not conflict with one another.

In the present experiment, we sought evidence bearing on developmental changes in the factors that contribute to the perceived similarity of speech patterns. A category generalization task was employed for this purpose. In the initial, pretraining phase of the task, children learned one exemplar (the standard) from each of two categories. The two categories were defined as the sounds made by specific puppets. The child was then presented with novel, test items and, for each test item, was asked which puppet would have made that sound. All stimuli were two syllables in length and of the structure CV-CV.

Table 1 shows an example of a stimulus set employed in the experiment. The two standards in the set, /nu-li/ and /ba-go/, differ maximally on each same-position phoneme. The test items are of three kinds and share either the first C, CV, or CVC with one standard and differ maximally from that standard on all remaining phonemes and from the other standard on all phonemes. Thus, each test item is maximally similar to one and only one standard, but the three test-item types differ in the kind and amount of overall similarity. Moreover, overall similarity is determined strictly by identity on one or more constituents; i.e., a test item which is similar overall to a particular standard shares one or more phonemes with that standard and none with the other standard. Thus, in the stimulus arrays, the two types of relations do not conflict with each other.

------------ ------------

Insert Table 1 about here

---------- ---------------

If subjects' classification performance is highly sensitive to the degree of wholistic similarity of test items and standards, then classification of CVC_ items should be more accurate than classification of CV__ items, which in turn should be more accurate than classification of C___ items (see the top panel of Figure 1). In other words, performance should be good when there are many units of correspondence and relatively poor when there are fewer shared units. However, it subjects decompose the stimuli into their constituent phonemes and if they selectively attend to the initial consonant, then they might perform equally well on all three types of test items (see middle panel of Figure 1). We expected this pattern of performance for second graders; previous research has shown that by this age, children are able to isolate individual phonemic components. The initial consonant identity in our stimuli provides a sufficient basis for partitioning all items into two groups. Classification by this strategy, if it is available, might, therefore, entail the least amount of effort. Alternatively, the implementation of this strategy might be based on some inclination to exercise strategies that have recently been added to the child's repertoire (perhaps through reading instruction). A third pattern of performance is also possible. If subjects analyze the stimuli into their constituent syllables, but do not decompose syllables into component phonemes, then classification of the CVC_ and CV__ items (the items that share an entire syllable with one of the standards) should be highly accurate, but classification accuracy for the C___ items should be relatively low (see bottom panel of Figure 1). We wondered whether kindergarteners, who are poor at phoneme manipulation tasks, might prove to be competent classifiers of speech sounds by constituent syllable identities. That is, their perception might not be wholistic in the sense that they are totally unable to decompose speech into its component units. Rather, kindergarteners might analyze speech into larger units than do older children. If so, they might be expected to show this third pattern of performance. However, if their perception is more wholistic in the sense of being closely tied to the amount of phoneme overlap among test items and standards, then their performance should reflect that pattern shown in the top of the figure.

Insert Figure 1 about here

TABLE 1

A Sample Stimulus Set from Experiment 1

Standards	nuli	bago
Test-item type		
1. C _ _ _	nə^tae	bə^tae
	naešə^	baešə^
2. C V _ _	nutə^	batə^
	nušae	bašae
3. C V C _	nulae	bagae
	nulə^	bagə^

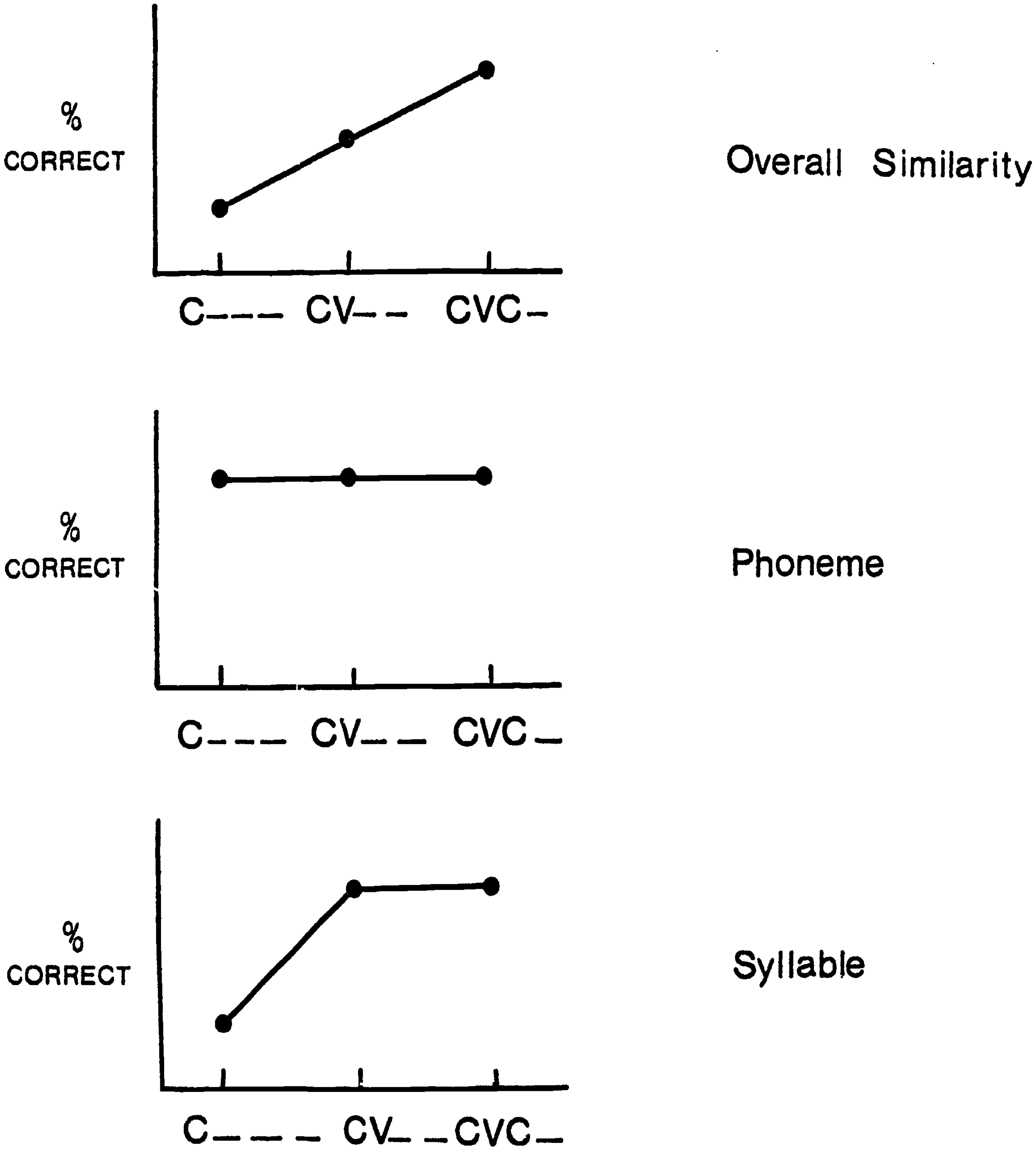


Figure 1. Predictions for classification performance in Experiment 1.

Method

Subjects

Twelve kindergarteners (Mean age = 5 years, 11 months; Range = 5,7 to 6,1; 6 males, 6 females) and twelve second graders (Mean age = 7,10; Range = 7,6 to 8,3; 8 males, 4 females) enrolled in a middle-class Indiana elementary school participated in the experiment. All subjects met the criterion of 10 correct consecutive responses in training and 67% correct classifications of the standards in testing. No gross speech or hearing disorder was reported for any child by parents at the time of testing and all subjects were native speakers of English.

Stimuli and Design

A pool of 6 consonants (/b, g, l, n, t, š/), each of which differed from one another by approximately the same amount according to adult perceptual similarity ratings collected by Singh, Woods and Becker (1972), was chosen for this experiment. Similarly, a pool of 6 vowels (/i, æ , a, ɚ , o, u/) was selected, such that the vowels were, in accordance with perceptual similarity estimates obtained by Singh and Woods (1971), approximately as dissimilar from one another as were the consonants. Two stimulus sets were then constructed from these consonant and vowel pools. Within a stimulus set, two standards (two CV-CV stimuli) were constructed by drawing from the consonant and vowel pools without replacement. For example, the standards in one test set were /nu-li/ and /ba-go/. Test items (also CV-CV stimuli) were then chosen such that they shared either the first C, CV or CVC with one, and only one, of the two standards; the remainder of a given test item consisted of a combination of items left in the consonant and vowel pools. Within a stimulus set, two different test items were related to either of the standards in one of the ways specified above (e.g., shared the initial C). Table 1 shows one of the stimulus sets.

The stimuli were produced in citation form by a female talker (Experimenter 1), who maintained approximately equal stress assignment between the two syllables of a given stimulus. The stimuli were recorded inside a sound-attenuated IAC booth using a high-quality microphone and Ampex (Model AG 500) tape recorder. Two audiotapes were prepared for each stimulus set. Each tape consisted initially of a randomized sequence of the standards for a particular stimulus set. This sequence was followed by the test items in the set. Each test item occurred twice; thus, there were, in total, eight test trials for each of the three item types. These 24 test trials occurred in random order, with the restriction that one of the standards intervened between every 4 test items (i.e., each training stimulus occurred 3 times within the test block) and that the order of standard presentation was alternated. Thus, a test block consisted of 30 stimulus presentations. Subjects were randomly assigned to one of the stimulus sets and the assignment of puppets to standards within a set was counterbalanced.

Procedure

Each subject was tested individually in a single session lasting no more than 45 minutes. The session included the auditory classification task (pretraining and testing), and, in the case of kindergarteners, an assessment of beginning reading ability (see below).

The experimental session began with pretraining. The subject was seated at a table facing two puppets. Experimenter 1 sat facing the child and Experimenter 2 to one side. The child was informed that each of the two puppets made a "special" sound that the child was to learn. The child was asked to pat the correct puppet on the head whenever he/she heard the puppet's sound. The standards (and test items in testing) were presented to subjects by audiotape on a portable Uher (Model 4200) tape recorder at a comfortable listening level. Experimenter 2 operated the tape recorder and stopped the tape whenever this was necessary for the child to complete his/her response on a trial. Experimenter 1 recorded each response from the child and informed him/her whether or not he/she was correct. After meeting the criterion of 10 correct consecutive responses on a maximum of 45 trials, the testing phase of the task was initiated.

In testing, the child was told that he/she was going to hear several more sounds made by the puppets. The child was told that each new sound (the test items) was made by one and only one of the puppets. The child was to indicate which puppet made the sound by patting one puppet on the head. Experimenter 2 operated the tape recorder, presenting one stimulus at a time. No feedback, only general encouragement, was given for the child's responses to test items. The child was given feedback on the trials for training items that were interspersed between the test items. A criterion of 67% correct responding on these trials was established for inclusion in the data analysis.

The final phase of the session for kindergarteners consisted of an assessment of "reading ability". Given previous findings concerning the relationship between phonemic analysis skills, the acquisition of spelling-sound rules and reading success (see Rozin & Gleitman, 1977; Treiman, 1980), we were interested in obtaining some measure of reading ability in kindergarteners. Of course, for second graders, developmental level and reading experience are necessarily confounded; however, in the event that kindergarteners' classification performance in the present study was similar to second graders', we wanted some indication of the extent to which this might be due to reading experience and/or ability vs. pre-existing analysis skills.

The reading assessment portion of the session began with a picture-letter matching task. On each of 10 trials, the child was shown 3 pictures and 1 alphabetic symbol which began the name of only one of the objects pictured. The child was asked to name each picture and was supplied with the intended name if that name was not volunteered. The child was then asked to point to the picture that "went best" with the letter shown to the side of the pictures. (On a given trial, the alphabetic symbol shown represented a sound in the target picture only.) Experimenter 1 recorded each child's responses. Only those subjects who responded correctly on 8 out of of the 10 trials advanced to the word identification part of the reading test.

The word identification test consisted of 20 real and nonsense words printed on 3 x 5" cards. There were equal numbers of one-syllable real words and one-syllable nonsense words. During this test, Experimenter 1 held up the cards with the test items one at a time in random order. Experimenter 2 transcribed all readings of each word offered by the child. Each child received the one-syllable real and one-syllable nonsense word lists in this same order. If a child could not read any of the first 4-5 words in the one-syllable real word list, the session was terminated. Before testing on the nonsense words, the child was told that he/she was going to see some "pretend" words that Experimenters 1 and 2 had made up and to try to say these aloud even though the words might sound silly.

Results and Discussion

In pretraining, kindergarteners reached the criterion of 10 consecutively correct associations of standards to puppets in 13.17 trials; the second graders met this criterion in 11.15 trials. During the test phase, the kindergarteners correctly classified the standards on 92% of the trials; the second graders did so on 98% of the trials. These differences were not statistically significant.

Group Classification of Test Items

Each subject's number of correct classifications of test items was submitted to an analysis of variance for a 2(Grade) x 3(Test Item) mixed design. The analysis revealed a main effect of Test Item ($F(2,44) = 9.91$, $p < .001$) and a reliable interaction between Grade and Test Item ($F(2,44) = 4.15$, $p < .025$). Figure 2 shows the mean proportion of correct classifications for the three test-item types at the two grade levels. Post-hoc analyses (Tukey's test, $\alpha = .05$, critical difference in proportion correct = .15) revealed that the second graders performed equally well on all test-item types. Consistent with our expectations then, the second graders were well able to decompose the bisyllabic utterances used in this experiment into phonemic units and to selectively attend to the initial consonant shared by all test items with the standards. The kindergarteners' performance, unlike that of the second graders, did depend on test-item type. Post-hoc analyses (Tukey's test, $\alpha = .05$) indicated that kindergarteners classified the CVC_ items more accurately than the CV__ or C___ items, but that the level of performance on these last two types did not differ. Therefore, a constituent syllable identity was apparently not particularly salient to these younger subjects. Rather, kindergarteners performed as well as the older children only when the test item was extremely similar overall to a standard (i.e., was identical on 3 out of 4 of its constituents) and they performed reliably more poorly than the older children both when the test item and standard shared a whole syllable and when they shared only the initial phoneme. Nevertheless, the kindergarteners' pattern of performance resembles that which would be expected for classifications on the basis of overall similarity, where overall similarity is determined by the number of shared phonemes between a given standard and test item.

Insert Figure 2 about here

Individual Patterns of Performance

The patterns of individual performances are consistent with the results of the group analyses. A child was scored as having consistently classified each test-item type if he or she correctly classified items of that type on at least 7 of the 8 possible trials ($\chi^2(1) = 4.50$, $p < .05$). As can be seen in Table 2, many more second graders than kindergarteners consistently classified items of all three types ($\chi^2(1) = 8.70$, $p < .005$) -- a pattern of performance consistent with the notion that older children are better able to attend to the phonemic structure of speech. Notice also in Table 2 that the majority of all subjects (83%) fall into one of three categories -- successful classification by initial consonant, successful classification only of test items maximally similar to one standard, or inconsistent classification of all test-item types. Very few subjects' pattern of performance then fits that expected by the hypothesis that the syllable is a perceptually more salient unit than the phoneme or that it is accessible earlier in development. Rather, the data suggest that the preferred pattern of responding is according to maximal similarity for kindergarteners and according to initial phoneme identity for second graders.

Insert Table 2 about here

Kindergarteners' Classification Performance and Reading Ability

The kindergartener's classification performance was related to success on the letter-name matching task (Pearson's $r = .82$, $p < .01$). Only 2 of the kindergarteners were able to read 5 or more of the 10 real words in the visual word identification task; 1 of these subjects correctly classified all test items and the other correctly classified only the test items that shared a syllable with the standards. These 2 early readers, unlike their classmates, appeared to be highly sensitive to the component phoneme and syllable identities among the stimuli.

Conclusion

The results of this experiment indicate that younger children's comparisons of speech sounds are more strongly influenced by overall similarity than are older children's comparisons. As a group, the kindergarteners correctly classified only test items that were highly similar to (i.e., shared 3 out of 4 phonemes with) the standards. Second graders, in contrast, classified all the

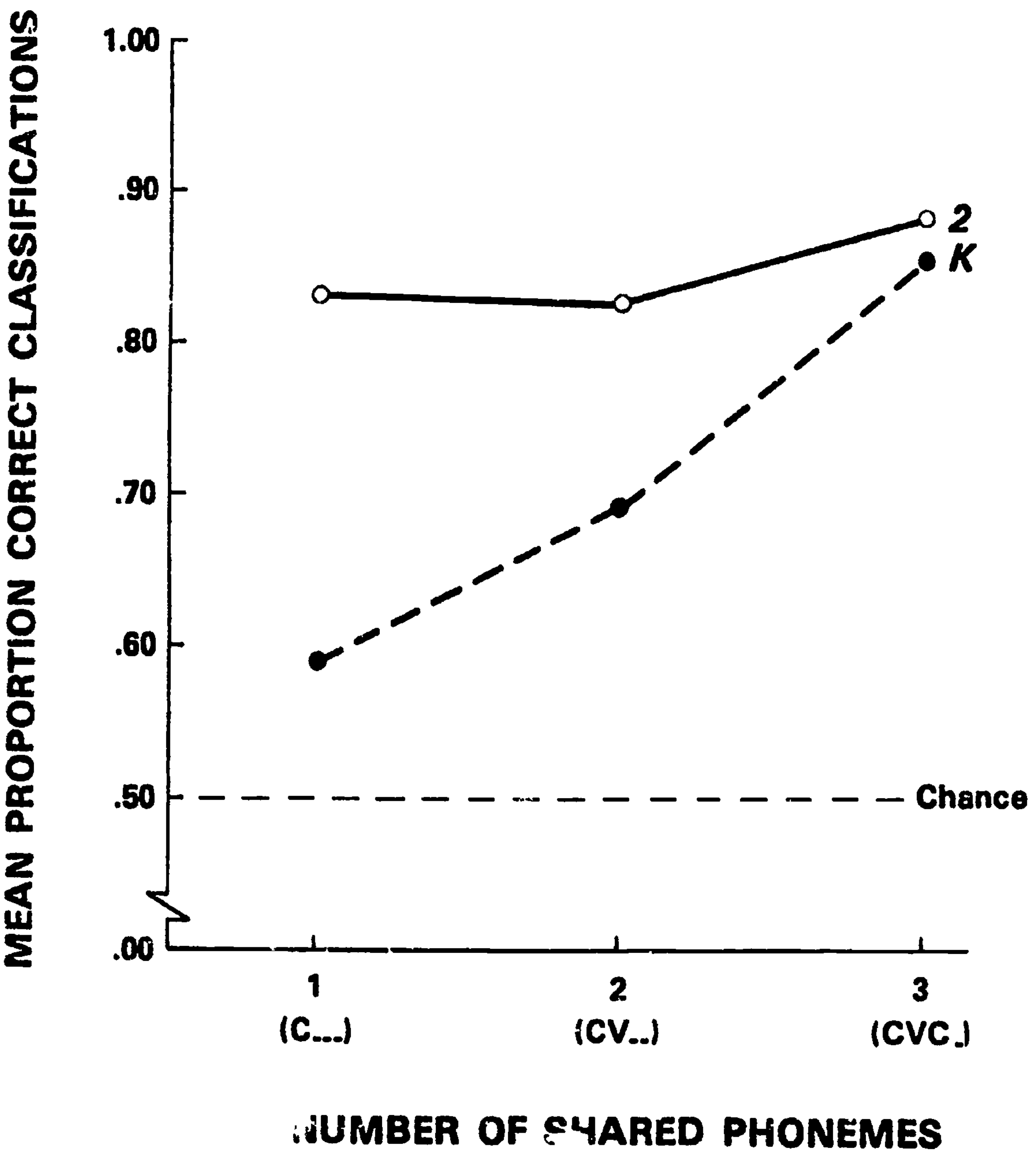


Figure 2. Group classification performance in Experiment 1.

TABLE 2

Individual Patterns of Classification Performance

in Experiment 1 (number of subjects)

	Consistent			Inconsistent
	Initial phoneme	Syllable	Maximal similarity	
	C _ _ _ C V _ _ C V C _	C V _ _ C V C _	C V C _	
Grade				
K	1	1	6	4
2	8	2	1	1
Total	9	3	7	5

test items correctly -- whether they shared one phoneme, a two-phoneme syllable, or three phonemes with the correct standard. Apparently, when there exist no competing similarity and phonemic relations between standards and test items, a single phoneme identity is sufficient for second graders to note that correspondence. This degree of correspondence does not seem salient, however, to kindergarteners.

One surprising result of this experiment was that a shared syllable appeared to have no special status in either the younger or older children's comparisons of speech sounds. According to previous theorizing (e.g., Gleitman & Rozin, 1977; Liberman et al., 1977; Rozin & Gleitman, 1977), syllables are perceptually primary units and relative to phonemes, they are accessible earlier in development. Therefore, in this task, one might have expected that syllable correspondences would seem particularly salient to kindergarteners. Yet, 10 out of the 12 kindergarteners tested failed to classify together items that shared their initial syllable. Note that failure with syllable-identity classifications was not due simply to a failure to do the task. All the kindergarteners were able to classify an identical test item with a standard and 75% of the kindergarteners correctly classified the test items that shared 3 of 4 constituent phonemes with a standard. These kindergarteners then were clearly attending to the test items and when similarity relations were maximal, kindergarteners were able to classify them. However, a shared syllable among test items and standards did not present sufficiently strong similarity relations for kindergarteners to classify such items easily and consistently. Moreover, a constituent syllable identity did not appear to assume any special role in classification performance at a later developmental level.

Experiment 2

One interesting possibility raised by Experiment 1 is that the syllable unit per se has no special status in determining children's judgments about the overall similarity of different utterances. Rather, what appears to be critical is the number of shared elements, and not necessarily whether these common elements occur together within the same syllable. To the extent that it is simply the number of common elements between two utterances that dictates whether or not they will be classified together, one would expect to find no advantage for the classification of items in which the common elements are members of the same syllable versus those in which the shared elements are in different syllables. Experiment 2 was designed to explore this hypothesis.

In this experiment, we further investigated the perceived internal structure of speech sounds for children by again asking them to classify nonsense sounds. Table 3 shows a sample stimulus set and illustrates the structure of the 4 different test-item types used in the experiment. All four test-item types share two phonemes (one consonant and one vowel) with one standard and no phonemes with the second standard. However, the test-item types differ in which two constituent phonemes are shared with a given standard. If the perceived similarity of speech sounds for children is simply determined by the number of corresponding phonemes (a possibility consistent with the results obtained with kindergarteners in Experiment 1), then classification of all the test items should be equally easy (or difficult). If the structure of the correspondences matters (i.e., whether or not the common elements are members of the same syllable), then the four test-item types should not be psychologically equivalent. In the first two test-item types (C1V1__ and __C2V2), the consonant and vowel combine to form one syllable. If whole syllable identities are partiaularly salient for children, these two types should be classified correctly more often than the remaining two types.

Insert Table 3 about here

This contrast between syllable vs. nonsyllable identity is not the only possible one of interest. Some researchers in speech perception and language acquisition have maintained that the beginnings of words are particularly salient (e.g., Cole 1973; 1981; Cole & Jakimik, 1980; Marslen-Wilson & Welsh, 1978; Marslen-Wilson & Tyler, 1980; Slobin, 1973; Tyler & Marslen-Wilson 1982; see also Foss & Blank, 1980; Grosjean, 1980; Salasoo & Pisoni, 1982; Williams, Blumberg & Williams, 1970). If a phonemic identity (in particular, a consonant identity) at the beginning of an utterance is a primary determinant of perceived similarity, then test items C1V1__ and C1__V2 should be classified correctly more often than the remaining two items.

Finally, it has also been proposed that the ends of words play a special role in the perception of spoken language (Cole & Jakimik, 1980; Slobin, 1973). If a phonemic identity at the end of an utterance (in particular, a vowel identity) is a major determinant of perceived similarity, then test items __C2V2 and C1__V2 should be classified correctly more often than the remaining two items. In other words, the four test-item types can be partitioned into the three orthogonal comparisons summarized in Table 4. The first comparison contrasts items sharing an entire syllable with the correct standard versus those sharing two phonemes that are parts of separate syllables. The second comparison contrasts test items sharing an initial consonant with those items that differ from the standard in the initial consonant. The third comparison contrasts test items sharing their final vowel with the correct standard versus those items which differ from the correct standard in the final vowel. The question of interest is whether any of these structural aspects of the correspondence between test items and standards is perceptually more dominant than the others. We expected, in contrast to the results of Experiment 1, that this more sensitive measure of the role of syllable identity in children's comparisons of speech sounds would show that phonemic identities united in one syllable result in increased perceived similarity relative to phonemic identities separated into distinct syllables.

Insert Table 4 about here

TABLE 3

A Saaple Stimulus Set From Experiment 2

Standards	šona	lə^tu
Test-item type		
1. C_1V_1 — —	šogae	lə^gi
	šobi	lə^bae
2. — — C_2V_2	gina	bitu
	bina	gitu
3. C_1 — — V_2	šaega	ligu
	šiba	laebu
5. — V_1C_2 —	bonae	gə^ti
	goni	bə^ti

TABLE 4

Summary of Orthogonal Comparisons in Experiment 2

Comparison type			
1. Syllable	C_1V_1 — —	vs.	C_1 — — V_2
	— — C_2V_2		— V_1C_2 —
2. Initial consonant	C_1V_1 — —	vs.	— — C_2V_2
	C_1 — — V_2		— V_1C_2 —
3. Final vowel	— — C_2V_2	vs.	C_1V_1 — —
	C_1 — — V_2		— V_1C_2 —

Method

Subjects

Twelve kindergarteners (Mean age = 5 years, 9 months; Range = 5,3 to 6,0; 3 males, 9 females) and twelve second graders (Mean age = 8,2; Range = 7,11 to 8,7; 7 males, 5 females), who were enrolled in an Indiana elementary school serving a middle-class population, participated in the experiment. All subjects met essentially the same criteria as specified for Experiment 1, including 10 correct consecutive responses in training and 75% correct classifications of the standards in testing.

Stimuli and Design

Two sets of CV-CV stimuli were constructed from the consonant and vowel pools described in Experiment 1. Each stimulus set consisted of 4 different test-item types; i.e., a test item was related to one of the two standards or training items of the set in 1 of 4 possible ways: A test item was identical to one of the standards in i) its initial syllable (C1V1__), ii) its second (or final) syllable (__C2V2), iii) its initial and final phonemic segments (C1__V2), or iv) its two medial phonemes (_V1C2_). The remainder of any given test item consisted of a combination of the items left in the consonant and vowel pools. Within each stimulus set, two different test items were related to either of the standards in one of the ways specified above (see Table 3).

Audiotapes of the two stimulus sets were prepared in the same way as specified for Experiment 1. The test items occurred twice within the test sequence of a tape in random order (i.e., 32 test trials), with the restriction that one of the training items intervened between every 4 presentations of a test stimulus. Thus, a test block consisted of 40 stimulus presentations. Subjects were randomly assigned to one of the stimulus sets and the assignment of puppets to standards within a set was counterbalanced.

Procedure

The procedure was identical to that of Experiment 1.

Results and Discussion

During pretraining, the kindergarteners and second graders learned to associate the two standards to the two puppets equally rapidly; the mean number of trials to criterion was 10.9 for kindergarteners and 10.1 for second graders. During testing, the kindergarteners correctly classified the standards on 96% of the trials; the second graders performed perfectly. This difference was not statistically significant.

Group Classification of Test Items

Each child's number of correct classifications of the four test-item types (C1V1__, __C2V2, C1__V2, _V1C2_) was submitted to a nested analysis of variance for a 2(Age) x 4(Test-Item) mixed design. The analysis revealed main effects of Age ($F(1,22) = 8.83$, $p < .01$), and Test-Item ($F(3,66) = 11.53$, $p < .001$) and a reliable interaction between Age and Test-Item ($F(3,66) = 5.03$, $p < .005$). Table 5 shows the mean proportions of correct classifications for each of the four test-item types at the two grade levels. Post-hoc comparisons (Tukey's test, α = .05, critical difference in proportion correct = .16) indicated that kindergarteners and second graders performed equally well on the test items identical to a standard in their initial syllable (C1V1__) and equally poorly on the test items identical to a standard in the medial vowel and consonant (_V1C2_). However, second graders correctly classified test items sharing the initial consonant *and* final vowel with a standard (C1__V2) and those sharing the final syllable (__C2V2) with a standard reliably more often than did kindergarteners.

Insert Table 5 about here

Two within-subject contrasts were also reliable in the second grade data (Tukey's test, α = .05): C1V1__ vs. _V1C2_ and C1__V2 vs. _V1C2_. Second graders' performance was high both when the test items shared the initial syllable with a standard and when it shared the initial consonant *and* final vowel with a standard. Performance was lowest when the test item shared only the medial vowel and consonant with a standard. In the kindergarteners' data, only the C1V1__ vs. _V1C2_ contrast was reliable. That is, these subjects classified correctly most often when a test item shared an initial syllable with a standard and least well when the test item did not share an initial syllable, initial consonant or final vowel with a standard. However, kindergarteners' classifications of C1V1__ items tended to be more accurate than their classifications of __C2V2 items.

The kindergarteners' and second graders' high level of performance on test items sharing an initial syllable with a standard could be due to either the *syllable* correspondence *or* the (initial) *position* of the correspondence (or both factors). The outcomes of the three planned orthogonal comparisons of performance on the test-item types suggest that whole syllable correspondences do not have any overriding importance and that it is instead the latter factor which is most important. (For all paired comparisons, $p < .05$ and the critical value for a difference in proportion correct is .11). At neither age level was the contrast between the two item types sharing a whole syllable and the two item types not sharing a syllable with the correct standard reliable (see Table 5). A syllable identity *per se* does not appear critical to perceived similarity. However, both kindergarteners and second graders correctly classified items sharing their initial consonant with a standard more often than items that differed from the correct standard in their initial consonant (see Table 5). A

TABLE 5

Mean Proportion Correct (and Standard Deviations) For the Four Test-Item Types in Experiment 2. (Shown at the bottom are the combined proportions correct that entered into the planned comparisons.)

	Grade	
Test-item type	K	2
1. C_1V_1 — —	.84 (.15)	.92 (.12)
2. — — C_2V_2	.68 (.20)	.86 (.23)
3. C_1 — — V_2	.75 (.24)	.97 (.06)
5. — V_1C_2 —	.65 (.12)	.70 (.18)
Comparison		
Syllable/Not S	.76/.70	.89/.84
Initial Consonant/Not IC	.80/.67*	.94/.78*
Final Vowel/Not FV	.72/.74	.92/.81*

*indicates reliable planned contrasts.

correspondence at the beginning of a speech sound -- be it a whole syllable or simply the initial consonant -- thus appears salient even to young children. Finally, the second graders, but not the kindergarteners, correctly classified test items that shared the final vowel with the correct standard more often than they did items that did not (see Table 5). Therefore, correspondences at the ends of speech sounds also appear to be important at least to the older children.

Taken together, these results suggest a developmental trend in those aspects of the internal structure of speech sounds that control their perceived similarity. Attention to the beginnings of utterances is developmentally prior to attention to the ends of utterances (cf. Williams *et al.*, 1970). In addition, correspondences in the ends of utterances apparently play a stronger role in perceived similarity than do correspondences involving whole syllables -- at least early in development. That is, the position of correspondences may be more important than the unit (phoneme vs. syllable) of correspondence. This interpretation contrasts with the hypothesis that the young child's perception of speech is structured by the wholistic unit of a syllable, but is consistent with the results of Experiment 1.

Individual Patterns of Performance

The patterns of individual performances support the notion that whole syllable correspondences have no special status in young children's perception of speech sounds. Table 6 shows the number of children correctly classifying each of the 4 types of test items on at least 7 of the 8 trials for that type. Note that the number of children succeeding is highest on C1V1__ and C1__V2 items -- the two item types that share their initial consonant with the correct standard. The high level of performance on item C1V1__ could be attributed to the syllable correspondence in initial position. However, the children succeeded on items containing a correspondence in only the initial consonant (C1__V2) just as frequently ($\chi^2(1)$ 1.00), whereas markedly fewer children succeeded with the item sharing a final syllable (__C2V2) with the correct standard ($\chi^2(1) = 10.39$, $p < .005$). The position of a correspondence between two speech sounds thus appears a more critical determinant of children's use of that correspondence than the *unit* of correspondence.

Insert Table 6 about here

Kindergarteners' Classification Performance and Reading Ability

The kindergarteners' performance on the classification task was not correlated with their ability to match letters to pictures. This result is not surprising since, in this experiment, all test items shared two phonemes with one of the standards and thus were relatively similar overall to that standard. Presumably, it is attention to the *phonemic* structure of speech that is most important for the decoding of alphabetically represented words. None of the

TABLE 6

Individual Patterns of Classification Performance

in Experiment 2 (number of subjects)

Grade	Test-item types classified consistently			
	C_1V_1 — —	— — C_2V_2	C_1 — — V_2	— V_1C_2 —
K	10	2	5	3
2	10	9	12	3
Total	20	11	17	6

kindergarteners in this experiment was able to read any of the words in the word identification task.

Conclusions

The results of this experiment suggest that there are age-related increases in attention to the internal, structural relations among speech sounds. The trend appears to be from reliance on initial to final phonemic correspondences. The surprising result was that whether a correspondence among utterances involves separate phonemes or phonemes united into one syllable does not matter greatly. Thus, although younger children's perception of speech might be characterized as wholistic in the sense that a greater amount of correspondence among speech patterns is required in order for them to be judged as similar, wholistic perception is not simply equivalent to syllable perception.

General Discussion

The results of the experiments reported here are consistent with previous research suggesting the existence of a developmental trend from the wholistic to analytic perception of speech sounds (see Treiman & Baron, 1981; Treiman & Breaux, 1982). Even when the nature of the stimulus arrays and the classification task were altered such that no conflict was present between overall similarity and component phoneme relations (as in Experiment 1), the kindergarteners' performance still depended strongly on overall similarity relations. Kindergarteners were able to classify correctly as well as second graders only when test items shared three out of four phonemes with the standards. Second graders' performance did not depend on the number of shared and irrelevant dimensions or phonemes in the stimulus array; apparently, they could selectively attend to one component identity (the initial consonant) among the test and training items and partition the stimulus array on this basis alone. However, when the minimum number of shared phonemes among test and training items in the classification task was increased and the number of irrelevant dimensions reduced (as in Experiment 2), the kindergarteners' performance improved and began to look more like that of second graders. For example, kindergarteners accurately classified items sharing an initial syllable and did so as well as second graders. In this sense then, the perceived similarity of speech sounds for kindergarteners is still more dependent on overall similarity relations.

In addition to providing support for previous research on children's classifications of speech sounds, the results of our investigation have served to specify what aspects of the internal structure of speech sounds are important for perceived similarity. Specifically, the position, as well as the number, of shared constituents (either phonemes or syllables) may be particularly salient; there appears to be a developmental trend whereby: 1) attention to the beginnings of sounds emerges prior to attention to the ends of sounds and 2) the ability to attend to final correspondences is stronger than when the correspondence involves a whole syllable (vs. separate phonemes). A surprising result of our experiments then was that the existence of a syllable correspondence per se did not seem to serve as a major determinant of the perceived similarity of speech sounds. Thus, younger children's greater reliance on overall similarity relations cannot be simply equated with syllable perception. Similarly, the greater tendency of older children to make more analytic classifications does not entail any special status for the syllable relative to the phoneme.

Our failure to observe any special status of the syllable in the perceived similarity of speech sounds does conflict with the wealth of empirical research indicating that the syllable, relative to the phoneme, is a much more psychologically accessible unit. As discussed earlier, young children experience substantial difficulty in tasks requiring explicit phoneme segmentation and manipulation. They perform substantially better in such tasks when the unit of segmentation is the syllable (see Rozin & Gleitman, 1977, for a review). The discrepancy between our results and those of these previous studies might be attributed to differences in the level of perceptual processing entailed in the classification and explicit segmentation tasks. The classification task only requires that the child decide which of two sounds (or standards) a target or test item is most like. In our implementation of the task at least, the child was not required to indicate how two sounds that were classified together were alike, nor was he/she actually required in any way to adopt a particular strategy to make a classification. In this respect, the classification task differs from explicit manipulation and segmentation tasks, where relatively greater conscious attention to units such as phonemes and syllables is necessary for successful performance. To the extent that the classification task requires a perceptual similarity match between a test item and two items in memory (the standards), it may more closely resemble the sort of situation where the child hears the word "telephone" and must decide whether he/she heard "telephone", "elephant", etc.

The processes underlying classification performance may be more similar to those involved in word identification and fluent speech perception than are those underlying the explicit segmentation and manipulation of speech sounds. Indeed, our findings concerning the importance of the number and position of correspondences to the perceived similarity of speech sounds is in keeping with recent accounts of the nature of auditory word identification (e.g., Marslen-Wilson & Welsh, 1978; Marslen-Wilson & Tyler, 1980; Tyler & Marslen-Wilson, 1982; see also Cole & Jakimik, 1980; Foss & Blank, 1980; Grosjean, 1980; Salasoo & Pisoni, 1982). According to these proposals, the acoustic-phonetic information corresponding to the beginnings of words plays a particularly important role in the word identification process. A similar characterization of spoken word recognition and lexical access in young children has been offered (Cole, 1981; Cole and Perfetti, 1980). In support of this is the finding that young children, like adults, are more accurate at detecting mispronunciations in word-initial vs. word-final position. Thus, the beginnings of words appear to be particularly salient in word identification and less attention is devoted to the analysis of subsequent acoustic-phonetic information. The importance of initial correspondences to perceived similarity early in development may serve then as part of the basis for this sort of word identification process.

The salience of initial correspondences may also serve as part of the foundation for the implementation of the sort of "operating principles" proposed by Slobin (1973) as characterizing language acquisition strategies. One procedure for discovering the formal and functional relations between linguistic elements, is, Slobin maintains, "Pay attention to the order of words and morphemes". Another is "Pay attention to the ends of words". Cole and Jakimik (1980) have also suggested that the isolation of a given word in fluent speech, and thus of its final phonemic segments, contributes to the identification of the subsequent word. However, as our results indicated attention to the final elements of spoken utterances may emerge later in development than attention to initial correspondences -- at least at the level of perception tapped in classification tasks. In any event, the findings reported here are generally in keeping with several proposals concerning the acquisition of spoken vocabulary. Moreover, they may be inuicative of how the comparison operations used to match encoded items with memory representations change with development and they are thus relevant to the development of spoken word recognition and lexical access.

The perception of the similarity of speech sounds then may involve relatively passive and unconscious encoding and pattern-matching procedures, such as those that have been used to characterize spoken word identification. In contrast, segmentation and manipulation tasks may depend on the deployment of additional, more strategic processes. Moreover, the sort of local factors which determine the ability to explicitly access and manipulate different linguistic units may not be precisely the same as those which influence the perceived similarity of speech sounds. For example, it has been suggested that the acoustic correlates of syllables may be relatively more salient and invariant than those of many phonemic segments (Gleitman & Rozin, 1977; Liberman _et al_, 1977). One such correlate may be the peak of acoustic energy associated with the vocalic nucleus of any syllable; amplitude fluctuations in the speech waveform may thus help to define syllable boundaries. This factor does not seem to be particularly important to perceived similarity, given our finding that initial _consonant_ correspondences are salient to children. Other types of factors, such as reading experience, have been implicated in the ability to consciously access phonemic and syllabic units, but these may also be less impor'ant at the level of processing where perceived similarity is determined. Our results concerning children's disregard for syllable correspondences thus point to the need for a more precise understanding of the processes mediating on-line wcrd recognition, the perceived similarity of speech sounds and the ability to explicitly manipulate the sound structure of language.

References

Aslin, R. N., Pisoni, D. B., & Jusczyk, P. W. Auditory development and speech perception in infancy. In M. Haith & J. Campos (Eds.), *Infancy and developmental psychobiology*. In Volume II of Paul H. Mussen's (Series Ed.), *Handbook of child psychology*, New York: John Wiley & Sons, 1983.

Bertoncini, J., & Mehler, J. Syllables as units in infant speech perception. *Infant Behavior and Development*, 1981, *4*, 247-260.

Bruce, D. J. The analysis of word sounds by young children. *British Journal of Educational Psychology*, 1964, *34*, 158-169.

Calfee, R. C., Chapman, R. S., & Venezky, R. How a child needs to think to learn to read. In L. W. Gregg (Ed.), *Cognition in learning and memory*. New York: Wiley, 1972.

Cole, R.A. Listening for mispronunciations: A measure of what we hear during speech. *Perception and Psychophysics*, 1973, *11*, 153-156.

Cole, R.A. Perception of fluent speech by children and adults. *Annals of the New York Academy of Sciences*, 1981, *379*, 92-109.

Cole, R. A., & Jakimik, J. A. A model of speech perception. In R. A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Cole, R. A., & Perfetti, C. A. Listening for mispronunciations in a children's story: The use of context by children and adults. *Journal of Verbal Learning and Verbal Behavior*, 1980, *19*, 297-315.

Elkonin, D. B. USSR. In J. Downing (Ed.), *Comparative reading*. New York: Macmillan, 1973.

Foss, D. J., & Blank, M. A. Identifying the speech codes. *Cognitive Psychology*, 1980, *12*, 1-31.

Foss, D. J., & Swinney, D. A. On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior*, 1973, *12*, 246-257.

Gleitman, L. R., & Rozin, P. The structure and acquisition of reading I: Relations between orthographies and the structure of language. In A S. Reber and D. L. Scarborough (Eds.), *Toward a psychology of reading*. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc., 1977.

Grosjean, F. Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics*, 1980, *28*, 267-283.

Jusczyk, P. W. Auditory versus phonetic coding of speech signals during infancy. In J. Mehler, M. Garrett & E. Walker (Eds.), *Perspectives in mental representation: Experimental and theoretical studies of cognitive processes and capacities*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1982.

Jusczyk, P. W. On characterizing the development of speech perception. In J. Mehler & R. Fox (Eds.), *Neonate cognition: Beyond the blooming, buzzing confusion*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1983.

Jusczyk, P. W. Rhymes and reasons: Some aspects of the child's appreciation of poetic form. *Developmental Psychology*, 1977, *13*, 599-607.

Liberman, A. M., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, *74*, 431-435.

Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology*, 1974, *18*, 201-212.

Liberman, I. Y., Shankweiler, D., Liberman, A. M., Fowler, C., & Fischer, F. W. Phonetic segmentation and recoding in the beginning reader. In A. S. Reber & D. L. Scarborough (Eds.), *Toward a psychology of reading*. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc., 1977.

Marslen-Wilson, W. D., & Tyler, L. K. The temporal structure of spoken language understanding. *Cognition*, 1980, *8*, 1-71.

Marslen-Wilson, W. D., & Welsh, A. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 1978, *10*, 29-63.

Mehler, J. The role of syllables in speech processing: Infant and adult data. *Philosophical Transactions of the Royal Society London*, 1981, *B295*, 333-352.

Morais, J., Cary, L., Alegria, J., & Bertelson, P. Does awareness of speech as a sequence of phones arise spontaneously? *Cognition*, 1979, *7*, 323-331.

Rosner, J., & Simon, D. The Auditory Analysis Test: An initial report. *Journal of Learning Disabilities*, 1971, *4*, 384-392.

Rozin, P., Bressman, B., & Taft, M. Do children understand the basic relationship between speech and writing? The mow-motorcycle test. *Journal of Reading Behavior*, 1974, *6*, 327-334.

Rozin, P., & Gleitman, L. R. The structure and acquisition of reading II: The reading process and the acquisition of the alphabetic principle. In A. S. Reber & D. L. Scarborough (Eds.), *Toward a psychology of reading*. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc., 1977.

Salasoo, A., & Pisoni, D. B. Sources of knowledge in spoken word classification. *Research on speech perception, Progress report No. 8*, Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University, 1982.

Savin, H. B. What the child knows about speech when he starts to learn to read. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationship between speech and reading*. Cambridge: MIT Press, 1972.

Savin, H. B., & Bever, T. G. The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior*, 1970, *9*, 295-302.

Shepp, B. E. From perceived similarity to dimensional structure: A new hypothesis about perceptual development. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization*. Hillsdale, N.J.: Erlbaum, 1978.

Singh, S., & Woods, D. R. Perceptual structure of 12 American English vowels. *Journal of the Acoustical Society of America*, 1971, *49*, 1861-1866.

Singh, S., Woods, D. R., & Becker, G. M. Perceptual structure of 22 pre-vocalic English consonants. *Journal of the Acoustical Society of America*, 1972, *52*, 1698-1713.

Slobin, D. I. Cognitive prerequisites for the development of grammar. In C. A. Ferguson & D. I. Slobin (Eds.), *Studies of child language development*. New York: Holt, Rinehart and Winston, Inc., 1973.

Smith, L. B., & Kemler, D. G. Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology*, 1977, *24*, 279-298.

Smith, L. B., & Kemler, D. G. Levels of experienced dimensionality in children and adults. *Cognitive Psychology*, 1978, *10*, 502-532.

Treiman, R. *The phonemic analysis ability of preschool children*. Unpublished doctoral dissertation, University of Pennsylvania, 1980.

Treiman, R., & Baron, J. Segmental analysis ability: Development and relation to reading ability. In T. G. Waller & G. E. MacKinnon (Eds.), *Reading research: Advances in theory and practice* (Vol. 3). New York: Academic Press, 1981.

Treiman, R., & Breaux, A. M. Common phoneme and overall similarity relations among spoken syllables: Their use by children and adults. *Journal of Psycholinguistic Research*, 1982, *11*, 581-610.

Tyler. L. K., & Marslen-Wilson, W. D. Speech comprehension processes. In J. Mehler, E. C. T. Walker and M. Garret (Eds.), *Perspectives on mental representation: Experimental and theoretical studies of cognitive processes and capacities*. Hillsdale, N.J.: Lawrence Erlbaum, 1982.

Williams, J. P., Blumberg, E. L., & Williams, D. V. Cues used in visual word recognition. *Journal of Educational Psychology*, 1970, *61*, 310-315.

Speech Perception: Some New Directions in Research and Theory*

David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

*Preparation of this manuscript was supported, in part, by NIH research grant NS-12179, NSF research grant BNS 83-05387 and contract F33615-83-K-0501 with the Air Force Office of Scientific Research, Air Force System Command. This is the text of an invited paper presented at the National Academy of Sciences, Committee on Hearing, Bioacoustics and Biomechanics (CHABA) meeting, October 19-20, 1983 in Washington, DC. A shorter version of this paper will appear in a special issue of the Journal of the Acoustical Society of America. I thank Howard Nusbaum, Paul Luce and Beth Greene for comments and suggestions on an earlier draft of this paper.

Speech Perception: Some New Directions in Research and Theory

This is an exciting time for researchers working in the field of speech processing. Many people feel that we are literally at the point of a major breakthrough in research and theory and that many of the old and long-standing problems in the field may be resolved within the next few years (Klatt, 1983). The solution to some of these problems could well have a number of practical implications such as the development of improved speech recognition systems, more sophisticated aids for the hearing impaired and a wide range of consumer products using speech I/O technology.

The increased optimism about speech processing was brought about, in part, by several different but closely related developments that have taken place in the last few years. First, the engineering and artificial intelligence communities realized that a final solution to the speech recognition problem could only be obtained by increased efforts at improving the front end performance of speech recognizers. Improved segmental recognition is one direct way to increase performance well beyond the levels obtainable with the currently available commerical technology. Many researchers working in this area now believe that such improvements will come about through new research efforts aimed at understanding how speech signals are coded at the auditory periphery.

Second, many investigators realized that our understanding of how human listeners perceive fluent continuous speech was quite impoverished compared to the voluminous research literature on the perception of isolated nonsense syllables and simple words. Moreover, it has become increasingly clear from the research carried out under the ARPA speech understanding project (Klatt, 1977), that a great deal more basic and detailed knowledge is needed about sentence-level phenomena in speech. The numerous acoustic-phonetic changes that are introduced at word boundaries and the operation of phonological rules in language create sources of variability in speech that are still poorly understood even today. These two areas of research on connected speech are assumed to be of great importance in eventually solving the recognition problem.

Third, several researchers have introduced new research strategies and techniques to study the different sources of variability in speech. Some of these efforts involve the use of large data bases of natural speech produced by multiple talkers which can be used to provide quantitative estimates of the probabilities of occurrence of various structures and phenomena in speech. These techniques are currently being used to develop new and vastly improved algorithms and decision strategies for machine recognition of speech (Bush, Hamilton & Hata, 1983; Chen, & Zue, 1983; Huttenlocker & Zue, 1983; Cole & Brennan, 1983). Other novel research efforts have involved more detailed analyses of speech spectrograms and new techniques for representing the auditory coding of speech as it is processed by the peripheral system. (Carlson & Granstrom, 1983; Goldhor, 1983; Seneff, 1983).

Finally, at least among some speech researchers, there has been an increased awareness that some of the seemingly unique properties of speech perception could be understood and accounted for by more detailed and careful experimental studies of complex nonspeech signals that preserve a number of the important dynamic/relational properties of speech. These recent nonspeech studies along with other findings in the last few years make it clear that it is necessary to draw a sharp distinction in perception between two modes of processing speech signals-- (1) an auditory-sensory-perceptual mode based on the psychophysical properties of speech signals, and (2) a speech mode based on the manner in which the human listeners respond to the phonetic-linguistic attributes of the speech signal. A detailed understanding of the processing carried out by the peripheral auditory system is clearly required to account for the first mode of processing and some important efforts have already begun in this very active area. However, much more research remains to be done on the contribution of the central auditory mechanisms, decisions processes and perceptual strategies in order to account for the way listeners respond to speech signals processed in the speech mode. The phonetic coding of speech signals relies, to a large extent, on perceptual, cognitive and linguistic considerations that go well beyond our current understanding of the auditory periphery.

(1) Basic Issues in Speech Perception

The field of speech perception is a very broad interdisciplinary area involving researchers from a number of disciplines including experimental psychology, linguistics, speech and hearing science, electrical engineering and artificial intelligence. Despite differences in approach and overall goals, there is fairly good agreement among researchers about what the basic issues are in the field today. In this section, I will first briefly review what I see as the five major theoretical issues in the field of speech perception. Some of these issues have been discussed in the past and continue to occupy a central role in current research on speech; others relate to new problems that will need to be pursued in the future.

a. Lack of Acoustic-Phonetic Invariance and the Segmentation Problem. For over thirty years, speech researchers have been relatively unsuccessful in identifying acoustic segments and properties of the speech waveform that uniquely match the units derived from perceptual analysis. The most obvious case is the failure to find acoustic units that map on to the phonetic segments or phonemes assumed from linguistic analysis of the intended message. A great deal of research has demonstrated that a single acoustic segment of the speech waveform often contains information about several neighboring linguistic segments; and, conversely, that the same linguistic segment is often represented acoustically in the speech waveform in quite different ways depending on the surrounding phonetic context, rate of speaking, talker and syntactic environment. When compared to words produced in isolation, the acoustic characteristics of individual speech sounds (i.e., phonemes) display even greater variability in connected speech because of the influence of the surrounding phonetic context. The lack of acoustic-phonetic invariance in speech is considered by all speech researchers to be the most important problem in the field; this was the major problem uncovered in the late 1940's after the invention of the sound spectrograph and it is still the central problem in the field of speech research today.

Closely related to the acoustic-phonetic invariance problem are several problems associated with the segmentation in speech. The context-conditioned variability of speech and the lack of any direct correspondence between the speech signal and linguistic-perceptual units such as phonemes or words presents enormous problems for the segmentation of speech into psychologically real and meaningful units that can be used in recognition. Research has demonstrated that it is extremely difficult, if not impossible, at least at the present time, to segment speech into acoustically defined units that are independent of adjacent segments and free from the contextual effects that occur in sentence environments. It has been difficult to use strictly physical (i.e., acoustic) criteria to determine where one word ends and another begins in fluent connected speech. Although some segmentation is possible according to acoustic criteria (see Fant, 1962), the number of acoustic segments is typically found to be much greater than the number of phonemes or words in an utterance (see also Cole, Rudnicky, Reddy & Zue, 1980).

b. Internal Representation of Speech Sounds. Until very recently, there was fairly good agreement among most investigators working on human speech perception that at some stage of perceptual processing the speech signal is represented internally as a sequence of discrete linguistic segments and features (see for example, Studdert-Kennedy, 1974, 1976). There was much less agreement, however, about the exact description of these features. Over the years, feature systems have been proposed based on distinctions in the acoustic domain, the articulatory domain and a combination of both. Recently, the trend has shifted to view these traditional feature descriptions with some skepticism, particularly with regard to the precise role these features and units actually play in on-going speech perception (Klatt, 1977, 1979, 1981). The argument here is that on reexamination, much of the original evidence cited in support of feature- or segment-based processing in perception is ambiguous and equally consistent with more parametric (i.e., acoustic/auditory) representations of the speech signal that do not make any a priori assumptions about intermediate or more abstract levels of analysis. Based on these arguments against segmental representations, a number of researchers have begun to look more closely at how speech signals are processed by the peripheral auditory system and what these much more detailed representations may contribute to resolving the issues surrounding of acoustic-phonetic invariance (see however, Pisoni, 1981).

c. Units in Speech Perception. A long-standing and closely related issue in speech perception has been the choice of a minimal unit of perceptual analysis. Because of channel capacity limitations in the auditory system, the rich sensory-based neural information output from the peripheral auditory system must be recoded into some reduced and more permanent form that can be used in perception and subsequent decision making. Many researchers have asked whether there is one basic or "natural" coding unit for speech. Over the years, numerous investigators have argued for the primacy of the feature, phoneme, syllable or word as their candidate for the basic perceptual unit in speech. Other researchers, motivated chiefly by psycholinguistic processing considerations, have argued for much larger perceptual units such as clauses and sentences (Bever, Lackner & Kirk, 1969; Miller, 1962). The continuing debate over the choice of a perceptual unit could be resolved, in my view, if a strict distinction were drawn over the level of linguistic analysis under consideration. In understanding spoken language, the size of the processing unit apparently varies from feature to segment to clause as the listener's attention is directed to different aspects or properties of the incoming linguistic message. When considered in this way, the arguments and continuing debate over the question of whether there is one and only one basic or primary unit of perception actually become irrelevant because there are, in fact, many different units used in speech perception and these units become more-or-less "primary" as a function of the level of processing required by the specific task presented to the listener.

d. Normalization Problems in Speech. In addition to the problems associated with the lack of acoustic-phonetic invariance and segmentation, another problem arises in connection with perceptual normalization of the speech signal. One aspect of this problem derives from physical and articulatory differences among talkers, specifically, the observation that the length and shape of the vocal tract differ quite substantially among different talkers. Moreover, the articulatory gestures used to produce individual phonemes and the strategies used to realize these in different phonetic environments also differ quite substantially among talkers. One consequence of these observations is that substantial acoustic-phonetic differences exist among talkers in the physical correlates of most, if not all, of the phonetic distinctions used in spoken language. While human listeners are able to ignore talker differences easily through perceptual compensation and talker normalization, currently available speech recognizers have found the same problem much more difficult to overcome.

Another related aspect of this problem concerns time and rate normalization. Research has shown that the durations of individual speech sounds are influenced quite substantially by an individual talker's speaking rate. Moreover, these durations are also affected by the locations of various syntactic boundaries in connected speech, by syllabic stress, and by the component features of adjacent phonetic segments in words (see Gaitenby, 1965; Klatt, 1975, 1976, 1979; Lehiste, 1970). In addition to these sources of variability, substantial differences have also been observed in the durations of segments in words when they are produced in sentence contexts compared to the same words spoken in isolation. Although human listeners display evidence of perceptual constancy in the face of enormous physical variation, the precise basis for their perceptual compensation is still unknown and remains a topic of intense study (see for example Miller & Liberman, 1979; Miller, 1981).

e. Interaction of Knowledge Sources in Speech. Spoken language understanding involves access to and use of a variety of sources of knowledge that a listener has available as a speaker and hearer of a natural language. In addition to the physical information encoded in the speech waveform, the listener has a great deal of detailed knowledge of the structure of his/her language which is actively used with the sensory input to develop a representation of the incoming message. A major issue in the field of speech perception concerns the precise representation of these knowledge sources and the extent to which these different sources of knowledge interact at different stages of perceptual analysis of speech (Reddy, 1976). For example, one of the major points of contention among current theories of auditory word recognition concerns whether words are recognized by an autonomous processing mechanism that is independent of other knowledge sources (e.g. Forster, 1979; Norris, 1982) or whether words are recognized through an interaction of top-down linguistic knowledge and bottom-up sensory information (e.g. Marslen- Wilson & Welsh, 1978; Cole & Jakimik, 1980).

(2) Auditory Modeling of Speech Signals in the Periphery

Over the last three or four years a great deal of new and exciting research has been reported on how the peripheral auditory system processes and encodes speech signals (see Carlson & Granstrom, 1982). The research on auditory modeling of speech signals comes from two different directions. On the one hand, a number of important physiological studies using animals have been carried out to describe, in fairly precise terms, how simple speech signals are coded in the peripheral auditory system (Kiang, 1980; Delgutte, 1980, 1981, 1982). These studies have examined auditory-nerve activity in response to simple speech signals such as steady-state vowels and stop consonants in CV syllables. The goal of this work has been to identify reliable and salient properties in the discharge patterns of auditory-nerve fibers that correspond, in some direct way, to the important acoustic properties or attributes of speech sounds (Sachs & Young, 1979, 1980; Young & Sachs, 1979; Miller & Sachs, 1983).

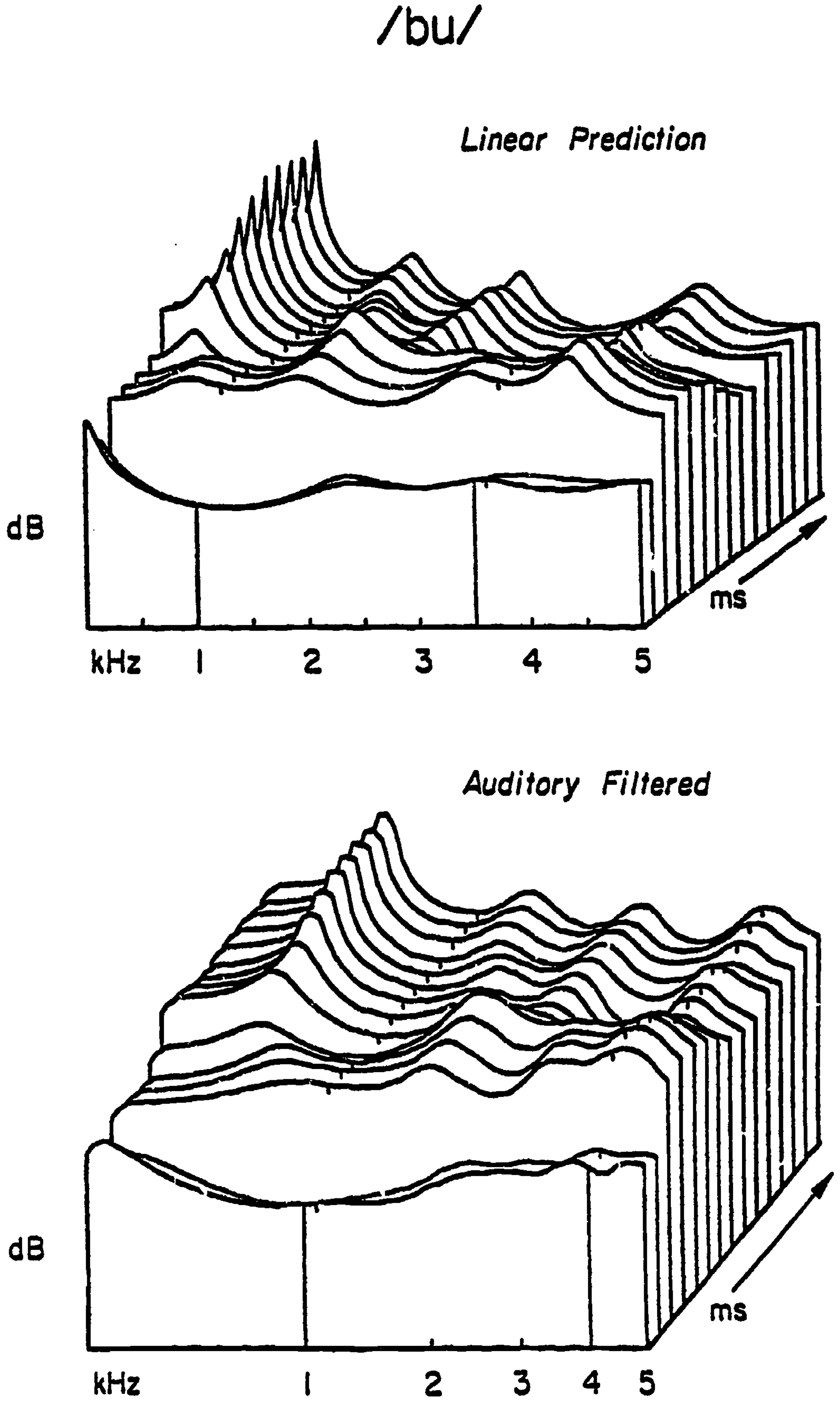
On the other hand, several researchers have begun to develop psychophysically based models of speech processing that explicitly incorporate well-known psychoacoustic data in their descriptions of the filtering that is carried out by the peripheral auditory system (Searle et al., 1979; Zwicker, Terhardt & Paulus, 1979; Kewley-Port, 1980, 1983). The goal of this line of research is to develop representations of the speech signal that take into account known psychophysical facts about hearing such as critical bands, upward spread of masking and the growth of loudness (Klatt, 1982; Zwicker et al., 1979).

The recent interest and extensive research efforts in developing new and presumably more appropriate and valid representations of speech signals such as "neurograms," "cochleagrams" or "neural spectrograms" derives, in part, from the assumption that a more detailed examination of these auditory representations should, in principle, provide researchers with a great deal more new and relevant information about the distinctive perceptual dimensions that underlie speech sounds (Stevens, 1980). Moreover, and perhaps most importantly, it has been further assumed that information contained in these so-called neuroacoustic representations will contribute in important ways to finally resolving the acoustic-phonetic invariance problem in speech (see for example Goldhor, 1983a, 1983b).

One recent approach has been developed along these lines in our laboratory in terms of representing speech signals with three-dimensional running spectral displays (see Kewley-Port, 1980, 1983; Kewley-Port & Luce, 1984). The interest in using this particular type of display to study the acoustic properties of stop consonants is based, in part, on the assumption that the perceptual dimensions for speech sounds should be represented in terms of what is currently known about transformations of acoustic signals by the peripheral auditory system. Sufficient physiological and psychophysical data are currently available in the literature about the operation of the auditory system to suggest that the peripheral auditory system can be modelled, to a first approximation, as a frequency analyzer with continuously varying neural signals as output (see Klatt, 1979; Plomp, 1964; Schroeder, Atal & Hall, 1979; Zwicker et al., 1979; Siebert, 1968; Searle et al., 1979; Carlson & Granstrom, 1982). In Kewley-Port's work, special emphasis has been placed on a very detailed examination of precisely how the distribution of spectral energy changes rapidly over time for stop consonants, particularly for the distinctive acoustic correlates of place of articulation in stops (see Kewley-Port, 1983).

In addition to looking at how the distribution of spectral energy changes over time in visual displays of speech, work has also been directed at several different ways of modeling the energy spectra that is derived from the filtering carried out by the auditory system. The spectra typically employed in speech analysis studies are often obtained from linear prediction (LPC) techniques which are used to compute formant frequencies (Atal & Hanauer, 1971; Makhoul, 1975; Markel & Gray, 1976). Because this type f analysis assumes constant bandwidths, the derived spectra are not really appropriate to model how the human auditory system filters these signals. Similarly, in analyses of speech, frequency is often displayed on a linear rather than log frequency scale despite what is known about how humans encode frequency in the auditory system. Some idea of the differences between these two types of representations can be seen by examining the three-dimensional running spectral displays shown in Figure 1.

Insert Figure 1 about here



Linear Prediction

Auditory Filtered

Figure 1. Two types of running spectral displays for the syllable /bu/ (see text).

A linear prediction (LPC) running spectrum for the syllable /bu/ is shown with a linear frequency scale in the top panel of Figure 1; a second running spectrum derived from a set of auditory filters using the Mel scale is shown in the bottom panel of this figure. Examination of the two displays reveals a number of important differences between the two representations. First, it is well known that the bandwidth of the frequency analysis performed by the auditory system increases with frequency. Various estimates of internal auditory bandwidth have been made in the past from psychophysical data and they range from one-half to one-tenth octave (see Moore & Glasberg, 1983). The ranges most often reported in the literature are shown in the hatched areas in Figure 2.

Insert Figure 2 about here

The bark scale which is based on the concept of critical bands is shown as an upper limit (Zwicker et al., 1979), while the one-sixth oct ve filtering which is similiar to Patterson's (1976) proposal is shown as the lower limit. The bandwidths assumed from the linear prediction filtering are shown in the bottom of this figure by the straight horizontal line. It should be obvious that compared to the other approaches, the LPC analysis which has been used quite extensively in speech processing work may not provide the most appropriate way of characterizing the filtering that takes place in the auditory system.

Using these three-dimensional running spectral displays, several important and long-standing questions have been examined in our laboratory over the last few years (see Kewley-Port, 1980, 1983; Kewley-Port, Pisoni & Studdert-Kennedy, 1983; Kewley-Port & Luce, 1984). These questions concern the acoustic analysis of stop consonants and the identification of invariant acoustic correlates for place of articulation. Using visual displays that show the running spectrum of the onset of CV syllables, a careful examination of the changes in the patterns of spectral energy from the release burst into the formant transitions has been carried out (Kewley-Port, 1983). The analysis of these displays indicated that important information about the identification of the consonant is contained in the dynamic, time-varying properties of the speech waveform. Based on these analyses, a number of perceptual experiments were performed with both natural and synthetic speech to verify the importance of these time-varying acoustic properties (see Kewley-Port, 1980). The results of these studies have demonstrated the existence of relational information in the signal that serves as reliable and <u>invariant</u> acoustic and perceptual correlates of place of articulation in stops. These results contrast sharply with the recent views of Stevens and Blumstein(1978) who have argued for the existence of "static" invariant cues to place of articulation in terms of the analysis of the gross shape of the spectrum (see also Blumstein & Stevens, 1979, 1980).

The recent work being carried out on the neural representation of speech in the auditory periphery and the efforts at developing realistic auditory models may be viewed as an important change in the direction of research efforts in the field. Within the last few years, investigators have gone back to the speech signal and carried out a more detailed and more sophisticated analyses in the hopes of gaining new insights into the types of coding that is performed by the peripheral auditory system. Following this orientation, the recent work in our laboratory on the primary phonetic recognition process has been aimed at developing a better representation of the initial sensory properties of speech

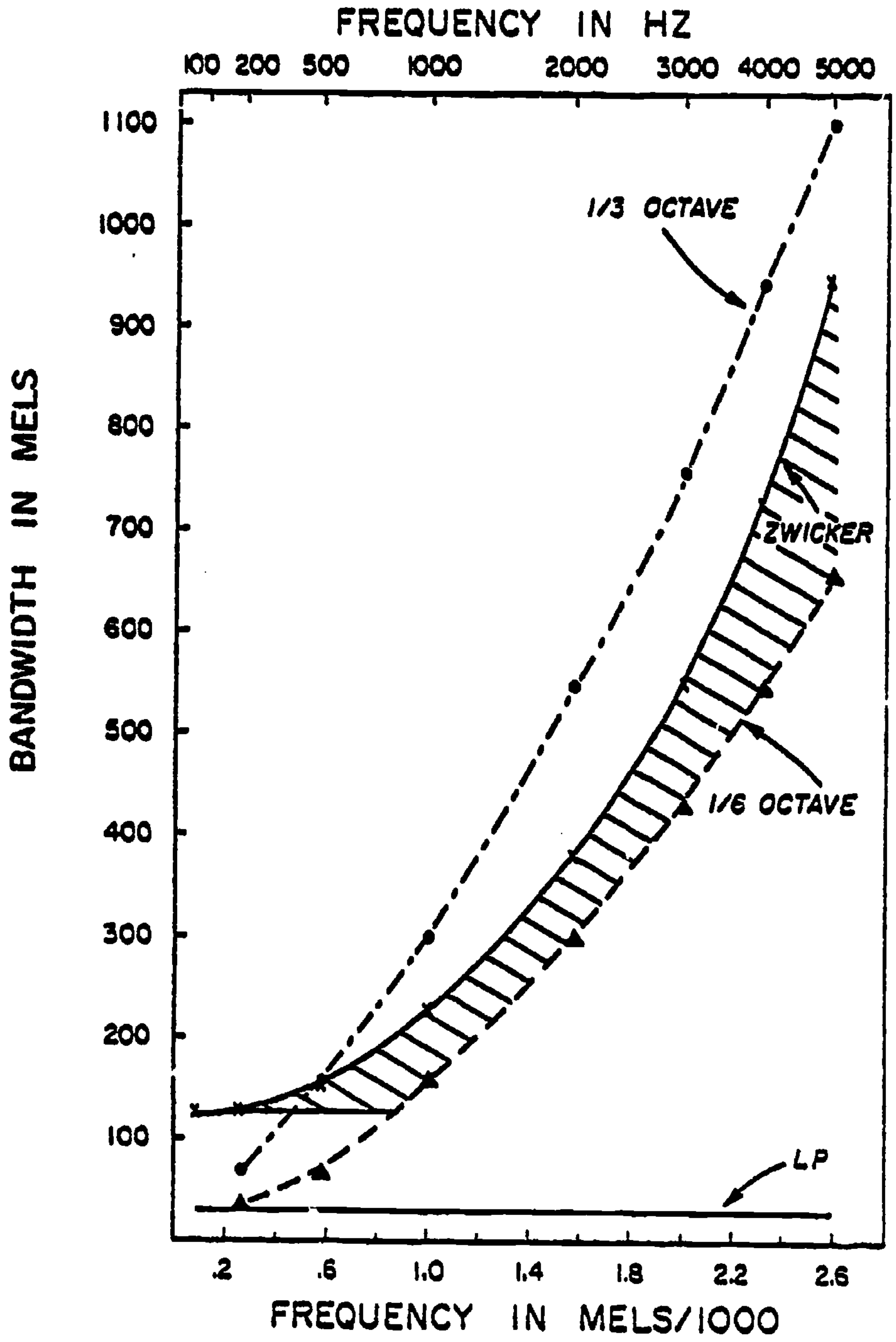


Figure 2. Four different filter bandwidths as a function of frequency are displayed using the technical Mel scale. Estimates of interval auditory bandwidths as reported in the literature are shown in the hatched portion of the figure.

sounds at the periphery. While there is good reason to be encouraged by this recent work demonstrating that the speech signal does contain a good deal of salient and reliable information that is encoded by the auditory system, it is also important to carry out other detailed studies of the central auditory mechanisms that integrate this initial sensory information. Research is also needed on the subsequent perceptual-cognitive decision processes that interpret the patterns of sensory stimulation in linguistically relevant ways.

(3) Spectrogram Reading and Acoustic-Phonetic Data-Bases

Over the last five years two important methodological developments have dramatically affected the way in which acoustic-phonetic research is carried out. The first development, owed largely to the efforts of Victor Zue at MIT, was the demonstration that speech spectrograms could be analyzed and interpreted at very high levels of accuracy. Through careful study of an "expert spectrogram reader" explicit rules were formulated to describe the strategies used to interpret spectrograms. In a detailed examination of Victor Zue's abilities, Cole et al. (1980) demonstrated that phonetic segments could be reliably identified from acoustic-phonetic information displayed in spectrograms of unknown utterances. Victor Zue's performance in analyzing spectrograms was shown to be comparable to a group of phoneticians who analyzed the same unknown utterances by careful listening and standard transcription techni ues.

While the initial report of Victor Zue's performance in analyzing spectrograms may be considered more-or-less as an "existence proof" that speech can be recognized from the visual information contained in spectrograms and that reliable cues to phonetic segments can be found in the speech signal, the results have already had several important consequences for research strategies in speech recognition. First, these findings have laid to rest a long-standing and apparently erroneous belief that speech spectrograms, particularly spectrograms of unknown and unfamiliar utterances, simply could not be read or analyzed (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1968). As a result, numerous researchers have started to examine spectrograms in very great detail as a rich and potentially valuable source of information about the phonetic content of the linguistic message. What was previously just a simple demonstration exercise in phonetics classes has now become the topic of intense formal study in specialized classes devoted exclusively to learning to interpret spectrographic displays of speech through the application of explicit rules and principles.

A related and perhaps a more far reaching consequence of these efforts has been the development of new feature-based descriptions of the acoustic correlates of speech sounds derived directly from the information contained in spectrographic displays (Cole, Stern & Lasry, 1983). The logic behind this approach is that rather than being thought of as impoverished or noisy, the speech signal may be considered as an extremely rich source of reliable information (Elman & McClelland, 1983). This information can be used to identify the acoustic correlates of speech sounds without having to resort to higher sources of knowledge such as morphology, syntax or semantics or elaborate hypothesis generation and prediction about constraints on vocal tract dynamics. In other words, a great deal of reliable information is present in the speech signal and, as the argument goes, this information has not been used effectively in past efforts at designing front ends of speech recognition systems. Previous work was largely unsuccessful because of the failure to capture the systematic regularities that exist between the phonemes or perceptual units of the linguistic message and their acoustic correlates as represented in spectrograms. Indeed, even the most successful speech understanding systems like HARPY were capable of correctly recognizing only about 45 percent of the phonemes in the speech signal (Lowerre & Reddy, 1979). Their successful overall performance came about primarily through the creative use of linguistic constraints and powerful search strategies applied to a precompiled knowledge-based network.

Taking a somewhat different approach to this problem, we have been studying how naive observers learn to identify spectrographic displays of isolated words without any formal knowledge of acoustic-phonetics or speech acoustics (Greene, Pisoni & Carrell, 1984). One of our goals was to examine the types of visual properties or attributes that naive subjects use in recognizing words in the absence of any explicit knowledge about the acoustic-phonetic properties of words and their visual correlates in spectrographic displays. The results demonstrated that naive subjects can learn to identify isolated (PB) words from several talkers at extremely high levels of accuracy. The subjects appear to base their performance on perceptual strategies that use well-defined criterial features that are present in the visual displays. Subjects in our study did not simply memorize unique visual patterns for each of the test words in the set because in subsequent tests we have shown that the subjects were able to generalize their knowledge of these visual displays to new talkers and novel words that they had never seen before (Greene et al., 1984). If naive and unsophisticated undergraduate observers with no knowledge of acoustics or phonetics can abstract regularities from spectrographic displays of speech and use these properties to identify novel isolated words, the visual properties or attributes of speech sounds obviously must be very salient. The properties or features for speech sounds displayed in spectrograms may therefore be extremely useful in developing new feature-based algorithms for use in improving the performance of front ends of speech recognizers which have had only limited success over the past few years (see Cole et al., 1983). Such an approach seems particularly well-suited to a situation where fine phonetic distinctions must be made among items in the vocabulary set. One example of this is the letters of the alphabet which are highly confusable (e.g. B, D, G, P).

Insert Figure 3 about here

The second recent development concerns the construction and systematic use of large data bases for research in acoustic-phonetics. The sources of variability in speech and the influence of different contextual factors have been found to be so enormous in scope that it is often difficult, if not literally impossible, to abstract regularities and general principles from a small number of measurements made by hand across a large number of different phonetic environments. This is especially critical when the hand analysis take a great deal of time. In the last few years, several efforts have been made to develop large data bases for use in acoustic-phonetic research so that precise quantitative measurements can be made on the effects of various phonetic environments and sources of variability (Shipman, 1983). Using techniques from artifical intelligence, very sophisticated data-base management procedures have been developed to permit a researcher to search and interrogate the data base for specific types of information that would otherwise be difficult or impossible to obtain with the conventional research techniques used in acoustic-phonetics (Shipman, 1982). Because of the availability of these large data bases and the data-base management procedures for extracting information from them, many research questions that seem trivial today would probably have never been conceived of only a few years ago. The ability to test experimental hypotheses

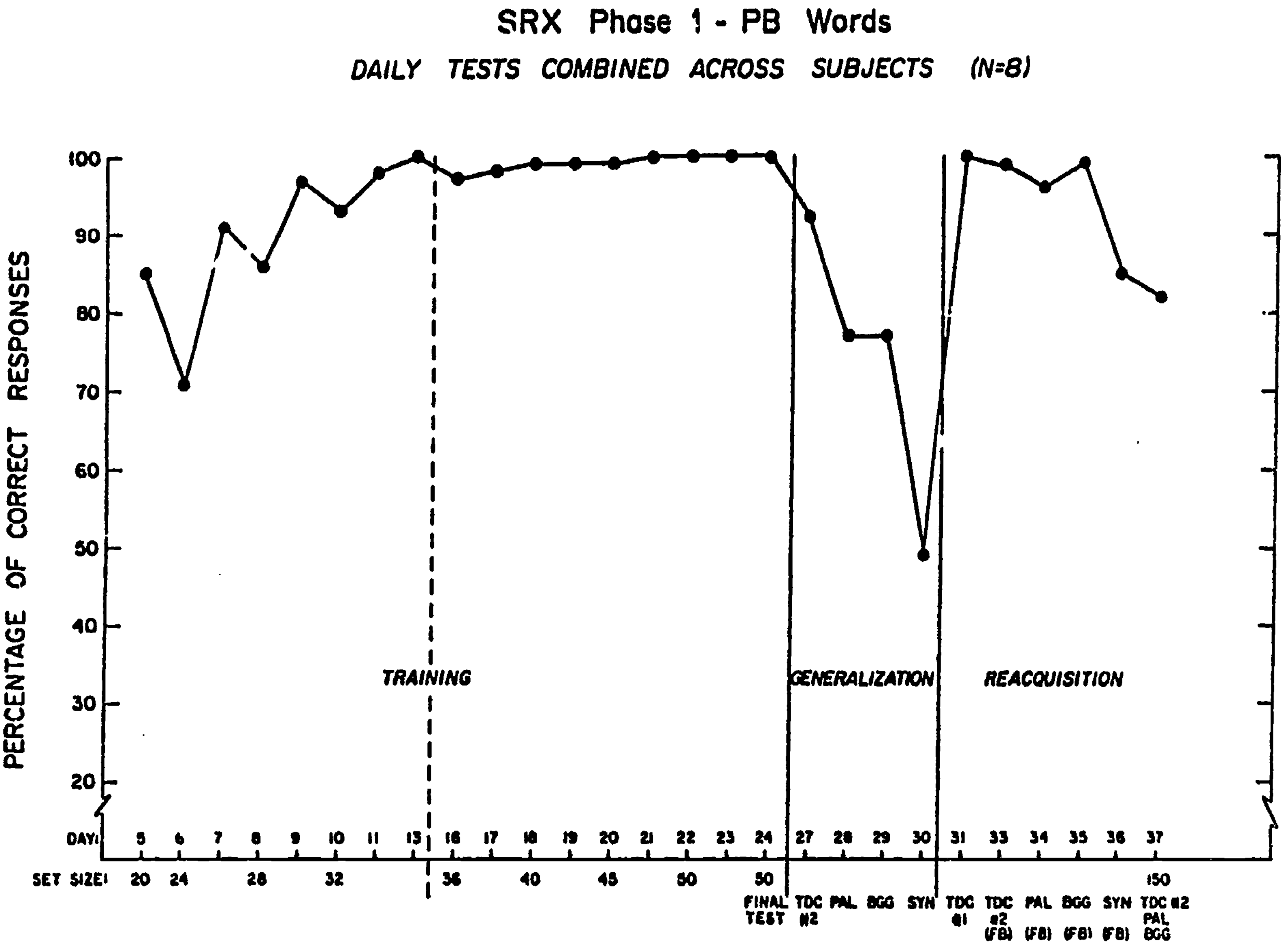


Figure 3. Daily test results from a spectrogram reading experiment in which subjects learned to recognize spectrograms of 50 isolated words. Test performance on unfamiliar talkers is shown in the generalization portion of the figure. The first point, TDC#2 is the talker used during training; PAL is a new male talker; BGG is a new female talker; SYN is a synthetic talker.

at a much faster rate allows speech researchers to make greater advances in accumulating new knowledge about the acoustic-phonetics of a particular language and the effects of different sources of variability on selected properties of the speech signal. These efforts have been particularly useful in detailing the operation of ; onological rules and the effects of local phonetic contexts on the acoustic properties of speech sounds across different environments (Zue, 1976; Lamel, 1983).

Other work using large data bases has shown that detailed knowledge about the structural properties of words in the lexicon and their relations in terms of "lexical density" can be used to constrain the potential search space used in word recognition and lexical access (Eukel, 1980; Landauer & Streeter, 1976; Crystal & House, 1982; Shipman & Zue, 1982). Although the use of specialized data-bases in acoustic-phonetic research is still at an early stage of development, it seems obvious that use of this detailed information will vastly improve the front end performance of the next generation of speech recognizers (Chen & Zue, 1983). More importantly, however, these recent developments demonstrate that the sources of variability in speech are finite, and that, at least in principle, the acoustic-phonetic invariance problem may finally be solved in the immediate future.

(4) Comparisons in Perception Between Speech and Nonspeech Signals

The study of speech perception differs in two important ways from the study of other aspects of auditory perception. First, the signals used to study the functioning of the auditory system have typically been short in duration, simple in spectral composition and discrete in time with well-defined stimulus onsets and offsets. Moreover, the experimental manipulations used in these studies have typically involved variations in only a single physical dimension. In contrast, speech signals display very complex spectral and temporal relations that change substantially over time as the articulators move rapidly from one set of gestures to another. Thus, it seems reasonable to suppose that the complexity of the spectral and temporal structure of speech and its variation may be an additional source of the differences observed in perception between speech and nonspeech signals (see Stevens, 1980).

Second, most of the research on auditory psychophysics that has accumulated over the last thirty-five years has been concerned with the discriminative capacities of the sensory transducer and the functioning of the peripheral auditory mechanisms. In the case of speech perception, however, there is good reason to feel confident that the relevant mechanisms used for phonetic coding of speech sounds are centrally located and intimately related to somewhat more general cognitive and decision making processes associated with the encoding, storage and retrieval of information in short- and long-term memory (see Pisoni. 1978). Moreover, except for a few isolated cases such as the research program being carried out by Watson (see Watson & Foyle, 1983), most experiments in auditory psychophysics have typically focused on experimental tasks and paradigms that involve discrimination rather than identification or recognition, processes typically thought to be most relevant to speech perception. Considered in this light, most researchers working on speech perception believe that a good deal of what has been learned from auditory psychophysics and complex auditory perception is probably only marginally relevant to the study of speech perception and to a detailed understanding of the underlying perceptual and linguistic mechanisms used in speech processing (Liberman et al., 1967; Stevens & House, 1969; Studdert-Kennedy, 1974, 1976).

Despite these obvious differences in approach and emphasis, a number of investigators have been quite interested in the differences in perception between speech and nonspeech signals. That such differences might exist was first suggested by the initial findings on categorical perception of speech by Liberman and his colleagues at Haskins Laboratories (Liberman et al., 1957). And, it was with this general goal in mind that the first so-called "nonspeech control" experiments were carried by Liberman et al. (1961) in order to determine the basis for the appparent distinctiveness of speech sounds in perception. In this nonspeech study, the spectrographic patterns of the /do/ and /to/ synthetic speech continuum were inverted in the frequency domain to produce a set of nonspeech patterns that differed in the relative onset time of the individual components. The results of perceptual tests showed very marked differences in discrimination between the speech and nonspeech patterns, a finding that was widely interpreted as support for the perceptual distinctiveness and "specialness" of speech sounds as auditory signals.

Numerous speech vs. nonspeech comparisons have been carried out since these early studies. For the most part, these experiments have revealed findings that were consistent with the earlier results of Liberman et al. Until quite recently, the experiments using nonspeech control signals failed to show the same types of discrimination functions that were observed with the parallel set of speech signals (see however Cutting and Rosner, 1974; Miller et al., 1976; Pisoni, 1977). Differences in perception between speech and nonspeech signals were assumed to reflect two different modes of processing-- a "speech mode" and an "auditory mode." Despite some attempts to dismiss this dichotomy, additional evidence continues to accumulate to support the existence of two different processing modes in speech perception.

The final picture is far from being resolved, however, because a number of problems have surfaced in comparing speech and nonspeech signals. Moreover, these problems have generated several important questions about the interpretation of results obtained in earlier studies using nonspeech control signals. The first criticism that can be raised about these nonspeech studies concerns whether the same psychophysical properties found in the speech stimuli were indeed preserved in the parallel set of nonspeech control signals. This criticism seems quite appropriate for the original /do/--/to/ nonspeech stimuli which were simply the inverted spectrographic patterns reproduced on the pattern playback speech synthesizer. A similiar criticism can be leveled at the well-known "chirp" and "bleat" nonspeech control stimuli used by Mattingly et al. (1971) which were created by removing the formant transitions and steady-states from the original speech contexts. Such manipulations, while nominally preserving the "speech cues," obviously result in marked changes in the spectral context which no doubt affects the detection and discrimination of the original formant transitions.

A number of these criticisms have been taken into account in the more recent experiments comparing speech and nonspeech signals (see Bailey, Summerfield & Dorman, 1977; Liberman, 1979) in which the stimulus materials remained identical across different experimental manipulations. The major difference simply involved instructions to subjects to code the stimuli differentially as either speech or nonspeech (see also Cross and Lane, 1962). However, while these more recent studies have tried to deal with some of the earlier ambiguities and criticisms, a number of important problems still remain. For example, subjects in almost all of these nonspeech experiments rarely, if ever, receive practice with the nonspeech control signals to develop the competence required to categorize them consistently before the experiment actually begins (however, see Pisoni, 1977; Pisoni, Carrell & Gans, 1983). With complex multidimensional signals it is often extremely difficult for subjects to attend to the relevant stimulus attributes or components that distinguish one signal from another presented in the experiment. A subject's performance with these nonspeech signals may therefore be no better than chance if he/she is not attending selectively to the criterial attributes that distinguished the original speech stimuli from which the nonspeech signals were derived (see, for example, the nonspeech data of Mattingly et al., 1971). Indeed, no knowing what to listen for in an experimental context may force a subject to selectively attend to an irrelevant or potentially misleading attribute of the signal itself. Alternatively, a subject may simply focus on the most distinctive auditory quality of the perceived stimulus without regard for the less salient acoustic properties which often are the most important in speech perception such as burst spectra, formant transitions or the duration of a silent interval within a stimulus complex. Since almost all of the nonspeech control experiments conducted in the past were carried out without the use of any initial discrimination training or feedback to subjects, an observer may have simply focused his/her attention on one aspect of the stimulus on one trial and an entirely different aspect of the stimulus on the next trial.

Recently, we completed several experiments comparing the perception of speech and comparable nonspeech control signals that were designed to deal with these criticisms (see Pisoni et al., 1983). The nonspeech signals were created with sinewave analogs that preserved the durations and temporal relations of the speech stimuli although the stimuli did not sound like speech to naive listeners. Examples of these are shown in Figure 4. The motivation for this nonspeech study was the finding reported by Miller and Liberman (1979) that the overall duration of a syllable influences the location of the identification boundary between the stop /b/ and the semivowel /w/. Miller and Liberman found that the duration of the vowel in an isolated CV syllable systematically influenced the perception of the formant transition cues for the stop-semivowel contrast. With short syllables, subjects required shorter transition durations to perceive a /w/ than with longer syllables. Miller and Liberman interpreted these contextual effects due to syllable duration as a clear demonstration that the listener "normalized" for speaking rate. That is, the listener adjusts his/her decision so as to compensate for the differences in vowel length that are conditioned by the talker's speaking rate. According to Miller and Liberman, the listener interprets a particular set of acoustic cues <u>in relation to</u> the talker's speaking rate rather than by reference to some absolute set of contextually invariant acoustic attributes contained in the signal itself (see also Eimas and Miller, 1980 for related findings with young infants).

The results of our comparisons between speech and nonspeech signals which are shown in Figure 5 demonstrated comparable context effects for the perception of a rapid spectrum change as a function of the overall duration of the stimulus for both speech and nonspeech signals. These findings therefore call into question the rate normalization account proposed by Miller and Liberman. Our findings demonstrate that context effects due to differences in stimulus duration are not peculiar to the perception of speech or to normalization for speaking rate. Context effects such as these may simply reflect general psychophysical and perceptual principles that affect the categorization and discrimination of all acoustic signals, whether speech or nonspeech (see also Jusczyk et al., 1983 for comparable findings with young infants using these nonspeech signals).

Insert Figures 4 & 5 about here

In another study, we used complex nonspeech stimuli to study the perceptual similarity of mirror-image acoustic patterns that resemble the formant transitions and steady-state segments of the CV ar'i VC syllables /ba/, /da/, /ab/ and /ad/. Using a perceptual learning paradigm, we found that subjects could learn to assign mirror-image acoustic patterns to arbitrary response categories more consistently than a similar arrangement of the same patterns based on spectro-temporal commonalities (Grunke and Pisoni, 1982). Our results demonstrated that subjects respond not only to the individual components or dimensions of these complex patterns but also to the entire patterns. Moreover, subjects make use of the patterns' internal organization in learning to categorize them consistently according to different sets of classification rules. The results of this study also suggested that differences in "mode of processing" are extremely important in controlling perceptual selectivity and in influencing how the individual components of a complex stimulus pattern are processed.

Differences in mode of processing can be quite dramatic depending primarily on whether the listener's attention is directed towards coding either the auditory properties of the signals or the phonetic qualities of the overall patterns (Schwab, 1981). In the former case, when subjects are expecting to hear nonspeech signals, the process appears to be more analytic, requiring the processing or "hearing out" of the individual components of a stimulus pattern; in the latter case, when subjects expect to hear speech signals, the process appears to be more holistic, insofar as the individual components may be combined or integrated together to form well-learned and highly familiar perceptual categories corresponding to phonological units in their language. The results of these nonspeech experiments therefore suggest that listeners probably do not isolate and then subsequently label or identify individual speech cues. Rather, it seems more likely that listeners respond to these attributes in relation to the overall context of the time-varying pattern (see for example Remez, Rubin, Pisoni and Carrell, 1981). Moreover, as Schwab (1981) has demonstrated recently with nonspeech signals, differences in mode of processing may entail the use of entirely different processing strategies with the same stimulus configuration. In one experiment, she found large backward masking effects for subjects who were instructed to process the signals as nonspeech auditory patterns instead of phonetic events. In another experiment, she found evidence for frequency masking among the components of a complex nonspeech pattern but only for the group of subjects who were instructed to process the signals as nonspeech events. No masking effects were observed for the subjects who were instructed to respond to the same signals as speech sounds. Thus, the same stimuli may be perceived quite differently depending on whether they are encoded by listeners as speech or nonspeech events.

(5) Word Recognition and Lexical Access

The study of word recognition and the nature of the lexical representations for words have been long-standing concerns of experimental psychologists (see Bagley, 1900), although these particular topics have not been studied routinely by investigators working in the mainstream of speech research (Cole & Rudnicky, 1983). Several reasons for this lack of attention to words can be identified

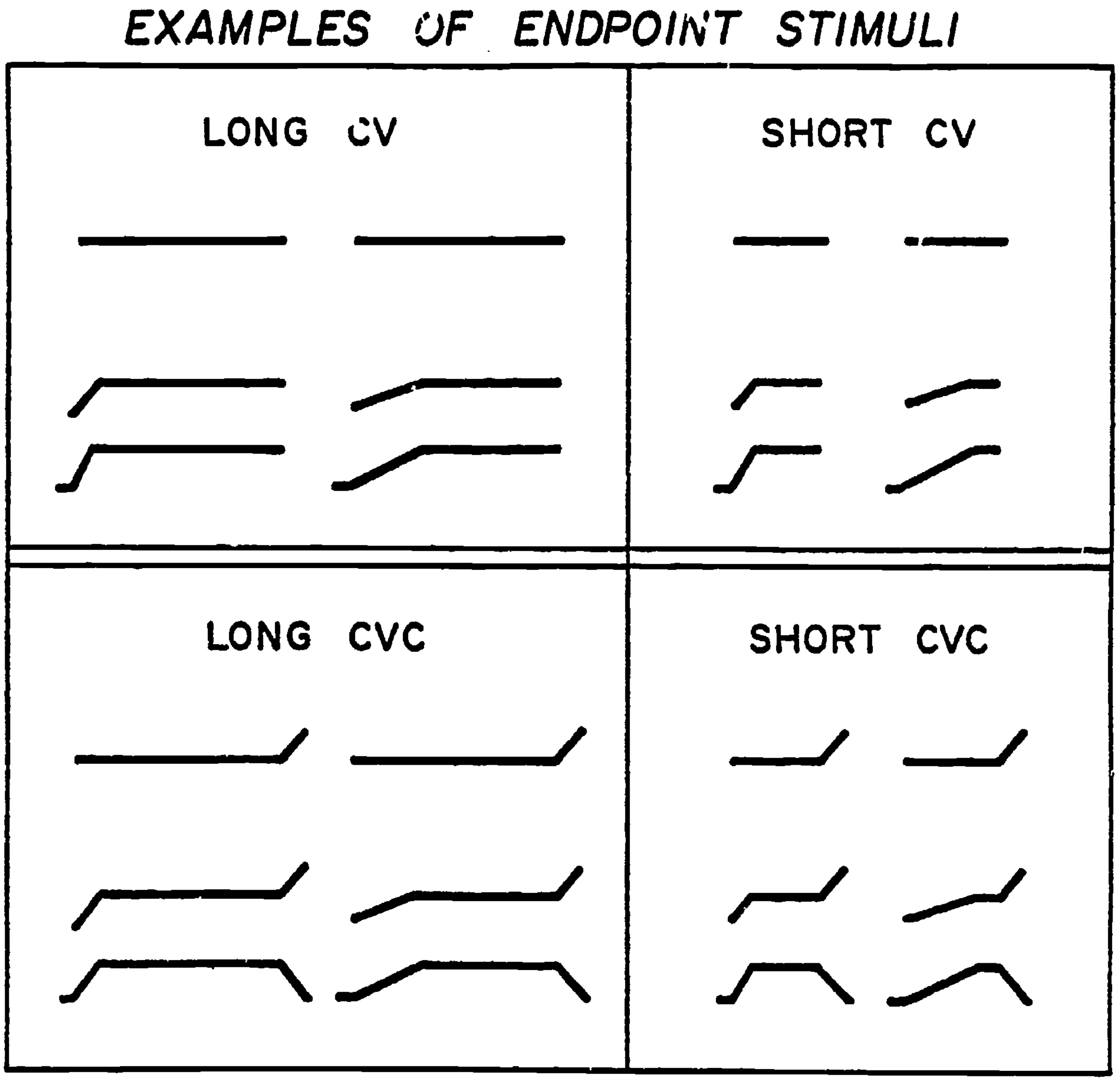


Figure 4. Schematic representations of the formant motions of the endpoint stimuli corresponding to [b] and [w]. The top panel shows long and short CV syllables; the bottom panel shows long and short CVC syllables which had final transitions added to the steady-state vowel.

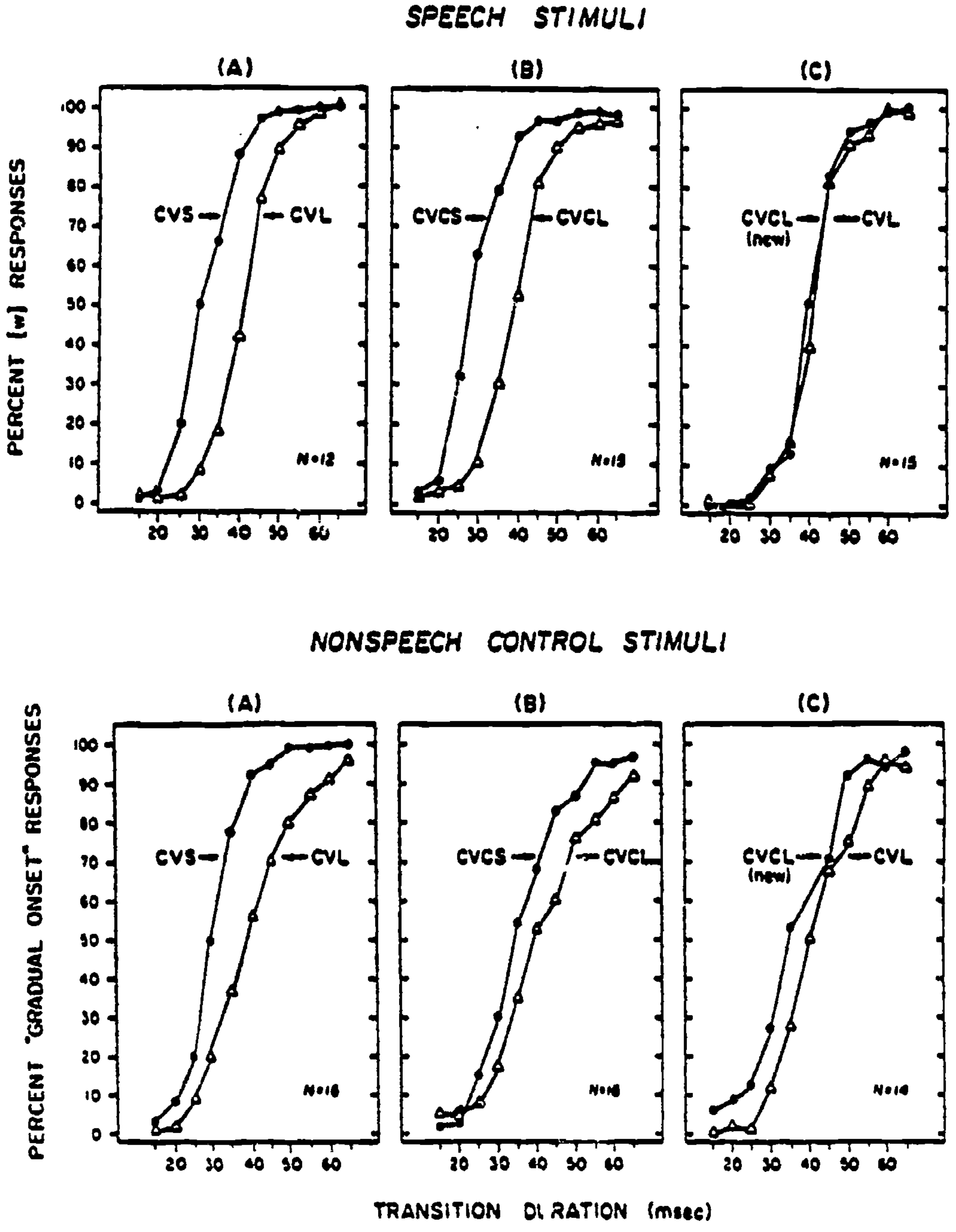


Figure 5. The upper figures display labeling data for synthetic CV and CVC syllables. Panels A and B display the data for CV and CVC syllables that differ in overall duration; filled circles are short stimuli; open triangles are long stimuli. Panel C displays data for CV and CVC syllables that were matched for overall duration. The lower figures display labeling data for nonspeech control stimuli that were modeled after the CV and CVC syllables shown in Figure 4. The three lower panels display the percent of "gradual onset" responses as a function of transition duration. The nonspeech data shown here are exactly parallel to the speech data shown above.

(see Nooteboom, 1981). First, much of what we know about word recognition has been obtained in the visual modality through studies of the reading process. Second, because of the concern for phoneme and feature perception ir. isolated nonsense syllable contexts, few speech researchers have been willing or interested in venturing out of their own narrow research areas to study word recognition. Moreover, few if any efforts were made by these researchers to generalize the findings on phoneme and feature perception to the study of the cues used to recognize words in isolation or in connected speech (for example, see reviews by Studdert-Kennedy, 1974, 1976; Darwin, 1976).

Several interesting and potentially important problems in speech perception involve the process of word recognition and lexical access. Moreover, these problems bear directly on issues surrounding the nature of the various types of representations in the mental lexical and the interaction of knowledge sources in speech perception. As we noted earlier in connection with units of perceptual analysis, there has been considerable recent interest in determining precisely what kinds of representations exist in the mental lexicon. Are words, morphemes, phonemes or sequences of spectral templates appropriate representations for lexical entries? Are words accessed in terms of their acoustic-phonetic or phonological structure? Why are high frequency words treated differently than low frequency words? These are a just few of the questions that people are trying to answer at this time.

One of the central and long-standing problems in speech perception concerns the interaction of the sensory information output from the auditory system with higher-level contextual information. A number of researchers such as Forster (1976) and Massaro and Oden (1980) have maintained that early sensory information is processed independently of higher-level context and knowledge and that the facilitation effects observed for word recognition in context are due to post-perceptual processes associated with readjusting decision criteria. Other researchers such as Marslen-Wilson and Tyler (1980) have argued strongly that context does influence the early sensory analysis of the input signal and that the various knowledge sources used in spoken language comprehension interact freely and quite substantially in supporting recognition (see also Reddy, 1976).

In our laboratory, we have been studying word recognition in sentence contexts and have compared recognition performance to the same words presented in isolation. Our experiments have used a variant of the "gating technique" which permits very precise control over the amount of signal duration of each word in a sentence (Grosjean, 1980). Subjects listened to truncated words in either meaningful or semantically anomalous sentences or in isolation. The subject's task was to identify the words in the sentence after each presentation. Initially, each target word in the sentence was completely replaced by envelope-shaped noise which preserved the amplitude and durations of individual segments but obliterated the spectral information in the signal. In consecutive presentations of a test sentence, 50-ms increments of the original speech waveform replaced either the initial or final segments of the noise until the entire original target word was presented on the last trial. Recognition points were determined and a detailed analysis of subjects' incorrect responses was undertaken.

Insert Figure 6 about here

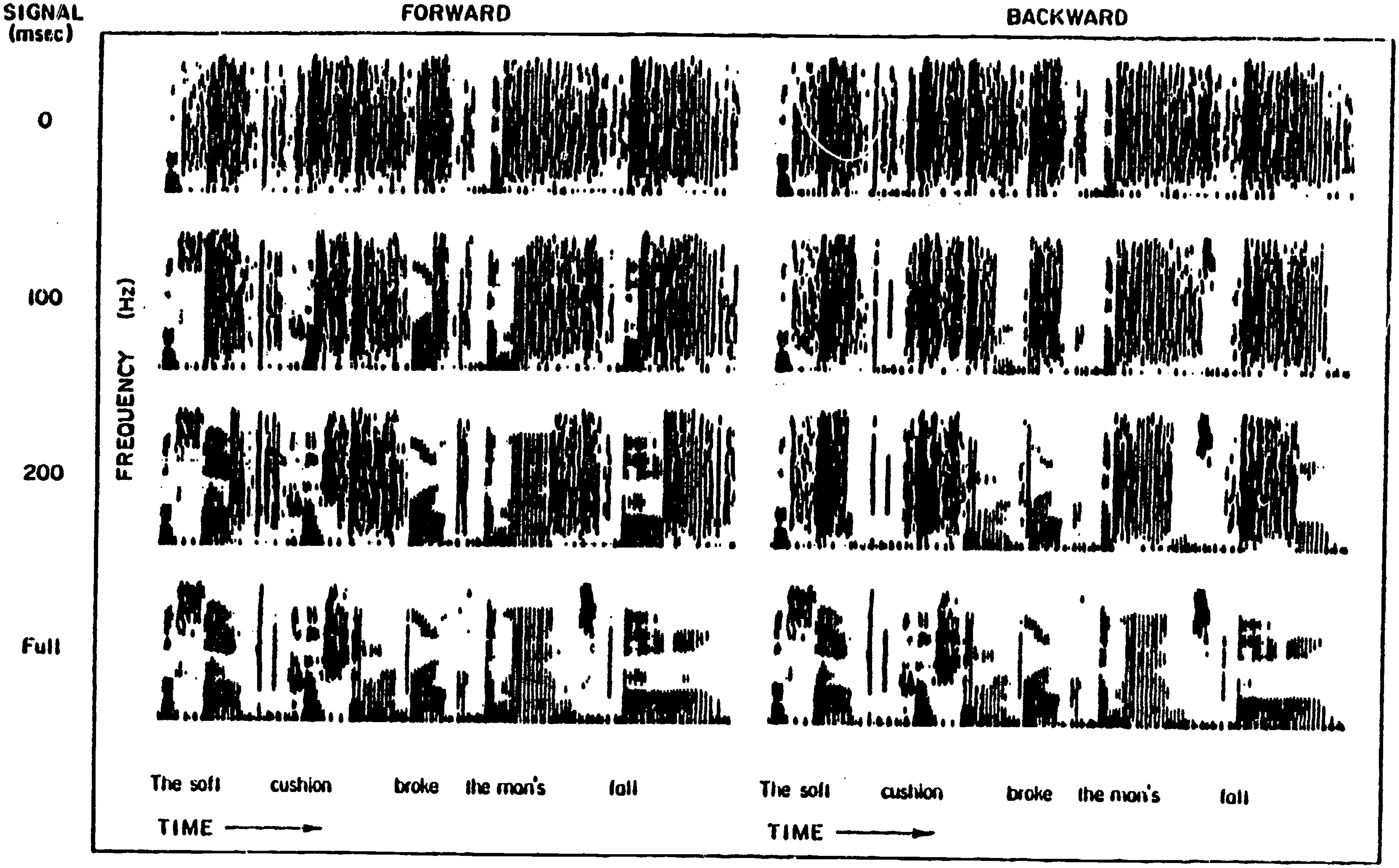


Figure 6. Sample speech spectrograms of forward-gated and backward-gated sentences. Increasing signal duration is shown left-to-right and right-to-left respectively for the target words of one meaningful test sentence.

In normal sentences, we found that subjects needed more stimulus information to recognize words from their endings than their beginnings. And, subjects needed substantially more stimulus information to recognize words in isolation than in context. Surprisingly, the difference in performance between information in the beginnings and endings of words was not observed in the semantically anomalous sentence contexts, suggesting an important early contribution of semantic knowledge to the word recognition process (see also Tyler & Wessels, 1983). The results of these experiments support a class of models in which word recognition processes produce a set of lexical candidates that are specified by both the early acoustic-phonetic input and the syntactic and semantic constraints derived from the sentence contexts. Bottom-up sensory input does, however, appear to have a greater weighting or relative priority in controlling lexical hypothesization from incomplete acoustic-phonetic input in the signal.

Insert Figure 7 about here

----

(6) Perception of Fluent Connected Speech

Most of the research in speech perception carried out over the last thirty-five years has been directed almost exclusively at trying to understand the perception of isolated speech sounds. The bulk of this work has examined phoneme perception in isolated contexts using simple nonsense syllables as stimulus materials. Although this approach is quite narrow in scope, the research strategy can be understood when one begins to consider the enormous complexity of spoken language perception and understanding, particularly the perception of fluent connected speech (Cole, 1980). Relative to the voluminous literature in speech perception on isolated phoneme perception, very little is actually known at this time about how the early sensory-based acoustic-phonetic information output by the peripheral auditory system is used by the higher and more central speech processing mechanisms in tasks that involve word recognition or language comprehension (Bagley, 1900; Cole and Rudnicky, 1983). And, very little is known about how changes in the segmental and/or suprasegmental structure of the speech signal affect intelligibility and subsequent comprehension of the linguistic message. Research on the perception of fluent speech obviously represents a sharp departure from the "mainstream" of traditional work in the field of speech perception (Cole, 1980). Understanding the perception of connected speech is an extremely important area of investigation in speech perception because spoken language comprehension is ultimately dependent on the initial sensory and perceptual analysis of the acoustic-phonetic input to the processing system.

Human speech perception and spoken language comprehension appear to take place very rapidly in close to real-time. A good deal of the perceptual processes and computational mechanisms that support such on-line activities operate automatically and therefore are unavailable to consciousness. Moreover, even under very adverse listening conditions, human observers are able to extract the linguistic message from the speech signal despite the fact that it may be distorted or parts of it entirely obliterated. The questions surrounding the perception of fluent connected speech differ substantially from the issues related to phoneme and feature perception because they inevitable involve the listener's cognitive system and a consideration of how various sources of linguistic knowledge interact to support perception and comprehension. For

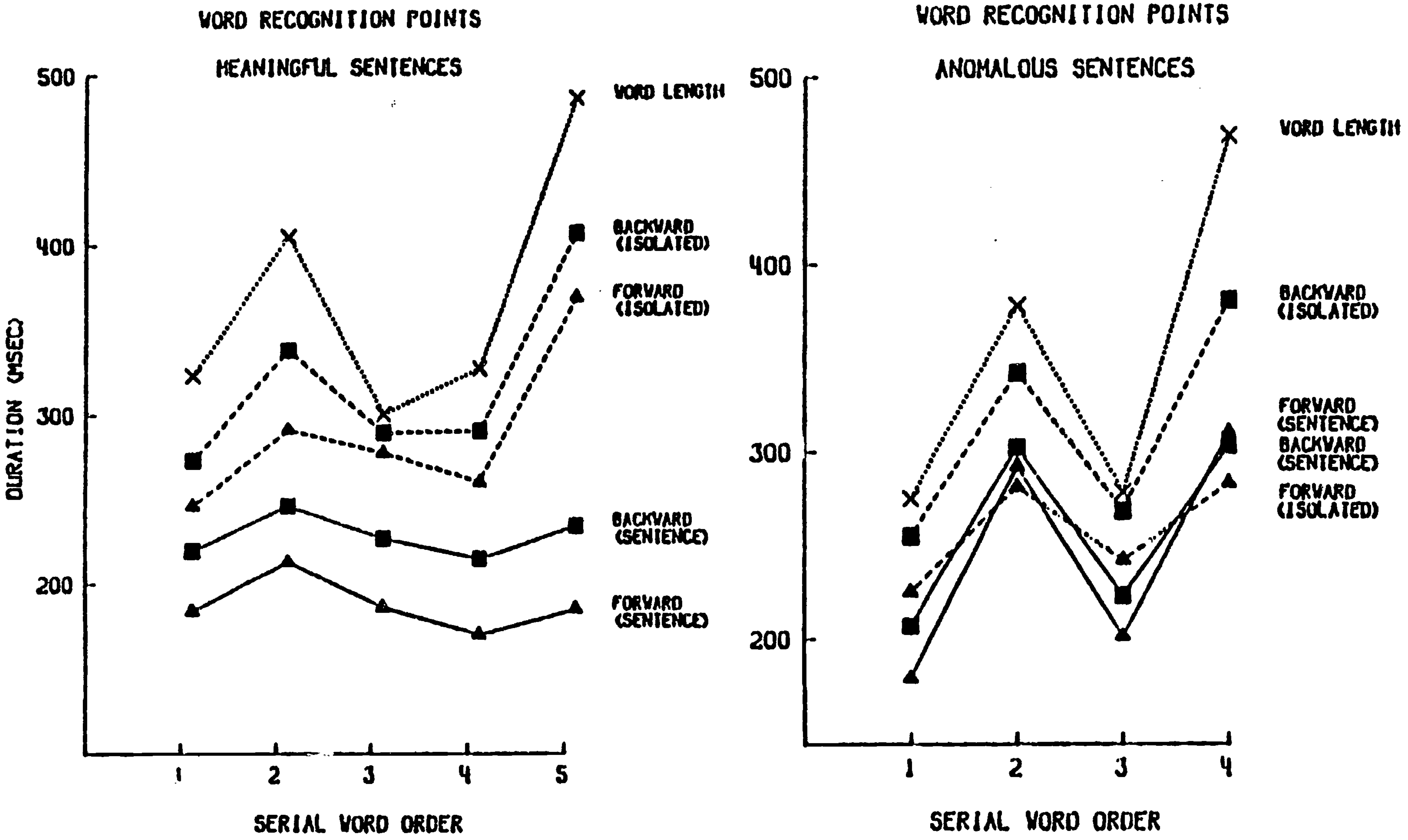


Figure 7. Identification points for words in meaningful and anomalous sentences expressed as msec of signal duration in each sentence position. Forward-gated and backward-gated words are shown both in sentence context and in isolation. The measured duration of each target word at each sentence position is also indicated.

example, what is the size of the scanning window over which low-level phonetic decisions are made? What depth of processing is required before a decision can be made about the segmental identity of the input signal? And, do all decisions of the speech processing system take pla:e immediately in real-time or are there processing delays at particular analytic levels of the system which wait for further information about the input signal and content of the message? These are just few of the questions that researchers are working on at this time (see Marslen-Wilson & Tyler, 1980).

Research on speech communication over the last thirty-five years, has shown that the linguistic message contains a great deal of redundant information. One of the goals of communication engineers was to identify the most important information in the speech signal and develop methods to encode this information in efficient ways for transmission. In a similar way, speech researchers have tried to identify the "minimal cues" for phoneme perception in the hope that once these cues could be identified the problems of speech recognition could be solved. Unfortunately, there is a great deal more to speech perception and spoken language understanding than simply discovering the minimal cues to phoneme perception. In studying the perception of fluent continuous speech, the problem becomes one of discovering the "islands of reliability" in the speech signal that listeners use to access various sources of knowledge. Examples of such information include the presence of stressed syllables, the beginnings and endings of words and the locations of various spectral changes indicating discrete shifts in the vocal tract source function (Jakimik & Hunnicutt, 1981; Bond, 1978; Nooteboom, 1981). Human listeners do not pay equal attention to all aspects of the speech signal; instead they use only those properties or attributes that are informative at any point in time to reach a decision (Marslen-Wilson and Tyler, 1980).

In our laboratory, we have begun to develop new experimental methodologies to study the listener's processing load and how it changes during the real-time analysis of spoken language. Using a measure of listening time, we have obtained reliable estimates of the cognitive or mental processes listeners use in perceiving and understanding long passages of connected fluent speech (Mimmack & Pisoni, 1982). In one procedure, listeners controlled the output of successive sentences of a passage in a self-paced procedure that allowed the recording of response latencies. We found that a number of important variables related to the structure of the linguistic materials affected the response latencies. Moreover, we found that subjects' listening goals also affected the latencies in listening to successive sentences. Subjects instructed to recall the passages showed much longer latencies than subjects instructed to simply understand the content of the passage.

In another study employing a noise detection procedure with connected speech (Salasoo & Pisoni, 1982), we found that function words (articles, conjunctions and prepositions) appear to have a special status compared to content words (nouns, verbs and adjectives). The results suggest that the mental lexicon may be structurally organized into two major classes of words and that these classes may be accessed differentially during the processing of fluent speech.

(7) Summary and Conclusions

In this report I have tried to summarize a number of the major theoretical issues in the field of speech perception and a few of the most promising and exciting areas of current research. It should be clear from my remarks that a great deal of research activity is going on in the field of speech perception today and that a great deal more is being learned about how humans perceive and understand spoken language. The approaches to studying these problems have become more diverse in the last few years as researchers in several related disciplines apply their knowledge and expertise to the central problems in the field. Moreover, a great deal more communication and interaction is taking place among speech researchers and hearing scientists, particularly on issues related to auditory modeling of the early stages of speech perception.

It is clear, at least to me, that new and important findings will come from continued research on how speech is processed in the auditory periphery. However, no one should be misled into believing that these new research efforts on processing speech by the auditory nerve will provide all the needed solutions in the field of speech processing (see also Klatt, 1982). On the contrary, a great deal more research is needed on central auditory mechanisms involved in pattern recognition and on word recognition and the perception of fluent connected speech. More basic research is needed on spectrogram reading and the development of large data bases that can be used to evaluate new hypotheses about the different sources of variability in speech. However, an important change in direction and attitude can be seen among speech researchers. Investigators appear to be directing their research efforts and attention toward much broader theoretical issues than just a few years ago, issues that encompass the study of more meaningful linguistic stimuli in somewhat more naturalistic contexts that require listeners to use several sources of knowledge to assign a linguistic interpretation to the sensory input. An important shift can be seen in the direction of research efforts towards a much greater concern for the differential contribution of context to the acoustic-phonetic realization of the speech signal. If nothing else, researchers appear to be very optimistic that a solution to the invariance problem is actually possible. It appears to be only a matter of time and more basic research into the complexity of the speech code.

References

Bailey, P. J., Summerfield, Q. and Dorman, M. On the identification of sine-wave analogues of certain speech sounds. Haskins Laboratories Status Report on Speech Research, 1977, SR-51/52, 1-25.

Bagley, W. C. The apperception of the spoken sentence: A study in the psychology of language. American Journal of Psychology, 1900-01, 12, 80-130.

Bever, T. G., Lackner, J. and Kirk, R. The underlying structure sentence is the primary unit of immediate speech processing. Perception & Psychophysics, 1969, 5, 225-234.

Blumstein, S. E. and Stevens, K. N. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. Journal of the Acoustical Society of America, 1979, 66, 1001-1017.

Blumstein, S. E. and Stevens, K. N. Perceptual invariance and onset spectra for stop consonants in different vowel environments. Journal of the Acoustical Society of America, 1980, 67, 648-662.

Bond, Z. S. Listening to elliptical speech: Pay attention to stressed vowels. Paper presented at the annual meeting of the Linguistic Society of America, Boston, December 1978.

Carlson, R. and Granstrom, B. Towards an auditory spectrograph. In R. Carlson & B. Granstrom (Eds.) The Representations of Speech in the Peripheral Auditory System. New York: Elsevier 1982, Pp. 109-114.

Cole, R. A. (Ed.), Perception and Production of Fluent Speech. Hillsdale, NJ: Erlbaum, 1980.

Cole, R. A., Rudnicky, A. I., Zue, V. W. and Reddy, D. R. Speech as patterns on paper. In R. A. Cole (Ed.), Perception and Production of Fluent Speech . Hillsdale, NJ: Erlbaum, 1980.

Cole, R. A. and Rudnicky, A. I. What's new in Speech Perception? The research and ideas of William Chandler Bagley, 1874-1946. Psychological Review, 1983, 90, 94-101.

Cole, R. A., Stern, R. M. and Lasry, M. J. Performing fine phonetic distinctions: Templates vs. features. Paper presented at the Symposium on Variability and Invariance of Speech Processes, MIT, October 8-10, 1983.

Cooper, F. S. Acoustics in human communication: Evolving ideas about the nature of speech. Journal of the Acoustical Society of America, 1980, 68, 1, 18-21.

Cross, D. V. and Lane, H. L. On the discriminative control of concurrent responses: The relations among response frequency, latency and topography in auditory generalization. Journal of the Experimental Analysis of Behavior, 1962, 5, 487-496.

Crystal, T. H. and House, A. S. Segmental durations in connected speech signals: Preliminary results. Journal of the Acoustical Society of America, 1982, 72, 705-716.

Cutting, J. E. and Rosner, B. S. Categories and boundaries in speech and music. Perception & Psychophysics, 1974, 16, 564-570.

Delgutte, B. Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. Journal of the Acoustical Society of America, 1980, 68, 3, 843-857.

Delgutte, B. Representation of speech-like sounds in the discharge patterns of auditory nerve fibers. Unpublished doctoral dissertation, M.I.T., 1981.

Delgutte, B. Some correlates of phonetic distinctions at the level of the auditory nerve. In R. Carlson & B. Granstrom (Eds.) The Representation of Speech in the Peripheral Auditory System. New York: Elsevier Biomedical Press, 1982.

Eimas, P. D. and Miller, J. L. Contextual effects in infant speech perception. Science, 1980, 209, 1140-1141.

Elman, J. L. and McClelland, J. L. Exploiting lawful variability in the speech wave. Paper presented at the Symposium on Variability and Invariance of Speech Processes, MIT, October 8-10, 1983.

Eukel, B. A phonotactic basis for word frequency effects: Implications for lexical distance metrics. Paper presented at the Acoustical Society of America, Los Angeles, 1980.

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

Forster, K. I. Accessing the mental lexicon. In R. J. Wales and E. Walker (Eds.), New Approaches to Language Mechanisms. Amsterdam: North Holland, 1976, Pp. 257-287.

Gaitenby, J. H. The elastic word. Haskins Laboratories Status Report on Speech Research, 1965, SR-2, 3.1-3.12.

Goldhor, R. A speech signal processing system based on a peripheral auditory model. Paper presented at the IEEE-ICASSP-83, Boston, April, 1983 (a).

Goldhor, R. The representation of speech signals in a model of the peripheral auditory system. Paper presented at the Acoustical Society Meetings, Cincinnati, Ohio, May, 1983 (b).

Greene, B. G., Pisoni, D. B., and Carrell, T. D. Recognition of visual displays of speech. Journal of the Acoustical society of America, 1984 (In Press).

Grosjean, F. Spoken word recognition and the gating paradigm. Perception & Psychophysics, 1980, 28, 267-283.

Grunke, M. E. and Pisoni, D. B. Some experiments on perceptual learning of mirror-image acoustic patterns. Perception and Psychophysics, 1982, 31, 3, 210-218.

Jakimik, J. and Hunnicutt, S. Organizing the lexicon for recognition. Paper presented at the 101st Meeting of the Acoustical society of America, Ottawa, Canada, May 18-22, 1981.

Jusczyk, P. W., Pisoni, D. B., Reed, M., Fernald, A. and Myers, M. Infants' discrimination of the duration of a rapid spectrum change in nonspeech signals. *Science*, 1983, *222*, 175-177.

Kewley-Port, D. Representations of spectral change as cues to place of articulation in stop consonants. *Research on Speech Perception Technical Report No. 3.* Bloomington, IN: Speech Research Laboratory, Indiana University, 1980.

Kewley-Port, D. Converging approaches towards establishing invariant acoustic correlates of stop consonants. Paper presented at the Symposium on Invariance and Variability of Speech Processes, M.I.T., October 8-10, 1983.

Kewley-Port, D. Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 1983, *73*, 1, 322-335.

Kewley-Port, D. and Luce, P. A. Time-varying features in voiced and voiceless stops produced at different speaking rates. *Research on Speech Perception. Progress Report No. 7.* Bloomington, Indiana: Speech Research Laboratory, Indiana University, 1981.

Kewley-Port, D., Pisoni, D. B. and Studdert-Kennedy, M. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, 1983, *73*, 5, 1779-1793.

Kiang, N. Y. S. Processing of speech by the auditory nervous system. *Journal of the Acoustical Society of America*, 1980, *68*, 3, 830-835.

Klatt, D. H. Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, 1975, *3*, 129-140.

Klatt, D. H. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 1976, *59*, 1208-1221.

Klatt, D. H. Review of the ARPA speech understanding project. *Journal of the Acoustical Society of America*, 1977, *62*, 1345-1366.

Klatt, D. H. Speech Perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phoentics*, 1979, *7*, 279-312.

Klatt, D. H. Lexical representations for speech production and perception. In T. Myers, J. Laver & J. Anderson (Eds). *The Cognitive Representation of Speech*. Amsterdam: North-Holland, 1981. Pp. 11-37.

Klatt, D. H. Speech processing strategies based on auditory models. In R. Carlson & B. Granstrom (Eds.). *The Representation of Speech in the Peripheral Auditory System*. New York: Elsevier, 1982, Pp. 181-196.

Klatt, D. H. The problem of variability in speech recognition and in models of speech perception. Paper presented at the Symposium on Variability and Invariance of Speech Processes, MIT, October 8-10, 1983.

Landauer, T., and Streeter, L. Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, 1973, *12*, 119-131.

Lehiste, I. Suprasegmentals. Cambridge, MA: M.I.T. Press, 1970.

Liberman, A. M. Duplex perception and integration of cues: Evidence that speech is different from nonspeech and similar to language. Proceedings of the Ninth International Congress of Phonetic Sciences, Copenhagen, 1979. Copenhagen: Institute of Phonetics, University of Copenhagen, 1980.

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

Liberman, A. M., Cooper, F. S., Shankweiler, D. and Studdert-Kennedy, M. Why are speech spectrograms hard to read? American Annals of the Deaf, 1968, 113, 127-133.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology, 1957, 54, 358-368.

Liberman, A. M., Harris, K. S., Kinney, J. A., and Lane, H. L. The discrimination of relative onset time of the components of certain speech and non-speech patterns. Journal of Experimental Psychology, 1961, 61, 379-388.

Lowerre, B. and Reddy, D. R. The Harpy speech understanding system. In W. A. Lea (Ed.) Trends in Speech Recognition. Englewood Cliffs: Prentice-Hall, 1979.

Markel, J. D. and Gray, A. H. Linear prediction of speech. New York: Springer-Verlag, 1976.

Marslen-Wilson, W. and Tyler, L. K. The temporal structure of spoken language understanding. Cognition, 1980, 8, 1-71.

Massaro, D. W. and Oden, G. C. Speech perception: A framework for research & theory. In N.J. Lass (Ed.), Speech and Language: Advances in Basic Research and Practice. Vol. 3, NY: Academic, 1980.

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

Miller, G. A. Decision units in the perception of speech. IRE Transactions on Information Theory, 1962, IT-8, 81-83.

Miller, J. D., Wier, C. C., Pastore, R., Kelly, W. J., and Dooling, R. J. Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. Journal of the Acoustical Society of America, 1976, 60, (2), 410-417.

Miller, J. L. Effects of speaking rate on segmental distinctions. In P. D. Eimas and J. L. Miller (Eds.) Perspectives on the Study of Speech. Hillsdale, N.J.: Lawrence Erlbaum, 1981. Pp. 39-74.

Miller, J. L. and Liberman, A. M. Some effects of later-occurring information on the perception of stop consonants and semi-vowels. Perception & Psychophysics, 1979, 25, 457-465.

Miller, M. I. and Sachs, M. B. Representation of stop consonants in the discharge patterns of auditory-nerve fibers. Journal of the Acoustical Society of America, 1983, 74, 2, 502-517.

Moore, B. C. J. and Glasberg, B. R. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. Journal of the Acoustical Society of America, 1983, 74, 3, 750-753.

Nooteboom, S. G. Speech rate and segmental perception or the role of words in phoneme identification. In T. Myers, J. Laver & J. Anderson (Eds.), The Cognitive Representation of Speech. Amsterdam: North-Holland, 1981, 143-150.

Patterson, R. D. Auditory filter shapes derived with noise stimuli. Journal of the Acoustical Society of America, 1976, 59, 640-654.

Pisoni, D. B. Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stops. Journal of the Acoustical Society of America, 1977, 61, 1352-1361.

Pisoni, D. B. Speech Perception. In W. K. Estes (Ed.), Handbook of learning and cognitive processes (Vol. 6). Hillsdale, N. J.: Erlbaum Associates, 1978. Pp. 167-233.

Pisoni, D. B. In Defense of Segmental Representations in Speech Processing. Paper presented at the Acoustical Society meeting, Ottawa, Canada, April, 1981.

Pisoni, D. B., Carrell, T. D. and Gans, S. J. Perception of the duration of rapid spectrum changes: Evidence for context effects with speech and nonspeech signals. Perception and Psychophysics, 1983, 34, (4) 314-322.

Plomp, R. The ear as a frequency analyzer. Journal of the Acoustical Society of America, 1964, 36, 1628-1636.

Reddy, D. R. speech recognition by machine: A review. Proceedings of the IEEE, April 1976, 64, 501-523.

Remez, R. E., Rubin, P. E., Pisoni, D. B. and Carrell, T. D. Speech perception without traditional speech cues. Science, 1981, 212, 947-950.

Sachs, M. B. and Young, E. D. Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. Journal of the Acoustical Society of America, 1979, 66, 2, 470-479.

Sachs, M. B. and Young, E. D. Effects of nonlinearities on speech encoding in the auditory nerve. Journal of the Acoustical Society of America, 1980, 68, 3, 858-875.

Schroeder, M. R., Atal, B. S. and Hall, J. L. Optimizing digital speech coders by exploiting masking properties of the human ear. Journal of the Acoustical Society of America, 1979, 66, 1647-1652.

Schwab, E. C. Auditory and phonetic processing for tone analogs of speech. Unpublished doctoral dissertation, SUNY at Buffalo, 1981.

Searle, C. L., Jacobson, J. F. and Rayment, S. G. Stop consonant discrimination based on human audition. *Journal of the Acoustical Society of America*, 1979, *65*, 3, 799-809.

Shipman, D. W. Development of speech research software on the MIT lisp machine. *Journal of the Acoustical Society of America*, 1982, 71, S103.

Shipman, D. W. Spirex: Statistical Analysis in the Spire Acoustic-Phonetic Workstation. *Proceedings of the IEEE ICASSP-83*, 1983, *3*, 1360-1363.

Shipman, D. W. and Zue, V. W. Properties of large lexicons: Implications for advanced isolated word recognition systems. *Proceedings of the 1982 IEEE International Conference on Acoustics, Speech and Signal Processing*. Paris, France, April 1982.

Siebert, W. M. Stimulus transformations in the peripheral auditory system. In P. A. Kolers and M. Eden (Eds.), *Recognizing patterns*. Cambridge, Mass.: MIT Press, 1968.

Stevens, K. N. Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America*, 1980, *68*, 3, 836-842.

Stevens, K. N. and Blumstein, S. E. Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 1978, *64*, 1358-1368.

Studdert-Kennedy, M. The perception of speech. In T. A. Sebeok (Ed.), *Current Trends in Linguistics* (Vol. XII). The Hague: Mouton, 1974.

Studdert-Kennedy, M. Speech perception. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics*. New York: Academic Press, 1976. Pp. 243-293.

Watson, C. S. and Foyle, D. C. Central factors in the discrimination and identification of complex sounds. Paper presented at the CHABA meetings, October 19-20, 1983, Washington, D.C.

Young, E. D. and Sachs, M. B. Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *Journal of the Acoustical Society of America*, 1979, *66*, 5, 1381-1403.

Zue, V. W. Acoustic characteristics of stop consonants: A controlled study. Technical Report No. 523, Lincoln Laboratory, M.I.T., May 1976.

Zwicker, E., Terhardt, E. and Paulus, E. Automatic speech recognition using psychoacoustic models. *Journal of the Acoustical Society of America*, 1979, *65*, 2, 487-498.

Linguistic experience and infant speech perception:

A re-examination of Eilers, Gavin & Oller (1982)*

Peter W. Jusczyk

Department of Psychology
University of Oregon
Eugene, Oregon 97403

and

Sandra L. Shea and Richard N. Aslin

Infant Perception Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

* Preparation of this manuscript was supported in part by grants from NICHHD to the first author (15795) and to the third author (11915 and 00309).

Linguistic experience and infant speech perception:

A re-examination of Eilers, Gavin & Oller (1982)

The first studies of speech discrimination in young infants (e.g., Eimas, Siqueland, Jusczyk & Vigorito, 1971) rekindled a longstanding interest in the role played by specific language experience in the development of speech perception. During the subsequent years of research on infant speech perception, investigations focused on whether these capacities were "learned" or "induced" directly from the specific language input experienced by the infant (see Aslin & Pisoni, 1980). One hypothesis was that infants might learn to discriminate and categorize phonemes in their native language by listening to their own vocal productions and then relating these vocalizations (as well as the vocalizations of other speakers) to a set of articulatory gestures (e.g., Liberman, Harris, Kinney & Lane, 1961). Sounds produced by the same set of articulatory gestures would become perceptually "equivalent", whereas sounds produced by different gestures would become perceptually "distinctive". In this way, phonetic categories would emerge from the mapping of perceptual cues onto articulatory categories. However, as the data from studies of infant speech perception began to accumulate, it became apparent that infants already possessed the capacity to discriminate many, if not all, speech contrasts soon after birth with little or no prior exposure to any specific language input (see reviews by Aslin, Pisoni & Jusczyk, 1983; Jusczyk, 1981).

In light of these empirical findings, the emphasis with respect to the role of specific language experience shifted somewhat. Instead of focusing on whether some capacities were acquired purely as the result of experience, many investigators turned their attention to the way in which specific language experience might modify pre-existing perceptual capacities. As is well known, a given pair of languages does not necessarily employ the same set of phonetic categories. Thus, it was of considerable interest to determine whether specific language experience exerts a significant effect on the perceptual capacities of adults. Several studies with adults demonstrated that cross-linguistic differences exist in the presence and location of perceptual boundaries for certain phonemic categories (e.g., Lisker & Abramson, 1970; Miyawaki, Strange, Verbrugge, Liberman, Jenkins & Fujimura, 1975; Williams, 1977). Thus, there was little doubt that specific language experience does affect underlying speech perception capacities. It now became important to ask when (at what age) and how (via what mechanism) these speech perception capacities are affected by specific language experience.

These two questions (when and how) are closely related. For example, if one hypothesized that specific language experience operated primarily at the sensory level, requiring little or no active processing by the listener, then one might expect the effects of specific language experience to begin shortly after birth. This passive model of early experience is essentially the one proposed to account for the dramatic effects of visual input on the neural properties in the cat and monkey visual cortex. According to this passive model, the frequency and distribution of sounds in the infant's listening environment would serve to tune or shape the basic perceptual capacities underlying speech discrimination and categorization. Alternatively, if one hypothesized that the listener must take a more active role in the processing of speech sounds, then one might not expect to see evidence of language specific differences in speech perception until the onset of a rudimentary semantic system (i.e., between 9 and 12 months of age). This more active model of the role of early experience in speech perception would incorporate the notion that the acquisition of a vocabulary (a specific correspondence between sound and meaning) would provide the primary impetus for adjusting the infant's underlying speech perception capacities to a specific language. Presumably, in the course of acquiring the words in a specific language, the infant listener would weight the available acoustic input to produce an optimal fit to the phonological structure of the language that is being learned (for a further elaboration of this view, see Jusczyk, in press, and Eimas, Miller & Jusczyk, in press).

At present, the available data on the development of speech perception are insufficient to choose between these two hypotheses. However, regardless of which hypothesis is correct, one would expect to find a lawful relation between the specific language input actually experienced by an infant and any developmental changes in perceptual capacities. For example, one would not expect infants from a given language community to have perceptual capacities for infrequently experienced speech contrasts that were superior to capacities of infants from a different language community whose linguistic input contained numerous instances of this speech contrast. For this reason, the findings reported recently in this journal by Eilers, Gavin, and Oller (1982; hereafter referred to as EGO) are quite puzzling. EGO reported the results of an experiment in which the discriminative capacities of two groups of infants, one from English-speaking homes and the other from Spanish-speaking homes, were tested on phonetic contrasts from English, Spanish, and Czech. EGO reported evidence for the superiority of the Spanish-learning infants on the Spanish contrast, a result which they attributed to specific experience with the native language. By comparison, EGO reported no evidence for a superiority of the English-learning infants on the English contrast, a result which they attributed to the Spanish-learning infants' greater opportunity to experience the English contrast in their bilingual environment.

Although there is a troublesome asymmetry in these findings with English and Spanish contrasts, the most perplexing aspects of the findings reported by EGO were those involving discrimination performance with the Czech contrast. Both the Spanish-learning and English-learning infants provided evidence of discriminating the Czech contrast despite the absence of any previous experience with this contrast in the listening environment. Moreover, the performance by the Spanish-learning infants was superior to that of the English-learning infants, a result which EGO attributed to the Spanish-learning infants' "generally richer phonological experience" (p. 301). The most puzzling aspect of these results with the Czech contrast, however, was the fact that the Spanish-learning infants evidently discriminated this contrast better than the Spanish contrast. EGO never commented on this finding, but inspection of the data presented in their Table 6 shows that the highest performance was attained by the Spanish-learning infants on the Czech contrast. Furthermore, the size of the difference in performance by the Spanish-learning infants on the Czech and Spanish contrasts indicates that this difference was statistically significant. Therefore, one possible conclusion to be drawn from these findings is that the kinds of language input experienced by Spanish-learning infants make it easier for them to learn Czech than to learn Spanish!

It is difficult to see how any existing account of the role of specific language experience in the development of speech perception could handle these findings as reported by EGO. For example, such an account would have to face the following problems: (1) the asymmetric superiority of the Spanish-learning infants compared to the English-learning infants in discriminating contrasts employed in the phonology of their native language, (2) the fact that, although both the Spanish-learning and English-learning infants discriminated the Czech contrast, the Spanish-learning infants showed superior discriminative capacities, and (3) the superior performance of the Spanish-learning infants on the Czech contrast compared to the Spanish contrast. Without a number of post hoc assumptions, it is doubtful that the types of models of specific language experience considered above could deal with these problems. Given the inadequacies of current models for explaining these findings, there appear to be two alternatives: (1) reject these models in favor of some more coherent view of the effects of specific language experience that could explain the results reported by EGO, or (2) offer a different interpretation of the data presented in EGO's report.

A consideration of EGO's account of the role of linguistic experience

Let us consider the first alternative: do EGO provide some framework for understanding the effects of specific language experience on the development of speech perception? Unfortunately, EGO did not devote much discussion to this issue in their report. Nevertheless, they did state that "the difference between the infant groups on the Spanish and Czech contrasts offers a further suggestion that language has a measurable effect on early speech discrimination skills" (p. 300). And later, they "speculate that the phonological superiority of the Spanish-learning infant ... is a function of a generally richer phonological experience" (p. 301). Presumably, it is the richer phonological experience of the Spanish-learning infants that provides them with an advantage in the discrimination tasks. But in what way is their experience richer? It certainly is not the case that the phonology of Spanish is intrinsically richer than that of English. Rather, according to EGO, the richer phonological experience of Spanish-learning infants is because "to varying extents, the Spanish-learning infants are reared in a bilingual community while the English-learning infants are primarily in a monolingual environment" (p. 301). EGO base this claim on anecdotal observations that Spanish-learning infants are more apt to hear English spoken on television, to accompany parents to public places where English is spoken, and to have older siblings who speak English than English-learning infants are apt to have parallel experiences with Spanish. Even if this claim were true, it could, at best, only serve to explain the results for the Spanish and English contrasts. The Czech results would remain a mystery because (1) the English-learning infants discriminated this contrast even though they lacked the richer "phonological" environment, and (2) with the additional exposure to Spanish, the Spanish-learning infants showed superior performance on the Czech contrast.[1] To say that somehow "the richness of the Spanish-learning infants' linguistic environment may then contribute to superior speech discrimination skills early in life, and may even generalize to acoustic events which are truly foreign" (p. 301) serves only to restate the results, not to explain them. Thus, EGO's discussion of their results does not provide a coherent account of the effects of specific language experience on speech perception. Is it possible, then, that there is some anomaly in their treatment of the data that led them to draw erroneous conclusions? We believe that there were several aspects of EGO's data analysis that were misleading and we shall devote the remainder of this paper to a re-examination of their results.

A re-evaluation of EGO's empirical findings

The most striking aspect of the data that EGO present is the rather dismal levels of performance attained by the infants on the three speech contrasts. The percentage correct scores attained by the two groups of infants on these 3 contrasts (a total of 6 comparisons) ranged frcm 52 to 62%. The best performance was on the Czech contrast by Spanish-learning infants, while performances on the remaining 5 comparisons were below 56% correct. Given that chance level performance on these tasks is 50%, one wonders why such low scores yielded significant evidence for discrimination on 5 of the 6 comparisons. Ordinarily, one would expect to find significant results for scores so close to chance level responding only if there were large numbers of subjects in each group. Yet, in EGO's study there were only 14 subjects per group. In this case, the difference appears to lie in EGO's unorthodox statistical treatment of the data. In particular, there are two aspects of the statistics that bear close scrutiny: (1) the use of a discriminative index (DI) in place of percentage correct (PC) scores, and (2) the use of the *z*-test in place of either a *t*-test or an appropriate nonparametric test in analyzing the data.

The Discriminative Index

The DI is the number of correct responses, or hits, minus the number of false positives (headturns on control trials) divided by the number of possible hits. In arguing that the DI is a better measure of discrimination performance than PC, EGO state that the DI "takes into account the actual number of opportunities an individual has for demonstrating discrimination within a trial block" (p. 296). There are three types of 10-trial blocks that can occur in EGO's experiment: (1) 5 change and 5 control trials, (2) 4 change and 6 control trials, or (3) 6 change and 4 control trials. EGO base their argument in favor of the DI on an example in which an infant receives 6 control and 4 change trials and makes no headturns. In this example, a PC measure would yield a score of 0.60 because the infant was "correct" in not responding to the 6 control trials. The DI score for the same block of 10 trials would be 0.0 (0 hits minus 0 false positives divided by 4 trials), a score which would have been obtained if headturns had occurred by chance. EGO argue that a score of 0.0, or chance, better describes the infant's performance in this situation than does a score of 0.60 because the PC score is greater than chance. Thus, EGO conclude that "DI scores offer an advantage over PC scores whenever there is a possibility for unequal numbers of control and change trials within trial blocks" (p. 296), and, for this reason, they go on to report their results based primarily on the analysis of DI scores.

There are several objections that can be raised with respect to EGO's decision to base their analyses on DI scores rather than PC scores. First, even assuming that EGO are correct in asserting that DI scores are superior for blocks with unequal numbers of control and change trials, it is not clear why it is necessary to include such unbalanced blocks unless one were going to analyze the results for individual trial blocks (which EGO did not report)., Moreover, it was not clear if all infants actually received an equal number (20-20) of change and control trials (i.e., blocks of 4-6, 5-5, 6-4 and 5-5) or whether these numbers were allowed to vary between 16-24 and 24-16. Second, the example that EGO chose to justify using the DI measure is not a very realistic one. All the infants were pretested with a /ba/-/da/ stimulus pair until they met a criterion of 9 correct out of 10 successive trials. Thus, it would seem highly unlikely during a subsequent block of 10 trials that an infant would fail to make any headturns. In fact, our experience with the headturning technique suggests that false positives (turns on control trials) occur as frequently as misses (no turns on change trials).

A more serious objection to EGO's use of DI scores is that these scores are simply not superior to PC scores for blocks with unequal numbers of control and change trials. To support our claim, consider the following example. Assume that an infant is given a block of 4 change and 6 control trials, but instead of not responding at all, the infant makes a headturn on all 10 trials. This would result in a PC score of 0.40, which is the same distance from chance (0.10) as the example in which no headturns were made (0.60 - 0.50 = 0.10). However, the DI score for an infant who always made a headturn would be -0.50 (4 hits minus 6 false positives divided by 4), a score which is considerably below the overall population mean of 0.0. This introduction of negative scores into the DI prompted us to take a closer look at the distributions of DI and PC scores for the different possible combinations of test trials.

Consider all possible combinations of responses on change and control trials; that is, the probability of making a headturn is 0.5 on any given trial, and this probability is independent of the presence of a change in the speech sound. Under these conditions, PC scores are normally distributed between 0.0 and 1.0, with a mean of 0.50 regardless of the ratio of control to change trials (see Table 1). By comparison, DI scores have ranges (from 1.67 to 2.5), standard deviations (from 0.41 to 0.62) and means (from -0.25 to 0.17) which change as a function of the ratio of control to change trials. Most importantly, note that only one of the means (the one for equal numbers of control and change trials) equal the value of the population mean (u=0.0). If one examines the frequency distributions of the PC scores for the three different types of trial blocks (see Figure 1), it is apparent that all are centered on the population mean of 50%, and that the summed frequency distribution across the trial blocks also yields a symmetrical distribution about the mean. In contrast, the DI scores produce different frequency distributions for the three types of trial blocks, and the resulting summed frequency distribution across blocks of trials is negatively skewed.

Insert Table 1 and Figure 1 about here

One of the results of this skewed DI distribution is the introduction of more bias in the DI scores than in the PC scores when P(headturn) is greater than 0.0 (see Table 2). It is expected that an index would yield a score at or near chance when the infant responded randomly. As can be seen in Table 2, the average PC score across three trial blocks is 50% at any level of response bias. In contrast, the average DI score across three trial blocks equals the population mean (u = 0.0) only when the infant never responds. Note that this example of an infant who never makes a headturn is exactly the condition that EGO chose as the raison d'etre for the DI. But, as stated earlier, their screening criteria were designed to select infants who would make a headturn on at least half of the trials. Moreover, if an infant made a headturn on half of the trials, the average DI score could be as high as 0.15, which is almost a full standard deviation above the population mean. Recall that EGO specifically state that DI is better than PC when there are unequal numbers of control and change trials. It is interesting, then, that the sample mean equals the population mean only in a block of 5 change and 5 control trials.

Insert Table 2 about here

Therefore, there appears to be little support for EGO's contention that DI scores represent an improvement over PC scores. Rather, the negatively skewed distribution of the DI imparts a bias whenever the probability of a headturn is greater than 0.0. If the probability of a headturn is very high, the DI score generates more significant differences from chance than the PC score. Hence, use of the DI increases the likelihood of committing a Type I error and erroneously rejecting the null hypothesis. In the present case, this would be manifest by falsely concluding that infants could discriminate a particular contrast, and thereby incorrectly supporting the claim that specific language experience affected discriminative capacity.

The Zm Statistic

Given the foregoing objections to the use of DI scores, one might argue that EGO could simply base their conclusions on the PC scores because they "showed the same pattern of results as the DI scores" (p. 298). However, the use of PC scores does not eliminate a serious problem in EGO's statistical treatment of both the DI and PC scores.

Table 1

DI and PC Scores for all Possible Combinations of Headturn Responses

6 change/4 control			5 change/5 control			4 change/6 control		
Change/Control	DI	PC	Change/Control	DI	PC	Change/Control	DI	PC
6/0	1.00	1.00	5/0	1.00	1.00	4/0	1.00	1.00
5/0	.83	.90	4/0	.80	.90	3/0	.75	.90
4/0	.67	.80	3/0	.60	.80	2/0	.50	.80
3/0	.50	.70	2/0	.40	.70	1/0	.25	.70
2/0	.33	.60	1/0	.20	.60	0/0	.00	.60
1/0	.17	.50	0/0	.00	.50	4/1	.75	.90
0/0	.00	.40	5/1	.80	.90	3/1	.50	.80
6/1	.83	.90	4/1	.60	.80	2/1	.25	.70
5/1	.67	.80	3/1	.40	.70	1/1	.00	.60
4/1	.50	.70	2/1	.20	.60	0/1	-.25	.50
3/1	.33	.60	1/1	.00	.50	4/2	.50	.80
2/1	.17	.50	0/1	-.20	.40	3/2	.25	.70
1/1	.00	.40	5/2	.60	.80	2/2	.00	.60
0/1	-.17	.30	4/2	.40	.70	1/2	-.25	.50
6/2	.67	.80	3/2	.20	.60	0/2	-.50	.40
5/2	.50	.70	2/2	.00	.50	4/3	.25	.70
4/2	.33	.60	1/2	-.20	.40	3/3	.00	.60
3/2	.17	.50	0/2	-.40	.30	2/3	-.25	.50
2/2	.00	.40	5/3	.40	.70	1/3	-.50	.40
1/2	-.17	.30	4/3	.20	.60	0/3	-.75	.30
0/2	-.33	.20	3/3	.00	.50	4/4	.00	.60
6/3	.50	.70	2/3	-.20	.40	3/4	-.25	.50
5/3	.33	.60	1/3	-.40	.30	2/4	-.50	.40
4/3	.17	.50	0/3	-.60	.20	1/4	-.75	.30
3/3	.00	.40	5/4	.20	.60	0/4	-1.0	.20
2/3	-.17	.30	4/4	.00	.50	4/5	-.25	.50
1/3	-.33	.20	3/4	-.20	.40	3/5	-.50	.40
0/3	-.50	.10	2/4	-.40	.30	2/5	-.75	.30
6/4	.33	.60	1/4	-.60	.20	1/5	-1.0	.20
5/4	.17	.50	0/4	-.80	.10	0/5	-1.2	.10
4/4	.00	.40	5/5	.00	.50	4/6	-.50	.40
3/4	-.17	.30	4/5	-.20	.40	3/6	-.75	.30
2/4	-.33	.20	3/5	-.40	.30	2/6	-1.0	.20
1/4	-.50	.10	2/5	-.60	.20	1/6	-1.2	.10
0/4	-.67	.00	1/5	-.80	.10	0/6	-1.5	.00
-	-	-	0/5	-1.0	.00	-	-	-
$\bar{x}$=	.17	.50		.00	.50		-.25	.50
s=	.41	.25		.49	.24		.62	.25
Range=	1.67	1.00		2.00	1.00		2.50	1.00

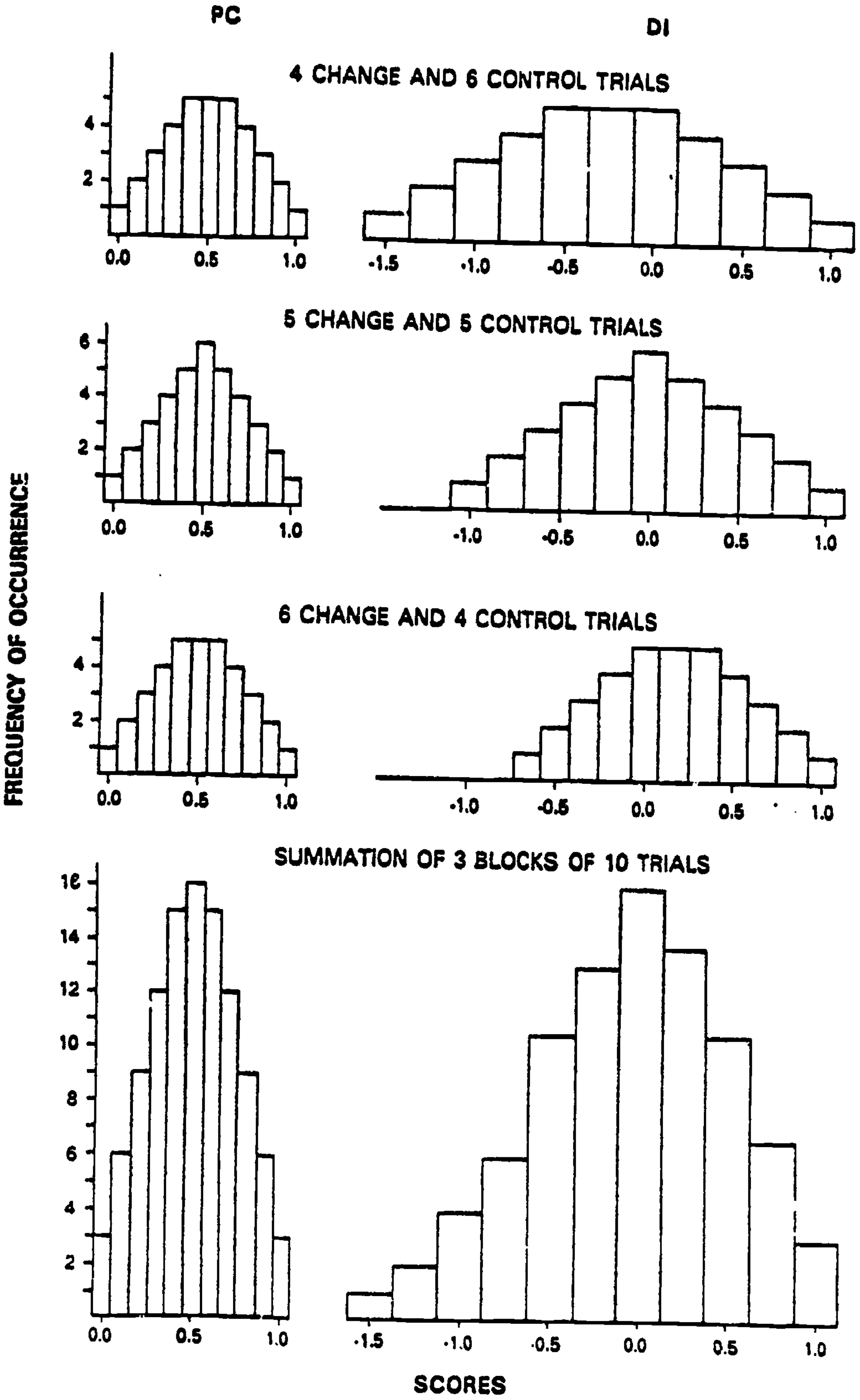


Figure 1. Frequency distributions for PC and DI scores for three blocks of 10 trials in which the ratio of change to control trials was varied. The bottom distributions illustrate the sums of the three types of trial blocks.

Table 2

Response Biases

	DI				PC			
P(Headturn)	6/4	5/5	4/6	$\bar{x}$	6/4	5/5	4/6	$\bar{x}$
1.00	.33	0.0	-.50	-.06	.60	.50	.40	.50
0.50*	.50	.20	-.25	.15	.50	.50	.50	.50
0.50*	.50	-.20	-.25	.02	.50	.50	.50	.50
0.0	0	0	0	0	.40	.50	.60	.50
Dist. $\bar{x}$	.17	0	-.25	-.03	.50	.50	.50	.50

*There are two alternatives for 50% responding when the trials are split between 5 change and 5 control; the infant either responds to 3 control and 2 change (-.20), or 3 change and 2 control (.20).

EGO used the Zm statistic to test the null hypothesis that the infants "as a group" exceeded chance expectation levels of responding on various speech contrasts. Unfortunately, this statistical test as applied to the data in their study is simply inappropriate because it grossly underestimates the amount of between-subject variance and thus inflates the probability of finding significant results. To demonstrate this point, let us consider the nature of the Zm statistic as used by EGO.

The Zm statistic requires that the population variance be known a priori. For this reason EGO employed a computer simulation to obtain an estimate of the population variance. The simulation was based on data from 10,000 pseudo-subjects who each received 40 trials in blocks of 10. The probability of each subject making a headturn was systematically varied from 0.0 to 1.0, and the resultant estimate of the population variance was greatest when the headturn probability was 0.5. EGO used this maximum variance estimate, divided by 14 (the number of subjects in each language group), to compute the DI and PC population parameters necessary to use the Zm test.

The major problem with EGO's use of the Zm statistic is that it violates the requirement of independent observations by treating the trial as the unit of analysis rather than the individual subject. Because the outcome of any given trial is either correct or incorrect regardless of the probability of a headturn, each trial can be considered a Bernoulli process and a sample of N trials conforms to the binomial distribution (Hays, 1972, pp. 178-193). However, in both EGO's computer simulation and their actual experiment, the use of the Zm statistic treats 40 trials from N subjects as if they were 40 x N trials from one subject. That is, if the 40 trials come from a single subject, or if 40 subjects provide a single trial, the observations are independent. But the 40 trials from one subject are not independent of each other when compared to the 40 trials from another subject. Thus, each subject's average level of performance (either absolute number or proportion of trials) must become the unit of analysis in any between-subject comparison and not the individual trial within-subjects.

To illustrate the way in which the Zm statistic, as employed by EGO, inflates the probability of obtaining a significant result, consider the following hypothetical example shown in Table 3. Two groups of 10 subjects each are tested on some speech contrast for 40 trials each (20 change and 20 control). The data are then analyzed to determine whether each group departs significantly from chance level responding. For this purpose both the Zm statistic and a conventional measure, the *t*-test, are used to evaluate the results. Using the *t*-test, the performance of Group I does not yield a mean that is significantly above chance, both because only 2 of 10 subjects performed above 50% correct and because the between-subject variance was quite high. In contrast, the performance of Group II does yield a mean that is significantly above chance, primarily because of the low between-subject variance. When the Zm statistic is applied to the results from Groups I and II (see lower portion of Table 3), both group means are found to differ from chance at the $p<.001$ level of significance. This finding for Group I is inappropriate because the between-subject variance is critical in determining whether a group mean differs from chance. It is not surprising, therefore, that a score of 54% correct responding was found to be significantly different from the 50% level expected by chance. In summary, not only did EGO's use of the Zm statistic systematically underestimate the between-subject variance, but it also violated the independence requirement by failing to recognize that the variance for trials collected within subjects is not equivalent to the variance for trials collected across subjects.

Insert Table 3 about here

A final analysis of the data

If we now examine the remaining data, based on the PC scores and the Tukey planned comparisons of the means, what evidence is there for an effect of early experience? First, consider only the data for the Spanish and English contrasts, the ones for which the strongest effects of specific language experience would be expected. Table 4 presents the PC scores on these contrasts for the Spanish- and English-learning infants. There are two points to be made with respect to these data. First, the largest difference between the two groups of language learners is 3.2% for the Spanish contrast, and this difference is not statistically significant according to the _t_-value (1.59) presented in EGO's Table 6. Second, the performance levels on all four conditions were remarkably poor, never attaining an accuracy rate of even 56% correct. Other studies using the headturning procedure (e.g., Aslin, Pisoni, Hennessy & Perey, 1981) have commonly used a criterion of 70% correct or greater for concluding that discrimination was present. EGO never commented on this low level of performance, but it seems highly unlikely that any of these scores would achieve statistical significance by a _t_-test or any other conventional statistical procedure. Thus, not only is there no evidence for an effect of language-specific experience, but it is highly doubtful that the infants gave evidence of discriminating either contrast.

Insert Table 4 about here

If we now consider the Czech contrast, EGO report PC scores of 56% for the English-learning infants and 62% for the Spanish-learning infants. It seems doubtful that the performance of the English-learners is significantly above chance. The data for the Spanish-learning infants may or may not exceed chance depending on the amount of between-subject variance. However, judging from the hypothetical example presented earlier (see Table 3), it is certainly plausible that this group mean of 62% was also not statistically above chance. Thus, at best, there is evidence that only the Spanish-learning infants discriminated the Czech contrast and, at worst, there is no evidence that either group discriminated any contrast -- Spanish, English or Czech. These findings are certainly not robust enough for EGO to make strong claims about the influence of

Table 3

Hypothetical PC Scores for Two Groups of Subjects

Subject	Group I	Subject	Group II
1	.50	11	.59
2	.50	12	.59
3	.50	13	.59
4	.50	14	.59
5	.50	15	.59
6	.50	16	.61
7	.50	17	.61
8	.50	18	.61
9	1.00	19	.61
10	1.00	20	.61
	$\bar{x}$=.60		$\bar{x}$=.60
	s=.2108		s=.0105
	t=0.15		t=30.3
	p=n.s.		p<.001

Zm test of both groups:

$x=.60$
$\sigma=.09$
$\sigma/\sqrt{n}=.0285$
$z=3.51$
$p<.001$

Table 4

PC Scores from Eilers, Gavin & Oller (1982)

		Language Group	
		English	Spanish
Phonetic Contrast	English	54.1%	54.3%
	Spanish	52.7%	55.9%

early language experience on speech discrimination. Moreover, EGO did not consider an alternative explanation of the relatively high level of performance obtained on the Czech contrast which involves the acoustic salience of the contrasts. Both the Spanish and English contrasts occurred in multi-syllabic utterances and in medial syllable position. By comparison, the Czech contrast occurred in single syllables and in syllable-initial position. This confound in the experimental design makes it difficult, if not impossible, to draw any conclusions about the contribution of early experience to speech perception in infants.

Conclusion

We undertook this re-examination of EGO's study because of their claim that "early experience DOES affect early discrimination, and further (since the stimuli were natural) that the effect may be of practical consequence in language learning" (p.289).[2] After reviewing EGO's theoretical and interpretive comments, their data and the statistical procedures used to analyze them, and certain methodological issues, we are forced to conclude that whatever factors are responsible for the pattern of results they obtained, it is extremely doubtful that linguistic experience is among them.

As we remarked at the outset, it is certain that linguistic experience does affect speech perception; otherwise, there would be no way to account for the cross-linguistic effects observed in adult listeners. In general, there are several ways in which such effects might occur. Exposure to a specific language might enable infants to somehow acquire or "learn" a contrast not previously discriminable using their initial repertoire of capacities. Experience with a specific language might serve to modify a set of innate perceptual capacities by tuning them to respond to certain values and/or by emphasizing certain portions of the signal. At present, the available data favor the latter view of the effects of experience. To cite one example, Werker and Tees (in press) have reported that between 6-8 months, English-learning infants are capable of discriminating phonetic contrasts that do not occur in English, but do occur in either Hindi or Salish. At a later date, by 10-12 months of age, the English-learning infants no longer show reliable evidence of discriminating the Hindi and Salish contrasts. Whether this shift in sensitivity is a result of passive stimulation or a consequence of active processing of the native language cannot as yet be determined.[3] Much empirical work remains before we have a clear understanding of the precise way in which language experience modifies the speech perception process.

Footnotes

1. EGO's use of the terms "phonological superiority" and "phonological experience" (p. 301) implies that these contrasts are treated as phonemic entities and cannot be discriminated simply on an acoustic basis. However, there is little empirical support for this claim (see Jusczyk, 1981; 1982 for a discussion of this issue). Moreover, the superior performance on the Czech contrast renders any phonological explanation extremely unlikely.

2. The parenthetical phrase in this quote deserves further comment. At one point EGO state "Moreover, the fact that the stimuli discriminated by the infants were natural rather than synthetic syllables suggests that early experience effects may offer concrete practical advantages to the infant in later language learning" (p.300). This seems to us to be an example _par excellence_ of a nonsequitor in logic. Apparently, EGO believe that discrimination of _natural_ speech contrasts provides stronger support for the interpretation that the infants' perceptual capacities are actually used in the native language environment (and not just in a laboratory testing session). However, this argument is severely weakened by the results from the Czech contrast which, of course, also consisted of natural and not synthetic speech sounds. Moreover, we suspect that EGO would have drawn a similar conclusion about the effects of early experience if they had used synthetic speech contrasts, as they did in an earlier report (Eilers, Gavin & Wilson, 1979).

3. MacKain (1982) makes a valid point when she argues that it is necessary to acquire better information for different languages about the frequency distributions of spoken tokens along various phonetic continua such as VOT, place of articulation, etc. Only in this way can one begin to evaluate the plausibility of a passive stimulation model. For example, even though some proportion of voiced tokens in English are actually prevoiced, it may be the case that a language, such as Thai, that employs prevoiced stops would have a much higher proportion of prevoiced utterances in a corpus drawn from its speakers.

References

Aslin, R. N. and Pisoni, D. B. Some developmental processes in speech perception. In G. Yeni-Komshian, J. F. Kavanagh and C. A. Ferguson (Eds.), *Child phonology: Vol. 2 Perception*. New York: Academic Press, 1980.

Aslin, R. N., Pisoni, D. B., Hennessy, B. L. and Perey, A. J. Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience. *Child Development*, 1981, *52*, 1135-1145.

Aslin, R. N., Pisoni, D. B. and Jusczyk, P. W. Auditory development and speech perception in infancy. In M. M. Haith and J. J. Campos (Eds.), *Infancy and the biology of development*. New York: Wiley, 1983.

Eilers, R. E., Gavin, W. J. and Oller, K. K. Cross-linguistic perception in infancy: early effects of linguistic experience. *Journal of Child Language*, 1982, *9*, 289-302.

Eilers, R. E., Gavin, W. J. and Wilson, W. R. Linguistic experience and phonemic perception in infancy: a cross linguistic study. *Child Development*, 1979, *50*, 14-18.

Eimas, P. D., Miller, J. L. and Jusczyk, P. W. On infant speech perception and the acquisition of language. In S. Harnad (Ed.), *Categorical perception*. Cambridge: Cambridge University Press, in press.

Eimas, P. D., Siqueland, E. R., Jusczyk, P. and Vigorito, J. Speech perception in infants. *Science*, 1971, *171*, 303-318.

Hays, W. L. *Statistics for the social sciences*. 2nd edition. New York: Holt, Rinehart and Winston, 1972.

Jusczyk, P. W. Infant speech perception: A critical appraisal. In P. D. Eimas and J. L. Miller (Eds.), *Perspectives on the study of speech*. Hillsdale, N.J.: Erlbaum Associates, 1981.

Jusczyk, P. W. Auditory versus phonetic coding of speech signals during infancy. In J. Mehler, E. Walker and M. Garrett (Eds.), *On mental representation*. Hillsdale, N. J.: Erlbaum Associates, 1982.

Jusczyk, P. W. On characterizing the development of speech perception. In J. Mehler and R. Fox (Eds.), *Neonate cognition: Beyond the blooming, buzzing confusion*. Hillsdale, N. J.: Erlbaum Associates, in press.

Liberman, A. M., Harris, K. S., Kinney, J. A. and Lane, H. The discrimination of relative-onset time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, 1961, *61*, 379-388.

Lisker, L. and Abramson, A. S. The voicing dimension: Some experiments in comparative phonetics. In Proceedings of the Sixth International Congress of Phonetic Scic..ces, Prague, 1967. Prague: Academia, 1970.

MacKain, K. S. Assessing the role of experience on infants' speech discrimination. Journal of Child Language, 1982, 9, 527-542.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J. and Fujimura, O. An effect of linguistic experience: The discrimination of /r/ and /l/ by native speakers of Japanese and English. Perception and Psychophysics, 1975, 18, 331-340.

Werker, J. F. and Tees, R. C. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behavior and Development, in press.

Williams, L. Voicing contrasts in Spanish. Journal of Phonetics, 1977, 5, 169-184.

II. SHORT REPORTS AND WORK-IN-PROGRESS

Contextual Variability and the Problem of Acoustic-Phonetic Invariance in Speech*

David B. Pisoni

Speech Research Laboratory
Department of Psycholcgy
Indiana University
Bloomington, Indiana 47405

*This is a draft of a paper to be presented at the Symposium on Invariance and Variability of Speech Processes, MIT, October 8-10, 1983. The work described in this paper was supported, in part, by NIH research grant NS-12179 and, in part, by contract F33615-83-K-0501 with the Air Force Office of Scientific Research, Air Force System Command. I thank H. C. Nusbaum, D. Kewley-Port and P. A. Luce for their comments and suggestions about the issues discussed in this paper.

Contextual Variability and the Problem of Acoustic-Phonetic Invariance in Speech

Most members of the audience are no doubt aware of the fact that the bulk of research on speech processes over the last thirty to thirty-five years has been concerned principally, if not almost exclusively, with features and phonemes in highly controlled experimental contexts using nonsense syllable test materials. I believe this observation applies equally well to research in both speech production and speech perception. Such a research strategy is not at all surprising when one begins to study complex phenomema such as speech and when one tries to understand the relations between spoken language and its acoustic-phonetic realization as speech. Researchers in any field of scientific investigation typically work on tractable problems that can be studied with existing methodology and experimental paradigms. However, relative to the voluminous literature on isolated phoneme perception, I think it is fair to say that very little is actually known at this time about how the acoustic-phonetic information in the speech signal is used by listeners in word recognition, sentence perception or understanding of fluent speech. Fortunately, the situation is changing rapidly as shown by the new research described in several of the papers presented at this symposium that have begun to examine in some detail the contribution that context and various knowledge sources make in speech perception.

In my view, there are several reasons for this change in research strategy and emphasis over the past few years. First, there are now many more researchers working on basic questions in speech than there were 10 or 15 years ago. Second, the cost of doing acoustic-phonetic research has dropped quite significantly with the wide availability of low-cost digital computers and signal processing techniques. The consequence of this is that many more major universities now have speech processing labs that are engaged in acoustic-phonetic research. Finally, with more interest in the field and more powerful tools available, many investigators have turned their research efforts to a much wider range of problems in speech. Moreover, a number of speech researchers have become more intimately concerned with issues in machine recognition, on the one hand, and problems associated with processing fluent continuous speech by humans on the other hand. Taken together, these developments have directed the focus of speech research efforts well beyond the domain of phoneme and feature perception in isolated environments to a number of other problems associated with understanding spoken language processing in more natural and realistic contexts. Much of this interest was no doubt motivated by the ARPA project which demonstrated, among other things, that a speech understanding system could be built (see Klatt, 1977).

Louis Pols' paper, "Variation and Interaction in Speech", which I have been asked to summarize and discuss here is a good example of this recent trend in research in acoustic-phonetics. While acknowledging the enormous amount of variability in speech and the contributions of multiple cues to perception of segmental phonemes, Pols nevertheless has set out to examine the contextual effects of sentence environments on phoneme perception using a deletion paradigm. In this procedure, parts of the speech signal are carefully removed via digital editing techniques and the effects of these manipulations on listeners' responses is observed. Instead of summarizing the details of Pols' experiments, which are described quite adequately in his paper, I will first focus my remarks on the major generalizations and implications of his findings for work in speech perception. Then, I will describe two recent studies from our own laboratory that bear very closely on the issues and problems raised in Pols'paper. These two studies deal with the role of context in speech processing and the search for acoustic-phonetic invariance.

Nonsense Syllables and Synthetic Speech

One of the issues that Pols raised in his paper was the almost exclusive reliance on the use of nonsense syllable stimuli in many perceptual experiments in speech. While much has been learned from these early studies about the minimal acoustic cues to phoneme perception (Liberman et al., 1967; Fant, 1973) and the interaction of these cues in certain well-defined although isolated environments (Liberman, 1982; Repp, 1982; Studdert-Kennedy, 1982), relatively little attention has been devoted in the acoustic-phonetic literature to word recognition, particularly word recognition in sentence contexts or to the processing of connected fluent speech (see however Cole & Jakimik, 1980; Marslen-Wilson & Welsh, 1978; Foss & Blank, 1980). Closely related to this point is the observation that much of what we currently know about speech cues comes from perceptual studies using highly simplified synthetic speech signals. There is reason to believe, especially from our own recent work at Indiana, that even high quality synthetic speech may be impoverished and less redundant than natural speech (Pisoni, 1982). Both criticisms raised by Pols are, in my view, quite valid and appropriate in this context when one looks over the acoustic-phonetic literature and examines the issues that a large number of researchers have focused their attention on over the years (e.g. Lane, 1965; Repp, 1983). In his recent investigations, Pols, as well as a number of other researchers, have begun to make deliberate and conscious efforts to move away from the exclusive use of synthetic speech stimuli and have focused their research efforts instead on natural speech. Likewise, recent efforts have been made by Pols and others to study the perception of meaningful stimuli such as words and sentence-length materials (rather than simple nonsense syllables) in more realistic or natural tasks that require recognition or identification of phonemes and words in context.

Moving to another issue, Pols argues that it is "unnatural" to ask naive listeners to make conscious and deliberate judgments and decisions about the acoustic-phonetic properties of nonsense syllables in tasks that require a speeded response or subjective evaluation. He appears to view this research with some skepticism. Of course, what is considered as natural to one researcher may be quite unnatural to another depending on one's goals and purpose in designing laboratory experiments. In modern psychological research, highly unnatural experimental tasks have often been used quite effectively to reveal important insights and to verify generalizations about underlying perceptual mechanisms and internal cognitive processes that are often obscured by more traditional procedures that may have strong face validity on intuitive grounds. What I think Pols is really getting at here by his remarks about naturalness is the issue of whether the findings on feature and phoneme perception using isolated nonsense syllables in controlled environments can be generalized above and beyond the specific experimental contexts and rigid procedures that are typically employed in these studies. This is also a legitimate and relevant concern to raise at this conference since, at least for a large number of studies, few, if any, efforts have been made to establish the relevance of these findings to some of the more general problems in the field, such as acoustic-phonetic invariance or variability of speech, topics that have been central to the field of modern speech research since its beginnings after W.W.II (Potter et al., 1946; Joos, 1948).

It seems to me that if we are ever to make substantial progress in solving these problems we must make a deliberate attempt to deal with these issues in our research and not get sidetracked by the intricate details of experimental paradigms and task-specific findings. It is very easy to do this and therefore avoid the problems you originally set out to study. (I know this quite well because I have been guilty of contributing research to this literature from time to time myself, as many of you know.) I think it is a reasonable conclusion to assert that many of these findings on feature and phoneme perception are not easily generalized to more general problems of acoustic-phonetic invariance or describing the many sources of variability and their interactions that are the hallmarks of speech. Moreover, few attempts are made to follow up on findings obtained in isolated contexts to determine if they generalize to other environments (see, however, Kewley-Port and Luce, 1983).

Context Effects in Speech

The role of context has been a major problem area not only in the field of speech but across almost all other domains of human and animal behavior. Pols' work in using sentence-level contexts to study word recognition and phoneme perception is clearly a step in the right direction, but much more work is needed to describe and account for the different sources of variability introduced by the large number of contextually relevant sources of information in speech. For example, we need to learn much more about the sources of variability and the contributions introduced by talker differences, variations in speaking rate, the local phonetic environment and the lingustic context including syntax, semantics and pragmatics. These are the major sources of variability known to exist in speech. Several years ago people thought quite optimistically that increased attention to prosody in speech research would somehow magically solve the problems of invariance and variability (Cohen and Nooteboom, 1975). Instead, this work has revealed even more sources of variability that need to be considered in dealing with these issues. The work on prosody in the last few years has been interesting and important in the overall effort to understand human speech processing but it has not proved to be the missing link that provides all the solutions to the old problems in speech. The problems are still with us today and have been since the earliest days of speech research.

Acknowledging that context plays an important role in speech processes unfortunately does not solve the problem of accounting for the types of modifications and conditioning that context produces in the acoustic-phonetic representation of the speech signal. What is needed now are more detailed and systematic studies that examine what these various sources of context do and how they operate in modulating the variability observed in the acoustic-phonetic structure of speech. In many cases, these seemingly troublesome sources of contextual variability are not something just added onto an "idealized" discrete segmental representation of speech as a linear sequence of phonemes. Instead, some context effects may be an inherent part of the signal itself, reflecting important properties of the speaker and the linguistic content of the message. We need further research on the role of context and its contribution to both speech production and speech perception.

The success of recent text-to-speech systems using phoneme-based synthesis techniques demonstrates to me that the sources of variability that need to be accounted for in speech are finite and knowable in principle, or nearly so, and that they play a central role in the generation of intelligible and natural sounding synthetic speech (Allen, 1981; Pisoni, 1982). Problems of detailing the differential contributions of context and associated variability observed in speech are difficult because they encompass a large number of variables operating over several domains of language. However, they are not beyond our reach given current methods of carrying out acoustic-phonetic research. With the recent efforts directed towards developing large data bases for measurement purposes and hypothesis testing (Crystal and House, 1982; Shipman, 1982), many of the seemingly difficult questions surrounding the operation of phonological rules or the effects of local phonetic contexts can be studied within a fairly manageable time frame. As many of you know, this was not always the case years ago when it took several minutes just to make a sound spectrogram of one utterance from a single talker. Because of advances in the available technology, some research questions in acoustic-phonetics that are trivial today would not even be conceived of only a few years ago. The ability to test hypotheses at a faster rate will allow researchers to make greater advances in accumulating knowledge about the acoustic-phonetics of fluent connected speech and the effect of different sources of contextual variability in speech (Shipman, 1983).

Effects of Speaking Rate on Phonetic Perception

In the last few years, a great deal of interest has been focused on the effects of speaking rate. The work of Port, Fitch and Miller among others who have studied the influence of speech tempo on the perception of phonetic segments is well-known to most of the audience (see Miller, 1981, for a review). This work has been interpreted as demonstrating a form of perceptual normalization whereby the listener "compensates" or "readjusts" his/her decision criteria in accordance with the systematic variability and changes introduced by different speaking rates. In one well-known perceptual study, Miller and Liberman (1979) reported that overall syllable duration influences the locations of the identification boundary between the stop /b/ and the semivowel /w/. More specifically, Miller and Liberman (1979) reported that the duration of the vowel in an isolated CV syllable systematically influenced the perception of the formant transitions cues for the stop-semivowel distinction. With short syllables, subjects required shorter transition durations to perceive a /w/ than with longer syllables. Miller and Liberman interpreted these results as a clear demonstration of perceptual normalization for speaking rate -- the listener adjusts his/her decision to compensate for the differences in vowel length that are conditioned by the talker's speaking rate. According to Miller and Liberman, the listener interprets a particular set of acoustic cues or attributes such as the duration of a transition for a /b/ or /w/ in relation to the talker's speaking rate rather than by reference to some absolute set of contextually invariant acoustic attributes in the signal itself. Although Miller and Liberman used isolated, synthetically produced nonsense syllables in a somewhat unnatural experimental setting, at least according to Pols' criteria, their findings were of some interest to us several years ago because of the strong claims Miller and Liberman made about the underlying perceptual mechanisms responsible for the observed compensation with these stimuli.

To evaluate their claims about rate normalization, we carried out several experiments comparing perception of speech and comparable nonspeech control signals (Carrell, Pisoni and Gans, 1980). The nonspeech control signals were created with sinewave analogs of the speech stimuli which preserved the durations and temporal relations, although these signals did not sound like speech to naive listeners. The results of these comparisons which are shown in Figure 1 demonstrated comparable context effects for the perception of the duration of a rapid spectrum change as a function of the overall duration of the stimulus for both the speech and nonspeech stimuli. Our findings therefore call into question the rate normalization account offered by Miller and Liberman by demonstrating clearly that context effects such as these are not peculiar to the perception of speech or to the normalization of differences in the talker's speaking rate. Rather, these types of context effects may simply reflect general psychophysical principles that affect the perceptual categorization and discrimination of all acoustic signals, whether speech or nonspeech (see Goldhor, 1983). Indeed, these context effects may not even be peculiar to the auditory modality but may reflect fairly general perceptual principles across all sensory modalities (e.g., Helson, 1964; Cutting, 1983).

Insert Figure 1 about here

<u>Invariance of the Consonant/Vowel Ratio</u>

Closely related to this work on speaking rate is a series of studies by Port on the consonant/vowel ratio, a presumed invariant cue to voicing of stops in syllable-final position (Port, 1981a,b; Port and Dalby, 1982). It has been known for some time that vowel duration and closure duration enter into a reciprocal relation in providing cues to the voicing feature of stops in this environment. Moreover, it has been shown that each of these cues is affected by changes in speaking rate (Fitch, 1981; Miller and Grosjean, 1981; Port, 1979). Recently, Port (1981a,b) has argued that vowel and closure duration should be considered together as a unitary property or attribute of the voicing feature that remains invariant over changes in speaking rate. Port claimed that the ratio of closure duration to vowel duration (e.g., the C/V ratio) should be considered as a "relational" cue to voicing that is contextually invariant. In a number of speech production studies, Port has reported that the C/V ratio appears to remain invariant despite variations in speaking rate, number of syllables in the test word and vowel tensity. Thus, although all three factors clearly affect the absolute durations of the vowel and closure intervals, the C/V ratios seem to remain constant due to the temporal compensation between the two cues.

In the studies on the effects of speech tempo reported by Port as well as the work of others, test words always appeared in fixed carrier sentences of the

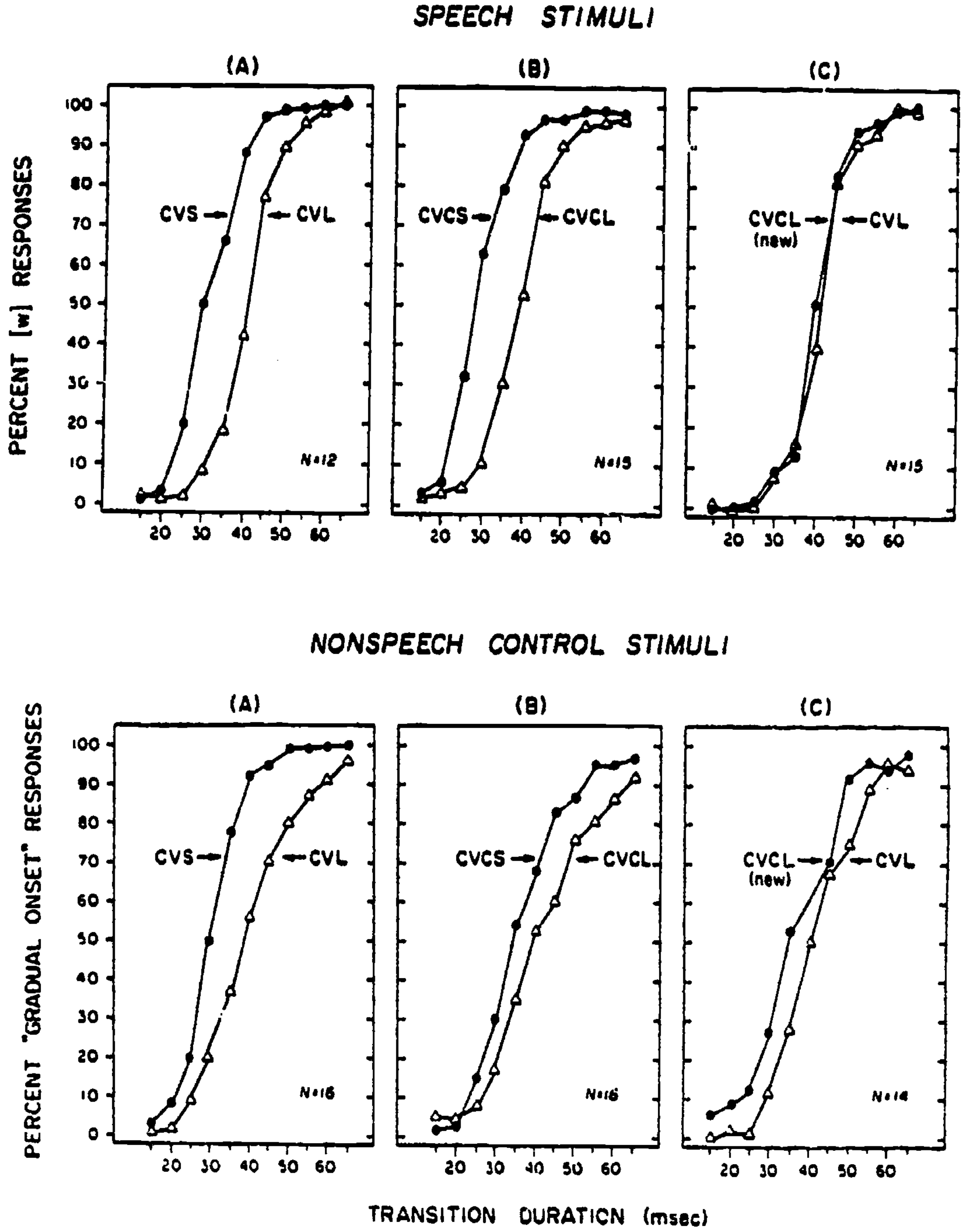


Figure 1. Labeling data for synthetic CV and CVC syllables (top panels) and nonspeech control signals (bottom panels). Panels A and B in the top display show percent of [w] responses as a function of transition duration in ms for CV and CVC syllables that differ in overall duration; the filled circles are the short (80 ms) stimuli, the open triangles are the long (295 ms) stimuli. Panel C shows data for CV and CVC syllables that were matched for overall duration. Panels A, B and C in the bottom display show percent of "gradual onset" responses for the matched nonspeech control signals. These functions are parallel to the speech labeling data shown in the top panel. (From Pisoni, Carrell and Gans, 1983).

form "He said the word ----- again." Recently, in our laboratory, Paul and Jan Luce (1983) carried out an important test of Port's invariance claims for the C/V ratio by examining the well-known effects of phrase-final lengthening on vowel and closure durations. In their production study, the Luce's were interested in determining if the C/V ratio would remain invariant from non-phrase-final to phrase-final sentence positions in which the vowel and closure durations are affected by local rate changes due to phrase final lengthening.

Insert Figure 2 about here

The C/V ratios from their study are shown in Figure 2 for CVC test words ending in bilabials. The results clearly demonstrate that the C/V ratios are larger for test words produced in non-phrase-final position compared to the same words produced in phrase-final position. However, the effect interacted with the voicing value of the final stop, the place of articulation of the stop and the immediately following local phonetic environment. Despite the contribution of these additional factors, the results demonstrate very clearly that the C/V ratio for syllable-final stops does not remain invariant across all environments.

I have described this study because I think it is a good example of how important it is to study the precise effects of known sources of variability in speech production and perception. The use of controlled carrier sentences in Port's earlier work is commendable since it avoids the biases introduced by reading lists of isolated words in citation form. However, the exclusive use of these "neutral" materials may be quite misleading when one proceeds to generalize the findings beyond the specific experimental contexts. What we need to do more of, as Luce and Luce have done so well in their experiment, is to study the contribution of various types of contexts and the specific form of the variability they introduce in the speech waveform. Much of the work on the effects of speaking rate has been done by asking talkers to consciously speed up or slow down their speaking rates when reading controlled experimental materials. The conditioning and variability of segmental durations observed under these explicit instructions to subjects may be quite different from the types of durational changes that occur more-or-less automatically as a consequence of naturally occurring sentence-level phenomena in speech. In short, we need more research on sentence-level effects in speech production and perception if we are ever to capture the regularities of speech beyond isolated nonsense contexts.

Conclusions

It is clear from Louis Pols' paper and the two recent studies that I summarized from our laboratory that the effects of context on variability in speech are enormous and quite diverse in scope. The problems associated with acoustic-phonetic invariance and variability in speech have not yet been solved, but I am encouraged by the change in direction and attitude of many researchers working in the field of acoustic phonetics. Research efforts appear to be directed toward much broader issues than just a few years ago, issues that involve the study of more meaningful linguistic stimuli in more naturalistic

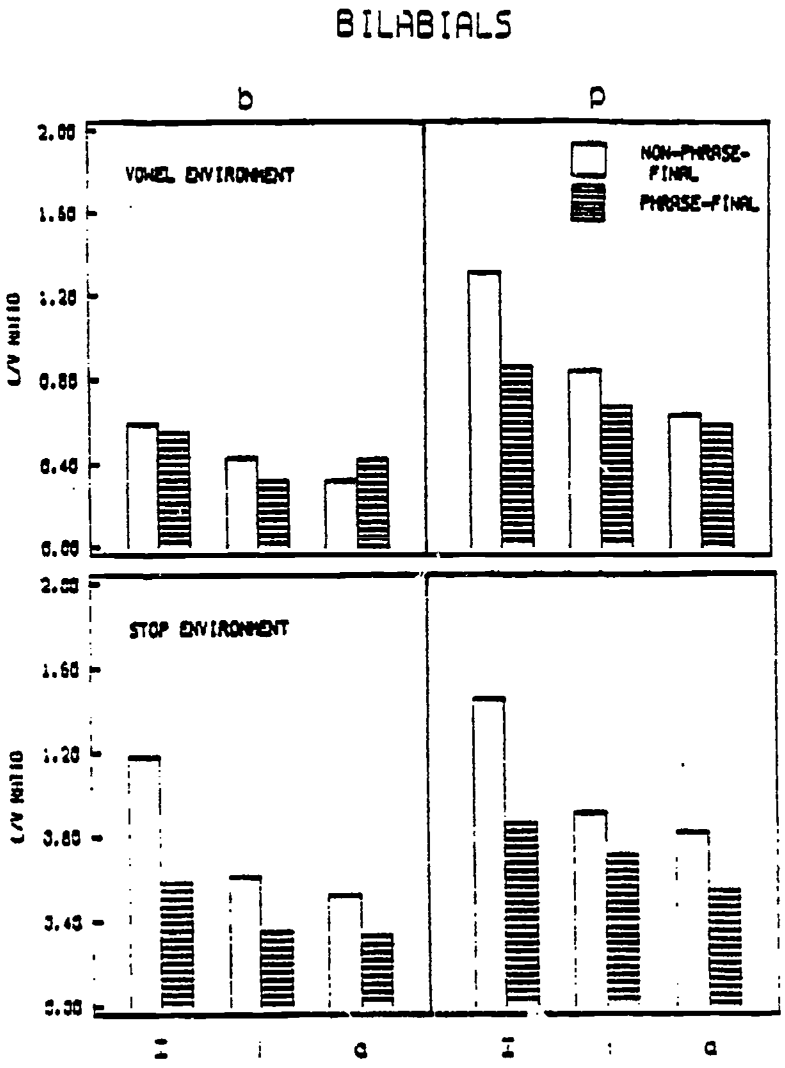


Figure 2. Mean C/V ratios for word-final bilabial stops. For each test word, vowel duration and closure duration for the final stop consonant were measured from a digital waveform display. Vowel duration was measured from the onset of periodic energy to a marked decrease in the periodic energy in the waveform. Closure duration was measured from this decrease in periodic energy to the onset of burst frication of the word-final stop. The upper panel shows the mean C/V ratios for test words ending in /b/ and /p/ produced in a vowel environment. The bottom panel shows the mean C/V ratios for test word ending in /b/ and /p/ produced in a stop enviroment. Mean C/V ratios are shown for each vowel /I/, /i/, /a/, separately. The open bars refer to the C/V ratios for test words produced in non-phrase-final position; the hatched bars refer to C/V ratios for test words produced in phrase-final position. Note the consistent difference in the C/V ratios across phrase-final and non-phrase-final environments (Adapted from Luce and Charles-Luce, 1983).

involve the study of more meaningful linguistic stimuli in more naturalistic tasks. These studies also employ experimental paradigms that require the listener's active deployment of phonological, lexical, syntactic and semantic knowledge in assigning an intepretation to the sensory input. In addition, there appears to be a more optimistic attitude about eventually understanding the role of context and the differential contribution of context to the acoustic-phonetic realization of the speech signal. Speech is a complex phenomenon and, as such, it seems to me to be very unlikely that we will find one simple unifying principle that will serve to rationalize and explain all of the different sources of variability observed across the articulatory, acoustic and perceptual domains.

References

Allen, J. Linguistic based algorithms offer practical text-to-speech systems. *Speech Technology*, 1981,*1*, 1, 12-16.

Carrell, T. D., Pisoni, D. B., and Gans, S. J. Perception of the duration of rapid spectrum changes: Evidence for context effects with speech and nonspeech signals. *Journal of the Acoustical Society of America*, 1980, *68*, S49.

Cohen, A., and Nooteboom, S. (Eds.) *Structure and process in speech perception*. Heidelberg: Springer-Verlag, 1975.

Cole, R. A., and Jakimik, J. A model of speech perception. In R.A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Erlbaum, 1980.

Crystal, T. H., and House, A. S. Segmental durations in connected speech signals: Preliminary results. *Journal of the Acoustical Society of America*, 1982, *72*, 705-716.

Cutting, J.E. Four assumptions about invariance in perception. *Journal of Experimental Psychology: Human Perception and Performance*, 1983, *9*, 2, 310-317.

Fant, G. *Speech Sounds and Features*. Cambridge, Mass.: The M.I.T. Press, 1973.

Fitch, H. L. Distinguishing temporal information for speaking rate from temporal information for intervocalic stop consonant voicing. *Haskins Laboratories Status Report in Speech Research*, SR-65, *1981*.

Foss, D. J., and Blank, M. A. Identifying the Speech Codes. *Cognitive Psychology*, 1980, *12*, 1-31.

Goldhor, R. The representation of speech in a model of the peripheral auditory system. *Journal of the Acoustical Society of America*, 1983, *73*, S4.

Helson, H. *Adaptation Level Theory*. New York: Harper & Row, 1964.

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

Kewley-Port, D., and Luce, P. A. Time-varying features of initial stop consonants in auditory running spectra: A first report. *Perception & Psychophysics*, 1984, *35*, 353-360.

Klatt, P. H. Review of the ARPA speech understanding project. *Journal of the Acoustical Society of America*, 1977, *62*, 1345-1366.

Lane, H. L. The motor theory of speech perception: A critical review. *Psychological Review*, 1965, *72*, 275-309.

Liberman, A. M. On finding that speech is special. American Psychologist, 1982, 37, 148-167.

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

Luce, P. A., and Charles-Luce, J. Temporal compensation and the consonant/vowel ratio: Contextual effects in speech perception. Research on Speech Perception: Progress Report No. 9.Bloomington, Indiana: Speech Research Laboratory, Indiana University, 1983.

Marslen-Wilson, W. D., and Welsh, A. Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology, 1978, 10, 29-63.

Miller, J. L. Effects of speaking rate on segmental distinctions. In P. D. Eimas and J. L. Miller (Eds.) Perspectives on the study of Speech. Hillsdale, NJ: Erlbaum, 1981.

Miller, J. L., and Grosjean, F. How the components of speaking rate influence perception of phonetic segments. Journal of Experimental Psychology: Human Perception and Performance, 1981, 1, 208-215.

Miller, J. L., and Liberman, A. M. Some effects of later-occurring information on the perception of stop consonants and semi-vowels. Perception & Pyschophysics, 1979, 25, 457-465.

Pisoni, D. B. Perception of speech: The human listener as a cognitive interface. Speech Technology, 1982, 1, 2, 10-23.

Pols, L, C. W. Variation and interaction in speech. Paper presented at the Symposium on Invariance and Variability of Speech Processes, MIT, October 8-10, 1983.

Port, R. F. Linguistic timing factors in combination. Journal of the Acoustical Society of America, 1981, 69, 262-274. (a)

Port, R. F. On the structure of the phonetic space with special reference to speech timing. Lingua, 1981, 55, 181-219. (b)

Port, R. F., and Dalby, J. Consonant/vowel ratio as a cue for voicing in English. Perception & Psychophysics, 1982, 32, 141-152.

Potter, R. K., Kopp, G. A., and Green, H. Visible Speech. New York: Van Nostrand, 1946.

Repp, B. H. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Bulletin, 1982, 92, 81-110.

Repp, B. H. Categorical perception: Issues, methods and findings. In N. J. Lass (Ed.), Speech and Language: Advances in Basic Research and Practice. New York: Academic Press, 1983.

Shipman, D. W. Development of speech research software on the MIT lisp machine. Journal of the Acoustical Society of America, 1982, 71, S103.

Shipman, D. W. SPIREX: Statistical analysis in the Spire acoustic-phonetic workstation. Proceedings of the IEEE ICASSP-83, 1983, 3, 1360-1363.

Studdert-Kennedy, M. On the dissociation of auditory and phonetic perception. In R. Carlson and B. Granstrom (Eds.), The Representation of Speech in the Peripheral Auditory System. New York: Elsevier, 1982.

Zue, V. W. Acoustic characteristics of stop consonants: A controlled study. Technical Report No. 523, Lincoln Laboratory, M.I.T., May 1976.

Converging Approaches Towards Establishing Invariant Acoustic Correlates of Stop Consonants*

Diane Kewley-Port

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

*This paper was presented at the Symposium on Invariance and Variablility of Speech Processes, MIT, October 8-10, 1983. This work was supported in part by NIH research grant NS-12179 to Indiana University. I thank D. B. Pisoni and M. Studdert-Kennedy for advice and comments on an earlier draft of this paper.

Converging Approaches Towards Establishing Invariant Acoustic Correlates of Stop Consonants

Many investigators are presently pursuing different approaches to the problem of determining invariant acoustic correlates of stop consonants. These approaches fall into three categories which are all represented at this conference. One is that of Blumstein and her colleagues who look at the problem of invariance from a formal linguistic approach. As Blumstein states in her introduction, this approach seeks to relate invariant acoustic properties directly to universal phonetic features as defined by linguists. Two other approaches are derived from an interest in modeling auditory processing of speech. Delgutte and Searle have been developing explicit physiological models of the auditory processing of speech sounds. Finally, Klatt (1982) and Kewley-Port (Kewley-Port, 1983; Kewley-Port and Luce, 1983) have based their work on psychophysical models of auditory processing. The inspiration behind much of this research has been Stevens' ongoing search for the invariant acoustic correlates of phonetic features (Stevens, 1967; 1975; 1980).

An examination of research from these three approaches suggests that they are converging towards a single description of the acoustic correlates for phonetic features associated with stop consonants. In this paper, two issues will be addressed. First, we will examine the similarities among these three approaches for specifying the acoustic correlates of the stop release burst, the onset of voicing and place of articulation. Second, we will examine Blumstein's claim that these acoustic correlates are, in fact, the invariant properties associated with universal phonetic features.

While a number of recent proposals have been made concerning the specification of the acoustic correlates of stops, the ones to be discussed here are found in the work of Blumstein (1983, Lahiri and Blumstein, 1981; Gewirth, Blumstein, Lahiri, and Kurowski, 1982; Mack and Blumstein, 1983), Delgutte (1981, 1983) and Kewley-Port (Kewley-Port and Luce, 1981; 1983). One important similarity between all three proposals is that the correlates of stop consonants are described by dynamic properties which are based on changes in energy or frequency over time. Naturally, the specific description of the correlates depends on the acoustic processing used: Delgutte based his work on the output of a physiological model, Kewley-Port used a psychophysical model, and Blumstein used a combination of hand-edited waveforms and LPC analysis. Besides proposing specific acoustic correlates, all three investigators experimentally tested whether the correlates were invariant over vowel context and, to a lesser extent, talkers. Other sources of acoustic-phonetic variation such as syllable position, speaking rate and language have not been examined in great detail by any investigator yet. It is clear, however, that all three approaches seek to define acoustic correlates of stops as invariant over several of the major sources of phonetic variation.

In order to discuss these proposals in detail, we will refer to some examples of stop-vowel syllables processed by Kewley-Port and Luce's psychoacoustic model (1981, 1983) shown on Fig. 1. Figure 1 displays the auditory running spectra of the syllables, /pi/, /da/ and /ku/. The spectral sections or frames, updated at 5 ms intervals, are 1/6 octave critical-band spectra displayed on a Mel scale. In this figure, the first frame was positioned during stop closure between 2 and 4 frames preceding the burst.

Insert Figure 1 about here

First, consider an acoustic correlate of the initial stop burst which is associated with the phonetic feature of manner of articulation (continuant versus non-continuant). Delgutte's correlate for detecting bursts is the presence of an "onset peak" in the discharge patterns of high-frequency channels. Kewley-Port's correlate is an abrupt increase in energy at high frequencies in the running spectra. Blumstein's correlate is an increase in relative level of energy over all frequencies. While the validity of these correlates was tested in different ways, all three were better than 90% successful in identifying the burst. Thus, all three approaches converge on the observation that an acoustic correlate of release bursts in stops is a detectable change in acoustic energy, a finding that was previously predicted by the acoustic theory of speech production. The primary difference in the approaches is that Blumstein examined energy changes over all frequencies while the other two approaches examined only high frequency energy. This difference may reflect differences in the experimental tasks. Dulgette and Kewley-Port examined both voiced and voiceless stops, while Blumstein examined only voiced stops contrasted with glides.

Both Delgutte and Kewley-Port have proposed an acoustic correlate for the onset of voicing in stops. These correlates were defined analogously to those for burst onset, except low frequencies were examined instead of high frequencies. Delgutte's results appear quite successful and superior to those obtained from Kewley-Port's psychophysical approach. Nonetheless, there is agreement that the acoustic correlate of voicing onset is an abrupt change in low frequency energy in the region of the first formant. While neither of the proposed correlates for the onset of the burst and the onset of voicing are surprising in light of previous research, the importance of the present work is that specific mechanisms for detecting these correlates in the human auditory system have now been experimentally tested.

Blumstein [in collaboration with Stevens (Stevens and Blumstein, 1978; 1981) and other colleagues] and Kewley-Port have been examining acoustic correlates for place of articulation for several years. In their recent formulation of these correlates, many similarities can be observed. First, both approaches define these correlates based on dynamic spectral properties, which for Blumstein is a change from her earlier static approach (Stevens and Blumstein, 1978; but also see Ohde and Stevens, 1983). Secondly, both rely on prior detection of the onset of both the burst and voicing to define a temporal course over which the correlates of place of articulation are assumed to occur. Delgutte is in agreement with this proposal, although he has not formally examined correlates of place of articulation.

In defining the details of the correlates of place of articulation, however, Blumstein's and Kewley-Port's approaches differ in several respects. Consider first the acoustic correlates for distinguishing labial versus alveolar place of articulation. Blumstein proposes that the tilt of all voiceless spectra compared

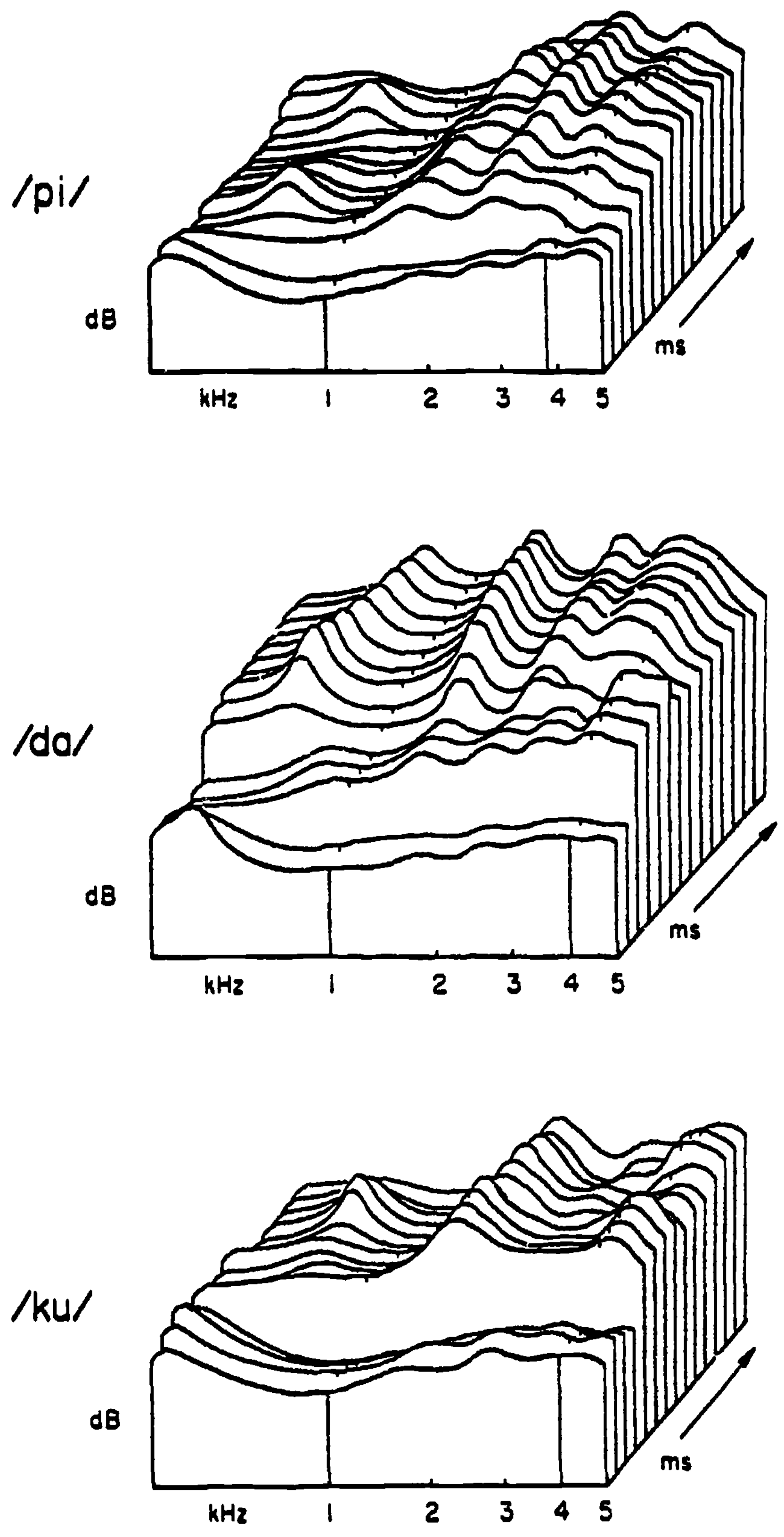


Figure 1. Critical-band running spectral displays for the syllables /pi/, /da/ and /ku/. Frequency in kHz is represented on the x-axis; relative dB is represented on the y-axis. The spectral frames are offset along the z-axis in 5 ms intervals.

to the tilt of the spectra of the first three pitch pulses is the basis of the correlates. Kewley-Port proposes that a difference in spectral tilt of up to 40 ms of voiceless frames (or 10 ms of voiced frames when the burst is voiced) is the basis of this distinction. Thus, Blumstein's approach requires comparison of spectra over a long temporal interval for the aspirated stops of English. That is, a typical stop with 50 ms VOT for a male talker would require about 80 ms of time for analysis (including the 3 pitch pulses). Although Blumstein has experimentally validated this correlate for aspirated English stops, it does not appear to be a good candidate for an invariant acoustic cue to place of articulation in speech perception. The results of speech perception studies have shown that place of articulation can be accurately identified within the first 20 to 40 ms after burst release (Ohde and Sharf, 1977; Tekieli and Cullinan, 1979; Blumstein and Stevens, 1980; Kewley-Port, 1980, Exp. 3B; Kewley-Port, Pisoni, and Studdert-Kennedy, 1983). Thus, Blumstein's correlate for long VOT stops requires analysis time intervals much longer than the 20 to 40 ms needed in perception. On the other hand, Kewley-Port's proposal for analyzing only the first 40 ms of voiceless spectra was made in light of these perceptual studies. Overall, however, the similarities between the two positions are more notable than their differences.

Regarding the velar place of articulation, Blumstein's recent papers have not presented a formal proposal for a velar acoustic correlate. Kewley-Port's proposed velar correlate is the presence of a prominent mid-frequency peak extending over at least 15 ms of the voiceless spectra. While Blumstein's discussion of compact spectral correlates for palatals and velars seems in accord with Kewley-Port's proposal, more research will be needed in this area.

Thus, it is obvious that a number of different converging proposals for defining acoustic correlates of stop consonants have developed from different research perspectives. Furthermore, the three approaches discussed above have specifically tested the extent to which these correlates are invariant over several sources of phonetic variation. There is, however, an important divergence between the claims of invariance made by our approach based on an auditory processing model and those of Blumstein based on a somewhat more formal linguistic approach. Specifically, Blumstein (1983) states that invariant acoustic correlates should "remain invariant across speakers, phonetic context, and languages." However, at least in our view, experimental research has not substantiated this claim. For example, consider the claim of invariance over speakers. Of the research reviewed here, only Kewley-Port's has seriously addressed the issue of defining high-, low- or mid-frequency ranges for talkers with large differences in vocal tract size (see Kewley-Port, 1983; Kewley-Port and Luce, 1983). While our proposed vocal tract normalization rule to locate a speaker's "mid-frequency" range appears to be a step in the right direction, many more speakers need to be studied. Blumstein's recent research on place of articulation has not specifically examined this problem since she has used only male speakers and did not examine velar correlates.

Perhaps more important from a linguistic point of view, recent work on invariants for stop consonants has not examined their validity over changes in syllable position or consonant manner class. The earlier studies of Blumstein and Stevens (1979) were, for the most part, unsuccessful in experimentally establishing invariant acoustic correlates for both syllable-initial and syllable-final stops (cf. Kewley-Port, 1983). In fact, it has been our view that it is quite unlikely that acoustic correlates for identifying place and manner of articulation will be physically the same for both initial and final stops. While the three proposals discussed above are very similar, they all depended on detecting the initial burst and voicing prior to identifying place of articulation. Since most syllable-final stops do not have bursts, it is obvious that the putative acoustic correlates simply cannot apply. Clearly, invariance over syllable position is essential to the linguist's definition of invariance over pnonetic context. Thus, these acoustic correlates, or even similar ones, cannot be invariant over phonetic context at least with the current definitions of invariance (see Port, 1983).

In conclusion, it appears that Blumstein's contention that there currently exits a theory of acoustic invariance which unites principles of phonology and models of speech processing is premature. On the other hand, models of auditory processing of speech in the nervous system are helping us to discover acoustic correlates which appear to be invariant over several sources of phonetic variation. This is obviously a step in the right direction (see Carlson and Granstrom, 1982). The converging knowledge and methodology from linguistics, hearing and speech sciences, and signal processing is enabling us to uncover invariant acoustic correlates and to further our understanding of human speech perception well beyond where we were just a few years ago.

REFERENCES

Blumstein, S. E. On acoustic invariance in speech. Paper presented at the Symposium on Invariance & Variability, MIT, 1983.

Blumstein, S. E., and Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. The Journal of the Acoustical Society of America, 66, 1001-1017.

Blumstein, S. E., and Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. The Journal of the Acoustical Society of America, 67 648-662.

Carlson, R. and Granstrom, B. (1982). The Representation of Speech in the Peripheral Auditory System. New York: Elsevier Biomedical Press.

Delgutte, B. (1981). Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. Unpublished doctoral dissertation, M.I.T, Cambridge, MA.

Delgutte, R. Analysis of French stop consonants with a model of the peripheral auditory system. Paper presented at the Symposium on Invariance & Variability, MIT, 1983.

Gewirth, L., Blumstein, S. E. Lahiri, A. and Kurowski, K. (1982). Perceptual and acoustic invariance for place of articulation in diffuse stop consonants. The Journal of the Acoustical Society of America, 72 (Supple. 1), S16.

Lahiri, A. and Blumstein, S. E. (1981). A reconsideration of acoustic invariance for place of articulation in stop consonants: Evidence from cross-language studies. The Journal of the Acoustical Society of America, 70 (Supple. 1), S39.

Mack, M. and Blumstein, S. E. (1983). Further evidence of acoustic invariance in speech production: The stop-glide contrast. The Journal of the Acoustical Society of America, 73, 1739-1750.

Kewley-Port, D. (1980). Representations of spectral change as cues to place of articulation in stop consonants. (Res. Speech Percept. Tech. Rep. No. 3). Bloomington: Indiana University, Department of Psychology.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. The Journal of the Acoustical Society of America, 73, 322-335.

Kewley-Port, D. and Luce, P. (1981). Time-varying features in voiced and voiceless stops produced at different speaking rates. (Res. Speech Percept. Prog. Rep. No. 7, pp. 197-214). Bloomington: Indiana University, Department of Psychology.

Kewley-Port, D. and Luce, P. (1983). Time-varying features of initial consonants in auditory running spectra: A first report. Perception and Psychophysics, 35, 353-360.

Kewley-Port, D., Pisoni, D. B. and Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. The Journal of the Acoustical Society of America, 73, 1779-1793.

Klatt, D. H. (1982). Speech processing strategies based on auditory models. In R. Carlson and B. Granstrom (Eds.), The Representation of Speech in the Peripheral Auditory System. New York: Elsevier Biomedical Press.

Ohde, R. N. and Sharf, D. J. (1977). Order effect of acoustic segments of VC and CV syllables on stop and vowel identification. Journal of Speech & Hearing Research, 20, 543-554.

Port, R. F. Phonetics as a signalling space. Paper presented at the Sympossium on Invariance & Variability, MIT, 1983.

Ohde, R. N. and Stevens, K. N. (1983). Effect of burst amplitude on the perception of stop consonant place of articulation. The Journal of the Acoustical Society of America, 74, 706-725.

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

Stevens, K. N. (1975). The potential role of property detectors in the perception of consonants. In G. Fant & M. A. A. Tatham (Eds.), Auditory analysis and perception of speech, (pp. 303-330). New York: Academic Press.

Stevens, K. N. (1980). Acoustic correlates of some phonetic categories. The Journal of the Acoustical Society of America, 68, 836-842.

Stevens, K. N. and Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. The Journal of the Acoustical Society of America, 64, 1358-1368.

Stevens, K. N. and Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas and J. Miller (Eds.), Perspectives on the Study of Speech, (pp. 1-38), Hillsdale, NJ: Erlbaum.

Tekieli, M. E. and Cullinan, W. L. (1979). The perception of temporally segmented vowels and consonant-vowel syllables. Journal of Speech & Hearing Research, 22, 103-121.

Identification of Speech Spectrograms:

Comparisons of Naive and Trained Observers*

Beth G. Greene, David B. Pisoni and Thomas D. Carrell

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

*This research was supported, in part, by BSRG S07 RR07031 and National Science Foundation grant NSF BNS 80-23099 to Indiana University, Bloomington. Preparation of this manuscript was supported by National Science Foundation grant NSF BNS 83-05387 and NINCDS grant NS 12179-07. An earlier version of these findings was presented at the 105th meeting of the Acoustical Society of America, Cincinnati, May 1983. We thank Jan Charles-Luce for her assistance in detailed data analysis.

Abstract

Two groups of subjects were presented with spectrograms of 50 words they had never seen before and were asked to provide a single monosyllabic English word for each display. One group had previously learned to identify a limited set of speech spectrograms after 15 hours of training using a study-test procedure which stressed holistic word identification. A subgroup of participants in a formal course on spectrogram reading at MIT served as a second group of subjects. These subjects learned specific acoustic-phonetic principles and knowledge-based strategies for interpreting spectrograms. Subjects in the first group correctly identified 30% of the possible phonetic segments in novel spectrograms. The second group of subjects correctly identified 40% of the possible segments in the same set of spectrograms given to the first group. When the data were scored for correct manner class of consonants only, the two groups did not differ significantly. Detailed descriptions of the identification results are presented. Implications of these findings for developing visual aids for hearing impaired persons and improved phonetic recognition strategies are also discussed.

Identification of Speech Spectrograms:

Comparisons of Naive and Trained Observers

In this paper we report the results of comparisons between two groups of subjects, each of which was trained to interpret speech spectrograms. One group, the IU subjects, were taught to recognize 50 spectrograms absolutely in a word recognition task. In previous papers we have reported that these subjects were able to generalize their knowledge from one talker to several other talkers using a limited set of words (Greene, Pisoni & Carrell, 1982, 1984). They were also able to generalize to a set of 50 spectrograms they had never seen before -- what we called the Novel word list. Since our earlier reports we had a unique opportunity to make several comparisons between these subjects and a group of individuals taught to interpret spectrograms at MIT by Victor Zue and his associates.

We would like to publicly acknowledge the full cooperation of Victor Zue and extend our appreciation to Victor and his colleagues at the MIT Speech Spectrogram Reading course for their help and cooperation. We would also like to publicly thank the course participants who volunteered to read the spectrograms that we will be discussing in this paper.

The Groups and the Training Procedures

Insert Figure 1 about here

The IU subjects came to the laboratory for one hour each day over a five week period. A study-test procedure was used for training. Each spectrogram represented a single isolated monosyllabic word spoken by a male talker. Spectrograms of 50 words were used during the first phase of this study. In each session, subjects learned to identify four spectrograms -- the "study" phase -- immediately followed by a "test" phase on all the spectrograms previously learned plus the newly learned set. Subjects were given only one rule to guide their perceptual learning: "Words that sound alike will look alike in spectrograms." Basically, these subjects memorized the set of spectrograms and the corresponding words for each display in a paired-associates format. Subjects consistently scored above 95% correct on all tests with the original 50 training words.

The MIT subjects participated in a formal one-week course on spectrogram reading taught by Victor Zue at MIT. Each day several hours were devoted to lectures on specific aspects of speech analysis, phonetics, acoustics, phonology, etc. Additional time was spent examining and analyzing spectrograms of connected speech spoken by one male talker. Members of the class were shown how to do segmentation and labeling on hard copy versions of spectrograms. Extensive discussions of how to do phonetic labeling were also conducted. Students learned to look for salient elements or attributes in the spectrograms and to label them

THE GROUPS

IU	MIT
--Study-Test Procedure	--Course on Speech Spectrogram Reading
--Holistic Word Recognition	--Principles of Acoustic-Phonetic Analysis
--1 Hour per Day 4 Weeks = 20 Hours	--6 Hours per Day 1 Week = 27 Hours
--Isolated Words	--Connected Speech
--Single Talker	--Single Talker

Figure 1. Training procedures used for each group.

consistently. The students did this by applying explicit acoustic-phonetic principles learned through course readings, lecture material and practice on previous spectrogram labeling exercises (see Cole, Rudnicky, Zue & Reddy, 1980; Potter, Kopp & Green, 1947; Zue, 1981).

When we refer to the MIT subjects as "trained observers," we mean trained to interpret spectrograms via explicit acoustic-phonetic principles. The IU subjects recognized spectrograms holistically and responded with an English word as the label. Thus, NAIVE in this context means untrained in acoustic-phonetic principles of segmentation and labeling; TRAINED means explicitly trained in acoustic-phonetic knowledge.

The Test and How Administered

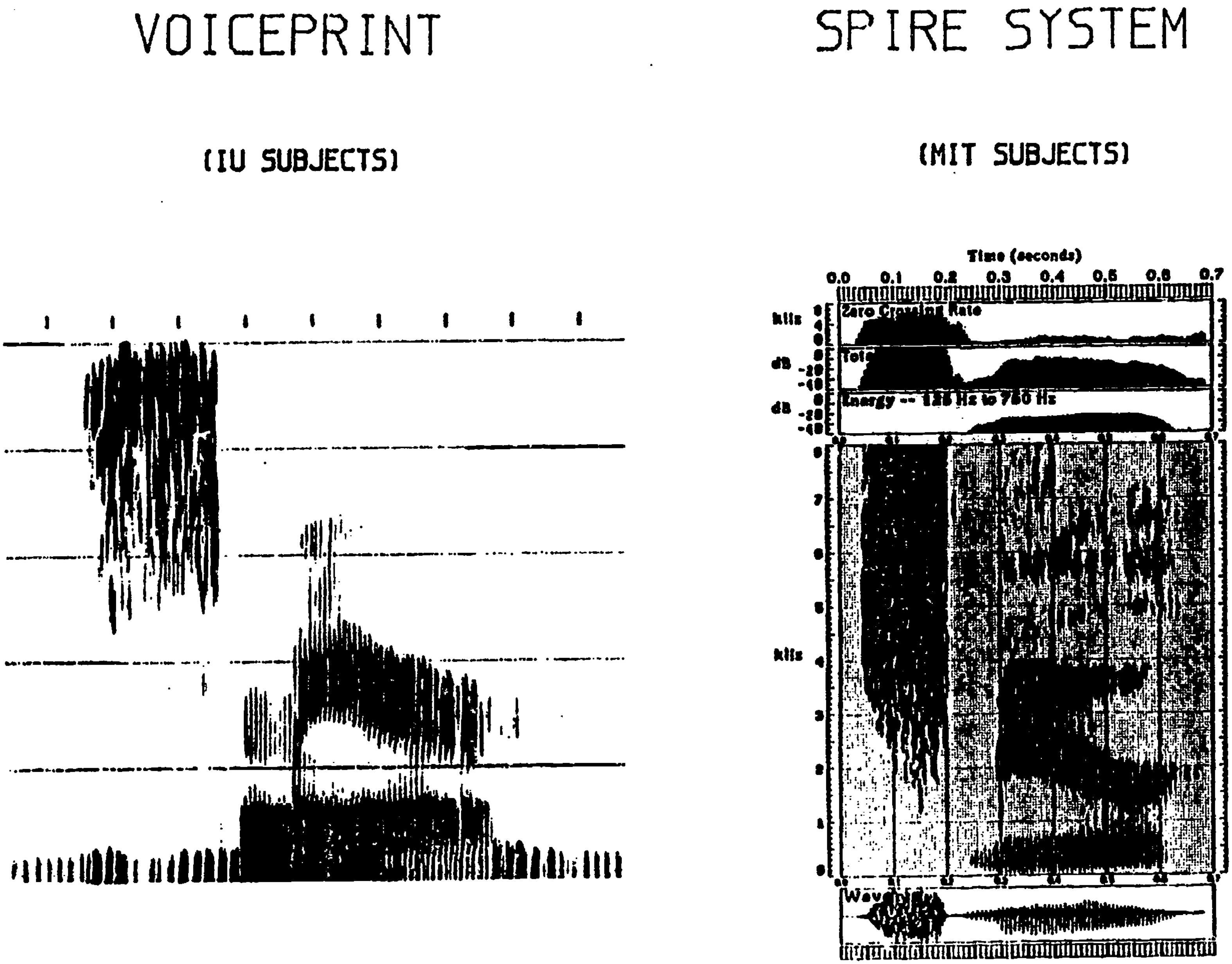
All the subjects were required to respond with a single real English word for each of 50 novel spectrograms. The 50 items were taken from a phonetically balanced (PB) list spoken by a male talker (Egan, 1948). The IU subjects saw the spectrograms on a 9" CRT screen. The output of a tape recorder was input to a Spectraphonics Speech Spectrographic Display (SSD) device (Stewart, Houde & Larkin, 1976) which was then interfaced to 6 CRT monitors. Subjects examined the visual display for approximately 25 seconds and recorded their responses on prepared response sheets. For each spectrogram, subjects were required to provide a single real English word and were told to guess if necessary.

The MIT subjects examined the same set of 50 novel spectrograms presented in a booklet. Each spectrogram was reproduced on the SPIRE system at MIT and xerox copies were made for each subject. Subjects were told to examine each spectrogram for 30 seconds to one minute and then write a single English word using standard orthography directly on the spectrogram.

Previously, these subjects had not been placed under time constraints for reading spectrograms. Several people reported that they needed to spend considerably more time analyzing each spectrogram in order to generate a single word as the response. In summary, the IU subjects provided responses to displays produced using the SSD whilc the MIT subjects responded to SPIRE hardcopy displays.

Insert Figure 2 about here

The left hand display on this figure is a 5 kHz Voiceprint spectrogram -- a display that is very close to the version shown on the SSD. The spectrogram produced by the SPIRE system is shown in the right hand panel. As shown in the figure, the SPIRE version displays 8 kHz and provides other information about the stimulus: the zero crossing rate, total energy spectrum, low frequency energy spectrum are shown above the spectrogram and the acoustic waveform is shown below the spectrogram. For our experiment, the SPIRE spectrograms displayed only 5 kHz of speech as in the SSD display we used with the IU subjects.



VOICEPRINT

(IU SUBJECTS)

SPIRE SYSTEM

(MIT SUBJECTS)

Figure 2. Examples of the 5 kHz Voiceprint and the 8 kHz Spire system displays. The word displayed is "smear."

Results and Discussion

Overall Comparisons of Exact Phoneme Correct. All subjects' responses were converted to broad phonetic transcription by a graduate student in linguistics and one of the experimenters (BGG). Discrepancies were resolved by mutual agreement. When the results were examined for exact phoneme correct identification summed across all subjects, vowels and consonants, we found that the IU subjects identified 30% of the phonetic segments correctly whereas the MIT subjects identified 40% of the segments correctly.

Insert Figure 3 about here

The overall percentage of correct phoneme identification is shown at the extreme left of the figure. The open bars represent the IU subjects; the filled bars represent the MIT subjects. As can be seen in the figure, the MIT subjects displayed consistently higher levels of exact phoneme correct responses than the IU subjects ($p < .005$). Moving from left-to-right across the figure, the MIT subjects correctly identified more phonemes overall (30% vs 40%), more vowels (23% vs 30%), more consonants (34% vs 40%) in initial (37% vs 41%) and final (32% vs 39%) position, and more consonants in clusters (28% vs 60%) in initial (27% vs 58%) and final (28% vs 61%) position. The two groups did not differ significantly on vowel identification or on identification of singleton consonants. The differences between the groups on identification of consonants in clusters was highly significant ($p < .0002$ for the second consonant in initial clusters, $p < .002$ for the first consonant in final clusters).

The most striking difference in performance between the two groups is the significant difference in identification of the phonemes in consonant clusters. The MIT subjects showed their best performance on these items correctly identifying the second consonant of an initial cluster 58% of the time and the first consonant of a final cluster 61% of the time. This difference is striking in our data but not too surprising. Consonant clusters can be recognized by durational cues -- the cluster is longer than a singleton consonant but each member of the cluster is shorter than it would be if it appeared by itself. When a spectrogram reader segments an utterance, common clusters such as initial /st-/ and final /-st/, initial stop/liquid combinations, etc. are often noted during the labeling process. The IU subjects were not told anything about two consonants together nor were they shown specific examples of these. Whatever clusters they identified, they probably recognized them as a unit (i.e., /-st/) from the original spectrograms they learned. It is clear that the MIT subjects simply knew more about the acoustic-phonetics of English and the phonotactics of the language.

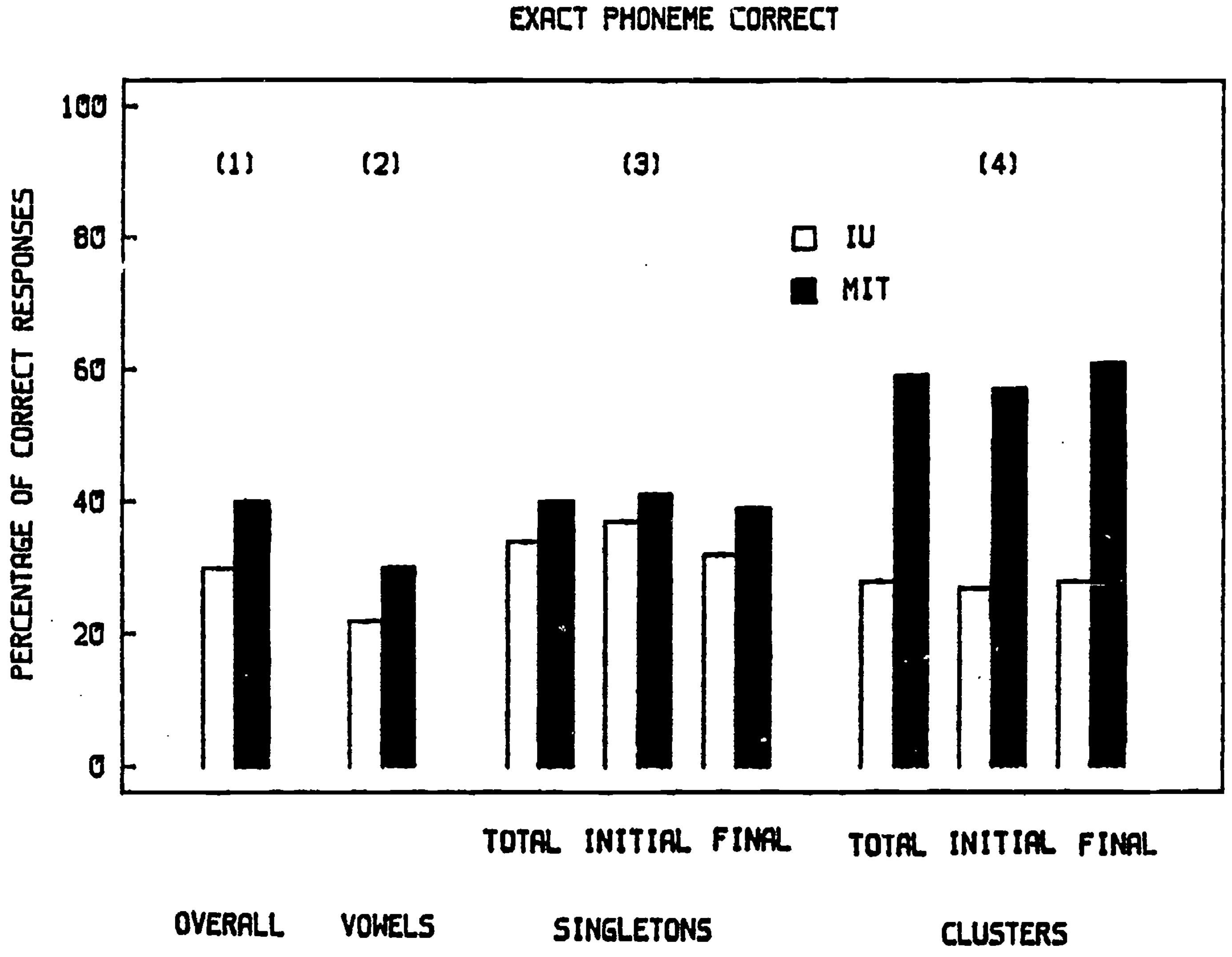


Figure 3. Overall percentage of correct phoneme identification responses.

Overall Comparisons of Manner Class Correct. In our earlier report we were less interested in examining the exact phonemes subjects identified correctly and more interested in how close they could come to the correct phoneme. Thus, we scored the data in a fairly liberal way -- manner class correct identification. If a subject wrote one of the stop consonants for a given stop consonant, it was scored as a correct manner class response. Applying this liberal scoring procedure to the present data produced an overall percentage of correct responses of 66% correct consonant manner class identification for both the IU and MIT subjects. Thus, the IU subjects and the MIT subjects showed comparable levels of manner class identification suggesting that a large number of gross features or attributes can be extracted from the visual display regardless of the kind of training provided. However, to further subdivide the consonant class into its specific members, more detailed analytic skill and knowledge is needed.

Exact Correct by Manner Class. We have examined the consonant data in greater detail to evaluate the saliency of different manner classes.

Insert Figure 4 about here

Figure 4 shows the exact correct phoneme identification as a function of manner class. The open bars represent the IU subjects; the filled bars represent the MIT subjects. The number of occurrences of phonemes in each class is written below each bar graph. MIT subjects showed superior performance for stops, fricatives, liquids and semivowels while the IU subjects performed better on the nasal consonants. Percentages for the IU subjects ranged from a low of 23% for the semivowels to a high of 48% for the liquids. MIT subjects ranged from a low of 30% for the nasals to a high of 52% for the liquids.

Manner Class Correct by Manner Class. We also scored the consonant manner class data for correct manner c ass identification rather than exact phoneme correct identification. This more liberal scoring procedure should capture the gross features for each of the respective manner categories.

Insert Figure 5 about here

Both groups of subjects showed higher percentages of correct responses for stops, fricatives, liquids and nasals. MIT subjects showed an increase on semivowels; performance of IU subjects did not change. In short, when examining spectrograms of isolated unknown words, subjects could identify the correct manner class over 60% of the time regardless of the ways in which they were

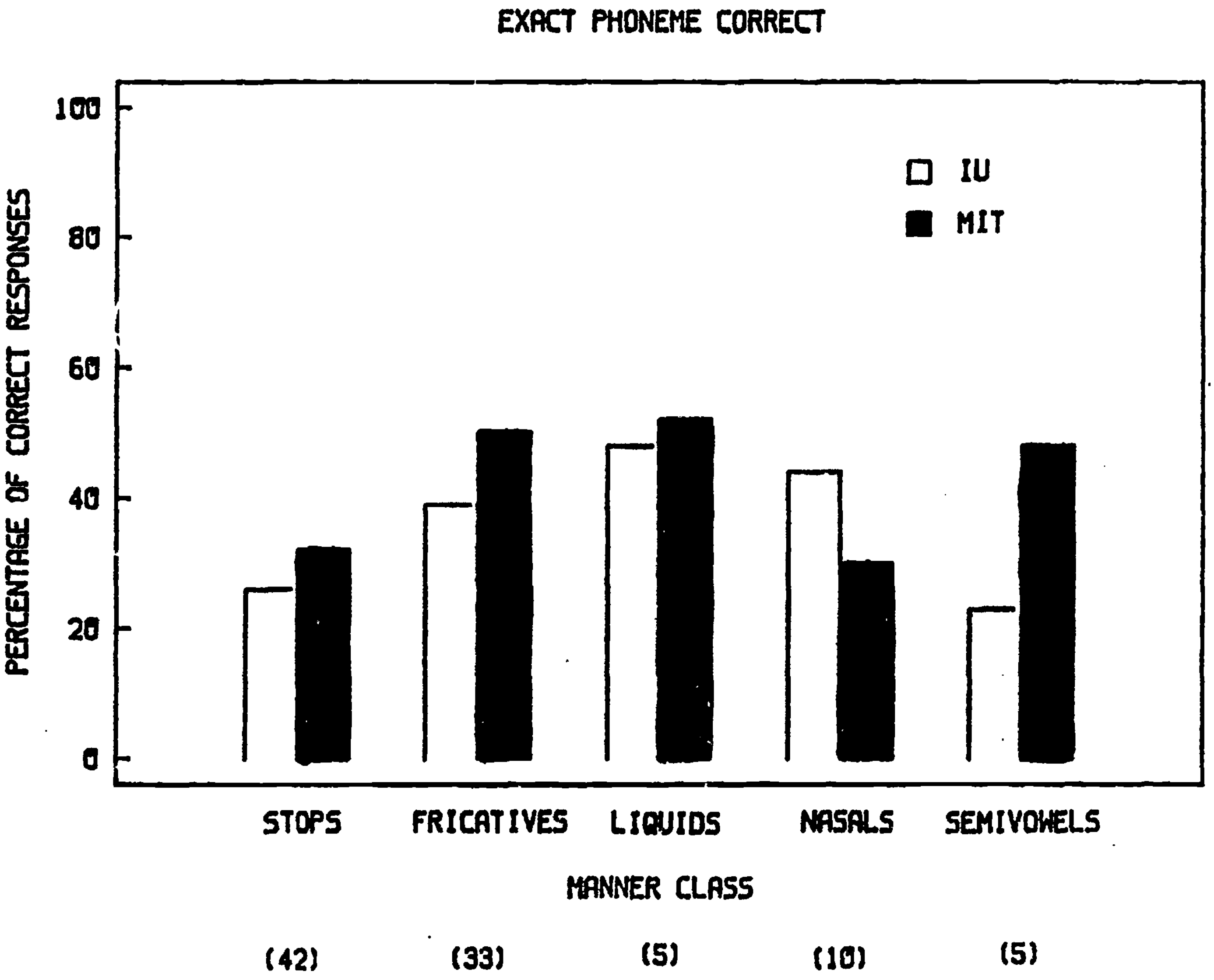


Figure 4. Percentage of correct phoneme identification responc..s as a function of manner class. The number of occurrences of phonemes in each class is shown below each bar graph.

277

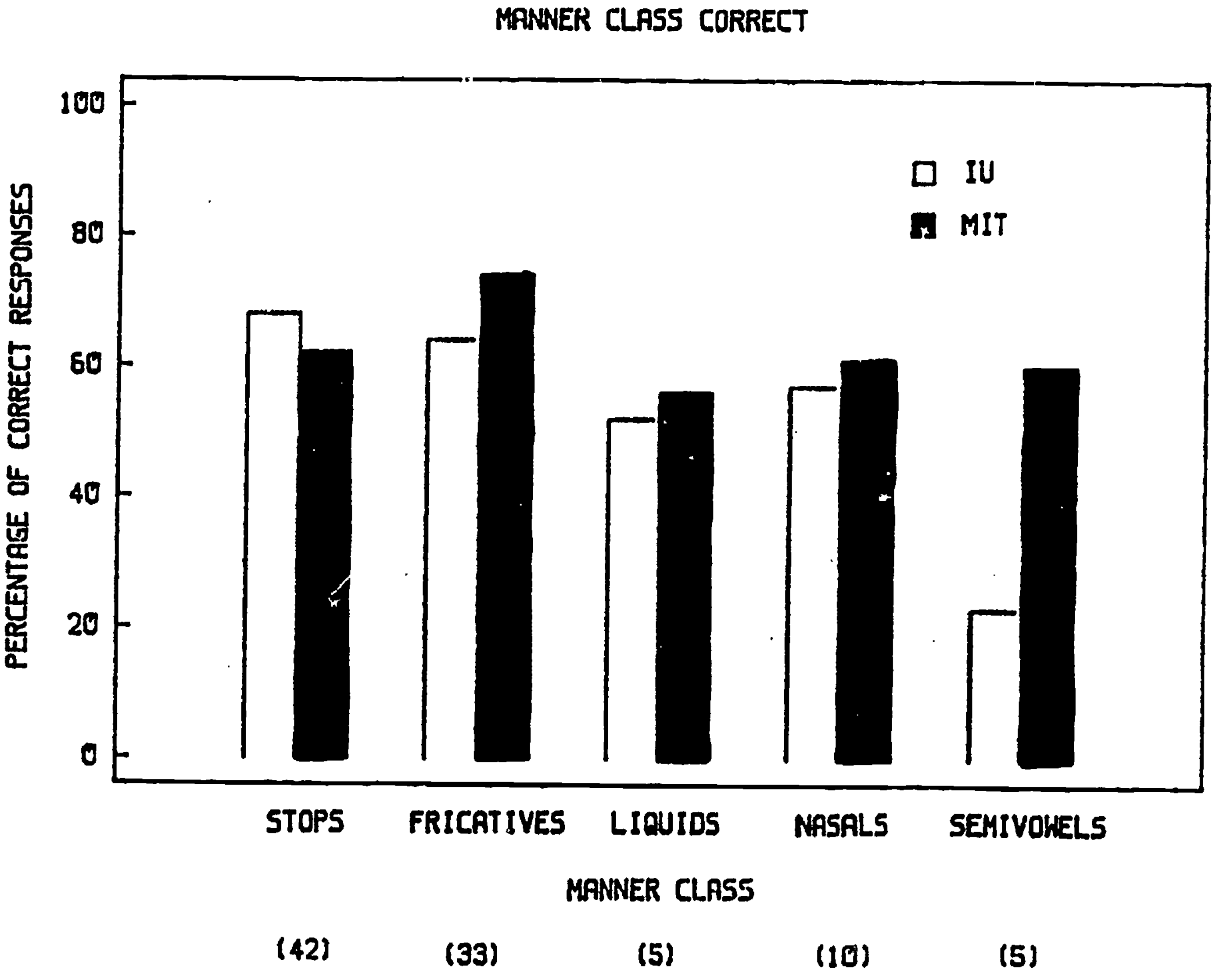


Figure 5. Percentage of correct manner class identification responses as a function of manner class. The number of occurrences of phonemes in each class is shown below each bar graph.

trained. The MIT subjects, trained using principles of acoustic-phonetic analysis, were superior to the IU subjects on exact phoneme correct identification. The IU subjects had no training on acoustic-phonetic analysis but they were still able to identify manner class correctly at levels of performance that were well above chance expectation.

Limitations

There are a number of important procedural differences between the study conducted with the IU subjects and the study conducted with MIT subjects. We would like to point these out here. Differences in the spectrographic displays, the training procedures and the subjects clearly limit the generality of these results.

Insert Figure 6 about here

First, the spectrograms used during training had different bandwidths and each was produced using different analysis equipment. While the two types of spectrographic displays are very similar, both qualitative and quantitative differences remain. Second, the procedures used to train each group were quite different. The IU subjects learned a limited set of spectrograms, each of which represented a single isolated word. The rate and extent of learning was under experimenter control. The goal of the IU training project was to learn to identify 50 spectrograms absolutely.

The MIT subjects learned explicit rules and analysis strategies to first segment and then label unknown samples of connected speech. These subjects were encouraged to provide multiple labels and eliminate poor candidates based their knowledge of English phonotactics and the effects of phonetic context. In contrast, the IU subjects were trained to recognize each spectrogram as a whole word. The procedures used in this experiment do not provide us with detailed information about the techniques or perceptual strategies employed by these subjects in identifying these novel spectrograms.

Finally, the IU subjects were paid an hourly wage to work in the lab^ratory. Participation in the experiment was their undergraduate work assignment for the semester. The MIT subjects spent several hours of their own time reading the spectrograms for this experiment. They all volunteered to do so in response to our request at the MIT summer course. Each MIT subject was professionally involved with speech research in academi?, business, industrial or R & D settings. The IU subjects were all undergraduate students at a large midwestern university.

		IU (N = 8)	MIT (N = 10)
1.	DISPLAY:	5 kHz	8 kHz
		SSD Device	SPIRE System
2.	TRAINING:	Study-Test Procedure	Acoustic-Phonetic Analysis
		(Implicit Training)	(Explicit Training)
		Holistic Word Recognition	A-P Analysis, Segmentation, Labeling
		Isolated Words	Connected Speech
		1 Hour per Day for 4 Weeks	6 Hours per Day for 1 Week
3.	Subjects:	Paid	Volunteers
		Students	Professionals in Speech Research

Figure 6. Some differences between the groups.

Summary and Conclusions

In summary, the results of this study demonstrate that both naive and trained observers can identify unknown spectrograms of isolated English words. Subjects trained in explicit acoustic-phonetic analysis techniques reported that the task was difficult because they usually worked with connected speech and were rarely, if ever, called upon to make an explicit response at the "word" level. On the other hand, the IU subjects were required to do something well beyond what they had been taught to do.

It is clear from these comparisons that salient and reliable cues are available in speech spectrograms and these cues can be used to identify unknown spectrograms regardless of the display or the training procedures used. As Victor Zue and Ron Cole have already noted in several papers and talks, the study of how human observers analyze visual display- of speech can provide valuable insights for development of algorithms for automatic recognition of speech by machine (Cole, Rudnicky & Zue, 1979; Zue, 1981; Zue & Cole, 1979). Moreover, such work may well suggest new ways to develop aids for hearing impaired persons who, by necessity, must rely on visual displays of the speech signal.

References

Cole, R. A., Rudnicky, A. I., & Zue, V. W. (1979). "Performance of an expert spectrogram reader," in Speech communication papers presented at the 97th meeting of the Acoustical Society of America, edited by J. J. Wolf & D. H. Klatt, New York: Acoustical Society of America.

Cole, R. A., Rudnicky, A. I., Zue, V. W., & Reddy, D. R. (1980). "Speech as patterns on paper," in Perception and production of fluent speech, edited by R. A. Cole; Hillsdale, NJ: Erlbaum.

Egan, J. P. (1948). "Articulation testing methods." Laryngoscope, 58, 955-991.

Greene, B. G., Pisoni, D. B., & Carrell, T. D. (1982). "Learning to recognize visual displays of speech: A first report." J. Acoust. Soc. of Am., 71, S96.

Greene, B. G., Pisoni, D. B., & Carrell, T. D. (1984). "Recognition of speech spectrograms," J. Acoust. Soc. Am., in press.

Potter, R. K., Kopp, G. A., & Green, H. C. (1947). Visible speech. NY: Van Nostrand, (reprinted 1966).

Stewart, L. C., Houde, R. A., & Larkin, W. D. (1976). "A real time sound spectrograph with implications for speech training for the deaf," Proc. IEEE ICASSP, Philadelphia, PA, Pp. 590-593.

Zue, V. W. (1981). "Acoustic-phonetic knowledge representation: Implications from spectrogram reading experiments," paper presented at the 1981 NATO Advanced Summer Institute on Automatic Speech Analysis and Recognition, Bonas, France, Pp. 203-222.

Zue, V. W., & Cole, R. A. (1979). "Experiments on spectrogram reading." Proc. IEEE ICASSP, Washington, D.C., Pp. 116-119.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 9 (1983) Indiana University]

Perceptual Evaluation of Synthetic Speech: Some Constraints on the Use of Voice Response Systems*

Howard C. Nusbaum, Eileen C. Schwab, and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

*The work described in this paper was supported, in part, by NIH research grant NS-12179 to Indiana University, and in part, by contract AF-F-33615-83-K-0501 with the Air Force Office of Scientific Research. The paper also appears in the proceedings of the Voice Input/Output Society, 1983.

ABSTRACT

As the use of voice response systems becomes more widespread in consumer products, industrial and military applications, and aids for the handicapped, it becomes important to understand how humans perceive and respond to synthetic speech. The selection of a specific voice response system for a particular application will depend on a wide variety of factors that affect the clarity and naturalness of the speech. This paper will describe a series of recent experiments that have examined the perception of synthetic speech produced by several text-to-speech systems. Our results demonstrate that important perceptual and cognitive limitations are imposed on the human observer when synthetic speech is used in psychological tasks ranging from phoneme perception to word perception to comprehension. We will report our attempts to overcome some of these limitations through training human observers to perceive synthetic speech more accurately. Finally, we will discuss how differences in perception and preference between text-to-speech systems are important for the design and use of voice response systems.

Perceptual Evaluation of Synthetic Speech:

Some Constraints on the Use of Voice Response Systems

Within the next few years, there will be an extensive proliferation of various types of voice response devices in human-machine communication systems. Such systems will no doubt be employed for a variety of commercial, industrial, and military applications. Unfortunately, at the present time, there has been relatively little basic or applied research on the intelligibility, comprehension and perceptual processing of synthetic speech produced by these devices. Moreover, there has been almost no research directed at understanding how human observers interact with this new technology. The prevailing assumption has been that simply providing automated voice response output and voice data entry will solve most of the human factors problems inherent in the user-system interface. Not only is this assumption untested, but we believe that it is simply false. In many cases, the introduction of voice response and voice data entry systems may create an entirely new set of human factors problems. To understand how the user will interact with these speech I/O devices, we will need to understand much more about how the human perceives, encodes, stores, and retrieves speech and how these basic information processing operations interact with the specific tasks the observer is required to perform.

FACTORS INFLUENCING RECOGNITION

We have been carrying out a program of research that is aimed, in part, at studying how listeners perceive and understand synthetic speech under a variety of task demands and conditions. There are three factors that must be considered when studying the perception of synthetic speech: (1) the specific task demands, (2) inherent limitations of the human information processing system, and (3) constraints on the structure and quality of the speech signal.

The first factor deals with the tasks that an observer is currently engaged in. Some tasks, such as deciding which of two known words was said, are relatively simple and require little conscious effort on the part of the human. Other tasks, such as trying to recognize an unknown utterance while engaging in an activity that already requires attention, are extremely complex. There is a substantial amount of research in the experimental psychology and human factors literature demonstrating the powerful effects of perceptual set, instructions, subjective expectancies, cognitive load, and response set on performance in a variety of perceptual and cognitive tasks. The amount of context and the degree of uncertainty in the task also strongly affect an observer's performance in substantial ways.

Some of the previous work conducted in the Speech Research Laboratory has allowed us to examine the effect of task demands on the perception of synthetic speech. Using synthetic speech produced by the MITalk text-to-speech system, we examined the effect of response set size on intelligibility (Pisoni & Hunnicutt, 1980). One task, the Modified Rhyme Test (MRT), consisted of presenting a monosyllabic word which subjects identified from six possible alternatives. Another task consisted of presenting a monosyllabic word which subjects identified from all known monosyllabic words. For the two tasks, correct word identification was 93.1% and 75.4%, respectively. Clearly, these two tasks produce different estimates of the intelligibility of the synthetic speech.

The second factor influencing recognition concerns the substantial limitations on the human information processing system's ability to perceive, encode, store, and retrieve information. Because the nervous system cannot maintain all aspects of sensory stimulation (and therefore must integrate acoustic energy over time), very severe processing limitations have been found in the capacity to encode and store raw sensory data in the human memory system. To overcome these capacity limitations, the listener must rapidly transform sensory input into more abstract neural codes for more stable storage in memory and subsequent processing operations. The bulk of the research on cognitive processes over the last 25 years had identified human short-term memory (STM) as the major source of the limitation on processing sensory input. The amount of information that can be processed in and out of STM is severely limited by the listener's attentional state, past experience, and the quality of the sensory input.

To investigate how limitations in STM affect the perception of synthetic speech we manipulated cognitive load and measured recall for synthetic and natural speech (Luce, Feustel, & Pisoni, 1983). In this experiment subjects had to retain a variable number of visually presented digits while listening to natural or synthetic speech. We found an interaction between the type of speech presented and the number of digits presented on digit-recall performance. Synthetic speech impaired recall of the visually presented digits more with increasing digit list size than did natural speech.

The third factor concerns the structure of the physical signal itself. Speech signals may be thought of as the physical consequence of a complex and hierarchically organized system of linguistic rules that map sounds onto meanings and meanings back onto sounds. At the lowest level in the system, the distinctive properties of the speech signal are constrained in substantial ways by vocal tract acoustics and articulation. The choice and arrangement of speech sounds into words is constrained by the phonological rules of language; the arrangement of words in sentences is constrained by syntax; and finally, the meaning of individual words and the overall meaning of sentences in a text is constrained by semantics and pragmatics. The contribution of these various levels of linguistic structure to perception will vary substantially from isolated words, to sentences, to passages of fluent continuous speech. In addition to linguistic structure, the ambient noise level of the environment in which the signal occurs will also affect recognition.

Using natural and synthetic speech, we have studied the effect of linguistic structure on word recognition. Subjects identified words in two types of sentences, syntactically and semantically correct sentences (Harvard Sentences) and syntactically correct but semantically anomalous sentences (Haskins Sentences). We found an interaction between the type of speech and type of sentence (Pisoni & Hunnicutt, 1980). With natural speech, subjects obtained 99% correct word identification for the Harvard Sentences and 98% correct for the Haskins Sentences. With MITalk synthetic speech, subjects obtained 93% correct identification of the Harvard Sentences and 78% correct identification of the Haskins Sentences. Obviously, the type of stimulus used to measure performance will have a significant effect on recognition.

IMPROVING INTELLIGIBILITY OF SYNTHETIC SPEECH

The human observer is a very flexible processor of information. With sufficient experience, practice, and specialized training, observers may be able to overcome some of the limitations on performance we have observed in our previous studies. Indeed, several researchers have reported a rapid improvement in recognition of synthetic speech during the course of their experiments.

Carlson, Granstrom and Larsson (1976) studied the recognition of synthetic speech produced from text for lists of sentences and found large improvements in performance throughout their experiment. By the end of the experiment, subjects had increased their performance from approximately 55% correct word identification to approximately 90% correct word identification. In a different experiment, Pisoni and Hunnicutt (1980) also reported consistent improvements in performance with synthetic speech with only one hour of exposure. These results suggest that practice may improve the recognition of synthetic speech.

However, it is possible that the improvements in performance were due to an increased mastery of experimental procedures rather than changes in processing of synthetic speech. For example, Pisoni (1981) used a lexical decision task and found that performance improved for <u>both</u> natural and synthetic speech stimuli. Slowiaczek and Pisoni (1982) ran subjects for five days in a lexical decision task. Again, it was found that performance improved for <u>both</u> synthetic and natural speech. Thus it is possible that the reported improvements in intelligibility of synthetic speech were actually due to an increased familiarity with the experimental procedures rather than an increased familiarity with the synthetic speech. In order to test these alternatives, we conducted an experiment to separate the effects of training on task performance from improvements in the recognition of synthetic speech (Nusbaum & Schwab, 1983; Schwab & Pisoni, 1983).

DESIGN

VOTRAX TRAINING EXPERIMENT

	Day 1	Days 2-9	Day 10
	TESTING	TRAINING	TESTING
	(Pre-test)		(Post-test)
Group			
1	VOTRAX	------	VOTRAX
2	VOTRAX	NATURAL VOICE	VOTRAX
3	VOTRAX	VOTRAX	VOTRAX

The basic design of our experiment is shown above. There was a pre-test on Day 1 of the experiment. This pre-test determined baseline performance for the Votrax Type-N-Talk text-to-speech system. We chose the low-cost Votrax system primarily because of the poor quality of its segmental synthesis. Thus, ceiling effects would not obscure any effects of training. On Day 1, all subjects listened to the synthetic speech and received no feedback on their performance.

On Day 10, there was a post-test to determine the improvements in recognition for the synthetic speech. Again, all subject listened to the Votrax Type-N-Talk synthetic speech but received no feedback about their levels of performance. Groups 1 and 2 were the control groups. The first group controlled for the effects of prior exposure to synthetic speech. This group had minimal exposure to the synthetic speech and the experimental procedures. Any improvements in performance on Day 10 should be due to the exposure to the synthetic speech on Day 1. The second group controlled for the effects of familiarity with the experimental procedures. This group also was only exposed to synthetic speech on the Day 1 pre-test. Any improvements in performance on Day 10 which were greater than the improvements for the first group would be due to the extensive practice with the experimental procedures since both groups had the same amount of experience with the synthetic speech. The third group of subjects listened to the synthetic speech throughout the experiment. If this group showed the same level of performance on Day 10 as the group trained with natural speech, then the reported improvements in performance with training could be due only to mastery of the experimental procedures. On the other hand, if we found that the group trained with Votrax had better levels of performance on Day 10 than the group trained with the natural speech, we would have evidence that we could selectively improve the processing of synthetic speech.

On the pre-test and post-test days, the subjects were given the MRT, isolated words and sentences for transcription. The word lists were taken from phonetically balanced (PB) lists and the sentences were both Harvard Sentences and Haskins Sentences. Subjects were given different materials on every day. During training, subjects were presented with spoken words, and sentences, and they received feedback indicating the correct response.

The results showed that performance improved dramatically for only one group -- the subjects that were trained on the Votrax synthetic speech during Days 2-9. We found that all groups improved significantly on the PB Lists and the Haskins Sentences, suggesting that even one exposure to synthetic speech on the pre-test improved performance. However, at the end of training, the Votrax-trained group showed significantly higher levels of performance than the other two groups with these stimuli. This effect is shown in Figure 1 which displays the pre-test and post-test performance for the transcription of isolated words by our three groups of subjects. The group that received no training (control group) and the group that was trained with natural speech (natural group) both show a significant increase (in identifying words produced by the Votrax system) from Day 1 to Day 10. However, the group trained with Votrax (votrax group) showed a considerably greater improvement.

Insert Figure 1 about here

Clearly, our results demonstrate that it is possible to produce dramatic changes in the recognition of synthetic speech with only moderate amounts of training. Moreover, these improvements were obtained with synthetic speech produced by a low-cost commercially available text-to-speech system. It is

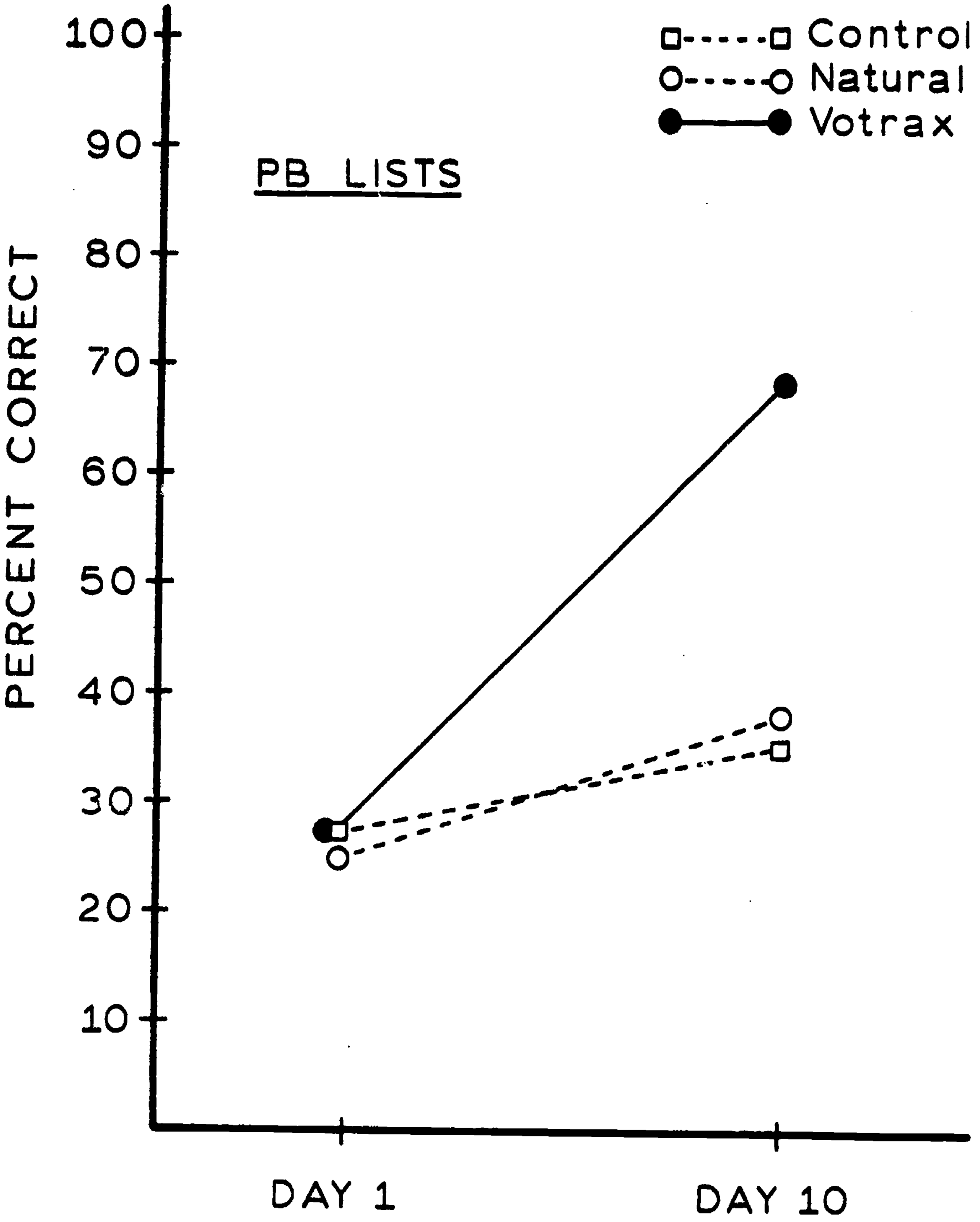


Figure 1. Mean accvracy for Lay 1 (pre-test) and Day 10 (post- test) for transcribing isolated synthetic words (PB Lists).

apparent that measuring performance with the output of a text-to-speech system with naive or unpracticed subjects may underestimate the performance characteristics of the system for certain applications. If a text-to-speech system is going to be used routinely by the same person, then it will be difficult to predict daily performance on the synthetic speech produced by that system from group recognition scores obtained with inexperienced subjects.

Another question we examined concerned the persistence of training once a listener has been acclimated to synthetic speech. If a text-to-speech system will only be used occasionally, then the effectiveness of training as a means of improving perception might be minimized. To investigate this possibility, we recalled subjects trained with the natural and Votrax speech six months after training was completed. In the six month follow-up, subjects were presented with synthetic speech and were tested using the same procedures employed for the pre-test and post-test. The subjects transcribed words from PB Lists, Harvard Sentences, and Haskins Sentences. They were also given the MRT. No feedback was given for any of these tasks and all of the words and sentences were different from the materials used six months earlier.

The results of the follow-up indicated no change in performance for the MRT. We found a significant decrement in performance for the subjects trained with Votrax for the PB Lists and the Haskins Sentences. However, these subjects still showed significantly higher levels of performance than the subjects trained with natural speech. Surprisingly, we found a significant improvement in accuracy with the Harvard Sentences for subjects trained with natural speech. It would appear that for this type of stimulus, subjects trained with natural speech were able to develop a strategy which improved accuracy even though isolated word recognition did not improve.

Our results from the six-month follow-up demonstrate that improvements in recognition produced by training are maintained long after training is finished. These results are even more interesting since the subjects had no exposure to synthetic speech in the intervening period. Thus, there was no opportunity for the subjects to bolster the effects of training by listening to synthetic speech.

The results of our training study suggest several important conclusions. First, the effect of training is apparently to improve the encoding of synthetic words produced by the Votrax Type-N-Talk. Clearly, subjects were not simply learning to perform the various tasks better, since the subjects trained on natural speech showed little or no improvement in performance. Moreover, training affected performance similarly with isolated words and words in sentences, and for closed and open response sets. This indicates that subjects in the group trained on synthetic speech were not learning special strategies; that is, they were not learning to use linguistic knowledge or task constraints to improve recognition. Rather, subjects seem to have learned something about the structural characteristics of this particular synthetic speech system that enabled them to perform better regardless of the task. This conclusion is further supported by the design of the training study. Improvements in performance were obtained on novel materials even though the subjects never heard the same words or sentences twice. In order to show improvements in performance, subjects must have learned something about the detailed acoustic-phonetic properties of the synthetic speech produced by the system.

In addition, subjects retained the training even after six months with no further contact with the synthetic speech. Thus, it appears that training produced a relatively stable and long-term change in the perceptual encoding processes used by subjects. Furthermore, it is likely that more extensive training would have produced greater persistence of the training effects. If subjects had been trained to asymptotic levels of performance, the long-term effects of training might have been even more stable.

Thus, it appears that even low-cost commercially available text-to-speech systems can be used to provide intelligible speech in certain applications. If the cost of training is minimal compared to the cost of voice response systems, then for some applications a low-cost text-to-speech system may provide acceptable performance even when the system is not used on a regular basis.

In addition to studying the contribution of the human observer to the perception of synthetic speech, we have continued to study the effects of stimulus quality on listener performance. In one experiment, we investigated the effects of speech rate, pitch contour, and sentence structure on the perception of synthetic speech produced by the Speech Plus Prose-2000 text-to-speech system (Slowiaczek & Nusbaum, 1983). The sentences were produced at two speech rates (150 or 250 words per minute), and with two pitch contours (monotone or inflected). In addition, we varied the length of the sentence (four or eight content words), and syntactic structure (active, passive, or center-embedded). An example of a short, active sentence is given in (1). An example of a long, center-embedded sentence is given in (2) below:

(1) The compulsive clerk organized the reports.

(2) The apathetic student the concerned dean
advised failed the English test.

We found significant effects of all our variables on recognition performance. Words in sentences presented at a slow rate were identified consistently better than words in the fast sentences. Also, in general, words in inflected sentences were identified more accurately than words in monotone sentences. However, pitch inflection produced by the Prose 2000 did not improve the perception of center-embedded sentences as much as it helped active and passive sentences. This suggests that (either) the pitch algorithm used by the Prose 2000 is not appropriate for center-embedded sentences, or inflected pitch contour does not aid in the perception of center-embedded sentences. As for the length effect, words in short sentences were identified consistently better than words in long sentences. And, for syntactic structure, we found that subjects performed best with active sentences and worst with center-embedded sentences.

The results of this experiment demonstrate that increasing syntactic complexity results in decreased word recognition for synthetic speech. However, the extent to which these variables affect word recognition appears to be related to the nature of the stimuli (i.e., the fact that the speech was synthetic). In previous research on the perception of natural speech, we have found extremely high levels of performance for speech produced at a rate of 230 words per minute. Other researchers have obtained extremely good performance for natural speech at speaking rates up to 275 words per minute and beyond (deHaan, 1977; Foulke &

Sticht, 1969; Garvey, 1953). Hence, at least for speaking rate, the effects found with synthetic speech cannot be generalized to naturally produced speech. These results may reflect fundamental differences in the mechanisms used to increase speaking rate in humans and text-to-speech systems. Nonetheless, the perception of words in sentences appears to be affected by such variables as length, rate, syntactic complexity, and pitch inflection. It should be stressed, however, that pitch produced only a relatively minor (albeit significant) effect on word recognition, at least in the sentence-length materials we used in this study. This suggests that the greatest improvement in performance in text-to-speech systems should come from future efforts to improve the segmental information generated by the system.

SUBJECTIVE EVALUATION OF SYNTHETIC SPEECH

In addition to the quality of the signal itself, another consideration with respect to the evaluation of synthetic speech concerns the user and the user's preferences and biases. If the person using the synthetic speech system cannot tolerate the sound of the speech or does not trust the information provided, the usefulness of this technology will be reduced. With this in mind, we have begun to develop a questionnaire to assess subjective ratings of synthetic speech. Some preliminary data have been collected using various types of stimulus materials and various synthetic speech systems. We have found that listeners' subjective evaluations of their performance generally correlates well with objective measures of performance. Also, we have found that the degree to which subjects are willing to trust the information provided by the synthetic speech is positively correlated with objective measures of performance. For the naive user, poor performance predicts low levels of belief in the messages, whereas high levels of accuracy predict a great degree of trust. In contrast, we found a different pattern of results for our listeners trained with synthetic speech. We administered the questionnaire to our subjects in the six month follow-up to the training experiment. Much higher levels of objective performance were obtained for our subjects trained with the Votrax system than for our subjects trained with the natural speech, and like our naive listeners, our trained subjects were fairly good at subjectively estimating their performance. The group trained with Votrax rated their performance at higher levels of accuracy on the synthetic speech than did the subjects trained with natural speech. However, we also found that the subjects trained with Votrax were _less_ likely to trust the information provided by the system than the subjects trained with natural speech. At least for this synthetic speech system, familiarity with the peculiarities of pronunciation reduces our subjects' willingness to rely upon what they hear.

FUTURE DIRECTIONS

We intend to continue our research on the effects of training on the perception of synthetic speech. In our original training experiment we used a wide variety of linguistic materials and tasks during the training phase. In the future, we will refine our procedures in order to determine what particular type of linguistic unit is most effective in improving performance. In addition, we are continuing work on our questionnaire for the subjective evaluation of synthetic speech. In the current version, many of the responses are in essay format. This is valuable for determining what characteristics of the speech are salient to our subjects. However, in the future we want to obtain ratings and forced-choice responses.

We also intend to expand our research into several new areas. We are currently developing new experimental procedures which will allow us to examine perception of synthetic speech in real time. One procedure uses a self-paced listening-time technique. In this procedure, subjects listen to prose passages and control the presentation rate of the sentences. With this procedure we can examine the effects of task demands on listening time. Another experimental procedure uses a word-monitoring technique in which subjects listen to sentences or passages and are required to make a response whenever a particular word is heard. With this procedure, we hope to examine the perceptual processing of synthetic sentences on a word-by-word basis.

We have also begun to examine different populations of subjects to determine the effects of linguistic experience on perception. Most of the current work on the perception of synthetic speech has used college-age, native-speakers of English with no speech or hearing disorders. It is quite possible that other populations of subjects would not perceive synthetic speech in the same manner. Therefore, we are beginning to test other groups of subjects such as: young children, foreign speakers, and hearing impaired adults.

CONCLUSIONS

In summary, our results have shown that the perception of synthetic speech is not as accurate as the perception of natural speech. Moreover, we have begun to localize the difficulties in the processing of synthetic speech: synthetic speech requires more processing capacity and effort than natural speech. Our results indicate that part of this is due to encoding of stimulus. Further, it is possible to reduce the difficulties in perception of synthetic words and sentences through training.

Our results on the perception of synthetic speech have important implications for the design, selection, and use of voice response systems in a variety of applications. Moreover, such results are quite relevant to the very basic questions surrounding the user-system interface involving natural language under various conditions of information overload -- particularly in situations requiring divided attention among several input modalities which must be monitored simultaneously while carrying out other complex cognitive tasks. Our experiments demonstrate that important interactions in performance occur among the signal, the observer's task, and the capacity of the observer.

References

Carlson, R., Granstrom, B., and Larsson, K. Evaluation of a text-to-speech system as a reading machine for the blind. Speech Transmission Laboratory, QPSR 2-3, 1976, 9-13.

deHaan, H.J. A speech-rate intelligibility threshold for speeded and time-compressed connected speech. Perception & Psychophysics, 1977, 22, 366-372.

Foulke, E., and Sticht, T. G. Review of research on the intelligibility, and comprehension of accelerated speech. Psychological Bulletin, 1969, 77, 50-62.

Garvey, W. D. The intelligibility of speeded speech. Journal of Experimental Psychology, 1953, 45, 102-108.

Luce, P. A., Feustel, T. C., and Pisoni, D. B. Capacity demands in short-term memory for synthetic and natural word lists. Human Factors, 1983, 25, 17-32.

Nusbaum, H. C., and Schwab, E. C. The effects of training on the intelligibility of synthetic speech: II. The learning curve for synthetic speech. Journal of the Acoustical Society of America, 1983, 73, S3.

Pisoni, D.B. Speeded classification of natural and synthetic speech in a lexical decision task. Journal of the Acoustical Society of America, 1981, 70, S98.

Pisoni, D. B., and Hunnicutt, S. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. In 1980 IEEE International Conference Record on Acoustics, Speech, and Signal Processing, 1980, 572-575.

Schwab, E.C., and Pisoni, D. B. The effects of training on intelligibility of synthetic speech: I. Pre-test and post-test data. Journal of the Acoustical Society of America, 1983, 73, S3.

Slowiaczek, L. M., and Nusbaum, H. C. Intelligibility of fluent synthetic sentences: Effects of speech rate, pitch contour, and meaning. Journal of the Acoustical Society of America, 1983, 73, S103.

Slowiaczek, L. M., and Pisoni, D. B. Effects of practice on speeded classification of natural and synthetic speech. Journal of the Acoustical Society of America, 1982, 71, S95.

Capacity-demanding encoding of synthetic speech in serial-ordered recall*

Paul A. Luce and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

*This research was supported by contract AF-F-33615-83-K-0501 from the Air Force Systems Command, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio. We thank Laurie Walker for her help in running subjects.

Abstract

Two experiments were performed to extend the recent finding of Luce, Feustel, and Pisoni (1983) that the perception of synthetic speech places increased demands on short-term memory. In the first experiment, serial-ordered recall for four types of experimental word lists was compared. The word lists were composed of (1) all natu'al words, (2) all synthetic words, (3) five natural words and five synthetic words (in that order), or (4) five synthetic words and five natural words (in that order). We predicted that recall of natural words in natural-synthetic lists would be poorer than recall of natural words in pure natural lists if encoding of the synthetic words decreases short-term memory capacity for the rehearsal of earlier presented natural words. We also predicted that synthetic words in synthetic-natural lists would be recalled better than synthetic words in pure synthetic lists. The results of this experiment failed to support our predictions. The voice in which the last five words was presented had no differential effect on recall of the first five words in the lists. In Experiment 2, we replicated our earlier finding that greater decrements in recall are observed for synthetic words compared to natural words in early list positions than late positions. The results of both experiments are interpreted as supporting an encoding-based explanation of capacity demands in short-term memory for synthetic speech.

Capacity-demanding encoding of synthetic speech in serial-ordered recall

Luce, Feustel, and Pisoni (1983) have recently reported a series of experiments that attempted to determine the locus of the demands placed on short-term memory by the perception of synthetic speech. In this study, we compared recall of short lists of natural and synthetic words under various task manipulations designed to vary the demands placed on short-term memory. In our first experiment, subjects were required to recall lists of natural and synthetic words presented at three different rates: one, two, and five words per second. We reasoned that capacity in short-term memory would be decreased by increasing the presentation rate. We predicted, furthermore, that as capacity decreased, recall of the synthetic word lists would be more severely impaired than recall of the natural word lists. This prediction was based on the hypothesis that synthetic speech may place increased demands on short-term memory processes. Although overall recall for the synthetic word lists was consistently poorer than recall for the natural word lists, no differential effect of presentation rate was observed, as shown by a nonsignificant interaction of voice (natural vs. synthetic) and presentation rate.

We argued that the failure to find effects of differential capacity demands for synthetic speech may have resulted from the failure to place sufficient demands on short-term working memory. Thus, in order to further reduce the capacity in short-term memory for encoding and/or rehearsing the synthetic word lists, we conducted a second experiment employing a memory preload technique (see Baddeley and Hitch, 1974). Prior to presentation of each natural or synthetic word list, subjects were visually presented with zero, three, or six digits on a CRT display which they were to recall prior to recalling the word list. We found that the number of subjects able to correctly recall all of the digits decreased as memory preload increased from three to six digits. Of special interest, however, was the finding that the number of subjects correctly recalling all of the digits dropped more from three to six digits when the digits preceded the synthetic word lists than when they preceded the natural word lists. This finding suggested that the synthetic word lists placed differential demands on encoding and/or rehearsal processes in short-term memory.

To gather stronger evidence for the hypothesis that synthetic speech places increased demands on short-term memory, Luce et al. conducted a third experiment in which subjects were again required to recall lists of natural and synthetic words. In this experiment, however, subjects were required to recall the words in the exact order in which they were presented. By forcing subjects to encode both item and order information, we reasoned that differential performance in recall of the natural and synthetic word lists would be manifested primarily in the primacy portion of the serial position curves for the natural and synthetic word lists. Specifically, we predicted that increased demands on encoding and/or rehearsal processes arising from synthetic speech would cause fewer items presented early in the synthetic lists to be transferred to long-term memory.

The results from the serial-ordered recall experiment are shown in Figure 1. As predicted, recall from the primacy portion of the synthetic word lists was poorer than recall from the primacy portion of the natural word lists, whereas differences in recall of the natural and synthetic speech were not so large for the recency portion of the lists. We therefore concluded that synthetic speech does in fact place increased capacity demands on encoding and/or rehearsal processes in short-term memory.

Insert Figure 1 about here

One possible explanation of our last finding is that encoding of the synthetic words in the later positions of the word lists interfered with rehearsal of the synthetic words presented earlier in the list, thus reducing performance for the primacy portion of the list. Support for this hypothesis comes from an earlier study by Rabbitt (1968; see also Nakatani, 1970). Rabbitt demonstrated that if naturally produced digits are presented for recall when the digits in later serial positions are embedded in noise, recall of items from the primacy portion of the digit list is impaired. Thus, embedding digits in noise results in encoding difficulties that interfere with rehearsal of previously presented digits, much like encoding difficulties encountered with synthetic speech may interfere with rehearsal of earlier presented words.

To directly test the hypothesis that encoding of synthetic words presented later in a list actively interferes with rehearsal of words in earlier positions, we conducted a recall experiment in which mixed lists consisting of both natural and synthetic words were presented for recall. Specifically, four types of lists were used. The lists were composed of: (1) ten natural words, (2) ten synthetic words, (3) five natural and five synthetic words (in that order), and (4) five synthetic and five natural words (in that order). Our predictions, based on the hypothesis stated above, were as follows: Compared to recall performance on the homogeneous lists of natural words, recall of the lists composed of five natural followed by five synthetic words should show a decrement in performance for the primacy portion of the curve, despite the fact that the first five words are natural. Likewise, recall of words from the primacy portion of the five synthetic-five natural word lists should be higher than recall of items from the primacy portion of the lists composed only of synthetic words. In short, regardless of the voice in which the first five words are produced, words from the later portion of the list should affect the primacy portion of the curve by either raising recall performance (when the last five items are natural) or lowering recall performance (when the last five items are synthetic), compared to the appropriate pure or homogeneous list controls.

Experiment 1

Method

Subjects. Ninety-six Indiana University undergraduates participated in partial fulfillment of an introductory course in psychology. Subjects reported no speech or hearing disorders at the time of testing and had no prior exposure to the synthetic speech employed in this study.

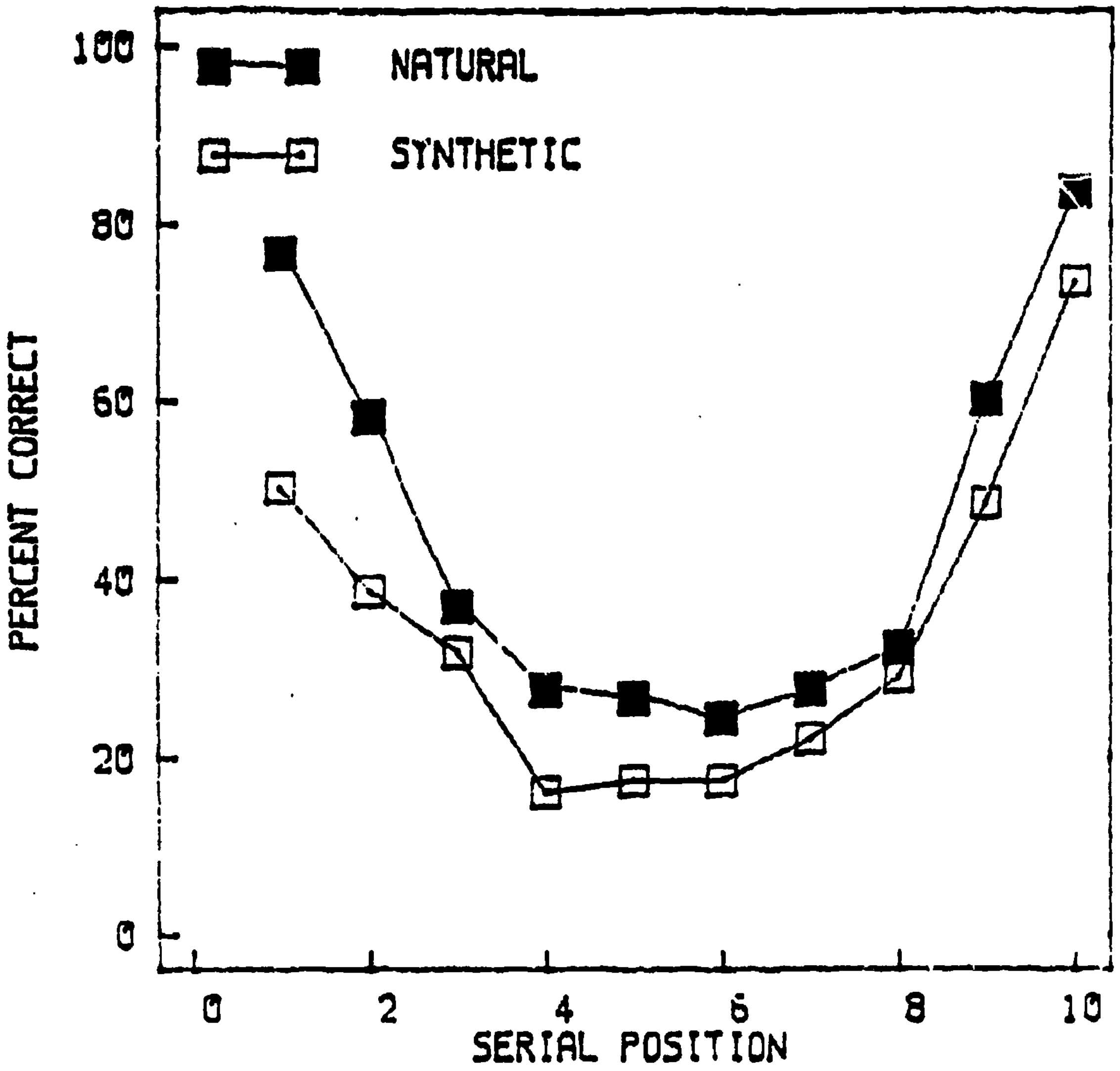


Figure 1. Serial position curves for natural and synthetic word lists. (From Luce et al., 1983.)

Stimuli. The stimuli consisted of 16 lists of ten words each. Four lists were composed of natural words only, four of synthetic words only, four of five natural words followed by five synthetic words, and four of five synthetic words followed by five natural words. All words were selected from the Modified Rhyme 'Iest (House, Williams, Hecker, and Kryter, 1965). The synthetic words were generated by the MITalk text-to-speech system (see Allen, 1976, and 1981, and Allen, Hunnicutt, Carlson, and Granstrom, 1979, for a description of the MITalk system).

The test words were first low-pass filtered at 4.8 kHz and digitized via a 12-bit analog-to-digital c.onverter. All stimuli were played back to listeners through a 12-bit digital-to-analog converter that was interfaced, after appropriate filtering, to matched and calibrated TDH-39 headphones. The words were presented at a comfortable listening level of 80 dB SPL. Presentation of the stimuli was controlled in real time by a minicomputer.

Procedure. Sixteen groups of six subjects were tested in a sound-treated room used for perceptual experiments. Each subject heard one list from each of the four conditions. Presentation of the lists was counterbalanced across sessions.

For each list, subjects first heard a 500-msec 1000-Hz tone warning them that a list was about to begin. Subjects then heard a list of ten words presented at a rate of one word every 2 sec. Immediately after presentation of the tenth word, a tone was presented to indicate the beginning of the recall period. The recall period lasted 90 sec and was terminated by another tone. Subjects were instructed to recall the words in the exact order in which they were presented and to leave blank any spaces on their answer sheets that corresponded to the words they were unable to recall.

Results

Serial position curves, collapsed across subjects, were obtained for each of the four types of lists. An item was scored as correct if and only if it was recalled in the same position in which it was presented in the list.

Insert Figure 2 about here

Figure 2 shows the serial position curves for the natural word lists and the natural-synthetic word lists. Serial position is plotted on the abscissa and percent correct recall on the ordinate. Filled symbols represent natural words and open symbols represent synthetic words. Squares connected by solid lines represent the pure lists and triangles connected by dashed lines represent the mixed lists.

A two-way analysis of variance revealed significant main effects of voice (F(1,95)=8.10, p<0.01) and serial position (F(9,855)=45.34, p<0.01). Clearly, overall recall of natural words was superior to recall of synthetic words, and

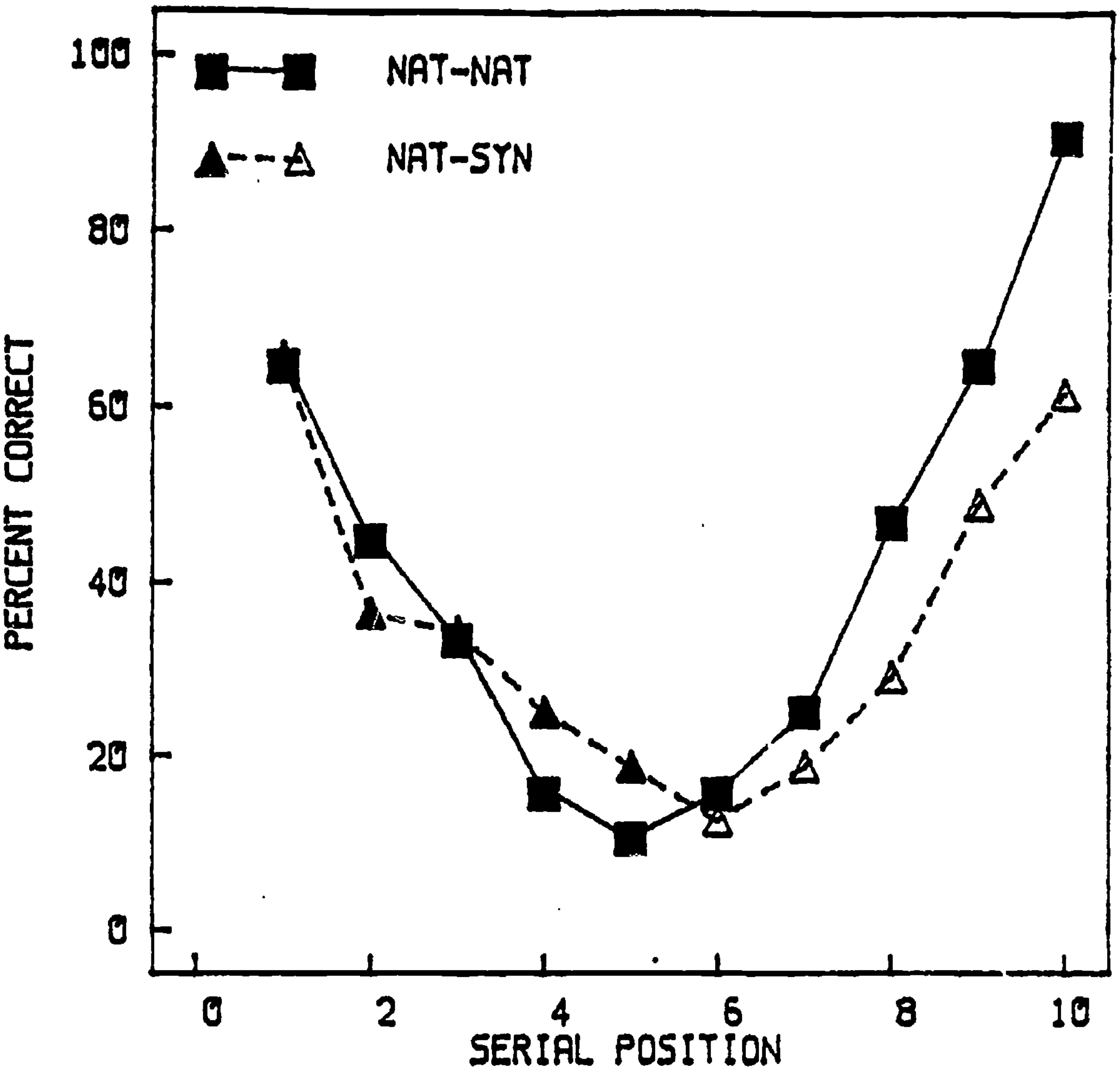


Figure 2. Serial position curves for natural and natural-synthetic word lists for Experiment 1.

recall of words from the primacy and recency portions of the lists was superior to recall of items from the middle serial positions. Contrary to our predictions, however, recall for the primacy portion of the mixed lists was not worse than recall for the primacy portion of the pure natural lists. Recall was, however, consistently worse for the synthetic words from the recency portion of the lists compared to the natural words. Both the lack of differential performance in recall for natural and synthetic words from the primacy portion of the lists and the presence of a difference for the recency portion of the lists resulted in a significant two-way interaction of voice and serial position (F(9,855)=4.66, p<0.01).

Insert Figure 3 about here

Figure 3 shows the serial position curves for the synthetic word lists (squares) and the synthetic-natural word lists (triangles). Natural words are again represented by filled symbols and synthetic words by open symbols.

As before, significant main effects of voice (F(1,95)=27.07, p<0.01) and serial position (F(9,855)=45.09, p<0.01) were obtained. Also, as in the previous condition, our predictions were not borne out. Namely, recall of synthetic words from the primacy portion of the synthetic-natural lists was not superior to recall of synthetic words from the primacy portion of the pure synthetic lists. However, for the recency portion of the lists, recall of natural words was consistently superior to recall of synthetic words. These two results again combined to produce a significant interaction of voice and serial position (F(9,855)=4.80, p<0.01).

Because the results for recall of the mixed lists failed to show the predicted effects of the encoding of words in later positions on rehearsal of words in earlier positions, we were interested in determining if we had replicated the original Luce et al. finding for the pure lists of natural and synthetic words. To do this, analyses of variance were performed on the data for the pure natural and synthetic word lists. The serial positions for these word lists are replotted in Figure 4.

Insert Figure 4 about here

Clearly, recall of the natural word lists was superior to recall of the synthetic word lists (F(1,95)=56.57, p<0.01). However, inspection of Fig. 4 shows no strong tendency for differences in recall to be greater for the primacy portions of the curves than the recency portions. In fact, one-way analyses of variance at each serial position revealed significant differences in recall between the natural and synthetic words at serial positions one (F(1,95)=13.26, p<0.01), two (F(1,95)=21.65, p<0.01), three (F(1,95)=7.26, p<0.01), six

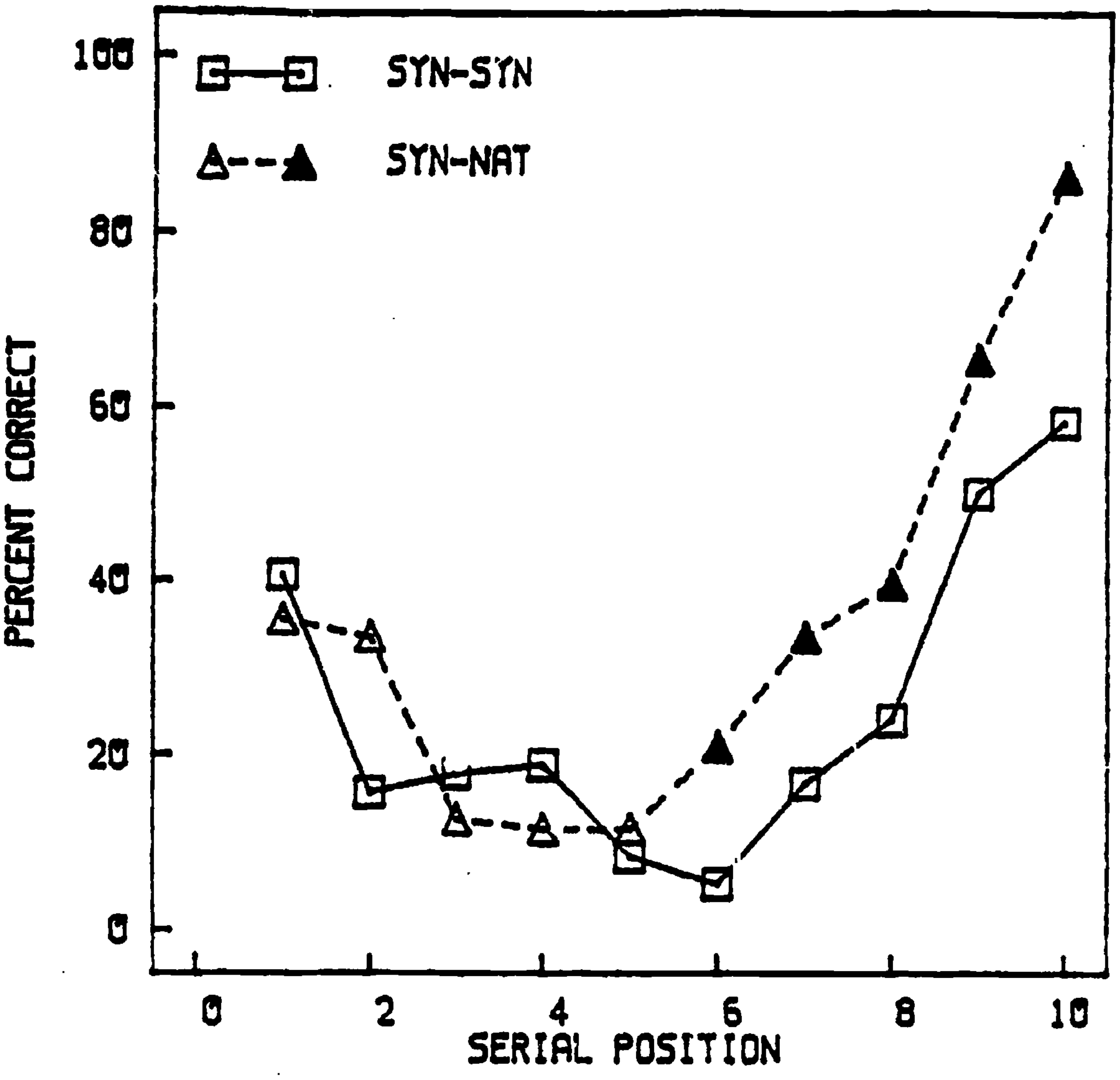


Figure 3. Serial position curves for synthetic and synthetic-natural word lists for Experiment 1.

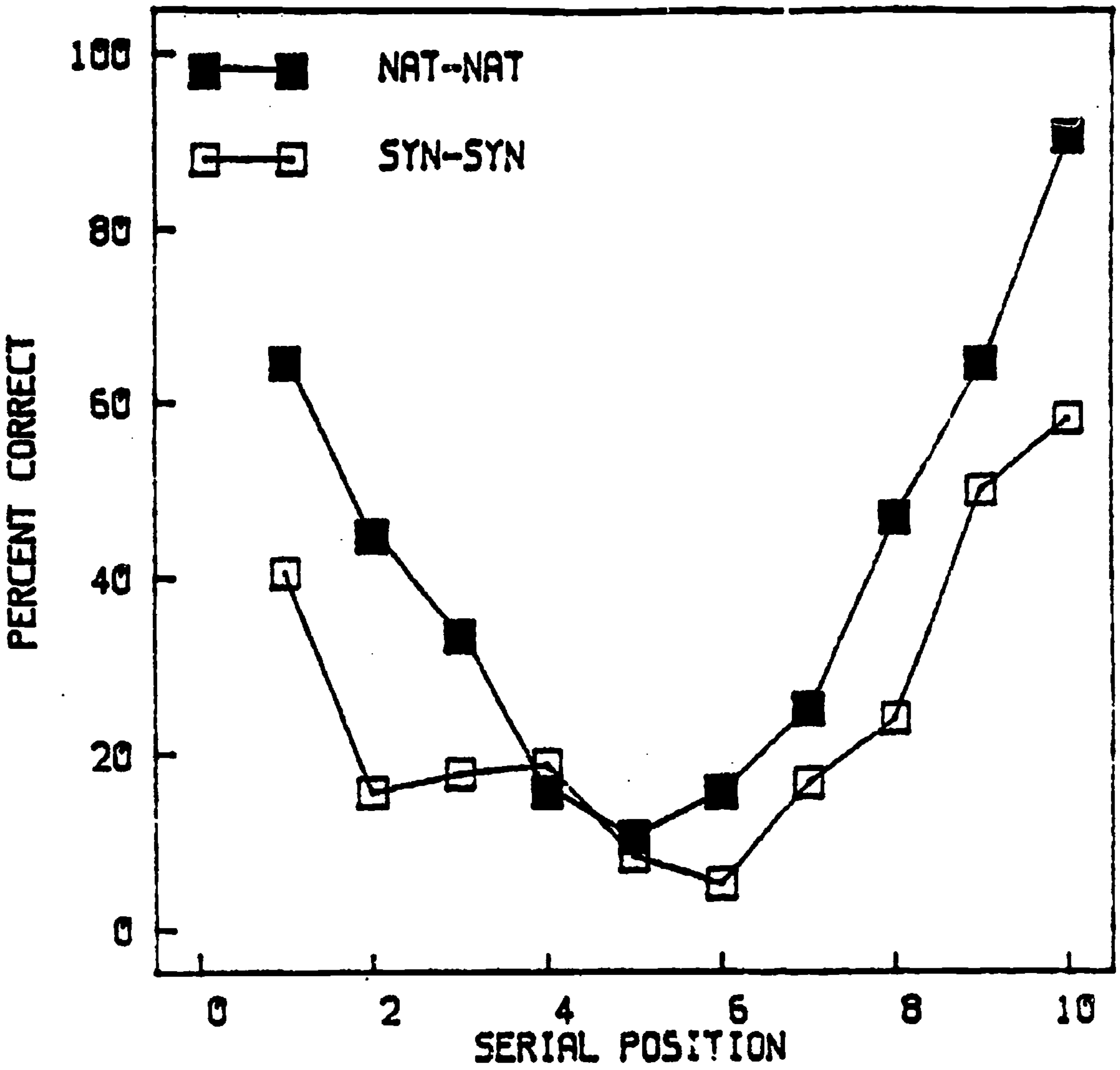


Figure 4. Serial position curves for natural and synthetic word lists for Experiment 1.

($\underline{F}(1,95)=5.84$, $\underline{p}<0.02$), eight ($\underline{F}(1,95)=10.65$, $\underline{p}<0.01$), nine ($\underline{F}(1,95)=4.85$, $\underline{p}<0.03$), and ten ($\underline{F}(1,95)=41.37$, $\underline{p}<0.01$). Clearly, differences in recall performance between natural and synthetic words were not primarily restricted to the primacy portions of the serial position curves. These results thus failed to replicate the original Luce et al. finding.

Discussion

The results from the present experiment tend to disconfirm our original hypothesis that the encoding of synthetic words in later serial positions interferes with the active rehearsal of words presented in earlier serial positions. Specifically, we failed to find two predicted effects. First, recall of natural words in the natural-synthetic lists was not worse than recall of natural words in lists composed entirely of natural words. If encoding of the synthetic words in the natural-synthetic lists interfered with rehearsal of the earlier presented natural words, we should have observed lower recall performance for the natural words from the natural-synthetic lists compared to the natural words from the primacy portion of the pure natural lists. Our findings revealed no such pattern of results. Second, recall of synthetic words in the synthetic-natural lists was not better than recall of synthetic words in lists composed entirely of synthetic words. If encoding of natural words interfered less with rehearsal or required less capacity than the encoding of the synthetic words, recall performance for the synthetic words from the synthetic-natural lists should have been superior to recall performance for the synthetic words from the primacy portion of the pure synthetic lists. Again, our results failed to support this hypothesis.

Particularly problematic, however, was the failure to replicate the original Luce et al finding that differences in recall between natural and synthetic word lists are greater for the primacy portion of the curves than the recency portion. One possible explanation of this failure to replicate may be that too few data points per subject were obtained in the present experiment to detect the effect obtained by Luce et al. In our original study, subjects received three lists of both natural and synthetic words. In the present experiment, subjects received only one pure list of natural words and only one pure list of synthetic words. Thus it is possible that the effect we reported earlier was not detected by the experimental design employed here.

In light of the present findings, we attempted a more rigorous replication of the Luce et al. result. In Experiment 2, we again presented subjects with natural and synthetic word lists for serial ordered recall. As in the original study, three lists of both natural and synthetic words were presented to each subject. However, the stimuli used in Experiment 2 comprised a subset of the original natural and synthetic words used by Luce et al. In a preliminary experiment, we obtained identification scores for each of the natural and synthetic MRT words from a separate group of 40 subjects. In this experiment, subjects were simply presented with each word and asked to write down what they heard. Only natural and synthetic words that were identified at a 98% level of accuracy or better were selected for use in Experiment 2. We used these words for two reasons: First, we were interested in determining if recall differences could be demonstrated when both the synthetic and natural words were equated for intelligibility under optimal conditions. Second, we were interested in determining if the effects observed by Luce et al. may have been due to an accidental preponderance of poorly identifiable synthetic words presented in the primacy portion of the lists to be recalled. Although assignment to list position was always random in all of the Luce et al. experiments, such a possibility could have occurred by chance. In short, we attempted to replicate and extend the original Luce et al. findings under somewhat more rigorous conditions.

Experiment 2

Method

Subjects. Forty-eight undergraduates participated in partial fulfillment of an introductory course in psychology. Subjects reported no speech or hearing disorders at time of testing and had no prior exposure to the synthetic speech employed in this study.

Stimuli. Three lists of natural and three lists of synthetic words that were identified correctly at a 98% level of accuracy or better in a preliminary experiment were used. The words employed constituted a subset of the words used in Experiment 1. Stimulus generation and presentation were identical to that in Experiment 1.

Procedure. The procedure was identical to that employed in Experiment 1 and in the earlier Luce et al. study.

Results and Discussion

Serial position curves for the natural and synthetic word lists are shown in Fig. 5. Natural words are represented by filled squares and synthetic words by open squares.

Insert Figure 5 about here

Analysis of variance revealed a significant effect of voice ($F(1,47)=9.66$, $p<0.01$). Thus, in spite of the fact that the natural and synthetic words were identified at equal levels of accuracy in isolation, synthetic words were recalled more poorly overall than natural words. To determine if these recall differences were larger in earlier list positions than later positions, one-way analyses of variance were again performed at each serial position. These tests revealed significant differences in recall at serial position two ($F(1,47)=6.88$, $p<0.02$), three ($F(1,47)=8.62$, $p<0.01$), and five ($F(1,47)=5.62$, $p<0.03$). No other differences at the .01 level or below were obtained.

Although Experiment 1 failed to replicate the original Luce et al. finding, Experiment 2 clearly supports that result: The differences in recall observed in

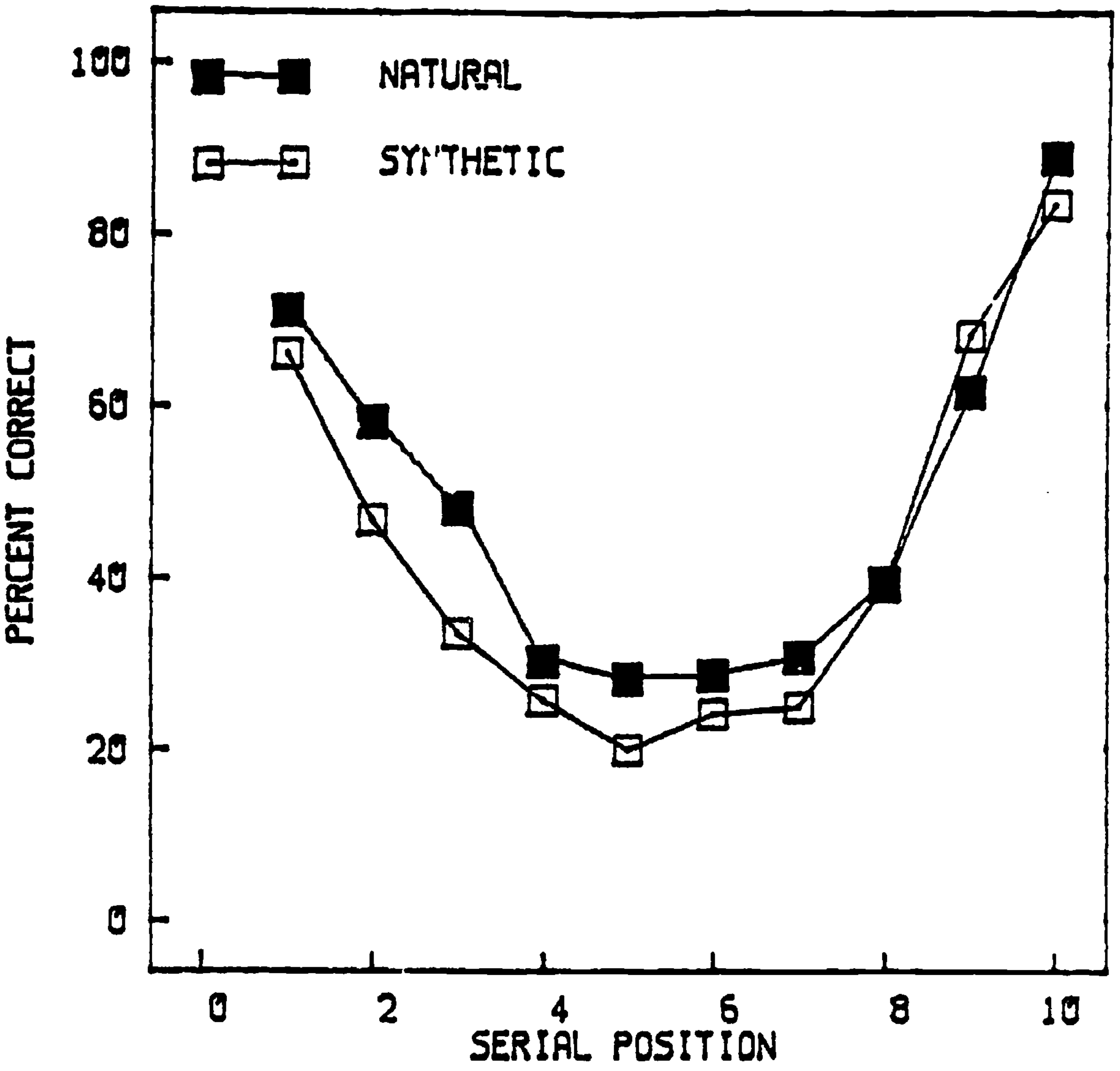


Figure 5. Serial position curves for natural and synthetic word lists for Experiment 2.

this experiment were restricted to the early serial positions (namely, positions two, three, and five). This finding is compelling given that the stimuli used in this experiment were previously matched for intelligibility. Indeed, the very demonstration of a significant difference in recall in this experiment lends further strong support to the original claim that synthetic speech does place increased demands on encoding and/or rehearsal processes in short-term memory. Thus, the process of encoding these items produces greater demands on short-term memory capacity when the stimuli are synthetic speech, even high quality synthetic speech.

General Discussion

The present experiments extend our earlier findings in a number of ways. First, the results from Experiment 1 in which mixed lists of natural and synthetic words were presented for serial-ordered recall call into question the hypothesis that encoding of synthetic words directly interferes with rehearsal processes in short-term memory. Instead, it appears that a purely encoding-based explanation is sufficient to account for the results from Experiment 1 as well as the results of Luce et al. According to this hypothesis, momentary demands produced by difficulties in encoding synthetic words reduce available capacity in short-term memory. As these demands accrue, previously presented words that are still being encoded may be lost from short-term memory or misencoded, thus producing relatively lower recall performance for synthetic words from the primacy portion of the list. However, no long term effects of encoding difficulties on rehearsal processes will be observed, as shown by our results for the mixed lists of words.

In a second extension of our findings, we replicated our original result for serial-ordered recall under more rigorous conditions. In this experiment, we demonstrated that recall deficits for synthetic words lists can be obtained even when the natural and synthetic words are equated for intelligibility in isolation. In addition, we also showed that the majority of the recall deficits incurred were due to lower performance at earlier serial positions in the list. Again, we believe this result is primarily the result of encoding difficulties that reduce available capacity in short-term memory.

In summary, these results further extend our knowledge about the nature of the demands perception of synthetic speech place on short-term memory. The results from Experiment 1 strongly suggest that encoding of synthetic items does not directly interfere with rehearsal. The results from Experiment 2, however, support the claim that the encoding of synthetic speech does place increased demands on short-term memory. In addition, Experiment 2 demonstrated that even when synthetic words are identified at very high levels of accuracy, significant decrements in performance may be observed under conditions of increased cognitive load where the observer is required to encode both item and order information and retain that information for a short period of time before retrieval.

References

Allen, J. Synthesis of speech from unrestricted text. Proceedings of the IEEE, 1976, 4, 433-442.

Allen, J. Linguistic-based algorithms offer practical text-to-speech systems. Speech Technology, 1981, 1, 12-16.

Allen, J., Hunnicutt, S., Carlson, S., & Granstrom, B. MITalk-79: The 1979 MIT text-to-speech system. In J. J. Wolf & D. H. Klatt (Eds.), Speech communication papers presented at the 97th meeting of the Acoustical Society of America. New York: Acoustical Society of America, 1979, 507-510.

Baddeley, A. D., & Hitch, G. Working memory. In G. H. Bower (Ed.), The psychology of learning and memory (Vol. 8). New York: Academic Press, 1974.

House, A. S., Williams, C. E., Hecker, M. H. L., & Kryter, K. D. Articulation-testing methods: Consonantal differentiation with a closed-response set. Journal of the Acoustical Society of America, 1965, 37, 158-166.

Luce, P. A., Feustel, T. C., & Pisoni, D. B. Capacity demands in short-term memory for synthetic and natural speech. Human Factors, 1983, 25, 17-32.

Nakatani, L. H. Evaluation of speech quality using immediate recall of digit strings. Bell Telephone Laboratories Technical Memorandum, 1970.

Rabbitt, P. Channel-capacity, intelligibility, and immediate memory. Quarterly Journal of Experimental Psychology, 1968, 20, 241-248.

The Representation of Synthetic Speech ın Precategorical Acoustic Storage*

Paul A. Luce and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

*This research was supported by contract AF-F-33615-83-K-0501 from the Air Force Systems Command, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio. We thank Laurie Walker for her help in running subjects.

Abstract

An experiment employing the suffix effect was performed to test the hypothesis that the perception of synthetic speech produces an impoverished representation in precategorical acoustic storage (PAS) that leads, in part, to difficulties in maintaining synthetic words in short-term memory. Subjects were presented with three lists of natural words and three lists of synthetic words. Each list was followed by one of three suffixes: a natural token of the word "ten," a synthetic token of the word "ten," or a tone. We predicted that if synthetic speech produces impoverished representations in PAS, synthetic words would be more easily displaced from PAS than natural words and a synthetic suffix would not overwrite PAS as effectively as a natural suffix. Although a large suffix effect was obtained for both natural and synthetic lists, no differential effect of natural or synthetic suffixes was obtained. In addition, synthetic and natural suffixes had equal effects in attenuating recall of the last item on the lists. These results indicate that the perception of synthetic speech does not produce a demonstrably impoverished representation in PAS. Furthermore, the synthetic speech employed in our experiment was sufficiently "speechlike" to act like natural speech in overwriting PAS.

The Representation of Synthetic Speech in Precategorical Acoustic Storage

Several recent experiments (Luce, Feustel, and Pisoni, 1983; Luce and Pisoni, 1984) have established that the perception of synthetic speech places increased processing demands on short-term memory. These studies have demonstrated that recall of short lists of synthetic words is inferior to recall of lists of natural words because of encoding and/or rehearsal difficulties encountered in attempting to maintain synthetic words in short-term memory. A question that has yet to be directly addressed, however, is whether synthetic speech places increased demands on the earlier stage of auditory sensory memory, called precategorical acoustic storage or PAS (Crowder and Morton, 1969).

PAS is a hypothetical construct invoked in the memory literature to account for a well-known phenomena referred to as the suffix effect (see Crowder, 1971, and Crowder and Morton, 1969). In a typical suffix experiment, subjects are presented auditorily with a list of digits that is just above memory span (seven or eight items). Subjects are then required to recall the digits in the exact order in which they were presented (i.e., serial-ordered recall is required). If the list is immediately followed by a redundant speech sound such as "zero" or "go" (called a suffix), recall performance for the last few items in the list is reduced, compared to control conditions in which the lists are followed by a nonspeech sound such as a tone or burst of white noise. This reduction in recall performance for the last few items has been attributed to the erasure of the contents of PAS by the suffix (Crowder, 1971; Crowder and Morton, 1969). When no suffix is appended to the list, subjects are purportedly able to sample information from PAS in order to recall the last item or items in the list. Appending a redundant suffix, however, writes over this information and recall performance is depressed.

The suffix effect provides a potentially effective means of determining the extent to which the perception of synthetic speech establishes an impoverished representation in auditory sensory memory that may subsequently produce the difficulties in encoding and/or rehearsal previously reported (Luce et al., 1983; Luce and Pisoni, 1984). To test this possibility, we conducted an experiment in which short lists of natural and synthetic words were presented for serial-ordered recall. Immediately following each list, one of three items was presented: a tone, a natural token of the word "ten," or a synthetic token of the word "ten." We predicted that if synthetic speech causes an impoverished representation to be established in PAS, one or both of two outcomes might be observed: First, synthetic words occurring at the end of the lists would be more easily displaced from PAS than natural words. And second, synthetic suffixes would overwrite PAS less effectively than natural suffixes. This last prediction was motivated by a study by Morton, Marcus, and Ottley (1981) that showed that the less "speechlike" the suffix, the smaller the effect of the suffix on recall. We were interested, therefore, in determining if subjects treat synthetic speech, at least at this initial stage of processing, as less like speech, perhaps due to the fact that phonetic cues are less redundantly encoded in synthetic speech than natural speech (Pisoni, 1982).

Method

Subjects. Seventy-two undergraduates participated in partial fulfillment of the requirements for an introductory course in psychology. Subjects reported no speech or hearing disorders at time of testing and had no prior exposure to the synthetic speech employed in this study.

Stimuli. The stimuli consisted of three lists of ten natural words and three lists of ten synthetic words selected from the Modified Rhyme Test (House, Williams, Hecker, and Kryter, 1965). The synthetic words were generated by the MITalk text-to-speech system (see Allen, 1976, and 1981, and Allen, Hunnicutt, Carlson, and Granstrom, 1979, for a description of the MITalk system). The suffixes consisted of the natural word "ten" and the synthetic word "ten," which was also generated by the MITalk system. A 500-msec 1000-Hz tone was used in the no-suffix condition.

The test words were first low-pass filtered at 4.8 kHz and digitized via a 12-bit analog-to-digital converter. After digitizing, all words, suffixes, and the tone were equated for peak amplitude. All stimuli were played back to listeners through a 12-bit digital-to-analog converter that was interfaced, after appropriate filtering, to matched and calibrated TDH-39 headphones. The words were presented at a comfortable listening level of 80 dB SPL. Presentation of the stimuli was controlled in real time by a PDP-11/34 minicomputer.

Procedure. Twelve groups of six subjects were tested in a sound-attenuated room used for perceptual experiments. Each subject heard all six lists of words. Each natural and synthetic word list was followed by either the natural word "ten," the synthetic word "ten," or the tone. Each subject therefore participated in each of six conditions: (1) natural word list--natural suffix, (2) natural word list--synthetic suffix, (3) natural word list--tone, (4) synthetic word list--natural suffix, (5) synthetic word list--synthetic suffix, and (6) synthetic word list--tone. Presentation of the conditions was counterbalanced across sessions. The words, including the suffixes, were presented at a rate of one word every two seconds.

The specific experimental procedure was as follows: Subjects first saw a sentence on CRT monitors in front of them with instructions to begin recalling the following list of words when they heard the natural word "ten," the synthetic word "ten," or a tone. A 500-msec 1000-Hz tone was then presented as a warning signal to indicate that a list was about to begin. Subjects then heard the list of ten words followed by a natural or synthetic token of the word "ten" or a tone, depending on what they were told prior to presentation of the list. Subjects then responded by writing the words they were able to recall on response sheets. Subjects were allowed to recall the words for a 60 sec period, the end of which was signalled by another tone. Subjects were instructed to recall the words in the exact order in which they were presented and to leave blank spaces on their response sheets that corresponded to words they could not remember.

Results and Discussion

Serial position curves for each condition were obtained by collapsing curves across subjects. An item was scored as correct if and only if it was recalled in the exact position in which it was presented. Serial position curves for the natural and synthetic word lists for each suffix condition are shown in Fig. 1. Serial position is plotted on the abscissa and probability of correct recall on the ordinate. Hourglasses represent lists followed by a tone, squares represent lists followed by the natural word "ten," and triangles represent lists followed by the synthetic word "ten."

Insert Figure 1 about here

Because we were interested in the effect of the suffix on recall of the last word presented, analyses of variance were performed on the last serial position alone. (Balota and Engle (1981) have shown that recall perfomance on the last word is relatively unaffected by subject strategies and thus provides the purest index of the effect of the suffix on PAS.) A two-way analysis of variance of Voice (natural word list and synthetic word list) by Suffix (natural word "ten," synthetic word "ten," and tone) revealed significant main effects of Voice ($\underline{F}(1,71)=7.66$, $\underline{p}<0.01$) and Suffix ($\underline{F}(2,142)=15.00$, $\underline{p}<0.01$). The main effect of voice type indicates that, for the last word on the lists, natural words were recall significantly better than synthetic words. The main effect of suffix type indicates that, for both the natural and synthetic word lists, recall of the last word on the lists was differentially affected by at least one of the three suffixes. Finally, a nonsignificant interaction of Voice by Suffix ($\underline{F}<1$) indicates that the effect of the suffix on recall of the last word was the same for both the natural and synthetic word lists.

Although the main effect of suffix type was significant, we do not know which particular condition was contributing to this effect. Separate one-way analyses of variance were therefore performed to determine which of the suffix conditions were contributing to the overall main effect. These tests revealed that for both the natural and synthetic words lists, the tone produced significantly higher recall performance for the last word ($\underline{F}(1,71)=15.00$, $\underline{p}<0.01$, for the natural words lists and $\underline{F}(1,71)=16.33$, $\underline{p}<0.01$, for the synthetic word lists). For both types of lists, however, the natural and synthetic suffixes produced equal decrements in recall ($\underline{F}(1,71)=1.71$, $\underline{p}>0.1$, for the natural word lists and $\underline{F}<1$, for the synthetic word lists).

Three findings are revealed by these results. First, we obtained a large and significant suffix effect for both types of word lists. Lists terminated by a tone produced higher recall performance for the last word than lists terminated by either token of the word "ten." We thus replicated the already well-established suffix effect. Second, both the natural and synthetic suffixes produced equal decrements in recall. The suffix effect for both list types was therefore equal for both the natural and synthetic suffixes. And, finally,

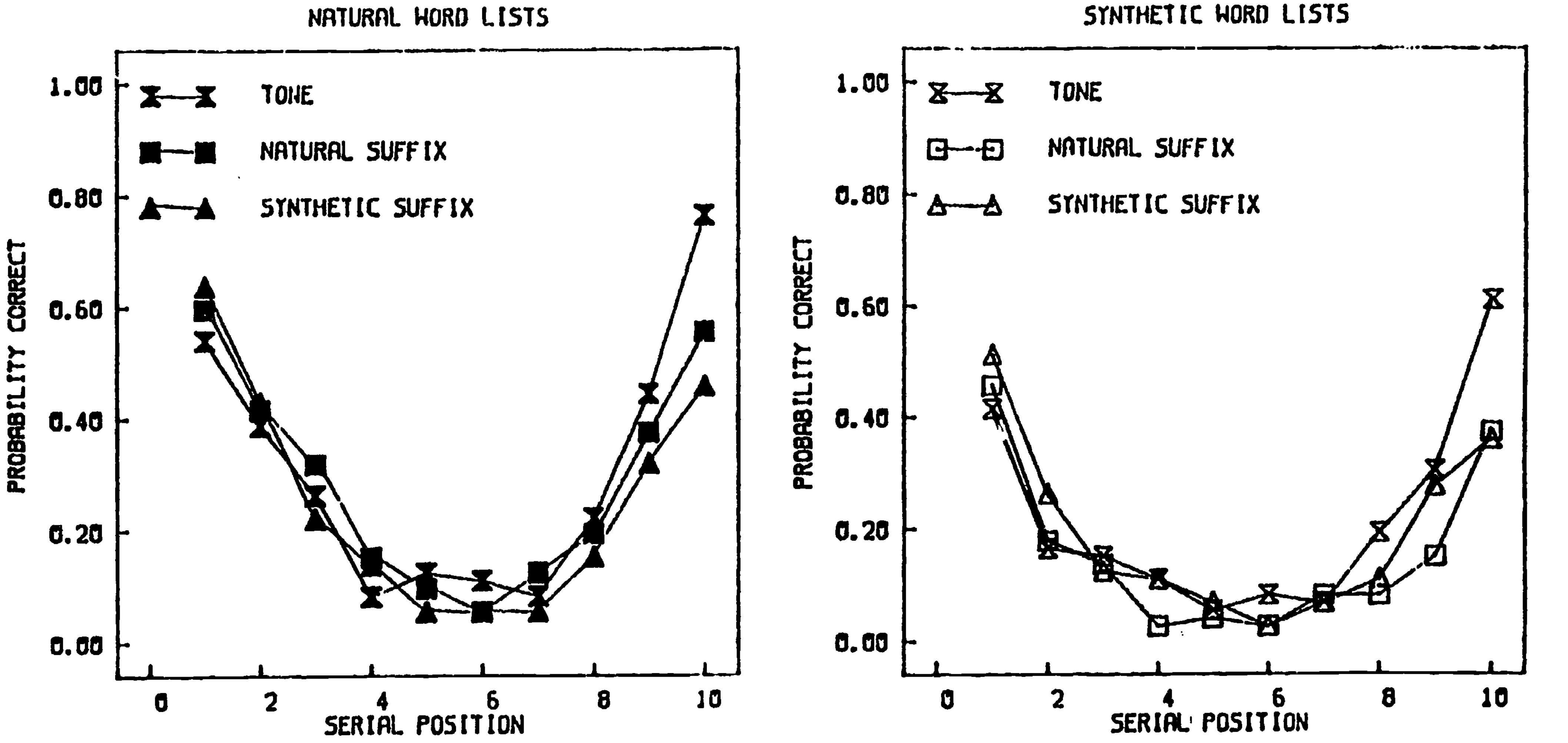


Figure 1. Serial position curves for the natural and synthetic words lists for each suffix type. Natural word lists are shown in the left-hand panel and synthetic words lists are shown in the right-hand panel.

although recall of the synthetic words from the final position in the lists was poorer overall than recall of the natural words, no differential effects of the suffixes were observed. The natural and synthetic suffixes had equal effects in reducing recall of the last item for both the natural and synthetic word lists.

General Discussion

The results of the present experiment indicate that the short-term memory deficits for synthetic speech demonstrated in our previous studies probably do not result from impoverished auditory representations of synthetic speech in PAS. Our results show that while both natural and synthetic suffixes do overwrite PAS, they do not have differential effects on the recall of the last word in the list. Moreover, natural and synthetic suffixes do not have differential effects when appended to natural or synthetic words lists. These results suggest that the representation of synthetic speech in PAS is not so impoverished as to be obliterated more easily than natural speech from auditory sensory memory. In addition, these results demonstrate that the synthetic speech employed in this study is sufficiently "speechlike" to act much like natural speech in producing the suffix effect.

One possible qualification of these conclusions should be noted in connection with the finding that the natural and synthetic suffixes produced identical decrements in recall. Previous research (Morton, Crowder, and Prussin, 1971) has shown that when the voice of the suffix differs from the voice of the word list to be recalled, the suffix effect is attenuated. It is surprising, then, that the synthetic and natural suffixes produced equal decrements in recall, regardless of the voice of the word list. This finding may indicate that synthetic speech does produce difficulties in extracting information from PAS that compensate for the predicted attenuation in the suffix effect when the voice of the word lists and the voice of the suffix are different. Further research is underway to determine the source of this result.

In summary, our results indicate (with one possible exception) that the difficulties in maintaining synthetic speech in short-term memory previously reported are probably not due to impoverished auditory representations for synthetic speech in the sensory memory system known as PAS. In addition, our results indicate that the synthetic speech employed in this study is sufficiently speechlike to elicit the same early phonetic processing elicited by natural speech. Thus, at least as revealed by the present experiment, synthetic speech may orient the listener to initiate phonetic processing in much the same way as natural speech does. These two findings demonstrate that the demands on short-term memory previously reported for synthetic speech probably arise in developing a phonetic representation (i.e., recogniqing phonemes from the waveform) or in maintaining an impoverished phonetic representation in memory and not in sampling auditory information from PAS.

References

Allen, J. Synthesis of speech from unrestricted text. *Proceedings of the IEEE*, 1976, *4*, 433-442.

Allen, J. Linguistic-based algorithms offer practical text-to-speech systems. *Speech Technology*, 1981, *1*, 12-16.

Allen, J., Hunnicutt, S., Carlson, S., & Granstrom, B. MITalk-79: The 1979 MIT text-to-speech system. In J. J. Wolf & D. H. Klatt (Eds.), *Speech communication papers presented at the 97th meeting of the Acoustical Society of America*. New York: Acoustical Society of America, 1979, 507-510.

Balota, D. A., & Engle, R. A. Structural and strategic factors in the stimulus suffix effect. *Journal of Verbal Learning and Verbal Behavior*, 1981, *20*, 346-357.

Crowder, R. G. Waiting for the stimulus suffix: Delay, decay rhythms, and readout in immediate memory. *Quarterly Journal of Experimental Psychology*, 1971, *23*, 324-340.

Crowder, R. G., & Morton, J. Precategorical acoustic storage (PAS). *Perception and Psychophysics*, 1969, *5*, 365-373.

House, A. S., Williams, C. E., Hecker, M. H. L., & Kryter, K. D. Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of th Acoustical Society of America*, 1965, *37*, 158-166.

Luce, P. A., & Pisoni, D. B. Capacity-demanding encoding of synthetic speech in serial-ordered recall. *Research on Speech Perception*, Progress Report No. 9, Indiana University, 1983.

Luce, P. A., Feustel, T. C., & Pisoni, D. B. Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, 1933, *25*, 17-32.

Morton, J., Crowder, R. G., & Prussin, H. A. Experiments with the stimulus suffix effect. *Journal of Experimental Psychology Monograph*, 1971, *91*, 169-190.

Morton, J., Marcus, S. M., & Ottley, P. The acoustic correlates of "speechlike": A use of the suffix effect. *Journal of Experimental Psychology: General*, 1981, *110*, 568-593.

Pisoni, D. B. Perception of speech: The human listener as a cognitive interface. *Speech Technology*, 1982, *1*, 10-23.

The Role of Fundamental Frequency and Duration in the Perception of Clause Boundaries: Evidence from a Speeded Verification Task.*

Paul A. Luce and Jan Charles-Luce

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

*This is a draft of a paper presented at the 105th meeting of the Acoustical Society of America, May 12, 1983, Cincinnati, Ohio. This work was supported by grants from NIMH, Research Grant No. MH-24027, and NINCDS, Research Grant No. NS-12179.

Abstract

The perception of clause boundaries is determined by at least three suprasegmental variables: fundamental frequency, duration, and amplitude [Streeter, J. Acoust. Soc. Am. 64 (1978)]. We investigated the relative contribution of FO declination and duration to clause boundary perception. Pairs of sentences were constructed so that both sentences of a pair were lexically identical up to a point at which a clause boundary occurred in one member of the pair; in the other member, the clause boundary occurred later. All sentences were truncated immediately after the last word shared by both sentences of a pair. Thus, the pairs of truncated sentences were identical except for the suprasegmental cues indicating a clause boundary. In a second manipulation, truncated sentences were produced in which FO and duration conflicted. In a speeded verification task, subjects decided if an auditorily presented truncated sentence matched a visually presented, intact sentence. The results of this experiment indicate that whereas FO may facilitate clause boundary perception, duration is the decisive cue.

The Role of Fundamental Frequency and Duration in the Perception of Clause Boundaries: Evidence from a Speeded Verification Task

Phrase-final lengthening of the last word in major syntactic units has been shown to be a possible perceptual cue to syntactic boundaries (Klatt,1975; Lehiste, Streeter, and Olive, 1976). In addition, the gradual declination of fundamental frequency (or FO) from the beginning to the end of syntactic units has also been shown to be a potentially important cue in the identification of syntactic boundaries (Maeda, 1976; O'Shaughnessy, 1976; Streeter, 1978). As FO declines it approaches a baseline frequency that characterizes the bottom of the speaker's pitch range (Pierrehumbert, 1980). It may be possible, then, for a listener to evaluate the distance of the FO declination from the baseline and use this information in identifying clause boundaries.

We were particularly interested in evaluating FO and duration as perceptual cues to sentence-initial subordinate clause boundaries. We hypothesized that FO declination may facilitate identification of a clause boundary, but duration will always be the decisive cue. To test this hypothesis, we employed a truncated sentence-verification task in an attempt to specify more precisely the relative importance of these two prosodic cues in the perception of clause boundaries.

Stimuli

Insert Figure 1 About Here

We first constructed ten sentence pairs, each containing an initial subordinate clause. An example of one of these pairs of sentences is shown in Figure 1. Each sentence pair was constructed so that both members of the pair were lexically identical up to a point (unexpectedly in this example). In one member of the pair, a clause boundary, indicated in this figure by a comma, preceded this last shared word, while in the other member, the clause boundary occurred after this last shared worc.

The stimulus sentences were randomized, read and recorded by a male talker, and digitized. For the control condition, each of the sentences was truncated using a digital waveform editor immediately after the word that preceded the last shared word. In our example, these sentences were both truncated after apartment, as indicated by the vertical line. This resulted in one truncated member of the sentence pair consisting of a complete initial subordinate clause, and the other truncated member consisting of an interrupted initial subordinate clause.

Insert Figure 2 About Here

EXAMPLE OF STIMULUS MATERIALS

1) WHEN MARY ARRIVED AT ED'S APARTMENT,| UNEXPECTEDLY HER FRIENDS GREETED HER WITH A BIRTHDAY PARTY.

2) WHEN MARY ARRIVED AT ED'S APARTMENT|UNEXPECTEDLY, SHE FOUND ED WITH ANOTHER WOMAN.

Figure 1. Example of stimulus materials used in Experiments 1 and 2. The vertical line indicates where the stimulus was truncated. Stimulus 1 was truncated at the clause boundary and stimulus 2 was truncated before the clause boundary.

EXAMPLES OF TRUNCATED STIMULI

CONTROL - AT CLAUSE BOUNDARY:

1) WHEN MARY ARRIVED AT ED'S APARTMENT,

CONTROL - BEFORE CLAUSE BOUNDARY:

2) WHEN MARY ARRIVED AT ED'S APARTMENT ... ,

CONFLICTING - AT CLAUSE BOUNDARY:

WHEN MARY ARRIVED AT ED'S APARTMENT,

CONFLICTING - BEFORE CLAUSE BOUNDARY:

WHEN MARY ARRIVED AT ED'S APARTMENT ... ,

Figure 2. Example of control and conflicting cue stimuli used in Experiments 1 and 2. The underlining indicates how the control stimuli were rearranged to produce the conflicting cue stimuli.

For the experimental condition, which we will refer to as the conflicting cue condition, all sentence pairs were truncated at the same point as the control stimuli. However, in the conflicting cue condition, the last words in the truncated sentences were exchanged within given sentence pairs. Examples of the control and conflicting cue stimuli are shown in Figure 2. For the control stimuli, At Clause Boundary indicates the sentences were truncated at the clause boundary, which resulted in a complete subordinate clause; Before Clause Boundary indicates the sentences were truncated before the clause boundary, which resulted in an interrupted subordinate clause. The underlining in this figure illustrates how the control stimuli were rearranged to produce the conflicting cue stimuli. For the conflicting cue stimuli, At Clause Boundary indicates that the last word of this sentence was excised from the control stimulus truncated at the clause boundary. Likewise, Before Clause Boundary for the conflicting cue stimuli indicates that the last word was excised from the control stimulus truncated before the clause boundary. Notice in this figure that the control stimulus that is truncated at the clause boundary is underlined by a single line and the control stimulus truncated before the clause boundary is underlined by two lines. For the conflicting cue stimuli, the number of lines under apartment indicates from which control stimulus the word was excised. Likewise, the number of lines under When Mary arrived at Ed's indicates into which control stimulus apartment was spliced.

Exchanging words within a given sentence pair resulted in two conflicting cues. In half of the conflicting cue stimuli, F0 was not sufficiently near the baseline to signal a clause boundary. However, the duration of the last word exhibited phrase-final lengthening, thus signalling a boundary. In the other half of the conflicting cue stimuli, F0 was near the baseline, signalling a clause boundary, but the last word was not lengthened, thus not indicating a boundary.

In addition to the ten minimal pairs of control and conflicting cue stimuli, four pairs of practice sentences were constructed. For the practice stimuli, however, the sentences were only truncated. We did not exchange the last words within sentence pairs.

Experiment 1

In our first experiment, we were interested in how well subjects could decide whether an auditorily presented truncated sentence was a complete clause or an interrupted clause. Subjects first read a list of the ten complete sentence pairs that would be presented in the testing phase of the experiment. For each of these sentences, a vertical line was drawn indicating where the sentence would be truncated during the auditory presentation (as in Figure 1). We instructed subjects to try to imagine how each sentence would sound if it were cut off at this line. At no time during the experiment were subjects told what specifically they should listen for.

After reading the experimental sentences, subjects were given practice. In the practice phase of the experiment, each subject was given a response booklet containing a list of the full sentence pairs for the practice stimuli. Subjects then heard a series of truncated sentences, each of which corresponded to one of the members of the sentence pairs in the response booklets. For the practice phase, the correct sentence in each pair was marked in the response booklets. Only control stimuli were presented for the practice phase of the experiment.

Insert Figure 3 About Here

In the testing phase of the experiment the procedure was the same as that in the practice phase except that the experimental sentences were presented and the answers were not provided. The subjects' task was to decide whether the truncated stimulus they heard was a complete subordinate clause or an interrupted clause. An example of one of the sentence pairs subjects read in their response booklets is shown under the heading "VISUAL PRESENTATION" in Figure 3. One of four truncated sentences was then presented auditorily, examples of which are shown under the heading "AUDITORY PRESENTATION" in this figure. Correct responses are indicated in this figure below each truncated stimulus. A response is considered correct in this example if one responds according to the durational cue.

Results

Insert Figure 4 About Here

Figure 4 shows the percent correct by duration as a function of whether the auditory stimulus presented was truncated at the clause boundary or before the clause boundary. The ordinate axis is labelled percent correct by duration because for the stimuli in which FO and duration conflicted, we had no a priori basis on which to score percent correct. We thus assumed that duration was the principle cue and scored the responses as correct if subjects used the duration of the final word as the basis for their response. The open squares represent the control stimuli and the filled triangles represent the conflicting cue stimuli.

The results clearly demonstrate that for both the control and conflicting cue sentences, duration of the final word is a powerful cue to initial clause boundary identification. Subjects correctly classified the control and conflicting cue stimuli an average of 94% of the time for both the At and Before Clause Boundary conditions. Although these results clearly indicate that duration is a powerful cue to clause boundaries, and, moreover, to the absence of clause boundaries, we found no indication that FO serves as an important cue in clause boundary identification. That is, we obtained no differences in performance between the control and conflicting cue conditions. We therefore conducted a second experiment in an attempt to look more closely at the contribution of FO to the perception of clause boundaries.

EXPERIMENT I

(TWO ALTERNATIVE FORCED CHOICE)

VISUAL PRESENTATION:

1) WHEN MARY ARRIVED AT ED'S APARTMENT, UNEXPECTEDLY HER FRIENDS GREETED HER WITH A BIRTHDAY PARTY.
2) WHEN MARY ARRIVED AT ED'S APARTMENT UNEXPECTEDLY, SHE FOUND ED WITH ANOTHER WOMAN.

AUDITORY PRESENTATION:

A) WHEN MARY ARRIVED AT ED'S APARTMENT,

CORRECT RESPONSE: "1"

B) WHEN MARY ARRIVED AT ED'S APARTMENT ... ,

CORRECT RESPONSE: "2"

C) WHEN MARY ARRIVED AT ED'S APARTMENT,

CORRECT RESPONSE: "1"

D) WHEN MARY ARRIVED AT ED'S APARTMENT ... ,

CORRECT RESPONSE: "2"

Figure 3. Examples of the visual and auditory stimuli used in Experiment 1.

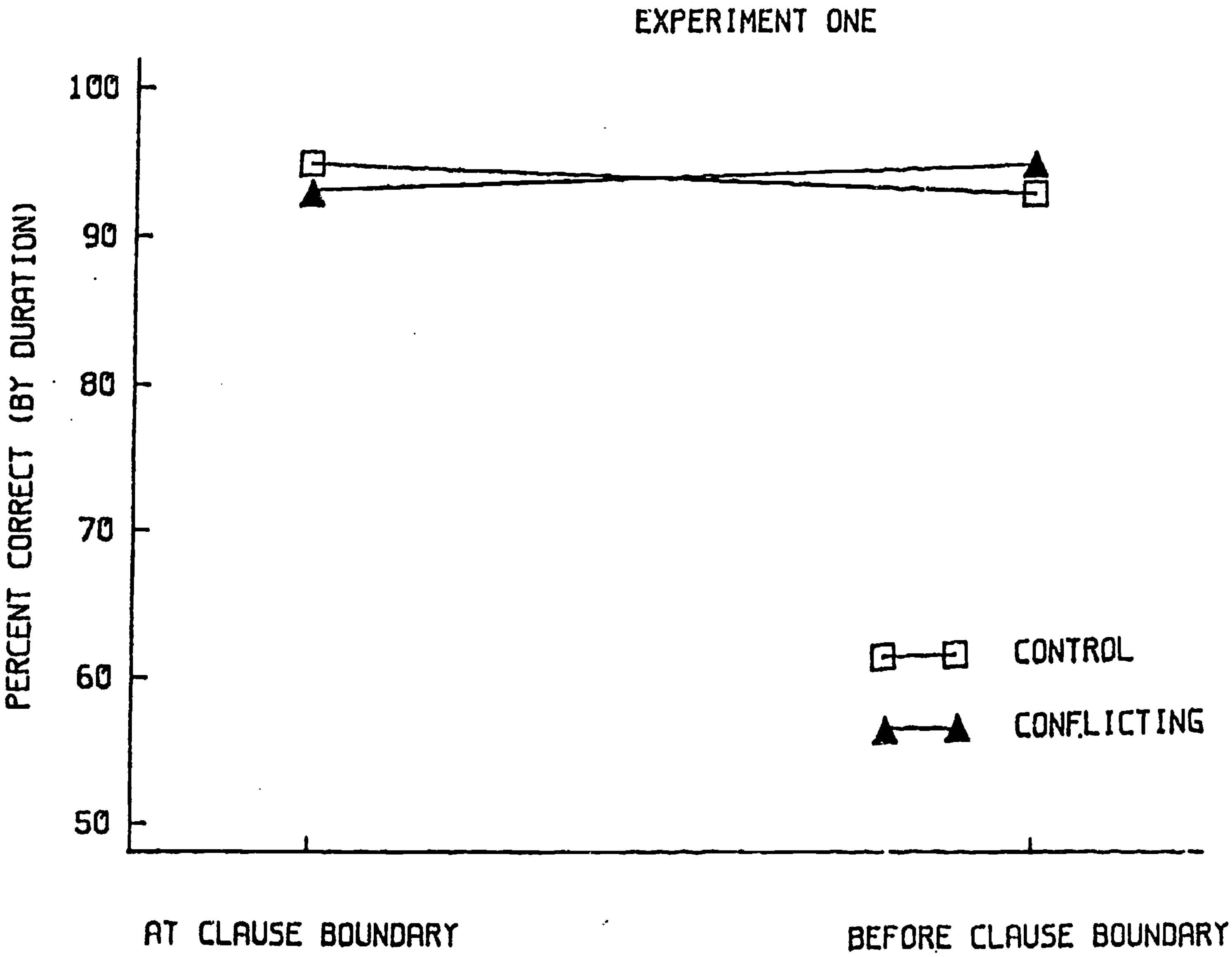


Figure 4. Percent correct for the control and conflicting cue stimuli for Experiment 1. Results are plotted as a function of whether the stimuli were truncated at the clause boundary or before the clause boundary.

Experiment 2

To obtain a more sensitive measure of subjects' processing of F0 information, we adapted the task from the first experiment in order to collect reaction times. The same practice and experimental stimuli that we used in the first experiment were used in the second experiment.

Insert Figure 5 About Here

As in the first experiment, subjects read the experimental sentences marked by a vertical line at the truncation point. However, the practice and experimental phases of the experiment were modified. In this new task, subjects were presented visually with a single, complete sentence from a given pair on a CRT screen. Approximately four seconds later they heard one of the four possible truncated sentences: a control stimulus truncated at the clause boundary or before the clause boundary or a conflicting cue stimulus truncated at the clause boundary or before the clause boundary. After presentation of the truncated sentence, the word RESPOND appeared on the CRT screen below the sentence. Subjects then responded by pressing the appropriately labelled button on response boxes in front of them. Subjects responded either "yes", the truncated sentence they heard corresponded to the visually presented sentence, or "no", the truncated sentence did not correspond to the visually presented sentence. Subjects were instructed to respond as quickly but as accurately as possible. The correct responses for each of the truncated stimuli are given in Figure 5. For the conflicting cue stimuli, we have again assumed that a correct response means subjects are deciding on the basis of the durational cue.

In the practice phase of the experiment, subjects received feedback. The procedure for the testing phase was identical to the practice phase, except no feedback was given and both control and conflicting cue stimuli were presented.

Results

Insert Figure 6 About Here

Figure 6 shows the percent correct by duration as a function of whether the sentence was truncated at the clause boundary or before the clause boundary for the HITS, or "yes" responses. The open squares represent the control stimuli and the filled triangles represent the conflicting cue stimuli. Although performance was somewhat attenuated from the first experiment, accuracy is still at or above 80% for all conditions. In addition, there was a significant main effect between the control and conflicting cue conditions, $F(1,36) = 6.4$, $p < .02$. The interaction, however, was not significant.

EXPERIMENT II

(YES/NO)

VISUAL PRESENTATION:

WHEN MARY ARRIVED AT ED'S APARTMENT, UNEXPECTEDLY
HER FRIENDS GREETED HER WITH A BIRTHDAY PARTY.

AUDITORY PRESENTATION:

1) $\underline{\text{WHEN MARY ARRIVED AT ED'S}}$ $\underline{\text{APARTMENT}}$,

CORRECT RESPONSE: "YES"

2) $\underline{\underline{\text{WHEN MARY ARRIVED AT ED'S}}}$ $\underline{\underline{\text{APARTMENT}}}$ ... ,

CORRECT RESPONSE: "NO"

3) $\underline{\underline{\text{WHEN MARY ARRIVED AT ED'S}}}$ $\underline{\text{APARTMENT}}$,

CORRECT RESPONSE: "YES"

4) $\underline{\text{WHEN MARY ARRIVED AT ED'S}}$ $\underline{\underline{\text{APARTMENT}}}$ ... ,

CORRECT RESPONSE: "NO"

Figure 5. Examples of the visual and auditory stimuli used in Experiment 2.

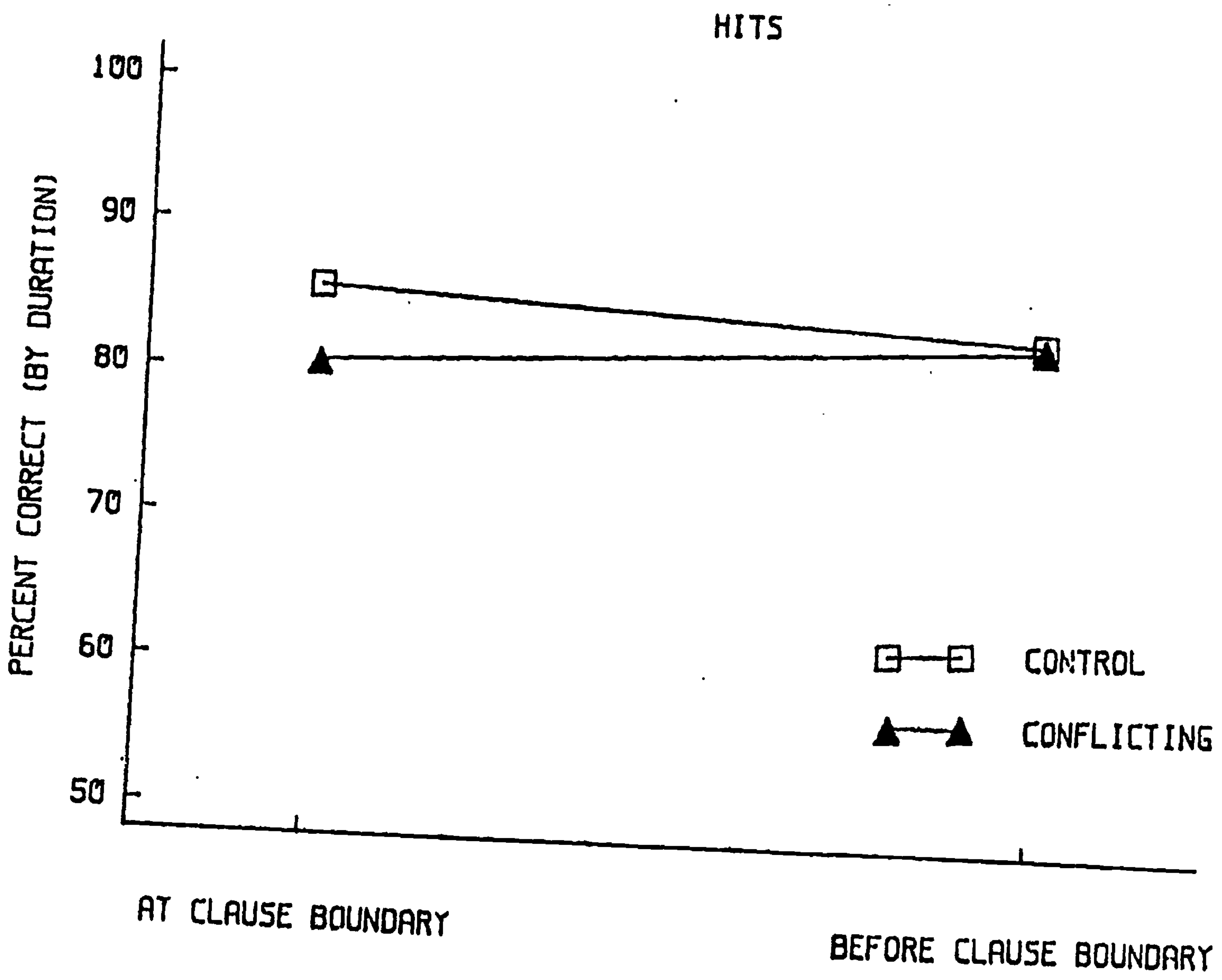


Figure 6. Percent correct for the "Hits" or "yes" responses for the control and conflicting cue stimuli. Results are plotted as a function of whether the stimuli were truncated at the clause boundary or before the clause boundary.

For the At Clause Boundary condition, Scheffe post hoc tests revealed significantly better performance for the control stimuli compared to the conflicting cue stimuli, S = 11.03, $p < .001$. We attribute this difference to the fact that, whereas in the control stimuli FO and duration both indicate a clause boundary, the conflicting cue stimuli contain an FO that is not yet close to the baseline and thus is perhaps not a highly informative cue. Subjects therefore must decide on the presence of a clause boundary on the basis of durational information alone for these conflicting cue stimuli.

In the Before Clause Boundary condition, however, we found no such difference in performance between the control and conflicting cue stimuli. We must ask, therefore, why the conflicting cue stimuli are not producing reduced performance relative to the control stimuli. We suggest that duration is providing a strong cue even to the absence of a clause boundary. However, in the conflicting cue stimuli, FO, which is at or near the baseline, is itself providing a powerful contradictory cue. These two cues thus appear to be establishing a salient contrast that enables increased identification performance for the conflicting cue stimuli. That is, the strong FO cue in the conflicting cue stimuli appears to contrast with the durational cue in such a way as to make the durational cue more salient.

Insert Figure 7 About Here

Figure 7 shows the reaction times for the HITS as a function of where the sentence was truncated. First, notice the large, significant difference in reaction times between the At Clause Boundary and the Before Clause Boundary conditions, $F(1,36) = 3.16$, $p < .0001$. This result demonstrates that subjects are slower in matching interrupted subordinate clauses to the visually presented sentences than in matching intact or uninterrupted clauses. This finding may imply some sort of extra facility on the part of our subjects to retain a representation of complete clauses, as opposed to incomplete clauses, during the execution of this task.

The second result we find here, and the more important finding for our purposes, is that conflicting cue stimuli produced significantly longer reaction times than the control stimuli, $F(1,36) = 4.75$, $p < .04$. Although the accuracy results showed that subjects are still basing their decisions on the durational cue, we now have evidence that misleading FO information does slow processing and/or decision times.

The conflicting cue stimuli truncated both at and before the clause boundary show slower reaction times than do the control stimuli. In addition, we obtained the same pattern of results for the reaction times that we obtained for the accuracy data: There was a larger difference between control and conflicting cue stimuli at the clause boundary than before the clause boundary. This result again suggests that an FO not close to the baseline is not a very informative cue. This is shown in large differences between the mean reaction times for the control and conflicting cue stimuli truncated at the clause boundary. In this condition, subjects apparently had only the durational information in the

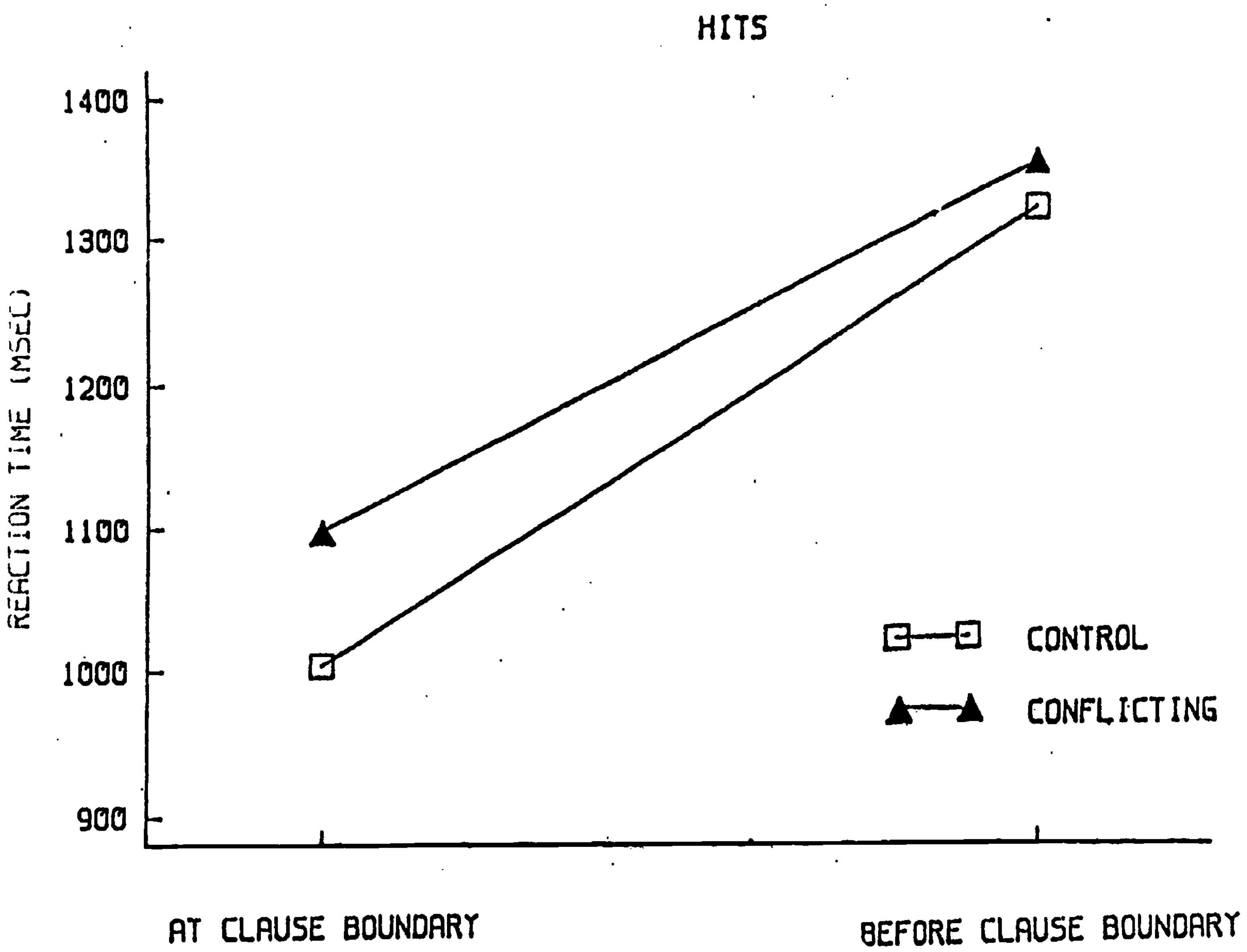


Figure 7. Reaction times for the "Hits" or the "yes" responses for the control and conflicting cue stimuli. Results are plotted as a function of whether the stimuli were truncated at the clause boundary or before the clause boundary.

conflicting cue stimuli upon which to base their responses, thus producing slower reaction times. On the other hand, in the Before Clause Boundary condition, for the conflicting cue stimlui,the FO at or near the baseline and the durational cue to the absence of a clause boundary provided a salient contrast, thus diminishing the differences in mean reaction times between the control and conflicting cue stimuli in this condition.

Conclusions

Overall, then, our results indicate that although FO declination is an important cue in signalling a clause boundary, duration is the most important cue in deciding the presence of a clause boundary. Moreover, our results suggest that FO declination is an important cue only at a local level. FO near the baseline is more informative as a cue to clause boundary identification than an FO which is not near the baseline.

We believe these results further our understanding of the relative contribution of FO and duration as major cues in the perception of clause boundaries, at least in initial subordinate clauses. FO can facilitate the perception of a clause boundary when it is near the baseline. But, phrase-final lengthening will always be a sufficient cue to the presence of a clause boundary (and, conversely, that lack of lengthening will be sufficient to cue the absence of a clause boundary).

References

Klatt, D. H. Vowel lengthening is syntactically determined in a connected discourse. Journal of Phonetics, 1975, 3, 129-140.

Lehiste, I., Olive, J. P., & Streeter, L. A. The role of duration in disambiguating syntactically ambiguous sentences. Journal of the Acoustical Society of America, 1976, 60(5), 1199-1202.

Maeda, S. A characterization of American English intonation. Unpublished doctoral dissertation, Massachusetts Institute of Technology, 1976.

O'Shaughnessy, D. Modelling fundamental frequency, and its relation to syntax, semantics, and phonetics. Unpublished doctoral dissertation, Massachusetts Institute of Technology, 1976.

Pierrehumbert, J. B. The phonology and phonetics of English intonation. Unpublished doctoral dissertation, Massachusetts Institute of Technology, 1980.

Streeter, L. A. Acoustic determinants of phrase boundary perception. Journal of the Acoustical Society of America, 1978, 64(6), 1582-1592.

Perception of Synthetic Speech by Children*

Beth G. Greene

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

*This is the final draft of a paper presented at the 105th meeting of the Acoustical Society of America, Cincinnati, May 11, 1983. This research was supported, in part, by NIH grant NS-12179 and NICHD grant HD-11915 to Indiana University at Bloomington.

Abstract

Many educational and commercial products that produce synthetic speech output are becoming widely available. However, there is little research on how adults or young children process (i.e., perceive, encode, retrieve) synthetic speech signals. Three experiments that studied children's perception of natural and synthetic speech were conducted using a high quality text-to-speech system. In one experiment, children 5-6 years old listened to natural and synthetic speech tokens in a task that required them to point to a picture, from among four alternatives, that corresponded to the word heard. Subjects consistently showed higher error rates when listening to synthetic speech than natural speech. In two additional experiments, children 8-10 years old listened to natural and synthetic digit strings in an immediate recall task. Performance deficits were observed for both synthetic and natural speech as the number of digits to be recalled increased. These results have implications for the design, selection and use of voice-response systems to be used in teaching and learning environments with young children and adults.

Perception of Synthetic Speech by Children

For several years we have been examining the perception of synthetic speech in the Speech Research Laboratory at Indiana University. We have used a wide variety of experimental paradigms and stimulus materials. It is often reported that synthetic speech, even good synthetic speech, is hard to understand, that it is foreign sounding and that it has a mechanical quality to it. Our investigations have dealt primarily with comprehension and intelligibility of synthetic speech. Considered together, our results have shown lower performance levels for synthetic speech than for natural speech although the differences are highly dependent on the type of synthetic speech used, the experience of the observer and the specific tasks he/she is asked to perform (see Pisoni, 1982 for a summary).

When subjects were required to identify a target word from a list of six phonemically similar choices, performance was better for natural speech than synthetic speech. When subjects were required to recall words in sentences, performance was better for words in natural sentences than synthetic sentences. This result was especially robust when the sentences were semantically anomalous (Pisoni, 1982).

Lexical decision studies have shown that subjects are roughly 150 msec slower to identify synthetic speech tokens as words and nonwords than natural speech tokens (Pisoni, 1982). A more recent series of experiments has shown that under a variety of experimental conditions, free and serial recall of lists of synthetic words was consistently poorer than recall of lists of natural words (Luce, Feustel & Pisoni, 1983). Our findings suggest that the perception of synthetic speech places increased processing demands on short-term memory and therefore adversely affects recall and subsequent processing of the linguistic input (Nusbaum & Pisoni, 1982).

The purpose of the present study was twofold: First, can school age children understand synthetic speech generated by a high-quality text-to-speech system and second, how does the child's perception of synthetic speech compare to perception of natural speech.

Speech synthesis chips have been put into talking toys, home computers and arcade videogames, cash registers, coke machines and other products too numerous to mention here.

There are now several inexpensive speech synthesis modules currently available for microcomputers. The home computer can be programmed to have voice output. We anticipate that schools will add voice input and output modules to their computers as the drive for computer literacy advances over the next few years.

Relatively few studies have examined children's perception of synthetic speech. In one of the first studies with elementary school children, Laddaga et al. (1981) used materials produced by several speech synthesis systems. These investigators examined the performance of first graders in recognizing the individual letters of the alphabet by name as spoken by various synthesis systems. Scores for the children ranged from 83% to 98% correct. In a second experiment, fifth graders listened to isolated words and were required to circle the word heard from among three acoustically similar choices. Performance ranged from 78% to 100% correct. Laddaga et al. concluded that "some form of synthetic speech was adequate for use in computer-assisted instruction in initial reading" (p.395).

The experiments we are presenting today complement our earlier work on perception of synthetic speech by adult observers. We have focused our attention on the children's responses rather than the evaluation the speech synthesis systems as Laddaga et al. had done. As mentioned earlier, there were two questions we wanted to answer: (1) can school-age children understand synthetic speech and (2) how does understanding of synthetic speech compare to natural speech.

The synthetic speech stimuli used in the three experiments we will report today was produced by the Telesensory Systems Prose 2000 text-to-speech system (Groner, Bernstein, Ingber, Pearlman & Toal, 1982). Previous research in our laboratory demonstrated that the speech produced by this system and its predecessor -- the MITalk system was highly intelligible to adult listeners.

Picture Vocabulary Task

The first experiment we carried out involved the recognition of isolated words. Items from the Peabody Picture Vocabulary Test were selected to serve as stimulus materials. Both natural and synthetic tokens of the items were produced using the recording, editing, and production facilities at our laboratory. The Peabody Picture Vocabulary Test is widely used in educational and clinical settings to evaluate verbal intelligence. The test provides for a very wide range of ability levels. The items are arranged sequentially by difficulty. A standard administration of the test provides two scores: A basal level and a ceiling level. These levels are operationally defined as: Basal - The last 8 sequential items answered correctly; Ceiling - the last 8 items in which 6 errors were made.

For our purposes in this experiment we began with the first item and continued to items that were appropriate for average 9 year old children. We presented the items in the same order in which they appear on the standardized test, with minoı changes to counterbalance the stimuli for the experiment. Pretesting eliminated those stimuli that the text-to-speech system had mispronounced or applied stress rules incorrectly. The stimuli were arranged sequentially from the easiest to hardest items, again slightly changed to permit counterbalanced conditions. Natural and synthetic tokens were presented in a mixed list condition. For each item, an accompanying picture card was used. Each card contained four different pictures, only one of which was the correct choice. Each item was introduced with the prompt "Show me ____."

Insert Figure 1 about here

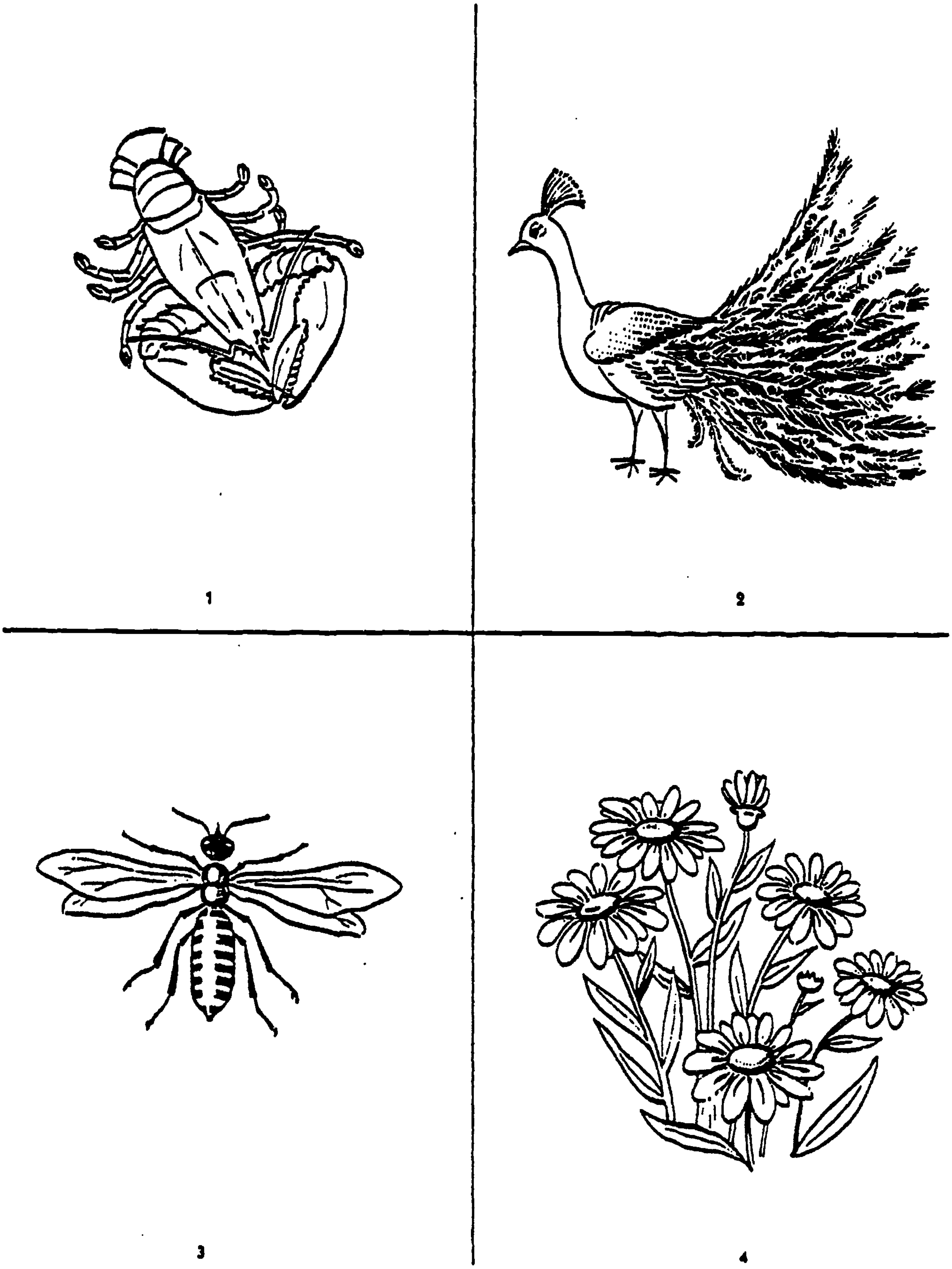


Figure 1. Sample item from the Peabody Picture Vocabulary Test: "Show me peacock."

In the example in Figure 1, the subject heard "Show me peacock". All test items were nouns, i.e., peacock, bush, key, or verbs, e.g., sewing, tackling, digging.

Insert Figure 2 about here

For the example shown in Figure 2, the subject heard "Show me sewing."

In this experiment, Kindergarten children heard 46 test items of this type and were required to point to the correct picture selected from among the four alternatives. An experimenter recorded the child's pointing response.

The results indicated that correct responses to natural stimuli exceeded the responses for synthetic stimuli. Figure 3 displays these results.

Insert Figure 3 about here

Kindergarten children pointed to the correct picture for 94% of the natural tokens and 8?% of the synthetic tokens ($t(32) = 7.62$, $p < .0001$). Only 2 out of 33 subjects made more errors on natural than synthetic speech ($p < .0001$). Most of the errors occurred towards the end of the list, a result that is not too surprising since the items on the PPVT are arranged from easy to hard. If we apply the scoring criteria used in the standardized administration conditions, that is, obtain the basal and ceiling score described earlier, 3 out of 33 subjects reached ceiling performance prior to completing the task.

The results obtained with kindergarten children, ages 5 and 6 1/2 years old, are consistent with previous results obtained from second grade children between 7 and 9 years old. Second graders showed higher percentages of correct responses overall for natural (98%) vs. for synthetic speech (94%). Second graders made fewer errors when listening to both natural and synthetic stimuli. And, like the kindergartners, the errors tended to be near the end of the list.

In summary, using a word recognition test, we found that kindergarten children understand synthetic speech produced by a text-to-speech system. However, their performance was at a significantly lower level than corresponding results for natural speech. The children had little difficulty responding to the synthetic speech as requested in this experiment. Several of the youngest children even tried to imitate the voice quality of the synthetic speech when they pointed to the picture.

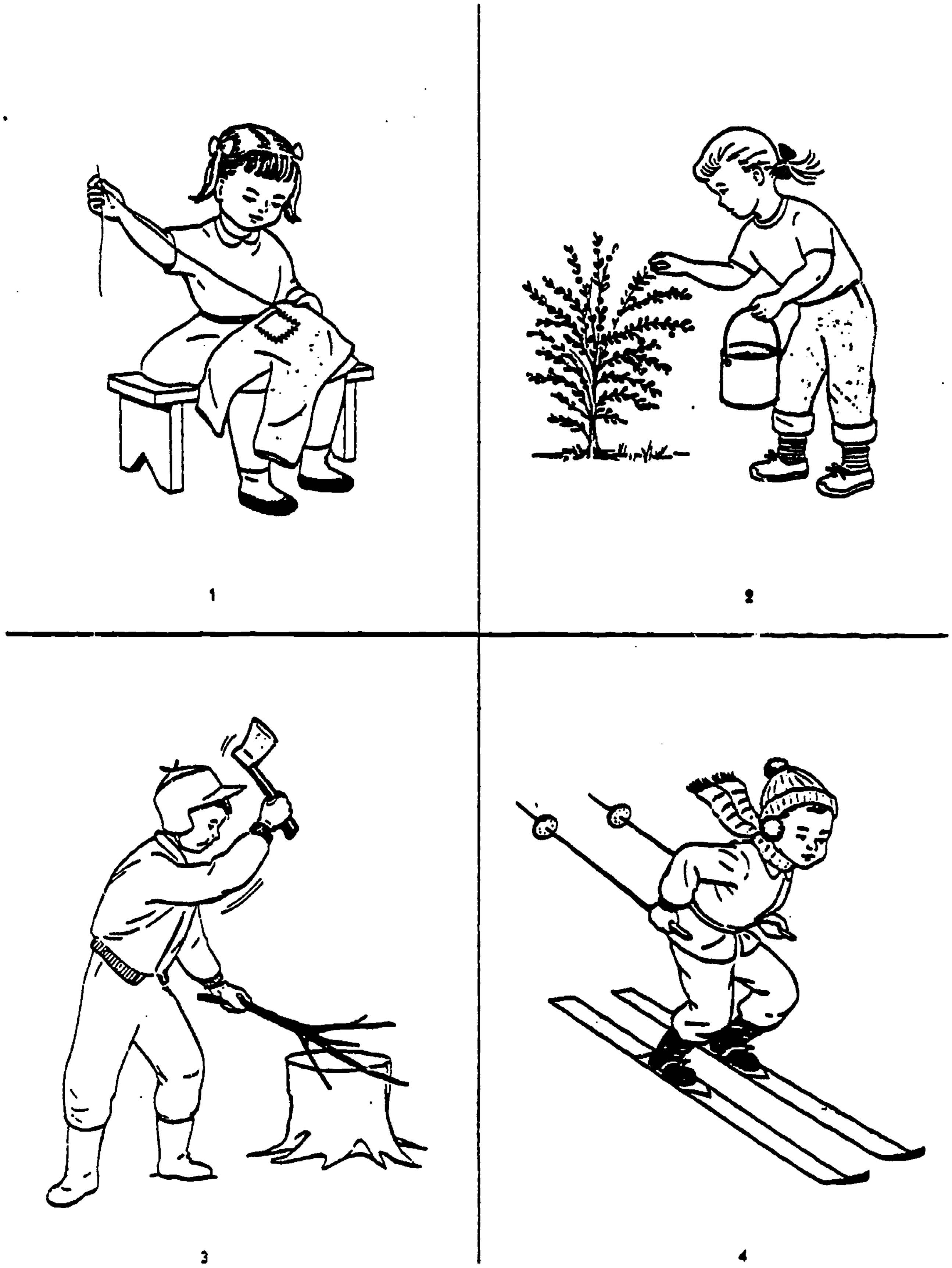


Figure 2. Sample item from the Peabody Picture Vocabulary Test: "Show me sewing."

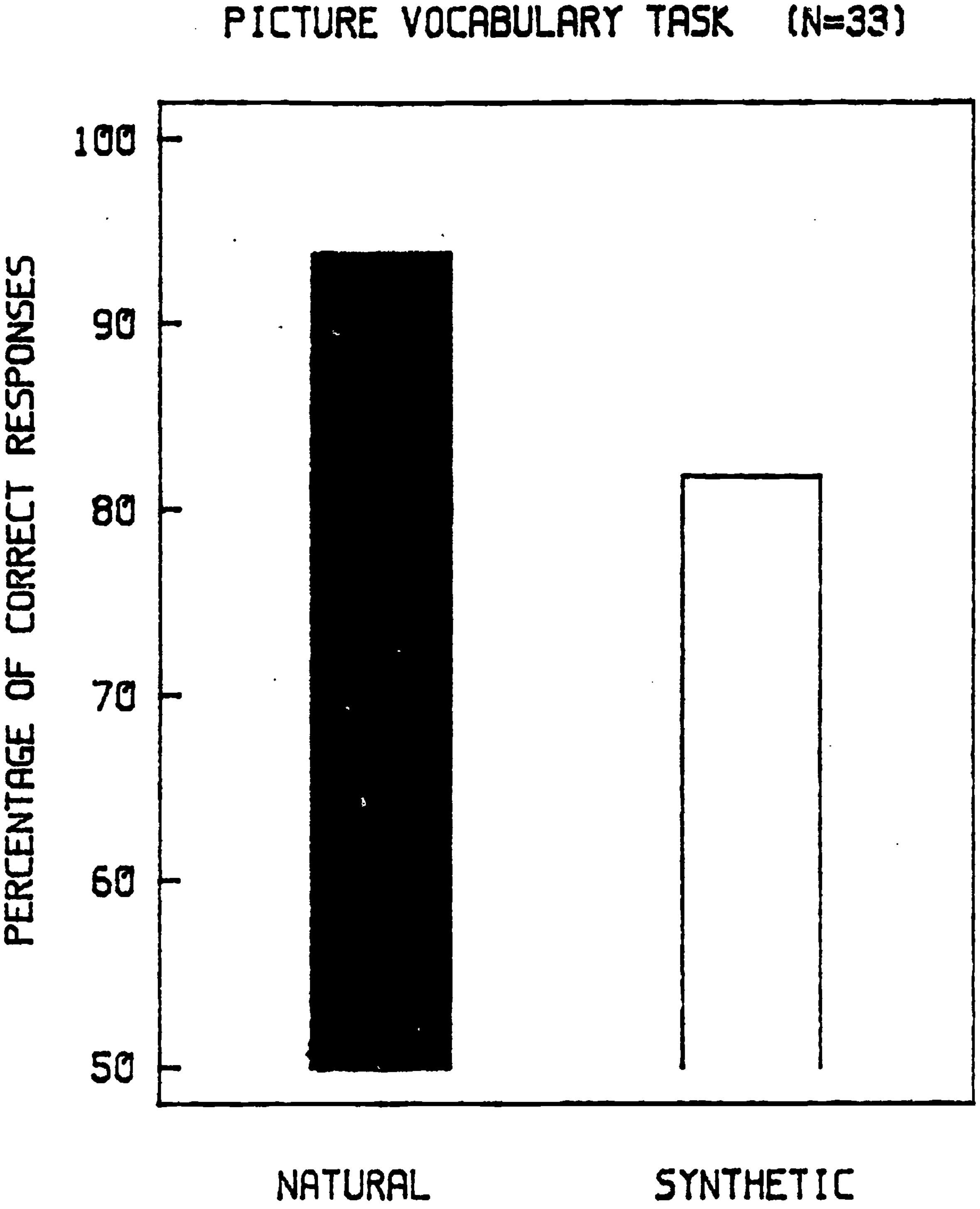


Figure 3. Percentage of correct responses for the Picture Vocabulary Task displayed for natural and synthetic speech.

Digit Span Task

We conducted two additional experiments in which the stimulus materials consisted of the digits one through nine. Sequences of digits were constructed ranging from two-digit sequences to nine-digit sequences. As in the Picture Vocabulary Task, there were both natural and synthetic tokens for each digit sequence. In the first digit-span experiment, children were required to listen to a digit sequence and repeat it aloud back to the experimenter. Sequences of two, three and four digits served as familiarization trials. Test stimuli consisted of sequences of five, six, seven, eight and nine digits. Fourth grade children between the ages of 8 1/2 and 10 served as the subjects for these experiments.

As expected the children showed decrements in performance as digit list length increased. We first scored the data for free recall. In this scoring procedure we counted items as correct if the subject recalled the items regardless of the order in which they were presented on the list. That is, if a list consisted of 5 digits and the subject repeated back all 5 of the digits presented, each item recalled would be scored as correct. Under this very liberal scoring procedure, subjects did quite well overall.

Insert Figure 4 about here

These data are shown in the left panel of Figure 4. When the list consisted of 5 digits, performance was practically error-free. However, when the list consisted of 9 digits the percentage of correct responses dropped to about 80% correct.

As shown here, percentage of correct responses decreased as the list length increased for both natural and synthetic lists. However, we did not find the expected difference between natural and synthetic digit strings.

We then applied stricter criteria for scoring the data. This procedure required than an item be scored as correct if and only if it was recalled in the exact ordinal position in which it was presented. That is, if a list consisted of 5 digits, the subject had to repeat all 5 digits in the exact order they were presented in to be scored as correct. The data are shown in the right panel of Figure 4. In contrast to the previous results, performance at list length 9 dropped to only about 30% correct. These results should be interpreted with caution, however, because the responses were given verbally by each subject. Children are likely to be unable to recall all the items in a list as the list gets longer. This is also true for adults.

In our next digit span experiment, a new group of children was required to write down the digits heard on prepared response sheets. Providing the subject with the appropriate number of blank spaces to write responses is the more traditional method for collecting serial ordered recall data. Results were again

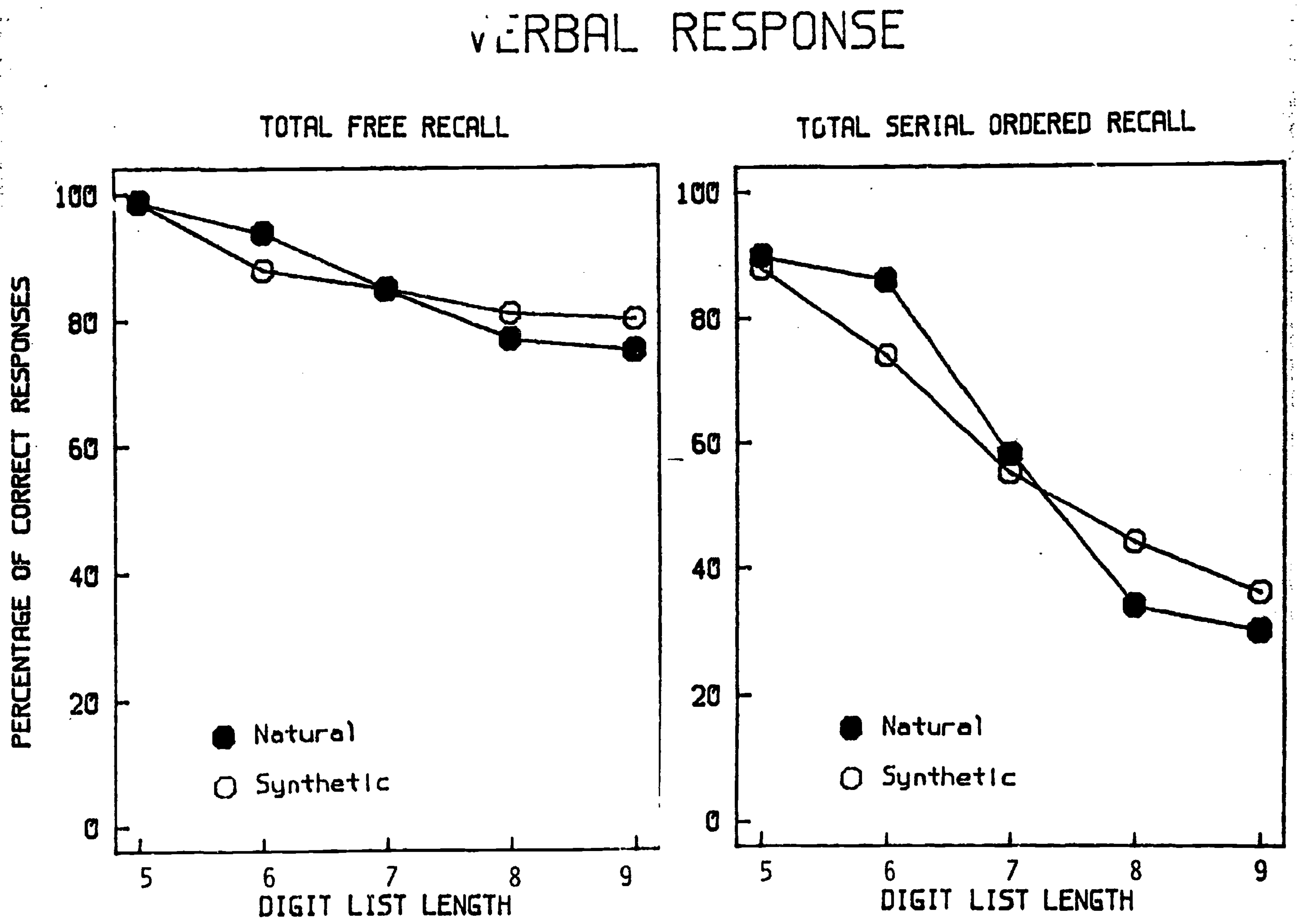


Figure 4. Percentage of correct verbal responses for the Digit Task at each list length. Data scored for free recall are displayed in the left hand panel of the figure. Data scored for serial ordered recall are displayed in the right hand panel.

scored using the two procedures already described -- free recall and serial ordered recall.

Insert Figure 5 about here

The overall findings were similar to those found with the verbal report procedure. The left hand panel of Figure 5 displays the data scored for free recall; the right panel displays serial-ordered recall scoring.

These digit list recall studies have not shown a decrement in performance due to synthetic speech as we expected from pilot data (Greene & Pisoni, 1982). Subjects do as well on natural as on synthetic speech. The use of a small highly constrained set of items, the nine digits, leads to a great deal of guessing, therefore inflating the results for the synthetic speech. The synthetic digits are quite intelligible and it is unlikely that they were misperceived or confused with each other. Furthermore, the very poor performance at the longer list lengths may simply represent a floor effect -- that is, the task was so hard that the differences in speech quality (i.e., natural vs. synthetic) could not influence performance. Both natural and synthetic digits were perceived well when list lengths were short. At longer list lengths, however, we found relatively poor performance for both natural and synthetic items. These two trends held up under both scoring procedures and across the two different response modes. We are confident in saying that fourth grade children understood natural and synthetic digits at comparable levels of performance. There is also good reason to suggest at this point that digits are not appropriate stimuli for this task since the stimulus set size is so constrained and so highly familiar even for young children.

Summary and Conclusions

In summary, the results from the Picture Vocabulary Task indicated that children performed in much the same manner as adults in listening and responding to synthetic speech. Decrements in recognition performance were observed when the input materials were generated using high-quality synthetic speech. Our experiments with natural and synthetically produced digits did not reveal the expected decrement for synthetic tokens as we had previously found in our earlier studies with adult listeners. We feel that the combination of the recall paradigm, difficulty of the task, and especially the limited stimulus set (i.e., the digits one through nine) were responsible for our failure to show reliable differences between natural and synthetic speech.

Our results on the perception of synthetic speech indicate that high-quality synthetic speech is appropriate for use in voice-response systems that may be used in teaching and learning environments with young children. Further research is currently under way in our laboratory assessing the differences in perception between natural and various kinds of synthetic and "processed" speech in a variety of experimental paradigms, with a wide range of stimulus material.

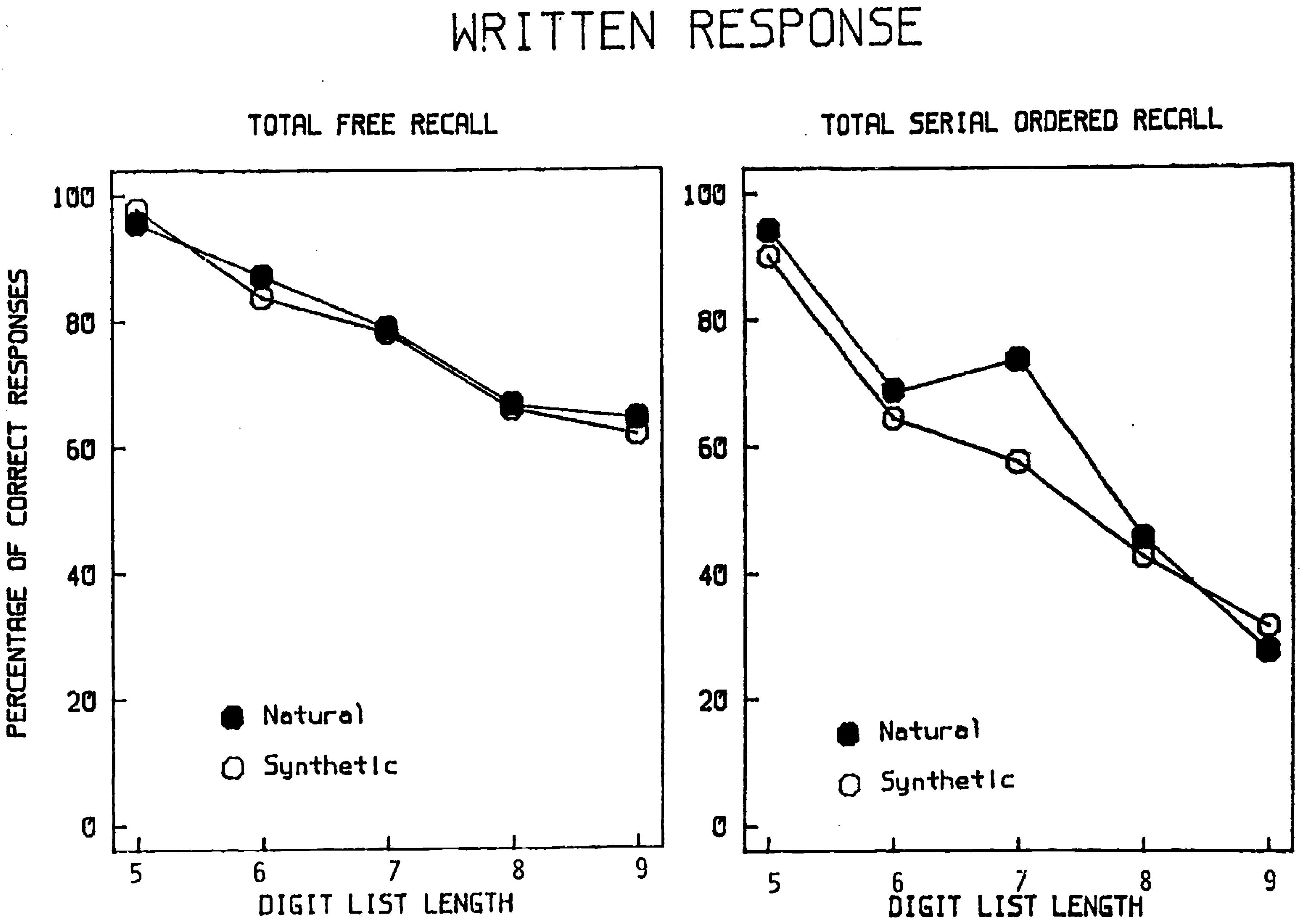


Figure 5. Percentage of correct written responses for the Digit Task at each list length. Data scored for free recall are displayed in the left hand panel of the figure. Data scored for serial ordered recall are displayed in the right hand panel.

References

Greene, B. G. & Pisoni, D. B. Perception of Synthetic Speech by Children: A First Report. In *Research on Speech Perception*. *Progress Report No. 8*. Bloomington, Indiana: Speech Research Laboratory, Indiana University, 1982.

Groner, G. F., Bernstein, J., Ingber, E., Pearlman, J., & Toal, T. A real-time text-to-speech converter. *Speech Technology*, 1982, *1*, 2, 73-76.

Laddaga, R., Sanders, W. R., & Suppes, P. Testing intelligibility of computer-generated speech with elementary-school children. In P. Suppes (Ed.) *University-level computer-assisted instruction at Stanford: 1968-1980*. Stanford, CA: Institute for Mathematical Studies in the Social Sciences, Stanford University, 1981.

Luce, P. A., Feustel, T. C., & Pisoni, D. B. Capacity demands in short term memory for synthetic and natural speech. *Human Factors*, 1983, *25*, 17-32.

Nusbaum, H. C. & Pisoni, D. B. Perceptual and cognitive constraints on the use of voice response systems. In L. Lerman (Ed.), *Proceedings of the Conference on Voice Data Entry Systems*. Sunnyvale, CA: Lockheed, 1982.

Pisoni, D. B. Perception of synthetic speech: Some contributions by the human listener. *Speech Technology*, 1982, *1*, 2, 10-23.

Perceptual Evaluation of Synthetic Speech:

Some Considerations of the User/System Interface*

David B. Pisoni, Howard C. Nusbaum,

Paul A. Luce, and Eileen C. Schwab

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

*This paper was presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Boston, 1983. This research was supported by NIH grant NS-12179, and by contract No. AF-F-33615-83-K-0501 with the the Air Force Systems Command, AFOSR through the Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio.

Abstract

With the rapid increase in the use of voice response systems in commercial, industrial, military, and educational applications, it has become important to understand how humans interact with devices that produce synthetic speech output. This paper describes a series of experiments that have examined the differences in perception of natural and synthetic speech. Our results suggest that important perceptual and cognitive limitations are imposed when synthetic speech is used in a variety of psychological tasks ranging from phoneme recognition to word recognition to spoken language comprehension. Moreover, these differences are manifest not only in terms of measures of response accuracy but also in estimates of the cognitive processing time required to execute responses to synthetic speech signals.

Perceptual Evaluation of Synthetic Speech:

Some Consideration of the User/System Interface

BACKGROUND

Within the next few years, there will be an extensive proliferation of various types of voice response devices in human-machine communication systems. Such systems will no doubt be employed for a variety of commercial, industrial, and military applications. Unfortunately, at the present time, there has been relatively little basic or applied research on the intelligibility, comprehension and perceptual processing of synthetic speech produced by these devices. Moreover, there has been almost no research directed at understanding how human observers interact with this new technology. The prevailing assumption has been that simply providing automated voice response output and voice data entry will solve most of the human factors problems inherent in the user-system interface. Not only is this assumption untested, but we believe that it is simply false. In many cases, the introduction of voice response and voice data entry systems may create an entirely new set of human factors problems. To understand how the user will interact with these speech I/O devices, we will need to understand much more about how the human perceives, encodes, stores, and retrieves speech and how these basic information processing operations interact with the specific tasks the observer is required to perform.

We have been carrying out a program of research that is aimed, in part, at studying how listeners perceive and understand synthetic speech under a variety of task demands and conditions. In conceptualizing the problems, we considered three factors that control an observer's performance: (1) inherent limitations of the human information processing system, (2) constraints on the structure and quality of the speech signal, and (3) the specific task demands.

In considering the first first factor, it is well known that there are substantial limitations on the human information processing system's ability to perceive, encode, store, and retrieve information. Because the nervous system cannot maintain all aspects of sensory stimulation and therefore must integrate acoustic energy over time, very severe processing limitations have been found in the capacity to encode and store raw sensory data in the human memory system. To overcome these capacity limitations, the listener must rapidly recode and transform sensory input into more abstract neural codes for stable storage in memory and subsequent processing operations. The bulk of the research over the last 25 years has identified human short-term memory (STM) as the major source of the limitation on processing sensory input. The amount of information that can be processed in and out of STM is severely limited by the listener's attentional state, past experience, and the quality of the sensory input. Many of the inherent limitations of STM can be overcome, however, by the redundancy of spoken language and the listener's access to multiple sources of knowledge. These factors contribute substantially to support perception of speech under very adverse listening conditions.

The second factor concerns the structure of the physical signal itself. Speech signals may be thought of as the physical consequence of a complex and hierarchically organized system of linguistic rules that map sounds onto meanings and meanings back onto sounds. At the lowest level in the system, the distinctive properties of the speech signal are constrained in substantial ways by vocal tract acoustics and articulation. The choice and arrangement of speech sounds in words is constrained by the phonological rules of language; the arrangement of words in sentences is constrained by syntax; and finally, the meaning of individual words and the overall meaning of sentences in a text is constrained by semantics and pragmatics. The contribution of these various levels of linguistic structure to perception will vary substantially from isolated words, to sentences, to passages of fluent continuous speech.

Finally the third factor deals with the specific task demands confronting the human listener. Human observers are capable of rapidly developing extremely specialized and efficient perceptual and cognitive strategies to maximize performance under different task demands. There is substantial data in the literature demonstrating the powerful effects of perceptual set, the role of different instructions, the effects of subjective expectancies, and the influence of long-term familiarity and practice on a variety of perceptual and cognitive tasks. Human observers are capable of varying the "depth of processing" that a stimulus receives depending on the requirements of the task. The amount of context and the degree of uncertainty in the task also strongly affect an observer's performance in processing information. In short, human observers are capable of adopting very different strategies depending on the needs and requirements of the tasks presented to them. The study of these strategies and an analysis of their requirements is crucial to evaluating the overall performance of observers using synthetic speech produced by voice response devices.

COGNITIVE PROCESSING TIME

In the time since our earlier report to the IEEE meeting in 1980 (Pisoni and Hunnicutt, 1980), we have continued to carry out detailed experimental studies comparing the perception of natural and synthetic speech in a wide variety of psychological paradigms. Our initial evaluation of the synthetic speech generated by the MITalk system relied entirely on performance measures involving response accuracy. Recently, we (Pisoni, 1981; Slowiaczek and Pisoni, 1981) completed a more sophisticated series of experiments that was aimed at measuring the time required to recognize natural and synthetic words and permissible nonwords. In carrying out these studies, we wanted to know how long it takes a human listener to recognize an isolated wcrd and how the process of word recognition might be affected by the quality of the initial acoustic-phonetic information in the signal. To measure how long it takes an observer to recognize isolated words, we used a lexical decision task. In this task subjects are presented with either a word or a nonword stimulus item on each trial. The listener is required to classify the item as either a "word" or a "nonword" as fast as possible by pressing one of two buttons located on a response panel.

The results showed that performance was more accurate for natural test items (98% correct) than for synthetic test items (78% correct). Moreover, this difference was present for both word and nonword test items. The mean reaction times for correct responses also showed significant differences between synthetic and natural test items. Subjects responded significantly faster to natural words (903 msec) and nonwords (1046 msec) than to synthetic words (1056 msec) and nonwords (1179 msec). On the average, reaction times to the synthetic speech took 145 msec longer than response times to the natural speech.

These findings demonstrate that the perception of synthetic speech requires more cognitive "effort" than the perception of natural speech. This difference was observed for both word and nonwords alike, suggesting that the extra processing does not depend on the lexical status of the test item. Thus, the phonological encoding of synthetic speech appears to require more effort than the encoding of natural speech.

In a more recent study, Slowiaczek and Pisoni (1981) used the same lexical decision procedure but gave the subjects five days of practice at the task. They found that although overall performance improved for all test items, the reaction time difference between natural and synthetic speech remained roughly the same. This is consistent with the conclusion that it is the perceptual encoding of the test items that is responsible for the reaction time difference. Furthermore, this result indicates that the processing of synthetic speech is a "data-limited" process; that is, the limitation may be in the structure of the synthetic speech itself.

In a third experiment, subjects were required to name synthetic and natural test items. The time required to make the naming response was measured together with its accuracy. As expected, subjects made more errors on synthetic speech. In addition, they were much slower to name synthetic test items than natural test items. Once again, subjects required more time to name synthetic words and nonwords than natural words and nonwords. These results demonstrate that the extra processing time needed for synthetic speech does not depend on the type of response made by the listener since the results were comparable for both manual and vocal responses. Our findings demonstrate that early stages of encoding synthetic speech require more computation than encoding natural speech.

SHORT-TERM MEMORY AND PROCESSING CAPACITY

Recently, we completed several experiments that were designed to study the effects of processing synthetic speech on the capacity of short-term memory (Luce, Feustel and Pisoni, 1983). In one experiment, on each trial, subjects were given two different lists of items to remember. The first list consisted of a set of digits presented visually on a CRT screen. On some trials no digits were presented and on other trials there were either three or six digits in the display. Following the visual list, subjects were presented with a spoken list of ten natural words or ten synthetic words. After the spoken list was presented, the subjects were instructed to first write down all the digits in the order of presentation and then write down all the words they could remember from the auditory list.

For all three visual digit list conditions (no list, three or six digits), recall of the natural words was significantly better than recall of the synthetic words. In addition, recall of the synthetic and natural words became worse as the size of the digit lists increased. In other words, increasing the number of digits held in memory impaired the subjects' ability to recall the words. But the most important finding was that there was an interaction between the type of speech presented (synthetic vs. natural) and the number of digits presented (three vs. six). This interaction was reve^led by the number of subjects who could recall all the digits presented in correct order. As the size of the digit lists increased, significantly fewer subjects were able to recall all the digits for the synthetic words compared to the natural words. Synthetic speech impaired recall of the visually presented digits more with increasing digit list size than did natural speech. These results demonstrate that synthetic speech required more short-term memory capacity than natural speech. As a result, it would be expected that synthetic speech should interfere much more with other cognitive processes.

In another experiment, we presented subjects with lists of ten words to be memorized. The lists were either all synthetic or all natural words. The subjects were required to recall the words in the same order as the original presentation. As in the previous experiment, overall, the natural words were recalled better than the synthetic words. However, a more detailed analysis revealed that in the second half of the lists, recall of synthetic and natural speech was the same. The difference in ordered recall performance between natural and synthetic speech was confined to the initial portion of the list. The first synthetic words heard in the list were recalled less often than the natural words in the beginning of the lists. This result suggests that, in the synthetic lists, the words presented later in each list interfered with active maintenance of the words presented earlier. This is precisely the result that would be expected if the perceptual encoding of the synthetic words placed an additional load on short-term memory, thus impairing the rehearsal of words presented in the first half of the list.

LISTENING COMPREHENSION

In our earlier experiments on comprehension of fluent passages of synthetic speech, listeners only had to understand the general ideas presented in the materials (Pisoni, 1984). Comprehension was measured with multiple-choice questions. If the subjects could understand only some parts of the passages, they could use previous knowledge to make inferences about the rest of the text. Thus, these experiments were not able to distinguish between information that was acquired by listening to the text and knowledge that the subjects might have had prior to the experiment. Recently, Luce (1981) conducted a more detailed examination of the comprehension of fluent synthetic speech. In this study, specific questions were designed to probe four different levels of comprehension. Surface structure questions were constructed to determine if listeners had heard a specific word in the spoken text. Low proposition questions queried specific details or facts explicitly stated in the passage. High proposition questions probed understanding of themes or messages in the text. Finally, inference questions required listeners to form a conclusion that had not been explicitly stated in the passage. These questions were presented visually following each passage.

One group of subjects heard synthetic passages; another group listened to natural versions of the same texts. Speed and accuracy of question answering were measured. Although subjects responded with comparable latencies to questions for natural and synthetic passages, there were significant differences in the accuracy of question answering. Subjects were less accurate in answering inference questions and high and low proposition questions for the synthetic passages. In contrast to our previous findings with multiple-choice questions, this indicates that comprehension of the natural passages was better than comprehension of the synthetic speech. However, there was a surprising result.

Subjects who heard the synthetic speech were more accurate at answering surface structure questions than the natural speech subjects. This indicates that the subjects who listened to the natural speech remembered fewer specific words from the passages than the subjects who listened to the synthetic speech.

The subjects who listened to synthetic speech the may have spent so much time and effort trying to understand each of the words in the text that they were unable to do anything else. Indeed, this effort to perceive and encode the words may have made them more memorable. On the other hand, the group that heard natural speech probably had no problems understanding the words in the passages so they could concentrate more effort on understanding the ideas of the passages. Previous research has shown that during sentence comprehension, the surface structure is quickly forgotten while the underlying concepts are retained.

The results of this experiment demonstrate that while listeners may understand the gist of simple synthetic and natural passages equally well, it is substantially harder to comprehend synthetic speech at more abstract levels. The reason for this difficulty may be that it is harder to encode synthetic words than natural words. This seems to be true even though the listeners should be able to use a great deal of prior knowledge to aid in word recognition.

PERCEPTUAL LEARNING

As we noted earlier, the human observer is a very flexible processor of information. With sufficient experience, practice, and specialized training, observers may be able to overcome some of the limitations we have observed in the previous studies.

Previous research using synthetic speech has provided some evidence to indicate that intelligibility can be enhanced by training. One early study reported by Carlson, Granstrom, and Larsson (1976) found that the intelligibility of a text-to-speech system significantly improved with practice listening to the synthetic speech. Similarly, Pisoni and Hunnicutt (1980) found that subjects' ability to comprehend synthetic passages and recognize synthetic words generated by MITalk improved with more experience listening to the text-to-speech system. In both of these studies, subjects were not given any explicit feedback about their performance.

More recently, Slowiaczek and Pisoni (1981) reported that performance in a lexical decision task for natural and synthetic speech improved with five days of practice. However, they found that the improvement occurred for both synthetic and natural stimuli (words and nonwords) suggesting that subjects may have been learning to perform the task better instead of enhancing intelligibility. In the previous experiments where improved intelligibility with synthetic speech was found with practice (Carlson et al., 1976; Pisoni and Hunnicutt, 1980), this improvement might be attributable to learning the tasks rather than learning to perceive the synthetic speech.

To test this possibility, Schwab, Nusbaum, and Pisoni (1984) conducted an experiment using three groups of subjects. All three groups were tested with synthetic speech produced by the Votrax Type-'N-Talk on the first day of the experiment (pre-test) and then ten days later (post-test). In the intervening period, one group received training with the Votrax synthetic speech, one group received the identical training procedures with natural speech, and one group received no training at all. Thus, since all groups were tested on Day 1 and Day 10 with synthetic speech but only one group was specifically trained on synthetic speech, we could separate any improvement in task performance from the effects of training on intelligibility. On the test days, the subjects were given the Modified Rhyme Test (MRT), synthetic words and sentences for transcription, and spoken passages of connected fluent speech followed by comprehension testing. The word lists were taken from phonetically balanced (PB) lists and the sentences were either syntactically and semantically correct (Harvard sentences) or they were syntactically correct but semantically anomalous (Haskins sentences). Subjects were given different materials on every day. During training, subjects were presented with spoken words, sentences, and passages, and they received feedback.

For the MRT (6-alternative forced choice) and the PB lists (open transcription), the results show that performance improved significantly for only one group on both tasks -- the subjects that were trained on the Votrax synthetic speech. Both of the control groups (the group trained on natural speech and the group that received no training) did not show any improvement in word recognition performance with the synthetic speech.

Our findings demonstrate that it is possible to increase intelligibility of poor-quality synthetic speech markedly with training and familiarity. However, our results also indicated that despite significant improvements in the intelligibility of synthetic words and sentences, no improvements occurred for comprehension of the passages. One possible reason for this was suggested by the comments of many of our subjects on a questionnaire administered after the experiment: The subjects devoted so much effort trying to perceive each word that it was extremely difficult to pay attention to the passage. Our results demonstrate that even though it may be possible to improve the intelligibility of synthetic words, the capacity demands imposed by perception of synthetic speech still remain and affect processes required for comprehension and understanding of the contents of the linguistic message.

CONCLUSIONS

Our results on the perception of synthetic speech obviously have important implications for the design, selection, and use of voice response systems in a variety of applications. Moreover, such results are quite relevant to the very basic questions surrounding the user-system interface involving natural language under various conditions of information overload -- particularly in situations requiring divided attention among several input modalities which must be monitored simultaneously while carrying out other complex cognitive tasks. Our experiments demonstrate that the important interactions in performance occur among the signal, the observer's task, and the capacity of the observer. Moreover, while our experiments have shown substantial effects of perceptual learning on word recognition, these improvements do not appear to generalize to more abstract levels required for comprehension of the linguistic message.

REFERENCES

Carlson, R., Granstrom, B. and Larsson, K. Evaluation of a text-to-speech system as a reading machine for the blind. Speech Transmission Laboratory, Quarterly Progress and Status Report 2-3, 1976.

Luce, P. A. Comprehension of fluent synthetic speech produced by rule. In Research on Speech Perception. Progress Report No. 7. Bloomington, Indiana: Speech Research Laboratory, Indiana University, 1981.

Luce, P. A., Feustel, T. C. and Pisoni, D. B. Capacity demands in short-term memory for synthetic and natural word lists. Human Factors, 1983, 25, 17-32.

Pisoni, D. B. Some measures of intelligibility and comprehension. In J. Allen (Ed.), Conversion of Unrestricted English Text-to-Speech. 1984 (In Press).

Pisoni, D. B. Speeded classification of natural and synthetic speech in a lexical decision task. Journal of the Acoustical Society of America, 1981, 70, S98.

Pisoni, D. B. and Hunnicutt, S. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. 1980 IEEE International Conference Record on Acoustics, Speech and Signal Processing, April, 1980. Pp. 572-575.

Schwab, E. C., Nusbaum, H. C. and Pisoni, D. B. Some effects of training on the perception of synthetic speech. Human Factors, 1984 (In Press).

Slowiaczek, L. M. and Pisoni, D. B. Effects of practice on speeded classification of natural and synthetic speech. In Research on Speech Perception. Progress Report No. 7. Bloomington, Indiana: Speech Research Laboratory, Indiana University, 1981.

III. Publications

Sinnott, J. M., Pisoni, D. B. and Aslin, R. N. Pure Tone Thresholds in the Human Infant and Adult. Infant Behavior and Development, 1983, 6, 3-17.

Pisoni, D. B. Perceptual evaluation of voice response systems: Intelligibility, Recognition and Understanding. Proceedings of the Workshop on Standardization for Speech I/O Technology. Washington, D. C.: National Bureau of Standards, 1982. Pp. 183-192.

Remez, R. E., Rubin, P. E. and Pisoni, D. B. Coding of the speech spectrum in three time-varying sinusoids. In C. W. Parkins & S. W. Anderson (Eds.), Cochlear Implantation. New York: New York Academy of Sciences, Vol. 405, 1983, 485-489.

Luce, P. A., Feustel, T. C. and Pisoni, D. B. Capacity demands in short-term memory for synthetic and natural word lists. Human Factors, 1983, 25, 1, 17-32.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A. and Schwab, E. C. Perceptual evaluation of synthetic speech: Some considerations of the user/system interface. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Boston, April 1983. Pp. 535-538.

McClaskey, C. L., Pisoni, D. B. and Carrell, T. D. Effects of Transfer of Training on Identification of a New Linguistic Contrast. Perception & Psychophysics, 1983, 34, 4, 323-330.

Aslin, R. N., Pisoni, D. B. and Jusczyk, P. W. Auditory development and speech perception in infancy. In P. Mussen (Ed.), Carmichael's Manual of Child Psychology, 4th Edition, Volume II: Infancy and the Biology of Development, M. M. Haith and J. J. Campos (Vol. II Editors). New York: Wiley and Sons, 1983. Pp. 573-687.

Kewley-Port, D. Time-varying features as correlates of place of articulation in stop consonants. Journal of the Acoustical Society of America, 1983, 73, 322-335.

Walley, A. C. and Carrell, T. D. Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. Journal of the Acoustical Society of America, 1983, 73, 1011-1022.

Pisoni, D. B., Carrell, T. D. and Gans, S. J. Perception of the Duration of Rapid Spectrum Changes: Evidence for Context Effects with Speech and Nonspeech Signals. Perception & Psychophysics, 1983, 34, 4, 314-322.

Kewley-Port, D., Pisoni, D. B. and Studdert-Kennedy, M. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. Journal of the Acoustical Society of America, 1983, 73, 5, 1779-1793.

Green, B. G., Craig, J. C., Wilson, A. M., Pisoni, D. B. and Rhodes, R. P. Vibrotactile identification of vowel spectra. Journal of the Acoustical Society of America, 1983, 73, 5, 1766-1778.

Green, B. G., Craig, J. C. and Pisoni, D. B. Vibrotactile communication of information about consonants: Vowels mask consonants. Perception & Psychophysics, 1983, 33, 6, 507-515.

Jusczyk, P. W., Pisoni, D. B., Reed, M., Fernald, A. and Myers, M. Infants' Discrimination of the Duration of a Rapid Spectrum Change in Nonspeech Signals. Science, 1983, 222, 175-177.

Nusbaum, H. C., Schwab, E. C. and Pisoni, D. B. Perceptual evaluation of synthetic speech: Some constraints on the use of voice response systems. Proceedings of the 3rd Voice Data Entry Systems Applications Conference. Sunnyvale, CA: Lockheed, 1983.

Nusbaum, H. C., Schwab, E. C. and Sawusch, J. R. The role of "chirp" identification in duplex perception. Perception & Psychophysics, 1983, 33, 4, 323-332.

Sawusch, J. R. and Nusbaum, H. C. Auditory and phonetic processes in place perception for stops. Perception & Psychophysics, 1983, 34, 6, 560-568.

Manuscripts to be published:

Pisoni, D. B. Some measures of intelligibility and comprehension. In J. Allen (Ed.), Conversion of Unrestricted English Text-to-Speech. 1984 (In Press).

Pisoni, D. B. Speech Perception: Research, Theory and the Principal Issues. In E. C. Schwab and H. C. Nusbaum (Eds.), Perception of Speech and Visual Form: Theoretical Issues, Models and Research. New York: Academic Press, 1984 (In Press).

Pisoni, D. B. Categorical perception of speech and nonspeech signals. In S. Harnad (Ed.), Categorical Perception. New York: Cambridge University Press, 1984 (In Press).

Pisoni, D. B. Contextual variability and the problem of acoustic-phonetic invariance in speech. In J. S. Perkell & D. H. Klatt (Eds.), Invariance and Variability of Speech Processes. Hillsdale, NJ: Lawrence Erlbaum Associates, 1984 (In Press).

Pisoni, D. B. Speech Perception: Some New Directions in Research and Theory. Journal of the Acoustical Society of America, 1984 (Special Supplement of Invited Talks Prepared for CHABA).

Schwab, E. C., Nusbaum, H. C. and Pisoni, D. B. Some effects of training on the perception of synthetic speech. Human Factors, 1984 (In Press).

Salasoo, A. and Pisoni, D. B. Sources of knowledge in spoken word identification. Journal of Verbal Learning and Verbal Behavior, 1984 (In Press).

Kewley-Port, D. and Pisoni, D. B. Discrimination of rise-time in nonspeech signals: Is it categorical or noncategorical? Journal of the Acoustical Society of America, 1984 (In Press).

Walley, A. C., Pisoni, D. B. and Aslin, R. N. Onset spectra and formant transitions in the infant's discrimination of place of articulation for syllable-initial stop consonants. Journal of the Acoustical Society of America, 1984 (In Press).

Greene, B. G., Pisoni, D. B. and Carrell, T. D. Recognition of speech spectrograms. Journal of the Acoustical Society of America, 1984 (In Press).

Kewley-Port, D. and Luce, P. A. Time-varying features of initial stop consonants in auditory running spectra: A first report. Perception & Psychophysics, 1984 (In Press).

Kewley-Port, D. Converging approaches towards establishing invariant acoustic-phonetic cues. In J. S. Perkell & D. H. Klatt (Eds.), Invariance and Variability of Speech Processes. Hillsdale, NJ: Lawrence Erlbaum Associates, 1984 (In Press).

Luce, P. A. and Charles-Luce, J. Contextual effects on the consonant/vowel ratio in speech production. Journal of the Acoustical Society of America, 1984 (In Press).

Dinnsen, D. A. and Charles-Luce, J. Phonological neutralization, phonetic implementation and individual differences. Journal of Phonetics, 1984 (In Press).

Nusbaum, H. C. Possible mechanisms of duplex perception: "Chirp" identification versus dichotic fusion. Perception & Psychophysics, 1984 (In Press).

Jusczyk, P. W., Shea, S. L. and Aslin, R. N. A Re-examination of Eilers, Gavin and Oller (1982). Journal of Child Language, 1984 (In Press).

IV. Speech Research Laboratory Staff, Associated Faculty and Technical Personnel:

(1/1/83 - 12/31/83)

Research Personnel:

David B. Pisoni, Ph.D.	Professor of Psychology and Director
Richard N. Aslin, Ph.D.	Professor of Psychology
Beth G. Greene, Ph.D.	Assistant Research Scientist
Diane Kewley-Port, Ph.D.	Research Associate
Michael R. Petersen, Ph.D.	Associate Professor of Psychology
Eileen C. Schwab, Ph.D.	Visiting Assistant Professor*
Rebecca Treiman, Ph.D.	Assistant Professor of Psychology
Cathy A. Kubaska, Ph.D.	NIH Post-doctoral Fellow
Howard C. Nusbaum, Ph.D.	NIH Post-doctoral Fellow
Thomas D. Carrell, M.A.	Graduate Research Assistant**
Jan Charles-Luce, B.A.	Graduate Research Assistant
Paul A. Luce, B.A.	Graduate Research Assistant
Peter Mimmack, M.A.	Graduate Research Assistant
Louisa M. Slowiaczek, B.A.	Graduate Research Assistant
Amanda C. Walley, B.A.	Graduate Research Assistant

Technical Support Personnel:

Debbie Acuff	Secretary
Robert H. Bernacki	Programmer/Research Assistant
Mary Buuck, A.A.	Research Assistant (Infant Laboratory)
Michael S. Dedina, B.A.	Research Assistant
Jerry C. Forshee, M.A.	Computer Systems Analyst
Thomas J. Jonas, B.A.	Programmer
Nancy J. Layman	Administrative Secretary
David A. Link	Electronics Engineer
Laura M. Manous, B.A.	Research Assistant
Kent Burns	Undergraduate Research Assistant
Lynn Krevitz	Undergraduate Research Assistant
Laurie Ann Walker	Undergraduate Research Assistant

*Now at ATT Consumer Products, Indianapolis, IN

**Now at Massachusetts Institute of Technology, Cambridge, MA