DOCUMENT RESUME

ED 318 054 CS 507 125

AUTHOR Pisoni, David B.; And Others
TITLE Research on Speech Perception. Progress Report No. 8, January 1982-December 1982.
INSTITUTION Indiana Univ., Bloomington. Dept. of Psychology.
SPONS AGENCY National Institutes of Health (DHHS), Bethesda, Md.; National Inst. of Mental Health (DHHS), Rockville, MD.
PUB DATE 82
GRANT MH-24027-07; NS-07134-04; NS-12179-07
NOTE 326p.; 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/PC14 Plus Postage.
DESCRIPTORS *Acoustic Phonetics; Auditory Discrimination; *Auditory Perception; Communication Research; Computer Software Development; Infants; *Language Processing; Language Research; Linguistics; Speech
IDENTIFIERS Indiana University Bloomington; *Speech Perception; Speech Research; Theory Development

ABSTRACT

Summarizing research activities from January 1982 to December 1982, this is the eighth 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, progress reports, and information on instrumentation developments and software support. The report contains the following 15 articles: "Acoustic-Phonetic Priming in Auditory Word Recognition: Some Tests of the Cohort Theory" (L. M. Slowiaczek and D. B. Pisoni); "Sentence-by-Sentence Listening Times for Spoken Passages: Test Structure and Listeners' Goals" (P. C. Mimmack and others); "Effects of Syllable Structure on Adults' Phoneme Monitoring Performance" (R. Treiman and others); "Controlled Perceptual Strategies in Phonemic Restoration" (H. C. Nusbaum and others); "Sources of Knowledge in Spoken Word Identification" (A. Salasoo and D. B. Pisoni); "Effects of Perceptual Load in Spoken Comprehension: Some Interactions with Comprehension Goals" (H. Brunner and others); "Cognitive Processes and Comprehension Measures in Silent and Oral Reading" (A. Salasoo); "Perceptual and Cognitive Constraints on the Use of Voice Response Systems" (H. C. Nusbaum and D. B. Pisoni); "Perceptual Anchoring of a Speech-Nonspeech Continuum" (H. C. Nusbaum); "Perceiving Durations of Silence in a Nonspeech Context" (H. C. Nusbaum); "Perception of Synthetic Speech by Children: A First Report" (B. G. Greene and D. B. Pisoni); "Context Effects in the Perception of English /r/ and /l/ by Japanese" (P. Dissosway-Huff and others); "An Activation Model of Auditory Word Recognition" (H. C. Nusbaum and L. M. Slowiaczek); "JOT: Improved Graphics Capabilities for KLTEXC" (B. Bernacki); and "EARS: A Simple Auditory Screening Test" (L. A. Walker). Lists of publications and of laboratory staff, associated faculty and personnel conclude the report. (SR)

ED318054

RESEARCH ON SPEECH PERCEPTION

Progress Report No. 8
January 1982 — December 1982

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-12179-07
and
National Institutes of Health
Training Grant No. NS-07134-04

CS507125

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. B. PISONI

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

BEST COPY AVAILABLE

RESEARCH ON SPEECH PERCEPTION

Progress Report No. 8

January 1982 - December 1982

David B. Pisoni

Principal Investigator

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-12179-07

and

National Institutes of Health
Training Grant No. NS-07134-04

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

Table of Contents

Introduction iii

I. Extended Manuscripts 1

Acoustic-Phonetic Priming in Auditory Word Recognition: Some Tests of the Cohort Theory; Louisa M. Slowiaczek and David B. Pisoni 3

Sentence-by-sentence listening times for spoken passages: Test structure and listeners' goals; Peter C. Mimmack, David B. Pisoni and Hans Brunner 27

Effects of Syllable Structure on Adult's Phoneme Monitoring Performance; Rebecca Treiman, Aita Salasoo, Louisa M. Slowiaczek and David B. Pisoni 63

Controlled Perceptual Strategies in Phonemic Restoration Howard C. Nusbaum, Amanda C. Walley, Thomas D. Carrell and William Ressler 85

Sources of Knowledge in Spoken Word Identification Aita Salasoo and David B. Pisoni 105

Effects of Perceptual Load in Spoken Comprehension: Some Interactions with Comprehension Goals Hans Brunner, Peter C. Mimmack, Alford R. White and David B. Pisoni 147

Cognitive Processes and Comprehension Measures in Silent and Oral Reading; Aita Salasoo 185

Perceptual and Cognitive Constraints on the Use of Voice Response Systems; Howard C. Nusbaum and David B. Pisoni 203

Perceptual Anchoring of a Speech-Nonspeech Continuum Howard C. Nusbaum 217

II. Short Reports and Work-in-Progress 247

Perceiving Durations of Silence in a Nonspeech Context Howard C. Nusbaum 249

Perception of Synthetic Speech by Children: A First Report Beth G. Greene and David B. Pisoni 261

ERIC

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

Context Effects in the Perception of English /r/ and /l/ by Japanese; Patricia Dissosway-Huff, Robert F. Port and David B. Pisoni 277

An Activation Model of Auditory Word Recognition Howard C. Nusbaum and Louisa M. Slowiaczek 289

III. Instrumentation and Software Development 307

JOT: Improved Graphics Capabilities for KLTEXC Bob Bernacki 309

EARS: A Simple Auditory Screening Test Laurie Ann Walker 319

IV. Publications 325

V. Laboratory Staff, Associated Faculty and Personnel 327

INTRODUCTION

This is the eighth annual report of research activities on speech perception, analysis and synthesis conducted in the Speech Research Laboratory of the Department of Psychology at 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 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 other colleagues.

We are distributing reports of our research activities primarily 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. 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, we 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 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.

I. EXTENDED MANUSCRIPTS

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

Acoustic-Phonetic Priming in Auditory Word Recognition: Some Tests of the Cohort Theory*

Louisa M. Slowiaczek and David B. Pisoni

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

*The work reported here was supported, in part, by NIMH grant MH-24027 and NIH grant NS-12179 to Indiana University. We would like to thank Paul A. Luce for his assistance in recording and preparing the stimuli and Thomas D. Carrell for his help with the experimental control programs. We also want to thank Howard Nusbaum for useful comments and suggestions throughout this study, particularly with regard to predictions of the Cohort theory.

Abstract

Cohort theory, as developed by Marslen-Wilson and Welsh (1978) proposes that a "cohort" of all lexical elements whose words begin with a particular acoustic-phonetic sequence will be activated during the earliest stages of word recognition. The present experiment was designed to determine the extent to which cohorts are activated by the phonemes at the beginning of a stimulus word. Subjects performed a lexical decision task in which they responded "word" or "nonword" to an auditory stimulus. Accuracy and response times were recorded. The stimuli were monosyllabic words and nonwords controlled for frequency. Each target item was preceded by the presentation of either a word or nonword prime. The prime was related to the target item in one of the following ways: 1) identical, 2) first phoneme was the same, 3) first and second phonemes were the same, or 4) first, second, and third phonemes were the same. Results showed that response times and error rates decreased when the prime and target item were identical. However, no facilitation was observed when the prime and the target had one, two or three initial phonemes in common. The results of the present investigation demonstrate that not all words that begin with the same acoustic-phonetic sequence are activated during word recognition. Thus, a major prediction of the Cohort theory was shown to be incorrect.

Acoustic-Phonetic Priming in Auditory Word Recognition:

Some Tests of the Cohort Theory

The perception of spoken language involves a complex interaction of different sources of linguistic information. To comprehend a sentence, one needs to analyze the phonetic, syntactic, prosodic, and semantic information contained in the speech signal. Word recognition and lexical access may be considered two of the preliminary stages in this comprehension process, and as such, they have received a good deal of attention in the literature in recent years. Word recognition refers to those processes by which a listener extracts (recognizes) the phonetic and/or phonological form of an utterance. Lexical access, on the other hand, involves processes which activate the meaning or meanings of words that currently exist in the listener's mental lexicon (Pisoni, Note 1).

Researchers interested in the structure of the mental lexicon and how listeners locate the appropriate word during language processing, have focused their investigations primarily on word recognition (i.e., recognition of the phonetic form of an utterance). Over the years, this research has revealed a number of phenomena related to recognition processes. Specifically, word frequency, word length, and context effects have all been discussed at length in the literature. A number of models have also been proposed to account for these phenomena.

Word frequency in the word recognition literature refers to how often a particular word occurs in the language on a statistical basis. The importance of word frequency for the process of word recognition is illustrated by the large number of articles which address themselves to this particular phenomena. The word frequency effect is the tendency for high frequency words to be responded to more quickly and more accurately than low frequency words (Rubenstein, Garfield, & Millikan, 1970; Stanners, Forbach, & Headley, 1971; Stanners, Jastrzembski, & Westbrook, 1975).

Recently, however, several investigators have begun to question whether the word frequency effect is a result of how often a word actually appears in the language (i.e., its experienced frequency). Landauer and Streeter (1973) suggested that word frequency may, in fact, be due to structural differences between high and low frequency words. They suggested two ways in which some words might be perceptually more salient than others. First, there might be a greater likelihood of confusing a particular word with other similar words. In a study testing this hypothesis, Landauer and Streeter found that common words were more confusable with a greater number of other words than rare words, when one letter in the original word was changed. Also, the average frequency of the similar words was higher for common original items. The second way words may be perceptually more salient involves the nature of the distribution of letters and/or phonemes which make up the word. In a second study, Landauer and Streeter found that common and rare words contained different distributions of phonemes and graphemes. From this observation, they concluded that assumptions of a general perceptual equivalence between common and rare words may be unwarranted

and that word frequency effects may be due to structural differences between words rather than experienced frequency.

Recently, Eukel (Note 2) has argued that word frequency effects may be explained in terms of the phonotactic density of a word; that is, the number of "neighbors" which similar sounding words have in the lexicon. Eukel asked subjects to subjectively estimate the relative word frequencies of real and nonsense words based on a magnitude estimation procedure developed by Shapiro (1969) and Carroll (1971). Subjects showed significant agreement in their judgement of frequencies of nonsense words, and showed significant correlations with lexical distance from English as measured by Greenberg and Jenkins (1964). The implication of Eukel's findings is that subjects judge word frequency by estimating the density of similar sounding words in the lexicon, rather than by referring to experienced frequency of occurrence in the language.

Recency has also been shown to be a factor which influences the word frequency effect. Scarborough, Cortese, and Scarborough (1977) found that although high frequency words were recognized faster than low frequency words in a lexical decision task, there was little evidence for a word frequency effect in a pronunciation task or an old-new recognition memory task. More importantly, in the lexical decision task, earlier presentations of words (priming) produced substantial reductions in the word frequency effect. Specifically, high and low frequency words differed by about 80 msec with regard to recognition latency for the first presentation. However, on the second presentation, the frequency difference was reduced to 30 msec. Scarborough et al. suggested that the modification of word frequency effects may be due, in part, to the recency with which words have occurred in the experimental context as compared to their experienced frequency of occurrence in the language.

From the previous discussion, it is obvious that regardless of what causes the word frequency effect, it is an important phenomenon in language processing. Word frequency must therefore be carefully controlled in lexical access experiments and accounted for in analyzing results or proposing new models of the word recognition process.

Although little research has examined the effects of word length on word recognition in great detail, results of several recent investigations report evidence of a consistent word length effect. Specifically, Grosjean (1980) found that as word length increases, the duration of the signal necessary for recognition also increases (i.e., the word's isolation time). One syllable words were isolated more rapidly than two-syllable words, which, in turn, were isolated more rapidly than three syllable words.

The importance of word length in word recognition may be illustrated by reference to the concept of critical recognition point (or optimal discrimination point) proposed by Marslen-Wilson (Note 3). The critical recognition point for spoken words in isolation is the point at which a particular word, starting from the beginning of that word, becomes uniquely distinguishable from any other word in the language beginning with the same acoustic-phonetic sequence (Marslen-Wilson, Note 3). If Marslen-Wilson's concept is correct, then a decision

about a nonword in a lexical decision task can be made at the point where the nonword sound sequence diverges from the possible English entries in the lexicon. A model which incorporates the concept of a critical recognition point would predict that decision time relative to the critical phoneme offset should be independent of where the critical phoneme is in the linear phonemic sequence, and of the total length of the sequence. Although decision time from critical phoneme offset did remain constant for nonwords in the study conducted by Marslen-Wilson, it is obvious that word length should influence word recognition since the critical recognition point will undoubtedly occur later in the sequence for longer words than for shorter words.

The presence of context effects in word recognition demonstrates that related context facilitates word recognition and unrelated context interferes with word recognition. Context effects have been found when the context consisted of word triplets which were related or unrelated in meaning (Schvaneveldt, Meyer & Becker, 1976), incomplete sentences (Schuberth & Eimas, 1977), and low, medium or highly constrained sentences (Grosjean & Itzler, Note 4). In general, context effects have proven to be quite robust in affecting both the speed and accuracy of word recognition in a number of different tasks.

In recent years, several models of word recognition have been developed which attempt to account for word frequency, word length and context effects. Each of these models contains certain assumptions about the representation of words in the mental lexicon and how people access words stored in the lexicon. One of the first models of word recognition was the Logogen model, developed by Morton (1979). This model uses the concept of a "Logogen" to account for how words are recognized. In particular, Morton was interested in accounting for both the word frequency effect and context effects in recognition of isolated words. To deal with these two diverse phenomena, Morton developed the concept of a Logogen. A Logogen is a counting device which uses both sensory input and higher levels of knowledge (i.e., context) to collect evidence that a word is present as a stimulus, appropriate as a response, or both. The amount of evidence needed to activate a particular Logogen determines its threshold. High frequency words are assumed to have lower thresholds than low frequency words, as the threshold of a logogen is permanently reduced each time the logogen is active. Contextual information serves to increment the logogen (or counter) thereby making that word more sensitive and more available for response. In the Logogen model, stimulus and context information combine directly via the logogen for a particular word. When the threshold for the logogen is reached, the word associated with the particular logogen is recognized. When a word is presented in context, less information is needed from the stimulus in order to recognize the word, thus, recognition occurs faster for words in context than for words in isolation. As a consequence, the Logogen model is able to account for frequency effects, and context effects in word recognition. However, the model does not take into account the internal structure of words (i.e., words can be phonologically and morphologically related to each other). Moreover, the model assumes that information is uniformly represented from the beginning to the end of a word.

To account for frequency and context effects, Forster (1976) proposed a search model in which lexical access involves search of a master file and three peripheral access files. The peripheral access files include orthographic, phonological and semantic/syntactic files which can be used to access an entry in the master file or lexicon proper. These files consist of bins containing entries of similar descriptions. The entries are listed according to their frequency of occurrence in the language such that high frequency words are located at the top of each bin. In this way, Forster's model predicts faster response times for high frequency words than for low frequency words, a terminating search for words versus an exhaustive search for nonwords, semantic relatedness effects between words, and various context effects in word recognition. The most controversial assumption made by Forster is that of autonomous levels of processing. His current search model does not support interactive processing between various knowledge sources but assumes, instead, that lexical access occurs using only sensory or bottom-up information contained in the stimulus input. Context effects occur later in the system by means of cross-referencing between the peripheral and master files. Thus, top-down knowledge can not be used to affect lower levels in the system which are involved in processing the acoustic phonetic input.

Recently, Marslen-Wilson and Welsh (1978) have proposed a model of word recognition known as the Cohort model. This model assumes two separate stages in the word recognition process. First, left to right sensory information in the signal is used to activate a particular cohort -- a group of word candidates which all begin with the same acoustic-phonetic sequence. Second, higher levels of knowledge and contextual information serve to deactivate those candidates which are not likely to be the particular word, given additional information. This deactivation process, therefore, reduces the size of the original cohort set. However, despite this deactivation process, Marslen-Wilson and Welsh state that an element in the pool of word candidates "may remain activated for a short period thereafter" (Marslen-Wilson & Welsh, 1978, p. 56). Recognition is assumed to take place when only one candidate remains active in the cohort.

The Cohort model assumes interaction of sensory and contextual information, as does the Logogen model. However, Marslen-Wilson and Welsh introduce the notion of deactivation of previously activated candidates into the word recognition process. According to Marlsen-Wilson and Welsh, a word is recognized at the point where it can be uniquely distinguished from all other words in the cohort. The point at which potential candidates diverge from others in the lexicon is called the "critical recognition point" of the word. Given a word's initial cohort and a sentential context, it can be determined which candidates will be deactivated due to inappropriate information. Thus, one can determine when the stimulus word diverges from the other candidates in the lexicon and when it will be recognized by analyzing the left to right acoustic information in the signal. A major assumption of Cohort theory is the activation of all words in the lexicon that begin with the same acoustic-phonetic sequence. This assumption predicts that at any time, a large number of acoustically similar words are active and could be recognized.

As evident from this brief review, several models of word recognition have been proposed, and a number of phenomena have been described in the word recognition literature. One of the primary questions in word recognition is how words are represented in the lexicon and hence how they might be activated during language processing. This question is the focus of the present investigation. In particular, we are interested in whether or not possible word candidates are activated during recognition of isolated words.

Cohort theory (Marslen-Wilson & Welsh, 1978) assumes that a "cohort" of all lexical elements which begin with a particular acoustic-phonetic sequence will be activated during the earliest stages of lexical access. However, Marslen-Wilson and Welsh do not specifically state what is meant by an acoustic-phonetic sequence in this model. Although never made explicit, one assumption is that an acoustic-phonetic sequence refers to the sequence of phonemes present at the beginning of a word. Although this assumption is implied Marslen-Wilson and Welsh state explicitly that phonemes are not necessarily the input to the word recognition system.

> The use here of phoneme categories should not be taken as a theoretical claim about human speech processing. In particular, I am not necessarily claiming that the input to the word-recognition system takes the form of a string of phonemic labels. (Marslen-Wilson, Note 3, p. 8)

> ...the use of a phonemic notation to specify the stimuli in the two experiments here, and to determine recognition points, etc., should not be taken as a theoretical claim. That is, I am not claiming that the input to the human word-recognition system takes the form of a string of phonemic labels. The reason that a phonemic analysis has been adequate here may mean only that phonemic labels tend to coincide reasonably well with the sequential patterning of informationally important variations in the speech signal. But as far as the cohort model is concerned, this sequential patterning could just as well be delivered to the word recognition system in the form of spectral parameters or auditory features as in the form of segmental labels (Marslen-Wilson, in press, p. 21).

The assumption that phonemes constitute an acoustic-phonetic sequence will result in different predictions than an assumption that an acoustic-phonetic sequence refers to morphemes or whole words or a sequence of power spectra (Klatt, 1977). The present experiment was designed to determine the extent to which cohorts are activated by the phonemes at the beginning of a stimulus word.

Given the robust effects context has on word recognition it seems appropriate to use this effect to learn more about the word recognition process. Several researchers have found context effects in word recognition when using a priming technique. This technique involves presenting stimulus items immediately before a target item and recording the influence, if any, such prior

presentations may have on response to the target item. As mentioned earlier, Scarborough, Cortese, and Scarborough (1977) found that frequency effects could be modified by repetition of the stimuli during the test sequence.

Recently, Jakimik, Cole and Rudnicky (Note 5) conducted a study in which subjects performed a lexical decision task with auditory stimulus items which were related phonologically and/or orthographically. Jakimik et al. found facilitation when successive words shared the same spelling. That is, the word "nap" was recognized faster when it was preceded by the word "napkin". However, there was no facilitation when successive words had the same pronunciation but differed in their spelling (i.e., there was no facilitation of "spy" when it was preceded by "spider"). This result suggests, therefore, that the lexicon contains an orthographic representation of words and that subjects access a word's orthographic representation in a lexical decision task.

To study the process of word recognition and, in particular, activation of a word's cohort structure, we took advantage of the data obtained regarding context effects in word recognition by using a priming technique in a lexical decision task. If a set of cohorts is activated during word recognition, then the presentation of a prime sharing certain selected properties with a target, should facilitate word recognition. That is, there should be a decrease in response time to a target item sharing properties with a previously presented prime item. To test this prediction we constructed several different primes. The prime items were words or nonwords that were related to target items (words and nonwords) in one of the following ways: 1) identical, 2) first phonemes are the same, 3) first and second phonemes are the same, or 4) first, second, and third phonemes are the same. In the identical prime condition, we predicted that the prime should facilitate recognition of the target item. With respect to the shared phoneme conditions, if cohorts are activated by the phonemes at the beginning of words, then we would expect to find facilitation of the target for each of the shared phoneme conditions, such that the condition involving three shared phonemes will be faster than the two-shared phonemes condition, which in turn, will be faster than the one-shared phoneme condition. Since most models of word recognition assume that nonwords are not stored in the lexicon, we expect that the presence of a nonword prime should not facilitate a word target, and a nonword target should not be facilitated by a word prime. Therefore, the response to nonword targets should be slower overall than to word targets, regardless of the priming condition used. However, Marslen-Wilson and Welsh (1978) have argued that reaction times to classify nonwords should depend on the point where the nonword diverges from all words in the lexicon beginning with the same sound sequence (i.e., the critical recognition point). If this assumption is correct, then nonword targets in which the critical recognition point occurs early in the sequence of sounds for that item should be classified faster than nonword targets in which the critical recognition point occurs later on in the sound sequence.

Method

Subjects

Forty-two undergraduate students were obtained from a paid subject pool maintained in the Speech Research Laboratory at Indiana University. Subjects were paid $3.00 for their participation in the experiment. All subjects were native speakers of English with no known history of hearing loss or speech disorders.

Materials

Ninety-eight monosyllabic words (49 high frequency and 49 low frequency) were obtained using the Kučera and Francis (1967) computational norms. The words were selected such that they included each of the following syllable types: 1) CVC 2) CCVC, 3) CVCC, and 4) CCVCC. In addition, ninety-eight nonwords were formed from the ninety-eight words by changing one phoneme in the word (e.g., best/besk).

Each of these one-hundred ninety-six target items were then paired with seven separate primes such that the primes were related to the target in the following seven ways: 1) identical, 2) a word with the same first phoneme, 3) a nonword with the same first phoneme, 4) a word with the same first and second phonemes, 5) a nonword with the same first and second phonemes, 6) a word with the same first, second and third phonemes, and 7) a nonword with the same first, second and third phonemes. Table 1 lists some examples of word and nonword targets with their corresponding prime conditions.

Insert Table 1 about here

A male speaker, seated comfortably in a sound attenuated IAC booth (Controlled Acoustical Environments, No. 106648), recorded the target and prime items on one track of a audio tape using an Electro-Voice D054 microphone and an AG500 tape deck. The stimulus items were produced in the carrier sentence "Say the word ----- please" to control for abnormal durations when words are produced in citation form in isolation. The stimulus items were then digitized at a sampling rate of 10 khz using a 12-bit analog-to-digital converter and then excised from the carrier sentence using a speech waveform editor (WAVES) on a PDP 11/34 computer. The targets and their corresponding primes were stored digitally as stimulus files on a computer disk for later playback to subjects in the experiment.

Table 1

Examples of Target Items and their Corresponding Primes

	PRIMES						
TARGETS	1	2	3	4	5	6	7
Word High							
black	black	burnt	/brɛm/	bleed	/blim/	bland	/blæt/
drive	drive	dot	/dalf/	drug	/drʌt/	dried	/draɪl/
Nonword High							
/blæf/	/blæf/	big	/bʌv/	blind	/blʌz/	blank	/blæʃ/
/praɪv/	/praɪv/	point	/poɪl/	print	/prɪl/	prime	/praɪk/
Word Low							
bald	bald	bank	/brɪl/	bought	/bɔʃ/	balls	/bɔlf/
dread	dread	cove	/dʌs/	drill	/drʌb/	dress	/drɛn/
Nonword Low							
/bʌld/	/bʌld/	bride	/braɪf/	bust	/bʌp/	bulb	/bʌin/
/drɪd/	/drɪd/	desk	/dist/	drag	/drʌs/	drip	/drɪs/

Procedure

Subjects were run in groups of four or less. The presentation of stimuli and collection of data were controlled on-line by a PDP 11-34 computer. Subjects heard the stimuli at 75 dB SPL with background noise at 45 dB SPL (re. .0002 dynes/cm2) over a pair of TDH-39 headphones. Subjects were asked to perform a lexical decision task for the one hundred ninety-six test items. The subject responded "word" or "nonword" as quickly and as accurately as possible after the presentation of each target stimulus item.

A typical trial sequence proceeded as follows: First, a cue light was presented for 500 msec at the top of the subject's response box to alert the subject that the trial was beginning. Then there was a 1000 msec pause followed by an auditory presentation of the prime item. The subject was not required to respond overtly in any way to the presentation of the prime. An interstimulus interval of 500 msec intervened between the prime and the presentation of the target item. The subject responded "word" or "nonword" to the presentation of the target item on each trial. Immediately following the subject's response, the computer indicated which response was correct by illuminating the feedback light above the appropriate response button. The subject's response (i.e., word vs. nonword) was recorded as well as response latency. Latencies were measured from the onset of the target item to the subject's response.

Six subjects were run in each of seven conditions for a total of forty-two subjects. Subjects received ninety-eight word and ninety- eight nonword targets, half of which were low frequency and half of which were high frequency. There was an equal number of words primed by each of the seven prime types. The distribution of primes for nonword targets was the same as for the word targets. The prime-target pairs were counterbalanced across the seven conditions. Presentation of prime-target pairs was randomized for each session, and subjects were never presented with the same target nor prime on any of the one hundred ninety-six stimulus trials.

Results

The data from the present experiment were analyzed with respect to two dependent measures: response accuracy (word vs. nonword) and response latency. Mean response times and error rates were calculated across subjects and conditions and subjected to an analysis of variance.

The main results are shown in Figure 1. The top half of the figure displays averaged response times, the bottom half shows percent errors for the four types of target stimuli as a function of seven prime types.

Insert Figure 1 about here

As typically found in lexical decision tasks, the lexicality (word-nonword) main effect was significant for the response time data ($F(1,41) = 42.78$, $p < .001$) and percent error data ($F(1,41) = 10.16$, $p < .002$). The mean response time for words was 968 msec and for nonwords 1041 msec.

In addition, as expected, there was a significant frequency effect for response times ($F(1,41) = 29.16$, $p < .001$) and error rates ($F(1,41) = 35.62$, $p < .001$). High frequency items were responded to faster and more accurately than low frequency items. The overall mean response time for high frequency target items was 990 msec and for low frequency items 1019 msec. A frequency by lexicality interaction was also observed. This interaction revealed that the frequency effect was different for word and nonword items. A simple post-hoc effects test confirmed that word targets were affected significantly by the frequency manipulation ($F(1,78) = 53.12$, $p < .01$) but nonword targets were not ($F(1,78) = 1.37$, N.S.). This result is illustrated in Figure 2. The top half of this figure displays averaged response times for high and low frequency word targets for each of the seven prime types, and the bottom half of the figure displays averaged response times for high and low frequency nonword targets as a function of the seven prime types. Note that the curve for high frequency words shows consistently faster response times than the curve for low frequency words. The lower half of the figure shows that this frequency effect was not observed for nonword targets.

Insert Figure 2 about here

The overall analysis of variance also revealed a main effect of prime type on response times ($F(6,246) = 34.75$, $p < .001$) but this effect was not found for error rates ($F(6,246) = .23$, N.S.). Figure 3 shows the response times of word and nonword targets averaged over frequency as a function of prime type.

Insert Figure 3 about here

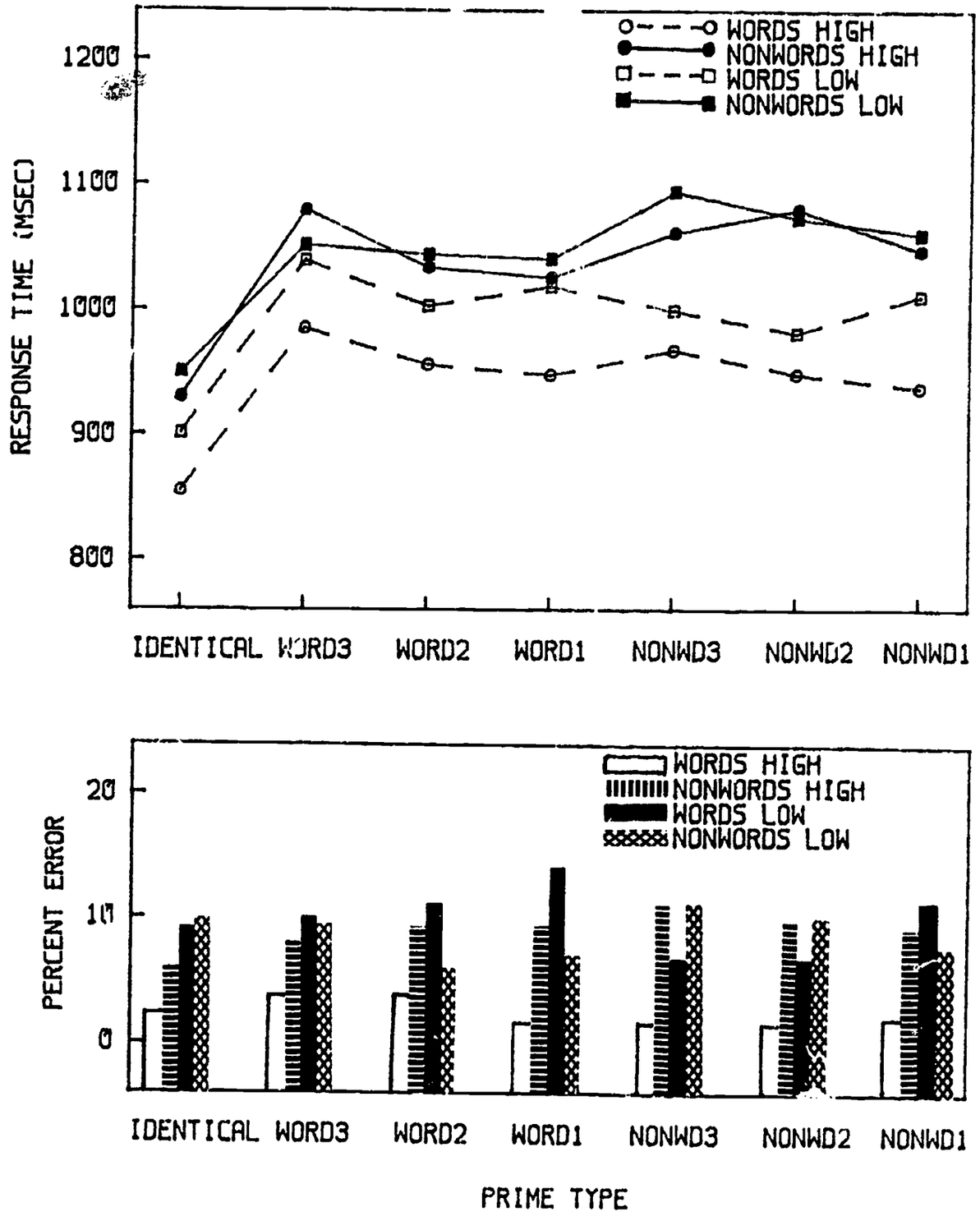
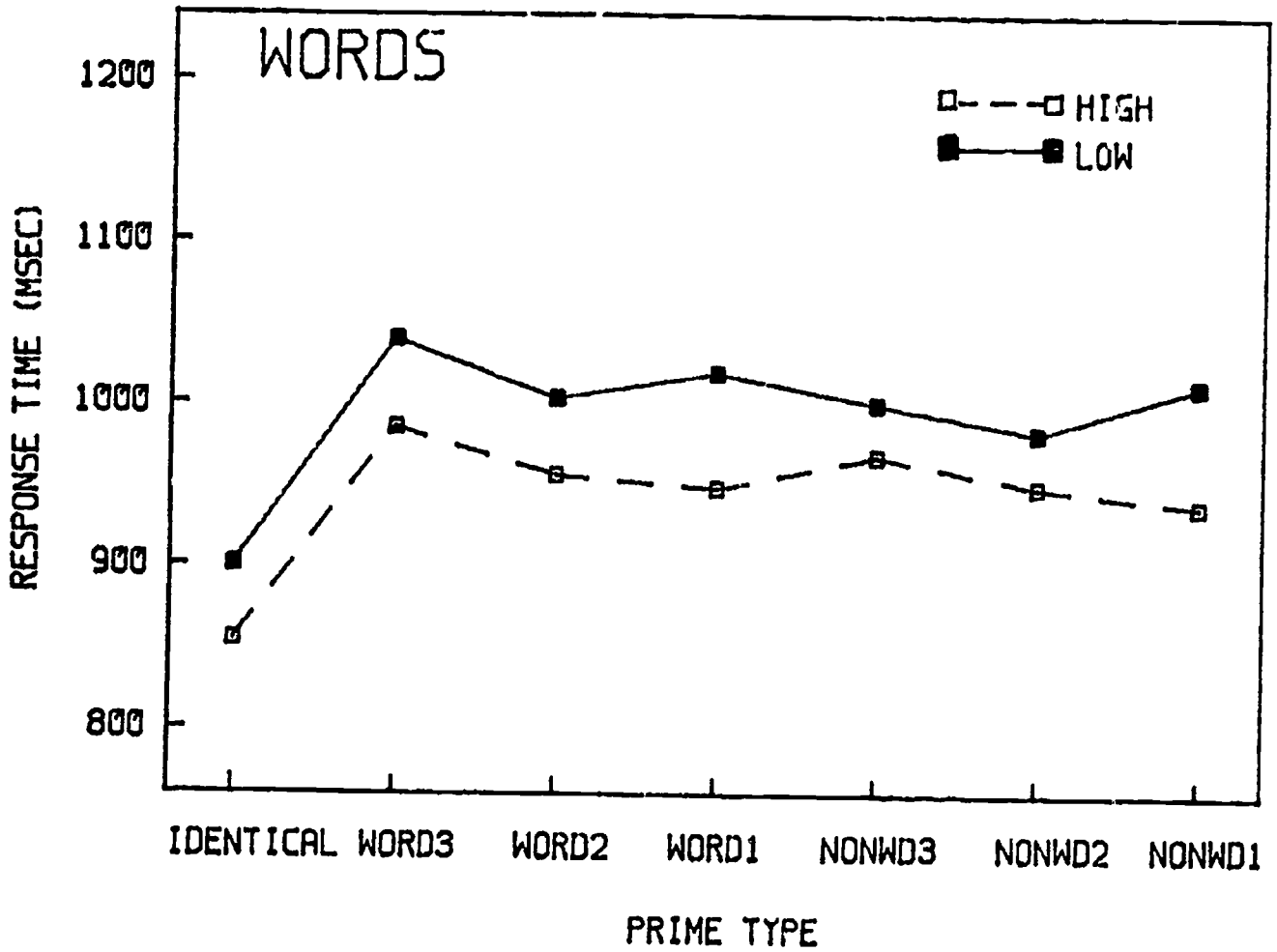


Figure 1. Response times (top panel) and error rates (bottom panel) of four types of target items (high frequency words, low frequency words, nonwords derived from high frequency words and nonwords derived from low frequency words) for the seven prime types (identical, word3, word2, word1, nonwrd3, nonwrd2, nonwrd1).



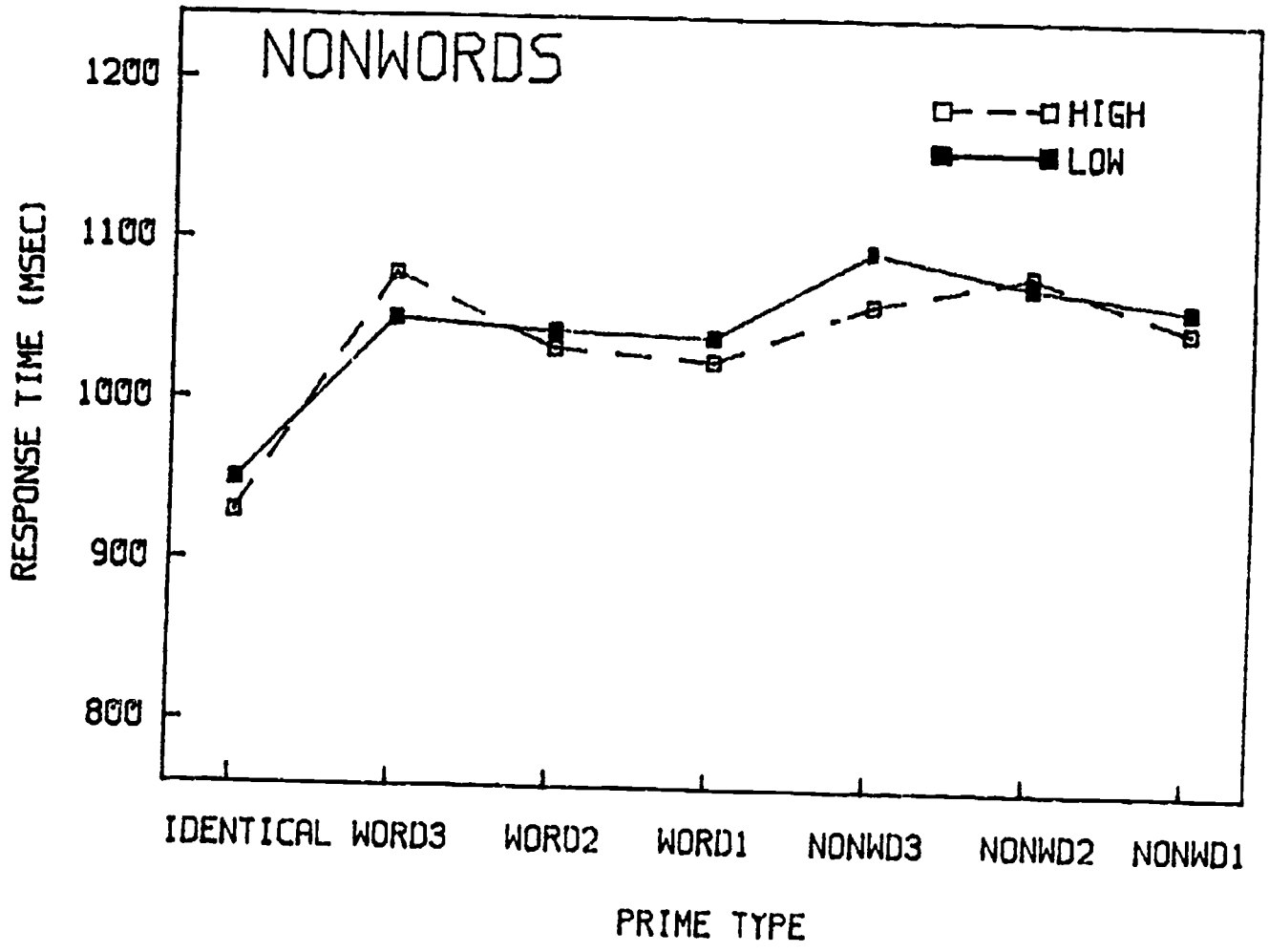


Figure 2. Response times for word targets (top panel) and nonword targets (bottom panel) for the seven prime types.

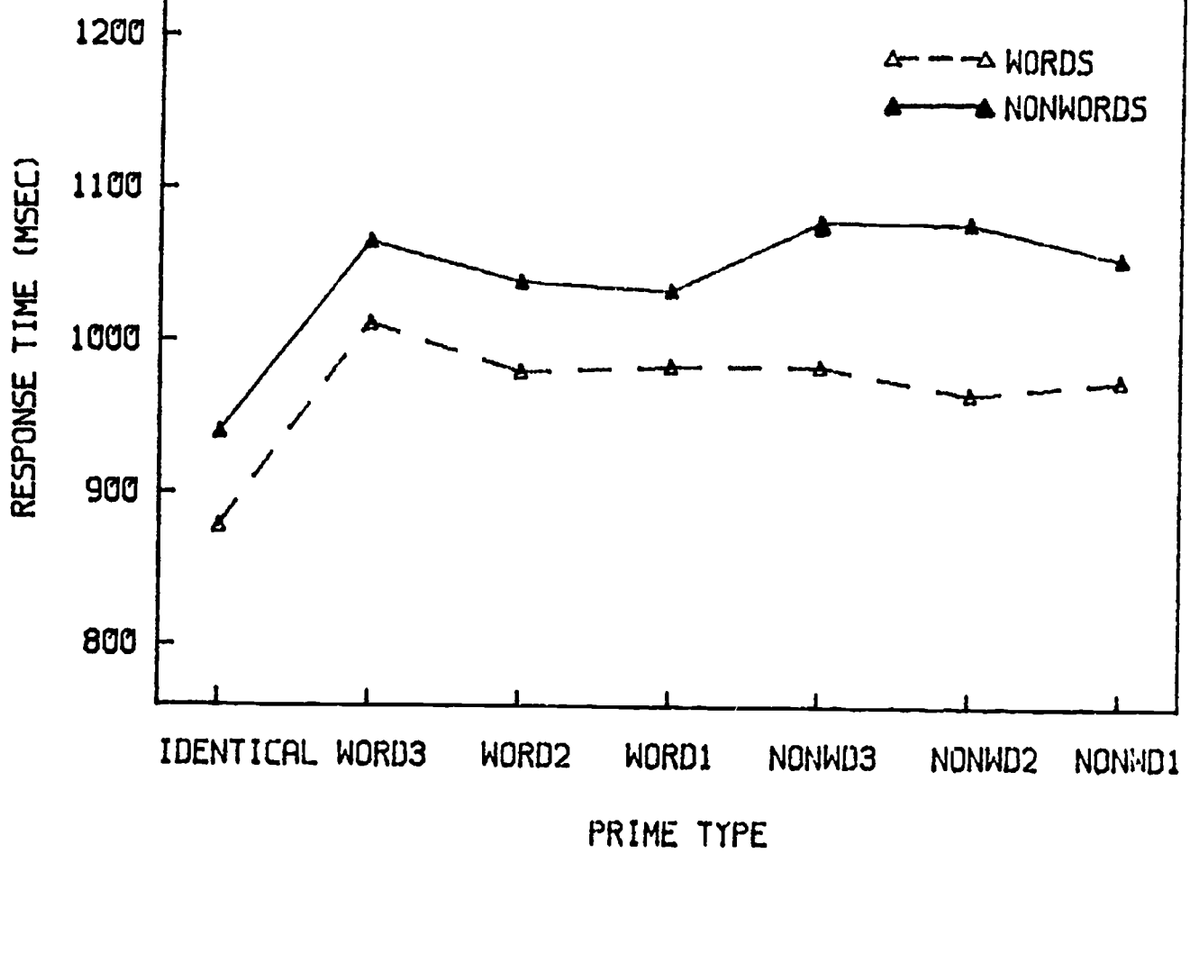


Figure 3. Response times for word and nonword targets averaged across frequency for seven prime types.

A simple effects test of this interaction revealed that both word and nonword target items were significantly different by prime type ($F(6,492) = 16.31$, $p < .01$, for words and $F(6,492) = 21.01$, $p < .01$, for nonwords). In the Newman-Keuls analysis, Identical primes were significantly different than all other prime types ($p < .01$). However, the shared phoneme prime types (Word3, Word2, Word1, and Nonword3, Nonword2, Nonword1) were not significantly different from each other. This result was found for identical word and nonword target items and can be clearly seen in Figure 4. In this figure, word and nonword prime types are collapsed and compared with identical prime type.

Insert Figure 4 about here

As shown here, the difference between identical prime type and word/nonword prime types holds for both high and low frequency word targets (left panel) and nonword targets (right panel).

Discussion

The results of the present investigation replicated several well-known phenomena found in the word recognition literature. Specifically, our analysis revealed a lexicality (word-nonword) effect that has been consistently found in lexical decision tasks. The increased response time to a nonword target when compared with word target items is a robust finding that has been incorporated into most contemporary models of word recognition. We replicated this effect in the present study.

Several theories have tried to account for this word-nonword effect by proposing that nonwords are not stored in the mental lexicon. Accordingly, a listener will only "recognize" a nonword as such after searching through the stored list of words without finding the target item. The Cohort model proposes a different process which predicts different results. According to Cohort theory, a nonword is recognized at the point where it diverges from all of the words in the cohort. This theory predicts that nonword items will only produce slower response times to the extent that the nonword item diverges from the word candidates later in the item. In the present investigation all targets were four or five phonemes in length. All nonword targets diverged from word candidates at the fourth phoneme, and all word targets were only distinguished from other candidates at the fourth or fifth phoneme (for twenty of the ninety-eight cases). For these stimulus items, Cohort theory would predict that nonword targets should show equal response times or possibly faster response times (for the twenty previously mentioned items) when compared to word targets. This was not the case. In terms of predicting response time our results do not appear to support

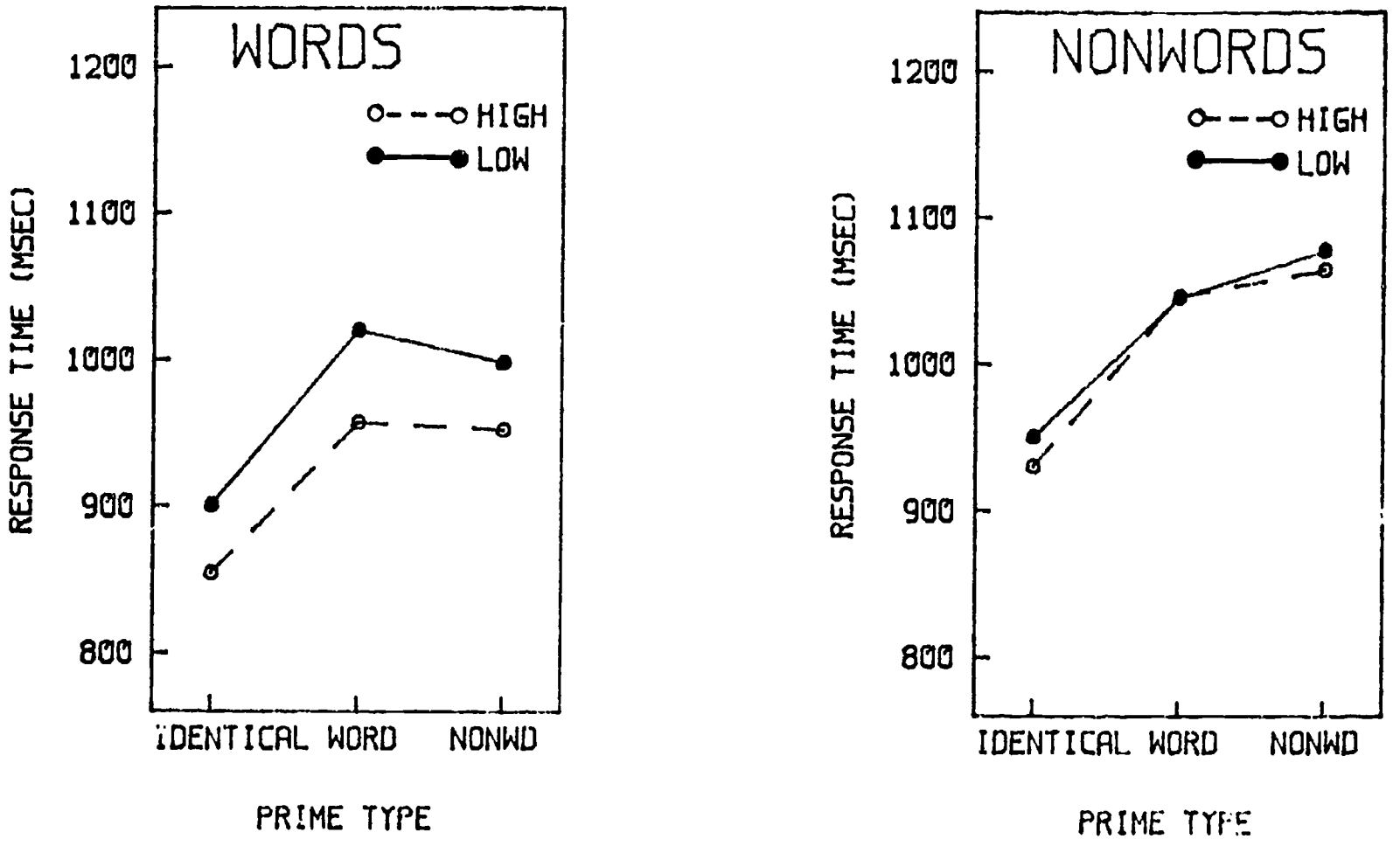


Figure 4. Response times for word targets (left panel) and nonword targets (right panel) for three prime types.

the concept of critical recognition point for nonword targets as proposed by Marslen-Wilson and Welsh (1978). Although Marslen-Wilson and Welsh would predict faster response time to nonword targets than to word targets for the stimulus items used in the present investigation, we found the opposite to be true. Our results revealed faster response times for word target items than for nonword target items, despite the fact that the critical recognition point occurred in nonword targets at the same point, or earlier, than in word targets.

In addition to a word-nonword effect, we also observed a strong frequency effect in the lexical decision task. High frequency target items were responded to faster and more accurately than low frequency items. However, subsequent analyses revealed that this frequency difference was only present for words and not for nonwords derived from high and low frequency words.

Two of the word recognition models summarized earlier have been successful in trying to account for word frequency effects in word perception. Forster (1976) proposed that words stored in the mental lexicon are ordered by frequency with high frequency words at the top of the list in each bin. The Logogen model (Morton, 1979) proposes that the frequency of a word modifies the threshold of the logogen for that word, such that high frequency words have lower thresholds than low frequency words. Surprisingly, the Cohort model (Marslen-Wilson and Welsh, 1978) does not attempt to account for word frequency despite its presence in word recognition experiments. Of course, it would be relatively easy to modify Cohort theory to deal with frequency by ordering members of the cohort by frequency.

Several current explanations of the effects of word frequency can account for the discrepancy in the frequency effect for words versus nonwords observed in the present study. The traditional view of word frequency (i.e., word frequency as "experienced" frequency) and frequency as recency (proposed by Scarborough, Cortese and Scarborough, 1977) suggest that nonwords should not show a frequency effect at all since one does not normally have experience with nonwords in using the language. Thus, these accounts of word frequency might predict the pattern of the word frequency effects obtained in this study.

On the other hand, the views of word frequency proposed by Landauer and Streeter (1973) and Eukel (Note 2) are less successful in accounting for the present results. Landauer and Streeter suggest that the phonotactic structure of high and low frequency words may be the factor which influences the response times to them, rather than their experienced frequency per se. In the present investigation, all nonword targets were derived from word targets by changing one phoneme in the word. In addition, the changed phoneme was balanced across each of four phoneme positions. Therefore, the structure of the nonword targets was closely related to the structure of the original word targets from which they were derived and should seemingly produce a frequency effect if Landauer and Streeter's structural hypothesis is correct. However, if the nonwords in the present study did not maintain the structure of their corresponding word targets, then Landauer and Streeter's explanation might account for the pattern of results observed here.

A similar situation arises when one considers Eukel's (Note 2) view of the effects of word frequency. According to Eukel, word frequencies are judged by estimating the density of similar sounding words in the lexicon. Eukel found that subjects' judgements of the frequency of nonsense words showed a significant correlation with their judged distance from English. For example, ARTY can be changed to ARMY, which can be changed to ARMS, by substituting one phoneme for each transformation. However, two substitutions are required to change ARTY to ARMS. Therefore, ARTY is closer to ARMY in the lexical space, based on a phoneme substitution matrix (Eukel, Note 2). The implication is that word frequency is judged by estimating the density of similar words in the lexicon (i.e., the frequency of ARTY would be judged relative to the frequency of ARMY). If Eukel's proposal is correct, and the nonwords in the present study were similar enough to the word targets from which they were derived, one would predict a frequency effect for nonwords. Such an effect was not observed.

The effects of the different prime types on response time were, of course, of primary interest to us in this investigation. We expected that response times to targets preceded by identical prime items would be faster than any of the shared phoneme prime types. This prediction was confirmed consistently and very strongly in our data. In addition, we expected to find facilitation of the shared phoneme conditions such that response times to three-shared phonemes would be faster than two-shared phonemes, which in turn, would be faster than one-shared phoneme. Our results were not consistent with this prediction. The six shared phoneme prime types did not differ significantly from each other. These results demonstrate that response times to target items were not primed by words with the same first, second or third phonemes as the target. The results suggest that the repetition effect found by Scarborough, Cortese and Scarborough (1977), as well as the priming effects seen in the present investigation for identical items, may be caused by semantic priming rather than some exact acoustical correspondence between the structure of prime and target items. In short, our specific predictions, derived in part from Cohort theory, were not confirmed.

The pattern of results we obtained in the lexical decision task appears to be in conflict with results recently reported by Jakimik, Cole and Rudnicky (Note 5). These investigators found that response times decreased when auditorally presented target items were preceded by items which shared the same initial spelling (e.g., "nap" preceded by "napkin"). However, they found no facilitation of response times to a target item when it was preceded by an item which did not share the same initial spelling (e.g., "spy" preceded by "spider"). An analysis of the orthographic relationship between the primes and target items used in the present study revealed that of the 1,372 primes used, only 155 of them were <u>not</u> orthographically related to the target (e.g., in the Word1 prime condition, "come" was used to prime the low frequency word target "quill"). However, despite the large number of primes and targets which were orthographically related, we did not replicate the priming effect reported by Jakimik et al.

An examination of the Jakimik et al. study revealed one major difference. Jakimik et al. included an unrelated priming condition as a control. In the present investigation, results from the shared phoneme conditions were compared

to an identical priming condition. Our results revealed no acoustic priming, whereas Jakimik et al.'s interpretation suggested the presence of acoustic priming. If an unrelated condition was incorporated into our design, we might very well find that this condition has the slowest response time of all. If this result was obtained, the interpretation would be quite different than the present one. Under such circumstances, it would appear that the shared phoneme conditions do facilitate recognition of the target item. However, the number of shared phonemes (i.e., three, two or one) would not be the critical variable as shown by the results of this investigation. Rather, such a result would suggest that only the first phoneme is critical in priming a target item. In the present study, we found no effects of acoustic similarity or overlap on response times (i.e., three, two and one phoneme priming did not differentially affect response times as we predicted). Further, this result could be used as evidence that an acoustic-phonetic sequence, as proposed by Marslen-Wilson and Welsh, is, in fact, constituted by only the first phoneme in the stimulus. Thus, a major prediction of the Cohort theory with regard to activation of words sharing an initial acoustic-phonetic sequence appears to be incorrect.

In summary, the results of the present investigation revealed both word frequency and word-nonword (lexicality) effects, as commonly reported in the word recognition literature. With regard to primed conditions, identical prime-target pairs resulted in decreased response times and error rates for the target item, as anticipated. However, no differences in acoustic priming were found when the target item shared one, two or three initial phonemes with the prime item. Our results, therefore, do not support a major assumption of cohort theory that a set of word initial cohorts are acitvated by the acoustic-phonetic information at the beginning of a stimulus item. This conclusion is based on the finding that the first three phonemes in the prime items did not differentially affect response times to phonetically related target items.

Reference Notes

1. Pisoni, D.B. In defense of segmental representations in speech processing. Paper presented at the meeting of the Acoustical Society of America, Ottawa, Canada, May 1981.

2. Eukel, B. A phonotactic basis for word frequency effects: Implications for automatic speech recognition. Unpublished manuscript, 1980.

3. Marslen-Wilson, W.D. Optimal efficiency in human speech processing.

4. Grosjean, F., & Itzler, J. Do the effects of sentence constraint and word frequency interact during the spoken word recognition process? Manuscript submitted for publication, 1981.

5. Jakimik, J., Cole, R.A., & Rudnicky, A.I. The influence of spelling on speech perception. Paper presented at the twenty-first annual meeting of the Psychonomic Society, St. Louis, Missouri, November 1980.

References

Carroll, J.B. Measurement properties of subjective magnitude estimates of word frequency. Journal of Verbal Learning and Verbal Behavior, 1971, 10, 722-729.

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

Greenberg, J.H. & Jenkins, J.J. Studies in the psychological correlates of the sound system of American English. Word, 1964, 20, 151-177.

Grosjean, F. Spoken word recognition processes and the gating paradigm. Perception & Psychophysics, 1980, 28 (4), 267-283.

Klatt, D.H. Speech perception: A model of acoustic-phonetic analysis and lexical access. Journal of Phonetics, 1979, 7, 279-312.

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

Landauer, T.I. & Streeter, L.A. 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.

Marslen-Wilson, W.D. Speech understanding as a psychological process. In J.C. Simon (Ed.), Spoken Language Generation and Understanding, Dordrecht: Reidel, in press.

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

Morton, J. Word recognition. In J. Morton and J.C. Marshall (Eds.) Structures and Processes, Cambridge: M.I.T. Press, 1979.

Rubenstein, H., Garfield, L. & Millikan, J.A. Homographic entries in the internal lexicon. Journal of Verbal Learning and Verbal Behavior, 1970, 9, 487-494.

Scarborough, D.L., Cortese, C. & Scarborough, H.S. Frequency and repetition effects in lexical memory. Journal of Experimental Psychology: Human Perception and Performance, 1977, 3 (1), 1-17.

Schuberth, R.E. & Eimas, P.D. Effects of context on the classification of words and nonwords. Journal of Experimental Psychology: Human Perception and Performance, 1977, 3, 27-36.

Schvaneveldt, R.W., Meyer, D.E. & Becker, C.A. Lexical ambiguity, semantic context, and visual word recognition. Journal of Experimental Psychology: Human Perception and Performance, 1976, 2, 243-256.

Shapiro, B.J. The subjective estimation of relative word frequency. Journal of Verbal Learning and Verbal Behavior, 1969, 8, 248-251.

Stanners, R.F., Forbach, G.B. & Headley, D.B. Decision and search processes in word-nonword classification. Journal of Experimental Psychology, 1971, 90, 45-50.

Stanners, R.F., Jastrzembski, J.E. & Westbrook, A. Frequency and visual quality in a word-nonword classification task. Journal of Verbal Learning and Verbal Behavior, 1975, 14, 259-264.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

Sentence-by-sentence listening times for spoken passages: Text structure and listeners' goals*

Peter C. Mimmack, David B. Pisoni and Hans Brunner

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47401

*This research was supported by NIMH research grant MH-24027 and NIH research grant NS-12179 to Indiana University in Bloomington.

Abstract

This paper describes a new method of measuring cognitive processing load while listening to spoken passages of text. In an analog of sentence-by-sentence reading tasks, subjects controlled the latency to output successive sentences of the passages. To test the measure's validity, a number of key factors that have proven to be reliable in studies of written texts were manipulated. Several of the factors pertained to the texts' structure, such as normal vs. random order of the parts of the text, height or centrality of the ideas, and narrative vs. expository genre. Another factor involved manipulation of the subjects' goals in listening. One group of subjects listened in order to answer questions (comprehension condition) while another group listened to produce verbatim recall protocols. In reading studies, latencies of recall subjects typically are higher than those of comprehension subjects, and latencies of subjects reading randomized texts are higher than of those reading normally ordered texts. Both results were replicated in this study. The height factor was not as important to these subjects as it was in earlier reading studies. The texts' genre was almost irrelevant. An unexpected factor was text length. Subjects produced shorter latencies for shorter texts: the data supported an interpretation that the beginning sentences of a text contain clues to the text's length.

Over the past 10 years there has been a profound shift in the level of analysis that psycholinguists deal with. Whereas the primary emphasis used to be on the sentence (Johnson-Laird, 1974; Levelt, 1978), it has now shifted to larger, text-level units such as paragraphs, stories, or expository discourse (Kintsch, 1974; de Beaugrande, 1980; Tannen, 1982b). These studies have investigated a number of issues with a variety of experimental techniques. Several researchers (Kintsch and Keenan, 1973; Cirilo and Foss, 1980; Haberlandt, 1980) have used sentence-by-sentence reading time to measure input processing. Others (Graesser, Hoffman, and Clark, 1980; Brunner and Pisoni, 1982; Hildyard and Olson, 1982) have used question-answering after input as a comprehension measure of what is derived from the text. Still others (Fredericksen, 1972; Kintsch, Kozminsky, Streby, McKoon, and Keenan 1975; Mandler and Johnson, 1977; Thorndyke, 1977) have had subjects produce recall protocols to measure what is retained over time. Reder (1980) provides a comprehensive review of the literature in the entire area.

A primary drawback of research on text comprehension is that it has been based almost entirely on studies of reading. What has been done with spoken passages has generally used post-input measures, either recall (Fredericksen, 1972; Rubin, 1978) or comprehension (Hildyard and Olson, 1982). Of course, the main reason for this is the ephemeral nature of spoken language. While writing is fixed in space and thus can be held for a controlled period of time (by experimenter or subject), the transitory nature of spoken language makes it much more difficult to control, and thus more difficult to do experimental work with. This is unfortunate, because it means that our conception of discourse processing is almost entirely based on reading, a self-paced language process which requires years of training and practice (Kavanagh and Mattingly, 1972).

The present research was directed at three goals. The first concerns our interest in developing a better measure of spoken language processing and comprehension. Levelt (1978) describes the lack of good "on-line" measures of speech processing at the sentential level. Since that review, Grosjean (1980) has developed a reliable measure through excising small segments of each word's waveform, and Wingfield and Nolan (1978) have measured the locations of where subjects stop a taperecorder while listening to compressed speech (speech with a higher information load per unit time than normal speech). However, both of these measures are fairly gross since they do not focus on response latencies.

This study is a preliminary step in the direction started by Wingfield and Nolan toward a more precise on-line measure of speech processing. The greater precision comes from using a computer-controlled digital waveform rather than the taperecorder's analog waveform. In this report, a procedure from research on reading has been borrowed: break a passage into units of sentence length, and allow subjects to determine the speed of presentation by pushing a button for each successive sentence.

At first glance, this approach may not seem entirely reasonable for use with spoken text, for while an entire written sentence can be presented at once and reading time measured, a spoken sentence is an acoustic stimulus presented over time and presumably processed during that time (Marslen-Wilson and Welch, 1978). Our common experience is that spoken conversation takes place very rapidly, with a minimum number of extended pauses between sentences. However, in conversation, speakers are exchanging information (both verbally and non-verbally if in person)

and they share common knowledge of each other or the topic and use other contextual factors to support understanding (Kay, 1977). It is quite a different situation not only to listen to a passage on what may not be a familiar topic spoken by someone who cannot be communicated with (whether on tape, TV, or radio) but moreover to listen to that passage for a particular reason. Therefore it does seem reasonable that a sentence-by-sentence listening task could produce meaningful and reliable results in an experimental situation. Given that such an on-line measure of spoken text processing is logically possible, initial studies with it should examine two major issues which have been explored in the reading literature, text structure and the reader's goals.

Text Structure

Texts encode a hierarchy of information. Some ideas are more central and thus are given more processing time at input (Cirilo and Foss, 1980; Cirilo, 1981) and are recalled better over time (Kintsch, 1974; Kintsch et al., 1975). Kintsch's (1974; Turner and Greene, Note 1) method of describing text structure in terms of semantic propositions accounts for these data nicely. It also accounts for sentence-by-sentence reading times better than simple measures based on the number of words in a sentence or text (Kintsch and Keenan, 1973). Maintaining an equal number of words but doubling the number of propositions in a sentence caused reading times to double as well.

A different approach has been used to describe the structure of stories in terms of episodic constituents (Thorndyke, 1977; Mandler and Johnson, 1977).

> An episode is defined from the point of view of a protagonist who is faced with a problem and tries to solve it. The problem is triggered by events in the beginning (B) of the episode. Then the hero reacts (R) to the problem, he or she formulates a goal (G) and attempts (A) to achieve the goal, producing a certain outcome (O). The episode concludes with an ending (E) .... (Haberlandt, 1980, p. 100)

Recall is best for the events which form the beginning of the episode, the actions which constitute the hero's attempt to achieve the goal, and the outcome of his efforts. Controlling for serial position, Haberlandt (1980) found that reading time is elevated above a story's mean for the beginning and ending segments of episodes, and he attributed the results to greater encoding for initializing or finalizing an episode as a unit.

The robustness of a story's intrinsic order has been examined in several studies which presented the constituents in scrambled order (Kintsch, Mandel, and Kozminsky, 1977; Rubin, 1978; Haberlandt, 1980). Haberlandt (1980) found that reading time for the beginning and ending segments was higher than the mean of sentences in the story even when those two segments were not the first and last sentences presented. Kintsch et al. (1977) found that overall reading times were longer for a story whose paragraphs were scrambled compared to normal order. However, recall time and recall protocols of scrambled stories were not significantly different from those of normal-order stories. Kintsch et al. concluded that subjects re-ordered the stories while reading them.

In contrast to the number of studies done explicitly on narratives, very little experimental work has been done on expository texts or essays. Recently, Olson, Mack and Duffy (1981) found that this genre produces a flat serial position curve in reading times for sentences, whereas narratives produce a downward sloping curve. The downward slope follows from the predictable structure of a narrative described above, whereas the flat curve for essays results from a less predictable, more open-ended format (Olson, Duffy, and Mack, 1980). Graesser et al. (1980) found that subjects' ratings of texts' "narrativity" was the most significant factor in a regression analysis of sentence-by-sentence reading time, accounting for 30% of the variance. Narrativity was defined by the experimenters as "events unfolding in time." Texts low in narrativity (i.e. expository passages) had significantly higher reading times than those high in narrativity.

In summary, four major findings dealing with the structure of texts have been revealed by reading time studies. The number of propositions in a sentence is more important than the number of words, episodic units are psychologically salient, reading times for randomly ordered passages are longer but the original order is recoverable, and expository texts produce longer reading times overall than narratives.

The Reader's Goals

The second major set of findings in text-processing may be summarized by stating that a person's initial goals in reading a passage markedly affect how the passage is read, what information is derived from it, and what is retained (Fredericksen, 1972; Aaronson, 1976; Aaronson and Scarborough, 1976; Graesser et al., 1980; Cirilo, 1981). Fredericksen (1972) had subjects listen to one passage and perform a task together four times in succession. All subjects wrote recall protocols at the end of four repetitions. At that time, subjects who had listened to the passage and recalled it each time produced better verbatim recall protocols with fewer inferences and elaborations than subjects who had listened each time only to solve the problem stated in the passage.

Using Rapid Serial Visual Presentation (RSVP), a technique developed by Forster (1970), to study sentence reading, Aaronson and Scarborough (1976) found a number of important differences in reading times between subjects instructed to read for the purpose of recall and subjects instructed to read for comprehension. Recall subjects spent more time reading, and their word-by-word response latencies reflected the surface syntactic structure of the sentences. On the other hand, comprehension subjects' reading times reflected the semantic structure of the sentence materials. Their reading times were longer for key content words in sentences and decreased with contextual redundancy. Thus, different classes of structural components affected reading times only as a function of the reading task presented to subjects and their consequent processing goals.

Graesser et al. (1980) also found elevated reading times under recall instructions in sentence-by-sentence reading of passages. The effect was marginally significant, however, probably because the 'recall' subjects were actually told they would have to write an essay about the passages. Regression

analysis showed that these essay subjects were more sensitive to macro-structural variables in the passage than were the question-answering subjects, who were presumed to be carrying out analysis appropriate for comprehension. These results indicate that these recall subjects were reading for gist meaning more than either the question-answering group or the subjects in Aaronson and Scarborough's recall group.

More recently, Cirilo (1981) had subjects read passages under instructions which emphasized either macrostructure (general comprehension) or microstructure (recall). The macrostructure instructions produced shorter sentence-by-sentence reading times than the microstructure ones. More importantly, however, the macrostructure reading times were affected by the propositional height or centrality of co-references in the text. Reading times in the microstructure condition were not only longer, but they were also affected more by the presence or absence and the distance of co-references in the text. Once again, reading times under recall instructions were longer and were affected more by superficial properties of text structure, while reading times under comprehension instructions were affected more by the underlying semantic structure of the text.

The Present Study

We were interested in examining how listeners process spoken texts. More specifically, we wanted to measure listening times and assess whether these would be affected by the same text variables and processing goals that affect reading times. The first variable we manipulated in this study was normal- vs. random-order presentation of the sentences in the texts, a between-subjects factor. The basic predictions for this factor were that sentence latencies for the random-order condition would be higher than in the normal-order condition and that there would be no decrease in latencies over serial position for the random-order narratives, whereas there would be a decrease for the normal-order narratives. As noted earlier, Olson et al. (1981) found that the predictable structure of normally ordered narratives created shorter reading latencies for sentences later in the passage.

The second variable was related to processing goals. One group of subjects was told they would have to answer questions about the content of the texts (comprehension condition); the other group was told they would have to recall the texts verbatim (recall condition). Recall subjects were expected to have higher sentence latencies than comprehension subjects. Moreover, we expected their latencies to be affected more by surface properties of the syntactic structure than comprehension subjects, whose listening times should be affected more by semantic factors and the macro-structure of the texts. These predictions follow directly from the findings of studies on the reader's goals that were discussed earlier.

In order to assess the effects of manipulating processing goals, several measures of surface structure and of semantic organization were obtained. Surface structure measures that were computed for each sentence included: the number of words, the number of propositions, the number of predicate propositions, the number of modifying propositions, the number of connecting propositions, the physical duration of the individual sentence. We also determined the overall length of each entire text. For a syntactic measure, the number of syntactic

units, i.e. clauses, verb phrases, noun phrases, prepositional phrases, etc. was computed.

The semantic measures that were selected for our analyses were based on Kintsch's (1974) propositional system. Kintsch divides a passage into meaning units consisting of a relational operator, or predicate, and arguments (Turner and Greene, Note 1). Kintsch and Keenan (1973) found that sentence reading times were related more closely to the number of propositions than the number of words in a text. They also found that the number of propositions recalled was a function of the propositions' height in the propositional structure, a measure of how 'central' an idea is to a story. Cirilo and Foss (1980) found that increasing a sentence's propositional height elevated reading times for that sentence, independent of its serial position. The height of individual sentences across texts was not controlled systematically in this study, although we expected to find an elevation of reading times for sentences with greater propositional height.

The propositional structure of a text could affect listening times in two other ways. First, the change in propositional height from one sentence to the next might cause an elevation in listening times rather than simply the absolute height of a single sentence. Propositional height decreases when an idea is elaborated on with details that are not central to the text as a whole. Height then increases when the elaboration stops and discussion of another central idea begins. The jump to a new idea should increase processing time partly because it is new or different information and partly because the listener recognizes the centrality (importance) or height of the new idea. To see the height effect alone, the effect of 'new ideas' would have to be controlled for. Indeed, Graesser et al. (1980) found that the number of new argument nouns in the propositions of a sentence was a significant factor in accounting for reading times. Thus, a count of the number of new argument nouns, as well as the number of new content words, was made for every sentence in each of the stories. The height in the propositional text base was determined by methods described in Kintsch (1974) and Kintsch and van Dijk (1978). An ordinal level of height was determined for every proposition, and a sentence's height was simply the mean of all its propositions' levels. The difference in height for a sentence was determined by subtracting the height of the previous sentence.

The second way for the propositional structure to affect input time is seen in the processing model of Kintsch and van Dijk (1978). As a reader or listener moves from sentence to sentence, his limited short-term memory capacity retains only a subset of the propositions from the previous sentence(s). New propositions are linked to those propositions via shared arguments. In the Kintsch and van Dijk model, each period of linking is called a cycle, and through the cycle a coherent representation of the text is formed. However, sometimes there is no argument overlap between propositions in the buffer and incoming propositions. In this case, long-term memory must be searched to provide an overlap, and this operation requires additional time. To determine a specific value for the extra processing time spent on each sentence, a count was made of how many cycles back one would have to search to find an overlapping argument and how many propositions one would have to retrieve to link the incoming propositions with the propositions in the buffer. This measure was approximately a combination of two factors which Kintsch and Vipond (1978) considered to contribute to the extra load, reinstatement searches and reorganizations.

Parameters for the Kintsch and van Dijk (1978) model also include the size of the processing cycle (number of propositions, clause, sentence, or whatever) and the number of propositions carried in the buffer between cycles. Kintsch and Vipond (1978) found that a buffer size of 4 propositions and a cycle size of one sentence fit their data best. To determine which propositions were carried, they used a 'leading edge' strategy, starting with the highest proposition in the buffer at the end of the cycle and taking it plus the next three connected to it, giving priority to the latest additions. But this strategy often brings in propositions which may not have had any new propositions linked to them for several cycles. Thus, we also used a strategy in which the first proposition chosen for the buffer was the highest one to which a new proposition (or one from memory) had been linked in that cycle. Because the criterion for choosing the first proposition was the recent linking of another proposition, we called this a 'recency' strategy. After choosing this proposition, three additional propositions were chosen as in the leading-edge strategy.

Linking a text's propositions together as a coherent structure may be considered a macro-process, a process very different from the micro-processes of reading or listening to words or parsing syntactic units. Graesser et al. (1980) found that macro-processing accounted for a larger proportion of the variance in reading times than did micro-processing, whether subjects read texts to answer questions or to write an essay. The single largest component of the macro-processing was narrativity, as judged by the subjects. Olson et al. (1981) found a significant difference between reading times for narrative passages and reading times for essays or expository texts. In order to assess whether there were differences in sentence-by-sentence listening times due to text genre, this factor was included in the design by using two fairy tales and two expository texts.

Method

Subjects and design

A total of 60 subjects participated in the experiment. Half of them heard the texts in normal order and half heard them in random order. Within each order, half of the subjects (15) were given comprehension instructions and half were given recall instructions. Subjects in the normal-order condition participated to fulfill a research requirement for an Introductory Psychology class. Subjects in the random-order condition were paid three dollars for participating. All subjects ran in the experiment one at a time.

Materials

Four passages were used in the study. Two were expository texts modified from articles in a newsmagazine. One of them (Dormitories) was short (12 sentences), the other (Locomotion) was long (19 sentences). The two narrative texts were modified from those used by Cirilo (1981) in order that the number of sentences, i.e. the number of output units for a subject, and their syntactic complexity, i.e. use of conjunctions, participial phrases, etc., approximated the expository texts (1). One narrative (Hanuman) was short (13 sentences) and the other (King's Ring) was long (20 sentences).

The author recorded the texts on audiotape with a professional-quality microphone (Electrovoice D054) and taperecorder (Ampex AG-500) in a sound-attenuated IAC booth. These analog recordings were then converted via a 12-bit A/D to digital form so that precise measurements of sentence duration could be made, so that precise "splicing" of each sentence from the whole story into its own unit for presentation could be done, and so that sentences could be presented directly from a computer, thus allowing millisecond accuracy in the collection of response latencies.

The random-order versions of the texts were created at the time of the experiment by the experiment control program. A random number generator determined which sentence would be presented at a given serial position for every text. Thus, every passage was presented in a unique random order to every subject in this condition.

Table 1 provides a description of the stories.

Insert Table 1 about here.

Despite containing more sentences, the narratives were physically shorter (63.15 secs) than the essays (79.05 secs). The 'duration' measure is the sum of the sentences and does not include natural pauses between sentences during the reading of the text because experimental output units ended precisely at the end of the acoustic stimulus. The longer durations for expository passages were partly a function of there being more words in these texts, but not entirely so. While the Dormitory passage was 30% longer than the Hanuman passage, it contained only 12% more words, and while the Locomotion passage was 23% longer than the Ring passage, it had only 4% more words. These differences can be seen directly in the fact that the average duration per word was longer in the expository texts (mean = 307 msec) than the narratives (mean = 263 msec).

Olson et al. (1981) have recently described the difference between these two text genres; narratives have a well-known structure so that the last half of the text is easy to predict and reading times should decrease. Essays do not typically have such a well-defined structure and reading times should therefore remain at text-initial level. Assuming that the two genres have approximately the same latencies at their beginnings, essays should take longer to process overall. The measurement data from the productions indicate that the talker was sensitive to the differential need at input and therefore spent more time speaking the expository texts than the narratives.

The fact that the speaker accounted for the different needs for the two genres is also shown by the correlations of number of words in a sentence with the time per word in production. One factor that would make a text easier to read or to listen to is the redundancy of its information. Redundancy would be expected to show up most clearly in the longer sentences of a passage, where the

Table 1

Description of Texts

Variables	1 Dormitory	2 Locomotion	3 Hanuman	4 King's Ring
		Length		
Duration	64.97	93.13	50.50	75.79
Number of sentences	12	19	13	20
Mean sentence duration (sec)	5.41	4.90	3.88	3.78
Duration range	2.5-9.45	1.2-9.97	1.6-5.87	1.8-7.77
		Words		
Number	202	319	180	307
Mean number per sentence	16.83	16.78	13.85	15.35
Range per sentence	9-32	5-30	5-21	6-28
Msec per word	322	292	280	246
Correlation of number of words and msec per word, by sentence	-.15	.42	-.36	-.01
		Propositions		
Number	91	158	78	138
Mean number per sentence	7.58	8.27	6.0	6.9
Range of number per sentence	3-14	3-16	2-10	4-14

greatest amount of information could be encoded if the passage was not redundant. The predicted effect is that word-by-word reading times would be shorter in long sentences of easy texts than in short sentences of easy texts. Of these four passages, a strong negative correlation of sentence length and time per word in production was found only in the Hanuman passage, the short narrative. Both length of text, in the Ring passage, and the expository genre, in the Dormitory passage, sharply attenuate the affect, and when these two factors are combined, in the Locomotion passage, they turn the correlation positive. When the entire text was long or of a less well-structured genre, the speaker maintained or slowed his rate of speaking on longer sentences rather than increasing it.

The final point to make in the description of the texts is that the number of propositions (Kintsch, 1974, Turner and Greene, Note 1) is greater for expository texts than for narratives: 17% more propositions in the Dormitory text than in the Hanuman story and 15% more propositions in the Locomotion text than in the Ring story. This is interesting because these percentage increases are closer to the increases in total duration than are the percentage increases in number of words. As Kintsch and Keenan (1973) found for silent reading times, propositions account for duration better than words do.

The questions to be answered in the comprehension condition were of a very specific design. Five questions were composed for each story. The first two simply probed memory for specific words in the passage to test recognition of surface structure. One word was from a high propositional level in the text base (as described by Kintsch, 1974), and the other was from a low propositional level. Two more questions probed memory for information contained in one or more clauses in the passage, in order to test integration and storage of information across a set of ideas. Again, one question was from a high propositional level and one was from a low level. Finally, one question could not be answered directly from information in the text, but required an inference derivable from the text. These questions were all answerable with a True or Yes response, and for every question, a corresponding False or No question was formed, often by just negating the affirmative. For the single words, an antonym not containing the root replaced the key word. Two sets of questions were created, each set contained half True or Yes answers and half False or No answers. Questions whose positive form was in the first set had their negative form in the second. Comprehension subjects were randomly assigned to receive one or the other set of questions.

Procedure

The experiment was controlled by a PDP-11/34 mini-computer. Subjects ran one at a time so they could control the latency to output each sentence. They sat in a booth with a CRT monitor mounted at eye level and a 7-button response box on the table in front of them. The stimulus materials were presented over TDH-39 headphones at 80 dB SPL (0dB = .0002 dynes/cm2) with a background of 50 dB white noise to mask the low-level transients at the onset and offset of each sentence. Instructions were presented on the CRT. However, eight sentences were output over the headphones to give subjects some experience with the experimental procedure before the presentation of a practice story. These auditory instructions actually contained no critical information about the procedure, but either reiterated what subjects had already read or explicitly described what

they were experiencing; e.g. "You now know that each sentence starts almost immediately after you push the button, and never before."

After all the instructions were presented, subjects were given a practice text to familiarize themselves with the task and the presentation method in an experimental situation. While the practice text's genre was expository, its topic was "The Myth of Santa Claus" and thus gave it a fairy-tale quality akin to the experimental narratives. Immediately before this text, and before each test passage, the sentence "New Story Coming Up !!!" was displayed on the CRT until the subject pressed button 4, which was marked "Continue." Then, after a one second pause, s/he heard the word "Ready", and after another one second, the first sentence of the passage was presented. The subject then pressed the Continue button when s/he was ready to hear each successive sentence.

After listening to the Santa Claus passage, subjects in the recall condition wrote down the story in as close to verbatim format as they could remember. Comprehension subjects were shown a set of questions on the CRT and pushed button 3 to respond False/No and button 5 to respond True/Yes. Subjects were then given a chance to ask questions about the procedure; they were told that four test passages would be presented in sequence, after which they would perform their respective task. The order in which the texts were presented was completely random.

Subjects in the random-order condition were not told that the texts' sentences had been randomized until after the experiment was completed. However, the sentences in the practice story were presented in random order, so that subjects were somewhat prepared for the main task.

The final task involved listening to a text as quickly as possible, neither for comprehension nor recall. This control measure was used to obtain simple reaction times in this task-situation. Subjects in the random-order condition heard this text's sentences in random order as well. This condition was included to remove as much extraneous variance from the data as possible. Before any analyses were performed on a subject's experimental latencies, the mean of the 12 sentences in this control text was computed and subtracted from all other sentence latencies.

Results

Manipulated Design

Sentence-by-sentence listening latencies were initially examined with a 2x2x2x2x10 ANOVA on the factors of Sentence Order (normal vs. random), Listening Goals or Instructions (comprehension vs. recall), Text Genre (expository vs. narrative), Text Length (short vs. long) and Serial Position. Each text had a different number of sentences and thus a different number of serial positions. To normalize for these differences, some sentences were paired and averaged to obtain one value. The choices of sentence pairs were made so as to leave the beginning and ending sentences unchanged if possible, because these positions are most sensitive to serial position effects. This pairing created 10 serial positions for every passage. For the Dormitory text, sentences 4-5 and 6-7 were

paired, for the Hanuman story sentences 3-4, 5-6, 7-8, for the Locomotion text all but number 11 were paired, and for the Ring story, all sentences were paired.

The main effect of Sentence Order was significant. Normal-order latencies (968 msec) were faster than those obtained in the random-order condition (1220) ($t(56) = 1.77$, $p < .05$, one-tailed). Figure 1 shows the difference of the two orders, including the breakdown by listening goals.

Insert Figure 1 about here.

Although the t value is not large, pairwise comparisons of the data at each serial position for every story across both conditions revealed that 66 of 80 points had normal-order latencies less than random-order ones (sign test $z = 5.81$, $p < .0000001$). This result was obtained for both instructional conditions. For the comprehension condition 31 of 40 comparisons were in the predicted direction ($z = 3.48$, $p < .001$), and for recall, 35 of 40, ($z = 4.74$, $p < .00001$). There was not a significant interaction between sentence order and serial position ($F(9,504) = 1.77$, $p > .05$).

The main effect of listener's goals was significant across orders (comprehension = 875 msec, recall = 1314 msec; $F(1,56) = 9.46$, $p < .005$) as well as within each order (normal order 810 vs. 1127; $t(56) = 1.94$, $p < .03$, one-tailed; and random order 940 vs. 1501; $F(1,28) = 5.74$, $p < .03$). See Figure 1. There was no interaction of order and instruction ($F < 1$). However, the two different listening instructions did produce very different serial position curves ($F(9,504) = 2.61$, $p < .006$). As shown in Figure 2, the serial position function for the comprehension subjects had a fairly flat curve, whereas the serial position function for the recall subjects had a downward slope.

Insert Figure 2 about here.

At first glance, these results appear to be the reverse of what would be predicted by Aaronson (1976). That is, serial position curves for listening time should be flat or rising for recall subjects and should drop for comprehension subjects. But Figure 3 resolves the conflict by showing a three-way interaction of Order by Instruction by Serial Position ($F(9,504) = 4.14$, $p < .0001$).

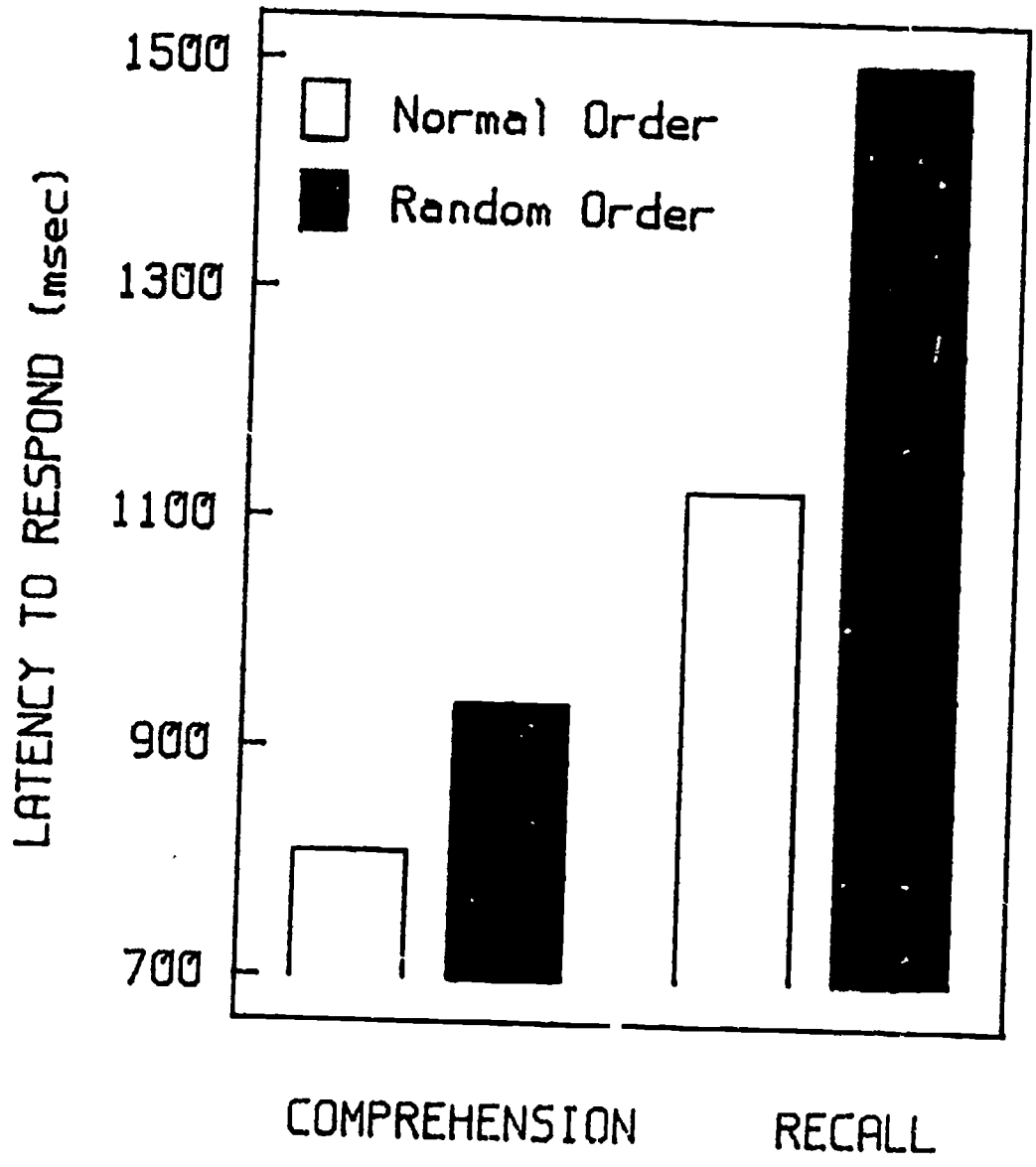


Figure 1. Main effects of text order and listeners' goals on sentence-by-sentence listening times.

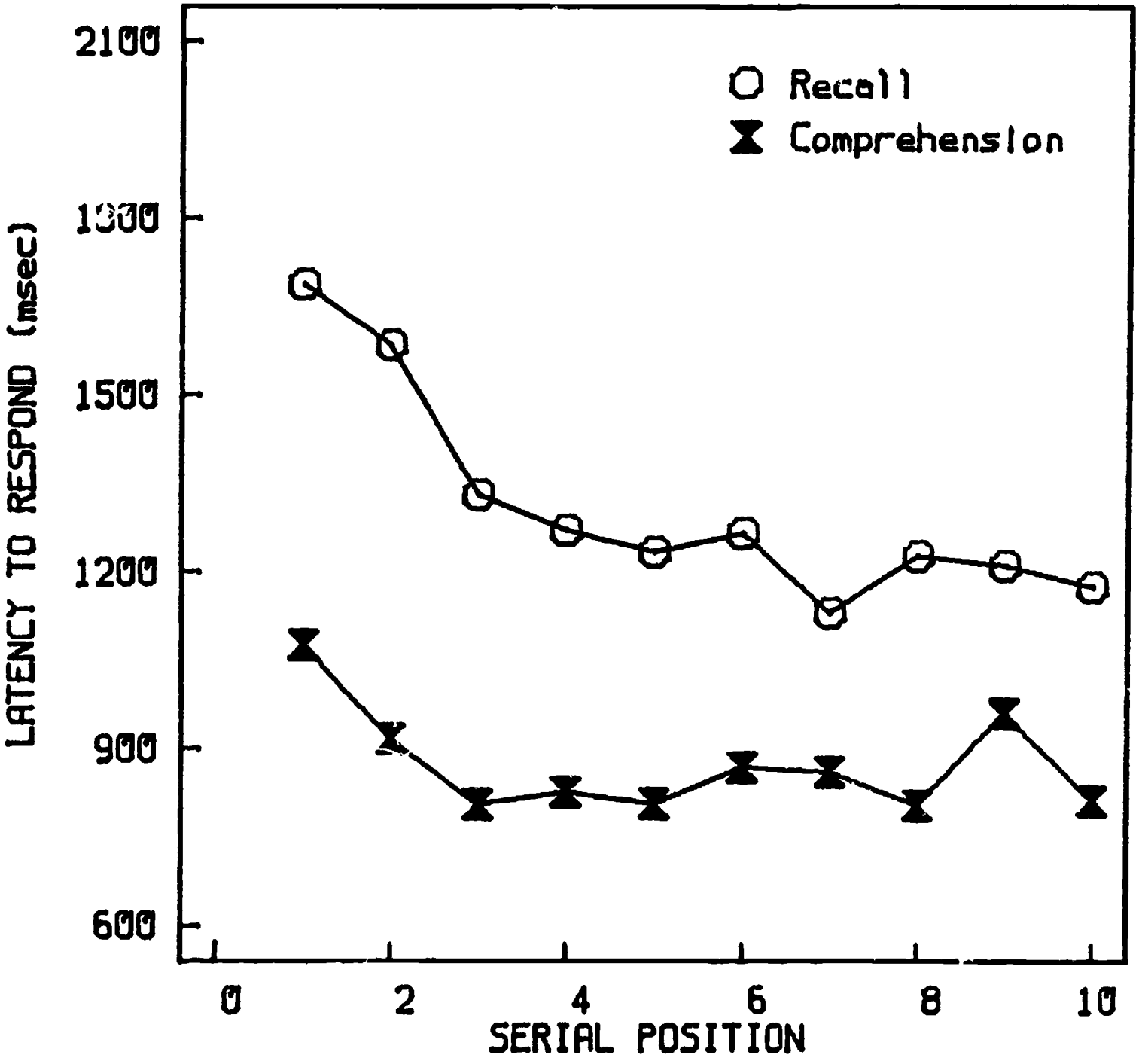


Figure 2. Interaction of listeners' goals and serial position on sentence-by-sentence listening times.

Insert Figure 3 about here.

In this figure, the slopes of the serial position curves for random-order subjects are reversed from the predictions: the comprehension curve is flat while the recall curve slopes down. The curves for normal-order subjects are in the predicted direction, albeit not significantly ($F < 1$). The reason for the lack of significance even for normal-order subjects is shown by their three-way interaction of Instruction by Text Length by Serial Position ($F(9,252) = 2.88$, $p < .003$) in Figure 4.

Insert Figure 4 about here.

Here, regardless of text length, the curves for recall subjects are fairly flat, showing only nonsystematic variations. Subjects listening for comprehension, however, show a different pattern of results. It can be seen that, whereas there is a decrease in listening latencies across serial position in short texts, latencies in long texts actually rise slightly with serial position. This interaction of instruction with text length was not predicted, but shows up in many places throughout these data.

On the other hand, effects on listening latencies were predicted for text genre but were not obtained. We predicted that the serial position curve for narrative texts would slope downward while the curve for expository texts would remain flat. This effect was not obtained either across or within Order conditions. The only systematic interaction was genre by serial position for normal-order subjects ($F(9,252) = 1.94$, $p < .05$). Figure 5 suggests that the interaction may be due to a U-shaped curve for the narratives contrasting with a possibly W-shaped curve for the essays.

Insert Figure 5 about here.

Trend analysis supported this interpretation. The quadratic trend was significant for the narratives ($F(9,261) = 14.96$, $p < .00001$), and the 4th power trend was significant for the essays ($F(9,261) = 188.8$, $p < .00001$). There is no theoretical justification for the expository trend, but since Haberlandt (1980) found a significant rise in latencies at the beginning and end of stories, this trend for narratives may be considered genuine.

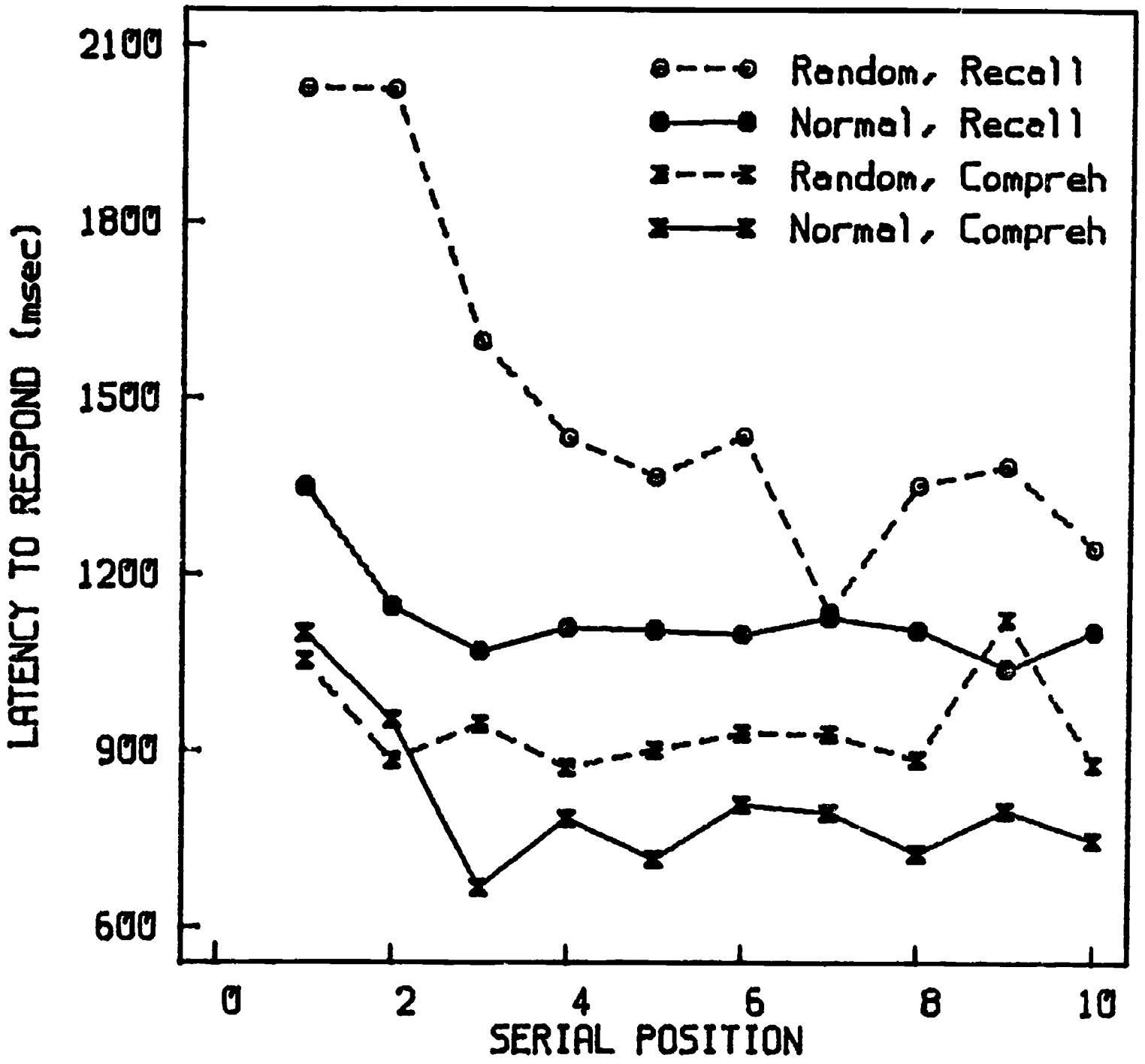


Figure 3. Interaction of text order, listeners' goals and serial position on sentence-by-sentence listening times.

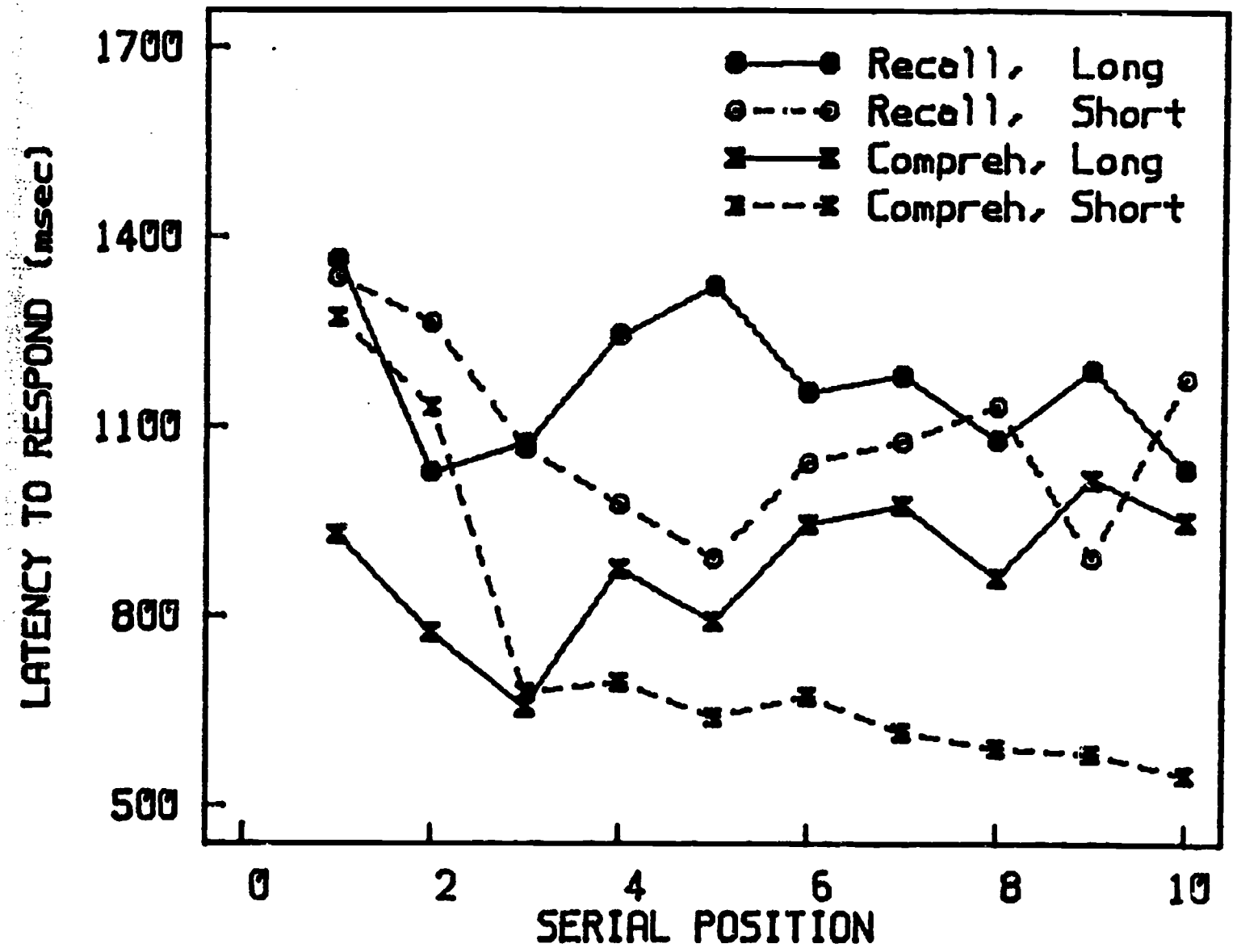


Figure 4. Interaction of listeners' goals, text length and serial position on sentence-by-sentence listening times for normal-order texts.

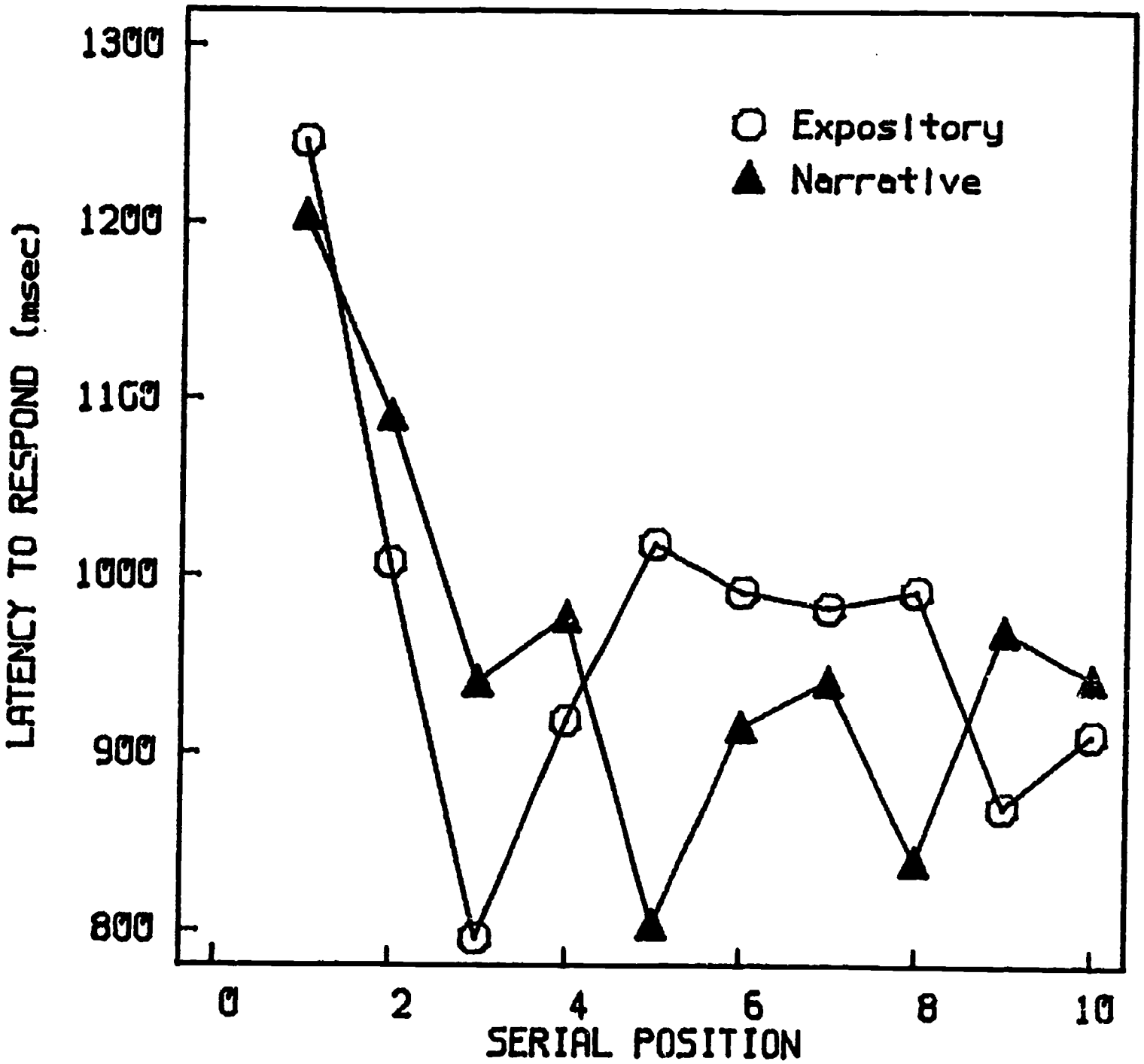


Figure 5. Interaction of genre and serial position on sentence-by-sentence listening times for normal-order texts.

The final main effect to be tested was Text Length. This factor was significant across Orders. Average latency for short texts (1044) was faster than the latency for long texts (1145) ($F(1,56) = 8.59$, $p < .005$). It was also significant for normal-order subjects (914 vs. 1023, $F(1,28) = 5.9$, $p < .02$), but only approached significance for random-order subjects (1174 vs. 1267, $F(1,28) = 3.17$, $p < .09$). The two lengths also displayed significantly different serial position curves ($F(9,504) = 6.06$, $p < .0001$), as shown in Figure 6.

Insert Figure 6 about here.

Subjects tended to spend more time on the beginning sentences of short texts than on the sentences in the middle or the end. In the contrast, they spent a constant amount of time across the serial positions of long texts. A breakdown of the data by Instructional condition shows that this interaction exists for both comprehension and recall, even though the exact form of the interaction is different enough to cause a three-way interaction ($F(9,504) = 2.40$, $p < .02$). Figure 7 shows that for comprehension subjects, latencies for short texts diverge from those for long texts, but for recall subjects, the curve for short texts falls well below that of the long texts in the first half, then meets the curve for long texts in the second half of serial positions.

Insert Figure 7 about here.

Post-hoc Analyses

Further analyses of the data were conducted by applying multiple regression techniques to all the factors described in the Introduction. The purpose of these analyses was to determine whether the sentence-by-sentence listening times would reveal the same text structures that word-by-word and sentence-by-sentence analyses of reading times have provided.

Separate regressions were run for subjects in the normal-order condition and for subjects in the random-order condition. Within each condition, separate regressions were run for subjects in the two different instructional sets, comprehension and recall. Since these factors were varied within subjects, the first step in each regression was to remove uncontrolled, between-subject variance by regressing z-scores of the subjects' overall performances onto the dependent measure. This was done by computing the mean response latency over the 64 experimental sentences for each subject, and then converting these 15 grand means into z-scores. In the normal-order condition, z-scores accounted for 29% of the variance in comprehension subjects' data and 40% of the variance in recall

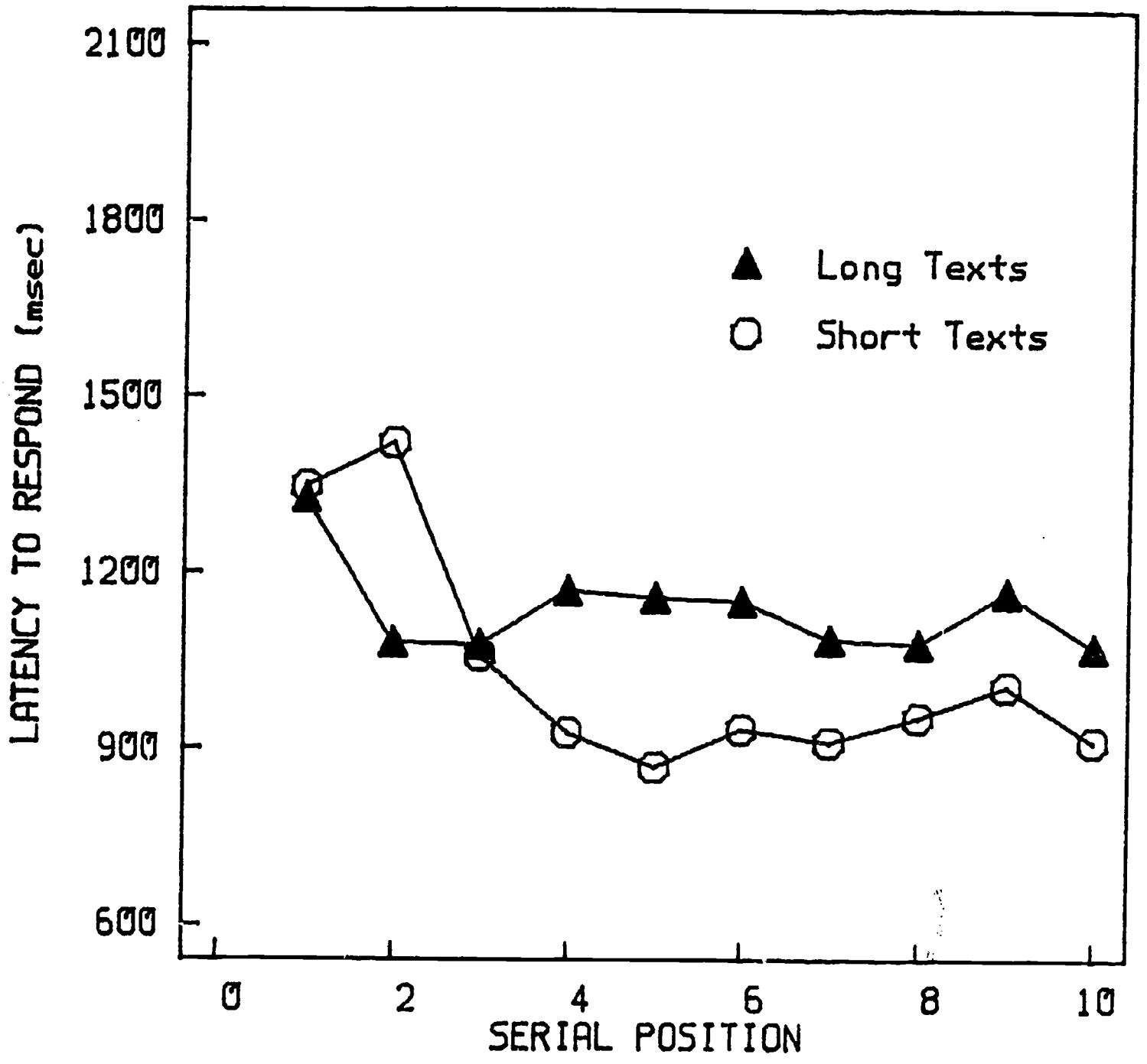


Figure 6. Interaction of text length and serial position on sentence-by-sentence listening times across presentation orders.

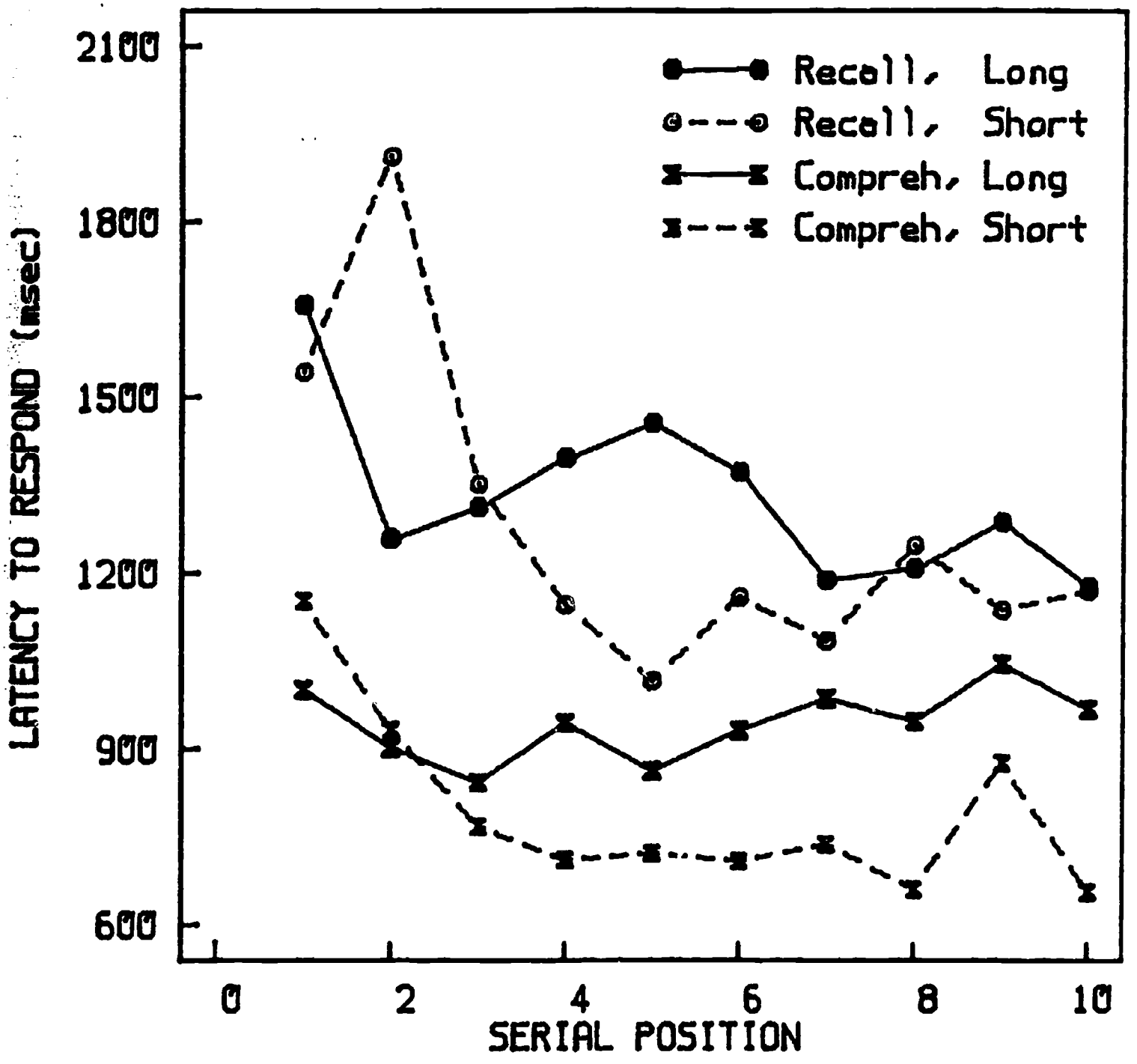


Figure 7. Interaction of listeners' goals, text length and serial position on sentence-by-sentence listening times across normal and random presentation orders.

subjects' data. In the random-order condition, z-scores accounted for 35% of the comprehension variance and 31% of the recall variance.

The proportions of variance accounted for by the significant independent variables are shown in Table 2.

Insert Table 2 about here.

In the normal-order condition, the instructional manipulation strongly affected which factors were loaded into the regression equation. Across all stories, the comprehension subjects were most sensitive to text length, syntax, the number of new argument nouns, and genre, in that order. Recall subjects, on the other hand, were sensitive to the Kintsch and van Dijk processing load as computed with the leading edge strategy, serial position, and the Kintsch and van Dijk processing load computed with the recency strategy.

Since two of the factors loading into the comprehension subjects' regression were bi-leveled, further regressions were performed on the two levels of text length (short vs. long) and genre (expository vs. narrative). These results are displayed in Table 3.

Insert Table 3 about here.

For the short texts, serial position is the dominant factor (cf. Figure 4), with genre, number of new argument nouns, and the propositional height of each sentence also playing a role. For the long texts, syntax entered the equation first, followed by number of new argument nouns, serial position, and two of the three sub-classes of propositions, modifiers and connectors. In sum, the two text lengths shared two factors, new argument nouns and serial position, but were very different on the remaining factors. For short texts, the whole-text factors of genre and propositional height entered the equation, but for the long texts, the more localized, surface factors of syntax and number of certain types of propositions in each sentence were important.

While genre was a small factor over-all for the comprehension subjects in the normal-order condition, the breakdown of the subjects by genre showed very different factors being correlated with the data for each genre. For expository passages, text length was the most highly correlated factor. Word length in production and the number of connecting propositions entered the equation next as very weak factors, and they were followed by two factors, new argument nouns and syntax, that had higher correlations. For narrative passages, three factors

Table 2

Variables which correlated significantly

with listening time data for normal-order subjects

Instructional set	Variable	F-to-	R^2
Comprehension	Text length	18.2	.013
	Syntax	6.0	.004
	New argument nouns	14.6	.010
	Genre	6.5	.005
Recall	Kintsch & van Dijk, leading-edge strategy	21.1	.013
	Serial position	11.9	.007
	Kintsch & van Dijk, recentist strategy	6.1	.003

Note. The F-to-enter statistic has no simple relation to F values that are normally tabled, so significance levels cannot be reported. For further see Dixon and Brown (1979), Appendix C2.

Table 3

Variables which correlate significantly with

listening time data for normal-order comprehension subjects

Division of texts	Variable	F-to-enter	R^2
Short	Serial Position	38.1	.066
	Genre	9.2	.015
	New argument nouns	8.6	.013
	Propositional height	4.2	.007
Long	Syntax	12.3	.014
	New argument nouns	7.5	.008
	Serial position	8.2	.009
	Modifying propositions	4.2	.005
	Connecting propositions	5.6	.006
Expository	Text length	28.1	.040
	Msec per word	8.3	.010
	Connecting propositions	4.2	.006
	New argument nouns	15.5	.020
	Syntax	9.5	.011
Narrative	Propositional height	8.8	.012
	Msec per word	9.6	.013
	Change in prop height	5.5	.008

entered the equation with approximately equal correlations. These were propositional height, word length in production, and change in propositional height from sentence to sentence.

The regressions for the random-order subjects did not include as many factors because the texts did not form coherent passages and thus did not have genuine values for propositional height, change in height, or either of the Kintsch and van Dijk processing measures. Furthermore, since the number of new argument nouns in each sentence is dependent on the nouns of previous sentences, this factor would have to be computed for every presentation of every text, and this was not done.

The most striking aspect of the regressions for the random-order subjects is the relatively few number of factors that were significant determinants of the dependent measure. For comprehension subjects, only text length (F-to-enter = 34.49 (2), proportion of variance = 2.27%) and sentence duration (F-to-enter = 6.35, proportion of variance = 0.42%) were included. No factors correlated highly enough with the recall subjects' data to be included. Analyzing the comprehension data separately by text length revealed the same lack of contributing factors. For short texts, serial position (F-to-enter = 5.1, proportion of variance = 0.8%) was the only factor (as it was the first and strongest factor for this cell of the normal-order data). For long texts, only sentence duration (F-to-enter = 7.5, proportion of variance = 0.8%) was included.

One last point about the random-order data is worth noting here. We observed a significant correlation of the recall subjects' data with the propositional height of each sentence as determined from the normal-order propositional structure (F-to-enter = 14.39, proportion of variance = 1.48%). This observation suggests that despite the random presentation order of the sentences, subjects were able to determine which sentences in the texts were the most central or highest in the propositional structure.

Discussion

In the introduction we outlined three broad purposes to this study: 1) testing a new experimental paradigm, 2) examining the structure of spoken texts, and 3) examining the effects of a listener's goals on listening performance. In this discussion, we will synthesize the findings in the latter two areas and validate the new paradigm by showing the similarity of the findings to those obtained in studies of written texts.

First of all, the fact that there is some structure to texts is indicated by the elevated listening times for random-order subjects over normal-order subjects. Whether that structure is semantic or prosodic or both, the normal order of texts was easier to listen to. This was also found in a study of written texts by Kintsch et al. (1977), who claimed that the texts were being re-ordered to the original state during input. That claim is supported in this study by the fact that random-order subjects in the recall condition were also sensitive to the propositional height of sentences as they would occur in the normal-order story. Apparently these subjects were trying to find the texts' natural structure and were spending additional time linking the text information to the central ideas. Still, most of the texts' structure was not discernible,

for very few factors correlated highly enough with the random-order subjects' data to be significant in the regression analyses. Those factors that were significant were only the very global ones at both the text level (text length, serial position) and sentence level (sentence duration).

Different text genres evidently have different structures (Olson et al., 1980). Narratives have a well-known structure so that input times should decrease over serial position. Essays have a less-defined structure, so input times would not be expected to decrease over serial position. In this experiment, serial position curves for both genres started much higher than they ended. The essay curve may have dropped because subjects knew they would be tested only after all texts were listened to: since it would be very hard to remember all the details, they may have spent more time at the beginning of each text to be sure to get the main ideas and structure of the passage.

Although both curves did start higher than they ended, the shapes of the two functions were significantly different. The curve for narrative texts was the same as that of an episodic unit (Haberlandt, 1980). To really test whether listening time is a function of episodes, narratives with more explicitly defined episodes would have to be used.

A structural effect that was not predicted but was found concerned the effects of text length. While this appears to be only a very global structure parameter, the faster listening times for the shorter texts strongly suggest that clues to length are encoded within the text structure. Thus, for shorter texts with less total information, subjects do not have to spend as much time integrating the entire text. Kintsch et al. (1977) have shown integration of text information during visual presentation. The results of the present study (see Figure 6) show that latencies at the beginning of short texts are as high or higher than beginning latencies for long texts. However, the former decrease over serial position while the latter stay roughly the same. These findings suggest that clues to text length are encoded at the beginning of the texts, so that subjects can quickly determine how much more is in the text and thus what pace they should set in listening to the sentences.

Three structural factors should be mentioned because they did not produce the strong effects that they have in earlier reading studies. The first is the number of propositions in a sentence. The total number of propositions did not show up in any regression analysis, although the number of connecting propositions and modifying propositions were significant factors for comprehension subjects listening to normally ordered texts that were either long or expository. The second factor is propositional height in the text. While it was a factor in the data for random-order recall subjects, it was not a factor in the main regressions for normal-order subjects. It did figure into two of the subdivided regressions, however. Subjects listening to texts in their normal order for the purpose of comprehension were sensitive to propositional height only if the texts were either short or narrative. The third factor is change in propositional height, which was only significant for comprehension subjects listening to normal narratives and here it was the least significant factor. Although none of these three factors played major roles in this listening study, the fact that they did correlate significantly does indicate that they were much more important than most of the other factors described in the Introduction, and therfore should not be abandoned.

The instructional manipulation of listening goals was very significant. Comprehension subjects spent less time listening to sentences than recall subjects, indicating that they were processing the material differently than the recall subjects. In a more detailed comparison of the two instructional conditions, Aaronson (1976) predicted that RSVP latencies for comprehension subjects would fall over serial position of a sentence while those for recall subjects would remain constant or rise over a sentence. Because other effects of instructions have been found in studies of text processing (Fredericksen, 1972; Graesser et al., 1980), we expected to find this difference over serial positions in a passage. However, the serial position effect was displayed only in the normal-order condition for short texts. For long texts, Figure 4 shows that latencies of comprehension subjects start downward in the first three serial positions, but then rise. This is further support for the interpretation that important clues to a text's length are encoded in the beginning sentences. For random-order texts, the prediction derived from Aaronson (1976) was actually reversed; the comprehension curve is fairly flat while the recall curve drops markedly (see Figure 3). The comprehension curve probably flattens due to the increased difficulty of the random ordering. The recall curve may fall because these subjects expend more effort at the texts' beginnings in order derive as much of the gist as possible before hearing a long list of poorly related information.

A major difference between subjects in the two instructional conditions is revealed in the regression analyses for normal-order subjects. Whereas the recall subjects were sensitive to the Kintsch and van Dijk (1978) macro-structure factors, the comprehension subjects were more sensitive to the general measures of text length and genre and the surface measure of syntax. At first glance, it would appear that these data are reversed from the prediction that comprehension subjects would be affected more by semantic factors while the recall subjects would attend more to surface factors. But since the recall subjects had to remember four different passages before writing about any of them, they may have adopted a strategy of listening for gist meaning with the intention of reconstructing each text at output (Fredericksen, 1972). This account is supported by Graesser et al's (1980) finding that subjects reading texts in order to "write an essay about them", i.e. derive the gist meaning, were more sensitive to macro-structure than to micro-structure. While verbatim recall was emphasized in the current study, it is nearly impossible to memorize the 64 sentences with one listening, so the gist-plus-reconstruction strategy is a very reasonable way to approximate rote memory.

Finally, the data shown in Table 3 suggest that for the normal-order comprehension subjects, there was a relation between long and expository texts, and another between short texts and narratives. The long and expository texts share three factors in their regression equations, syntax, new argument nouns, and number of connecting propositions. This is more overlap than in any other pair of regressions, and the factors are simple counts of sentence components. On the other hand, while short texts and narratives share only one factor, it is propositional height, a semantic factor which is determined by a sentence's relation to all others in the text. Thus, there is a fairly tight relation of long and expository texts in that both are processed in terms of their micro-structure, and there is a corresponding, though looser, relation of short and narrative texts in terms of macro-structure processing.

These pairwise relations in the comprehension subjects' data are also seen in Figure 4, where short texts have a downward-sloping serial position curve and long ones have a flat curve. This difference in the curves was exactly what is predicted, but not obtained, for the narratives and expository texts respectively. Thus, from the serial position curves and the regression analyses, it appears that either short texts or narratives are fairly easy to process and can be listened to for their meaning, but long or expository texts are more difficult and comprehension subjects attend more to sentential or surface factors.

Summary and Conclusions

The sentence-by-sentence listening measure proved reliable by revealing effects of text structure and listener's goals that have been reported in the reading literature. Subjects listening to randomly ordered texts had longer latencies than subjects listening to normally ordered texts, indicating a lack of semantic structure which could facilitate processing of individual sentences. Subjects listening in order to recall the passage spent longer on each sentence than subjects listening to answer questions. The Kintsch and van Dijk (1978) processing model predicted latencies for the recall subjects, just as it predicted 'readability' for subjects in Kintsch and Vipond's (1978) study. An unexpected effect was the overall length of each passage: short texts revealed shorter average latencies for each sentence. This factor has not been manipulated in reading studies, so no comparison can be made. However, the strength of the effect leads us to predict that it would be found in reading as well. This prediction is supported by the relationships of long texts to expository texts and short texts to narratives in the latencies of comprehension subjects. Each of the pairs shared their own regression factors, and the length factor behaved as the genre factor was predicted to behave in the interactions with serial position and listening goals. Genre has proved a significant factor in reading studies, so the close relation of genre and length in this listening study suggests that the length effect will be revealed in reading as well.

A number of predictions were either not supported or only weakly supported. Latencies for comprehension subjects were expected to drop across serial position but they did not. Similarly, latencies for narrative texts were predicted to drop over serial position, but they did not either. The number of propositions and the height of propositions were expected to predict listening times for all conditions, but they did so only for a few. The effects of text genre, number of propositions and propositional height have been firmly supported in the reading literature, so the weak support here is surprising. A couple of reasons may be offered for why the effects may actually exist in listening but were not revealed in this study.

First, only four texts were used, so it is possible that some of these results are text-specific. One of the passages, the Hanuman text, is a fairy tale of India, and although the storyline is very typical, some of the elements are unusual. For example, many subjects could not correctly say the hero's name (Hanuman) even though they had heard it six times in the story. Testing the sentence-by-sentence listening time measure on a broader sample of texts will certainly be necessary to validate it.

Second, the output unit of a sentence may be too gross to pick up some of these effects. As shown in Table 1, the sentences' durations ranged from 1.2 seconds to 9.9 seconds. Ninety percent of the sentences were over 2.5 seconds in length, and subjects apparently came to expect long sentences because listening times for the short ones were much higher than predicted by the sentence length factor. Thus, the mandatory output unit was highly variable and appeared to add some noise to the data. Wingfield and Nolan (1980) let subjects choose the size of the output unit themselves by giving them complete control over where the acoustic stimulus started and stopped in a passage. Until the technology to give subjects complete control of the digital stimuli is available in our lab, clauses may prove a better unit than sentences. Clauses are significant language processing units for listeners and readers (Jarvella and Herman, 1972; Hurtig, 1978; Wingfield and Nolan, 1980) and Cirilo and Foss (1980) have used clauses as output units for readers. Furthermore, measurements of the production data in this study showed pauses averaging about half a second at clause boundaries (Klatt, 1976). Since long sentences often have several clauses, the clausal unit would allow closer analysis of the on-line computational processes carried out by the listener during sentence comprehension. Thus, the propositional measures of the texts might be revealed more clearly, and extraneous effects of sentence length would be removed.

Reference Note

1. Turner, A., and Greene, E. The construction and use of a propositional text base (Tech. Rep. 63). Boulder, CO: U of Colorado, Institute for the Study of Intellectual Behavior, 1977.

References

Aaronson, D. Performance theories for sentence coding: Some qualitive observations. Journal of Experimental Psychology: Human Perception and Performance, 1976, 2(1), 42-55.

Aaronson, D, and Scarborough, H. Performance theories for sentence coding: Some quantitative evidence. Journal of Experimental Psychology: Human Perception and Performance, 1976, 2,(1), 56-70.

de Beaugrande, R. Text, discourse, and process. Norwood, N.J.: Ablex, 1980.

Chafe, W. Integration and involvement in speaking, writing, and oral literature. In D. Tannen, (Ed.), Spoken and Written Language. Norwood, N.J.: Ablex, 1982.

Cirilo, R.K. Referential coherence and text structure in story comprehension. Journal of Verbal Learning and Verbal Behavior, 1981, 20, 358-367.

Cirilo, R. and Foss, D. Text structure and reading time for sentences. Journal of Verbal Learning and Verbal Behavior, 1980, 19, 96-109.

Dixon, W.J. and Brown, M.B. (Eds.) Biomedical Computing Programs P-Series, 1979, Los Angeles: U. of California Press.

Fredericksen, C. Effects of task-induced cognitive operations on comprehension and memory processes. In R. Freedle and J. Carroll, (Eds.), Language comprehension and the acquisition of knowledge, N.Y.: Wiley, 1972.

Forster, K. Visual perception of rapidly presented word sequences of varying complexity. Perception and Psychophysics, 1970, 8, 215-221.

Graesser, A., Hoffman, N., and Clark, L. Structural components in reading time. Journal of Verbal Learning and Verbal Behavior, 1980, 19, 135-151.

Grosjean, F. Spoken word recognition processes and the gating paradigm. Perception and Psychophysics, 1980, 28(4), 267-283.

Haberlandt, K. Story grammar and reading time of story constituents. Poetics, 1980, 9, 99-116.

Hildyard, A. and Olson, D. On the comprehension and memory of oral and written discourse. In D. Tannen, (Ed.), Spoken and written language, Norwood, N.J.: Ablex, 1982.

Hurtig, R. The validity of clausal processing at the discourse level. Discourse Processes, 1978, 1, 195-202.

Jarvella, R.J. and Herman, S.J. Clause structure of sentences and speech processing. Perception and Psychophysics, 1972, 11, 381-384.

Johnson-Laird, P. Experimental psycholinguistics. Annual Review of Psychology, 1974, 25, 135-160.

Kay, P. Language evolution and speech style. In B. Blount and M. Sanches, (Eds.), Sociocultural determinates of language changes, NY: Academic Press, 1977.

Kavanagh, J.F. and Mattingly, I.G. (Eds.) Language by eye and by ear. Cambridge, Ma: MIT Press, 1972.

Kintsch, W. The Representation of meaning in memory. Hillsdale, N.J.: Erlbaum, 1974.

Kintsch, W. and van Dijk, T. Toward a model of text comprehension and production. Psychological Review, 1978, 85, 363-394.

Kintsch, W. and Keenan, J. Reading rate as a function of the number of propositions in the base structure of sentences. Cognitive Psychology, 1973, 5, 257-274.

Kintsch, W., Kozminsky, E., Streby, W., McKoon, G., and Keenan, J. Comprehension and recall of text as a function of content variables. Journal of Verbal Learning and Verbal Behavior, 1975, 14, 196-214.

Kintsch, W., Mandel, T.S., and Kozminsky, E. Summarizing scrambled stories. Memory and Cognition, 1977, 5(5), 547-552.

Kintsch, W. and Vipond, D. Reading comprehension and readability in educational practice and psychological theory. In L. Nilsson, (Ed.), Memory: processes and problems, Hillsdale, N.J.: Erlbaum, 1978.

Klatt, D. H. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. Journal of the Acoustical Society of America, 1976, 59(5), 1208-1221.

Levelt, W.J.M. A survey of studies in sentence perception. In W.J.M. Levelt and G.B. Flores d'Arcais (Eds.), Studies in the perception of language, N.Y.: Wiley, 1978.

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

Mandler, J. and Johnson, N. Remembrance of things parsed: Story structure and recall. Cognitive Psychology, 1977, 9, 111-151.

Olson, G., Duffy, S., and Mack, R. Applying knowledge of written conventions to prose comprehension and composition. New Directions for Teaching and Learning, 1980, 2, 67-84.

Olson, G., Mack, R., and Duffy, S. Cognitive aspects of genre. Poetics, 1981, 10, 283-315.

Reder, L.M. The role of elaboration in the comprehension and retention of prose: A critical review. Review of Educational Research, 1980, 50(1), 5-53.

Tannen, D. (Ed.) Spoken and written language. Norwood, N.J.: Ablex, 1982a.

Tannen, D. (Ed.) Coherence in spoken and written discourse. Norwood, N.J.: Ablex, 1982b.

Thorndyke, P. Cognitive structures in comprehension and memory of narrative discourse. Cognitive Psychology, 1977, 9, 77-110.

Wingfield, A., and Nolan, K. Spontaneous segmentation in normal and time-compressed speech. Perception and Psychophysics, 1980, 28 (2), 97-102.

Footnotes

1). Thanks to Randolph Cirilo for providing the stories from his study.

2). See Note, Table 2.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

Effects of Syllable Structure on Adults' Phoneme Monitoring Performance*

Rebecca Treiman
Aita Salasoo
Louisa M. Slowiaczek
and
David B. Pisoni

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

*This research was supported, in part, by NSF Grant BNS 80109892, NIMH Grant MH-24027, and NINCDS Grant NS-12179 to Indiana University in Bloomington.

Abstract

It has been proposed that the English syllable consists of an onset and a rime. The onset contains an initial consonant or consonant cluster; the rime contains the vowel and any following consonants. The two experiments reported here tested the hypothesis that the onset behaves as a cohesive unit in a phoneme monitoring task. Experiment 1 found that when subjects monitor for the fricatives /f/ and /s/ in syllable-initial position, they display increased response times and errors with CCV (consonant-consonant-vowel) syllables as compared to CVC and CV syllables. Similar results were obtained in Experiment 2, in which subjects monitored for stop consonants in the same three types of syllables. These results provide evidence for the psychological reality of the onset. They further suggest that the onset acts as a unit in the perception of speech.

Linguistic and behavioral evidence suggests that the syllable has a hierarchical internal structure. According to linguists such as Fudge (1969) and Halle and Vergnaud (1980; Vergnaud & Halle, Note 1), the English syllable consists of an onset and a rime. The onset, which is optional, contains one or more consonants. The rime contains an obligatory peak, or vowel nucleus, and an optional coda consisting of one or more consonants. Word-final syllables may end with an appendix of inflectional suffixes, but this unit will not concern us here. Distributional evidence supports this account of syllable structure. Virtually any onset can co-occur with any rime, but there are severe constraints on which codas can occur with which peaks (e.g., Fudge, 1969). Stress phenomena also provide evidence for this account of syllable structure. In English, rules of stress assignment refer only to the rime (Chomsky & Halle, 1968). The onset of the syllable is irrelevant.

Analyses of errors in spontaneous speech suggest that onsets and rimes function as units at some level of the speech production process. In Spoonerisms, for example, entire onsets are often exchanged, as in the example sweater drying -> dreater swying (Fromkin, 1971). Errors such as brake fluid -> blake fruid (Fromkin, 1971), in which the onset is divided, are less frequent (Mackay, 1970). While errors of the latter kind show that individual phonemes can function separately, their scarcity has been taken to support the hypothesis (MacKay, 1972) that syllables are first specified for production in terms of their initial consonant group plus remainder. Only late in the production process are initial clusters recoded into their constituent phonemes. Studies of blend errors, in which two roughly synonymous words combine to produce an unintended form (e.g., start + go -> sto), provide additional support for the claim that consonant clusters often behave as units in speech production (Mackay, 1972). Blends such as so, in which the initial consonant cluster of the first word is divided, have been found to be less frequent than blends in which the consonant cluster remains intact (MacKay, 1972).

Further behavioral evidence for the division of syllables into onset and rime units comes from studies of word games. Existing word games, such as Pig Latin, divide syllables at the onset/rime boundary (Hockett, 1967). In addition, adults learn novel word games that retain the onset and the rime more readily than games that divide these units (Treiman, submitted). For example, a rule by which two syllables are broken at the onset/rime boundary and blended into a new syllable (e.g. /krɪnt/ + /glʌpθ/ -> /krʌpθ/) is relatively easy to learn. (See Table 1 for a key to the phonetic notation used in this paper.) Rules by which the syllables are broken after the initial consonant (/krɪnt/ + /glʌpθ/ -> /klʌpθ/) or after the vowel (/krɪnt/ + /glʌpθ/ -> /krɪpθ/) are more difficult. Treiman's (submitted) recent findings suggest that when learning word games--a task that requires a fairly high level of metalinguistic awareness--adults often treat onsets and rimes as separate units. They can bring these units to consciousness and actively manipulate them when performing various tasks.

Children's ability to analyze spoken syllables into phonemes is also affected by syllable structure. Treiman (1980) asked 5-year-olds to judge whether spoken syllables began with a target phoneme. The fricatives /s/ and /f/ served as targets. Syllables that began with the target had one of three structures: CV (consonant-vowel), as in /sa/; CVC, as in /san/; or CCV, as in /sna/. The postulated structures of these three types of syllables are displayed in the form of three tree diagrams in Figure 1.

Insert Figure 1 about here

As the figure shows, the initial consonant of a CCV syllable is at a lower level in the hierarchy than are the initial consonants of CVC or CV syllables (which are equivalent). That is, the initial consonant of a CCV syllable is embedded within the onset. The initial consonant of a CVC or CV syllable is the onset. Consistent with this account, Treiman (1980) found that children were more likely to miss the target phoneme when it began a CCV syllable than when it began a CVC or a CV syllable. Error rates to CVC and CV syllables were indistinguishable. Treiman (1980) therefore suggested that children have difficulty analyzing onsets into their constituent phonemes. Barton, Miller, and Macken (1980) have recently made a similar proposal.

The experiments reported here examined the role of the onset for adult listeners. While the evidence reviewed above suggests that the onset functions as a cohesive unit for adults in the production of speech and in the learning of word games, its role in speech processing has not been studied. The present experiments employed a phoneme monitoring task in which subjects were asked to judge as quickly as possible whether spoken syllables began with a target phoneme. As in Treiman's (1980) earlier study with children, CV, CVC, and CCV stimuli were used. If the onset functions as a unit--a unit that takes some additional time to analyze into its constituent phonemes--response times to phoneme targets in CCV syllables should exceed response times to those same targets in CVC and CV syllables.

Experiment 1

Method

Subjects. Thirty-three students from Indiana University participated to fulfill an introductory psychology course requirement. All subjects were native English speakers with no known history of hearing loss or language disorder.

Stimuli. The fricatives /f/ and /s/ served as target phonemes. The syllables that began with the target--the experimental syllables--were the same ones previously used by Treiman (1980). They were constructed in groups of three: a CV syllable, a CVC syllable, and a CCV syllable. The items in a group overlapped in their phonemic composition. Samples are shown in Table 1.

Insert Table 1 about here

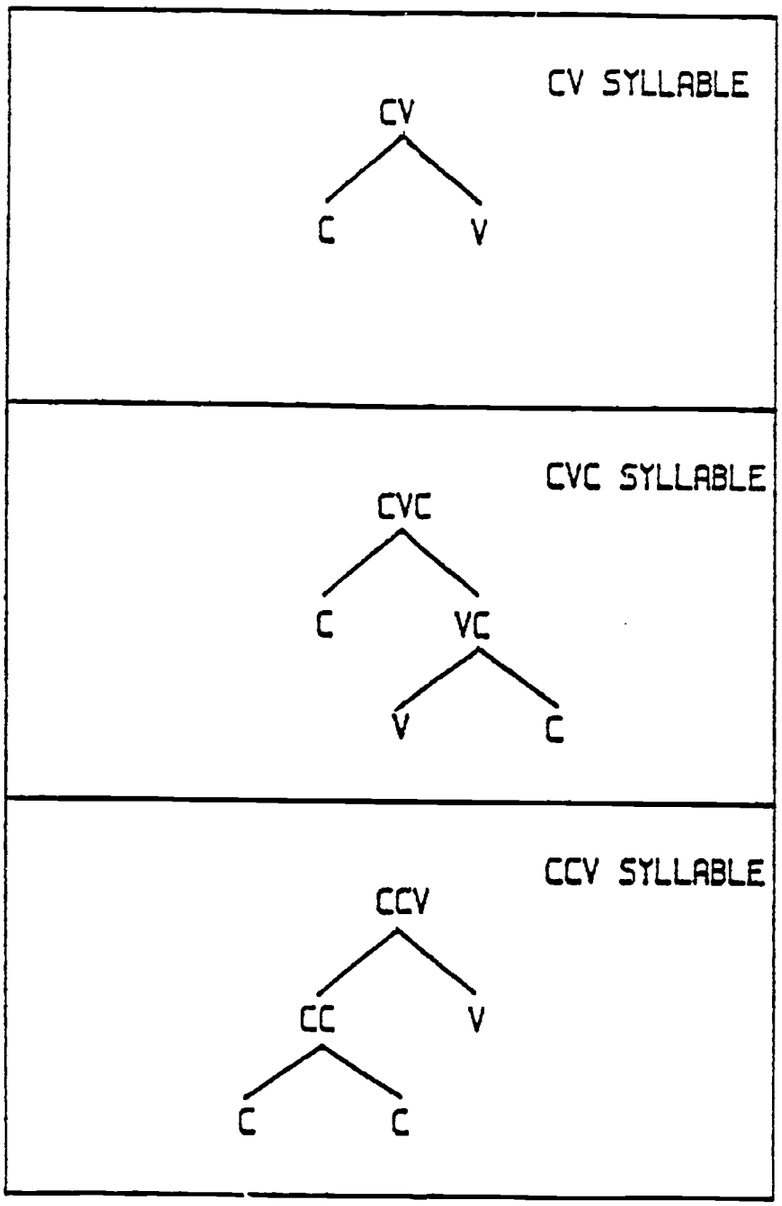


Figure 1. Postulated structures of CV, CVC, and CCV syllables.

Table 1

Sample Stimuli for /s/ Target

	Syllable Structure		
	CV	CVC	CCV
Experimental	/sa/	/san/	/sna/
	/so/	/son/	/sno/
	/si/	/sik/	/ski/
Filler	/tɔ/	/van/	/kwa/
	/ri/	/nik/	/glɔ/
	/he/	/nep/	/gru/

Key to phonetic notation:

i	beet	u	boot
ɪ	bit	o	boat
e	bait	ɔ	bought
a	hot	ɵ	bath
ʌ	but		

Four groups of experimental syllables were constructed for the /f/ target. Eight groups were constructed for the /s/ target, and were divided into two sets of four groups each. Added to each set of twelve experimental syllables--one /f/ set and two /s/ sets--were twelve filler syllables. These were syllables that did not begin with the target phoneme. The fillers were counterbalanced with regard to syllable structure, as shown in Table 1. All stimuli were phonologically legal in English, and most were not real words. Those stimuli that were real words were evenly divided among the CV, CVC, and CCV categories.

Three sets of practice syllables were also constructed, one with the target /f/ and two with the target /s/. Each set of practice syllables contained three experimental syllables and three filler syllables, which were different from those used in the experiment proper.

A male speaker read the syllables in citation form from printed cards. The stimuli were recorded using a professional microphone (Electro-Voice Model # D054) and tape recorder (Ampex AG-500) in a sound-attenuated IAC booth. The stimuli were then low-pass filtered at 4.8 KHz, digitized at a 10 KHz sampling rate via a 12-bit A/D converter, and stored digitally on a PDP-11/34 computer for later use during the experiment.

Procedure. The entire experimental procedure was run in real-time by a PDP-11/34 computer, which presented stimuli, recorded subjects' responses, and provided feedback after each trial. For each stimulus set, an experimental block of 120 trials was generated by repeating each of the 24 syllables in the set five times. The experimental session consisted of three blocks of trials, one /f/ block and two /s/ blocks. Each block of test trials was preceded by a practice block of 18 trials, which consisted of three repetitions of each of the six practice syllables. The practice block used the same target phoneme as the subsequent test block.

Experimental sessions were conducted with groups of three to six subjects. Subjects listened to the syllables over TDH-39 headphones. At the start of each block of trials, subjects were told the target phoneme for that block. The experimenter pronounced the target phoneme in isolation and in sample syllables. The letter name was not used. Subjects were required to detect as quickly and accurately as possible whether or not each syllable began with the previously specified target phoneme. On each trial, subjects responded "Yes" or "No" by pressing the appropriately labeled button on a response-box in front of them.

Each trial proceeded as follows: A 500-millisecond cue light on the subject's response box marked the beginning of a trial. Following a one second pause, the syllable was presented at 77 dB SPL (re. .0002 dynes/cm^2) through the subject's headphones. After the subject entered his or her response, the feedback light above the correct response was illuminated for one second. The next trial sequence began after a three second interval.

Subjects' "Yes"/"No" responses and latencies were collected by a computer-controlled experimental program. Latencies were measured from the beginning of the spoken presentation of a syllable. In each session, the order of target blocks was random. Within each block of 120 trials, the stimuli were presented in random order. All subjects heard a total of 54 practice syllables and 360 test syllables. The experimental session lasted about one hour.

Results and Discussion

For each subject, both the latencies for the correct responses and the error data were collapsed across the five repetitions of each stimulus. The error data were expressed as percentages. These data are shown in Figure 2. While the figure shows the results for /f/ and /s/ targets separately, there were no significant differences in response latencies or errors between the two target phonemes ($F(1,32) < 1.0$ for both). In subsequent analyses, therefore, the data were collapsed across both phonemes.

Insert Figure 2 about here

The latencies for correct detections were submitted to a repeated measures analysis of variance with syllable structure and syllable group as fixed variables. A main effect of syllable structure was found ($F(2,62) = 46.97$, $p < .0001$). The mean response times were 963 milliseconds for CV syllables, 968 milliseconds for CVC syllables, and 1005 milliseconds for CCV syllables. Planned comparisons showed that the latencies for CVC syllables did not significantly exceed those for CV syllables ($F(1,64) = 1.17$, $p > .25$). This result was expected, since both CV and CVC syllables begin with singleton consonant onsets. Latencies for CCV syllables, however, significantly exceeded the average latencies for CV and CVC syllables ($F(1,64) = 97.15$, $p < .0001$). That is, subjects were slower to respond correctly to a target phoneme when it was part of an initial cluster than when it was not. A main effect of syllable group was also found ($F(11,352) = 2.33$, $p < .01$), as was an interaction between syllable group and syllable structure ($F(22,704) = 7.79$, $p < .0001$). These effects are apparently due to phoneme-specific differences among the various syllable groups.

The error data were submitted to the same analysis as the latency data. A main effect of syllable structure again emerged ($F(2,64) = 19.12$, $p < .0001$). The mean error percentages were 5.0%, 6.6%, and 10.9% for CV, CVC, and CCV syllables, respectively. Planned comparisons showed that more errors were made to CVC than to CV syllables ($F(1,64) = 16.16$, $p < .001$). Also, errors on CCV syllables significantly exceeded the mean number of errors on CV and CVC syllables ($F(1,64) = 22.18$, $p < .001$). As with the latency data, a main effect of syllable group ($F(11,352) = 4.64$, $p < .0001$) and an interaction between syllable group and syllable structure ($F(22,704) = 8.42$, $p < .0001$) were found in the error data, reflecting phoneme-specific differences.

Responses to the filler items were also examined. Analyses of variance revealed main effects of syllable structure for response latencies ($F(2,64) = 4.92$, $p < .01$) and for error percentages ($F(2,64) = 12.46$, $p < .001$). Since the structure of the filler items was not controlled, these effects are confounded with phoneme-specific differences and are, therefore, not interpretable.

In view of previous suggestions that adults have difficulty rejecting filler items that begin with phonemes that are similar to the target phoneme (Newman &

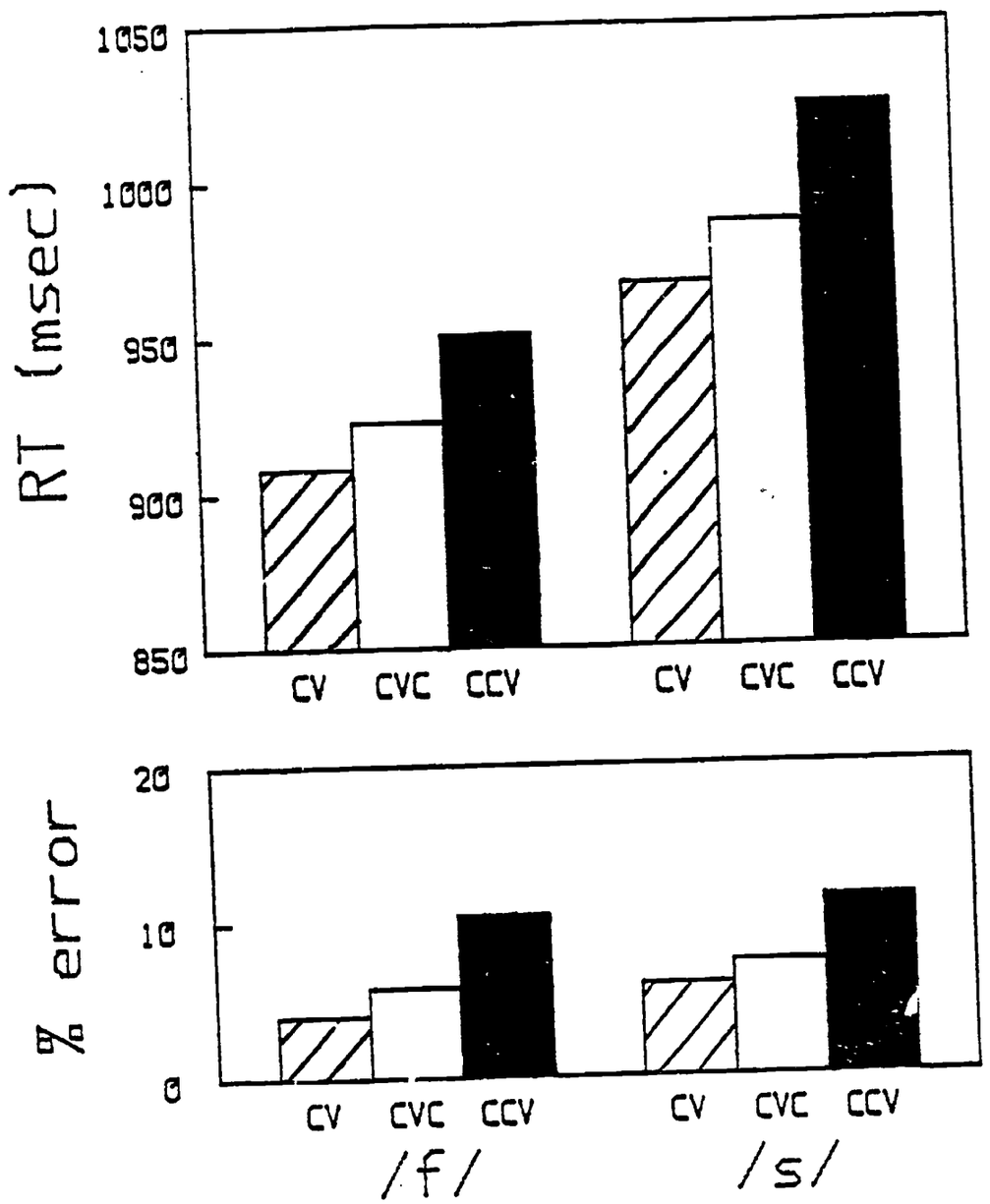


Figure 2. Latencies for correct responses (top panel) and percent errors (bottom panel) in detection of syllable-initial /f/ and /s/ targets in CV, CVC, and CCV syllables.

Dell, 1978), similarity effects were examined using two different indices of phonetic similarity. Correlation coefficients were calculated between these indices and both response time and error data. The first index of phonetic similarity was a linguistic one--the number of distinctive features in the Chomsky and Halle (1968) system that differed between the phonemes in question. As number of different features increased, latencies and errors tended to decrease, but these correlations were not significant (for response times r = -.246, $p > .1$; for errors r = -.252, $p > .1$).

The second index of phonetic similarity was a psycholinguistic measure obtained by Singh, Woods, and Becker (1971). These investigators asked subjects to rate the subjective similarity of pairs of consonants followed by the vowel /a/ using a seven-point scale. A rating of 1 indicated a high degree of similarity; a rating of 7 indicated extreme dissimilarity. As ratings increased, response latencies tended to decrease (r = -.335, $p < .07$). Apparently, subjects less quickly rejected filler items whose initial phonemes were more similar to the target phoneme than filler items whose initial phonemes were more similar. The correlation between error rates and similarity ratings was not significant (r = -.102, $p > .3$).

In summary, Experiment 1 found both increased response latencies and error rates to target phonemes in syllable-initial consonant clusters as compared to syllable-initial single consonants. Latencies to CV and CVC syllables were indistinguishable, although more errors were made on CVC than CV syllables. These results are similar to earlier findings reported by Treiman (1980). In that study, children made more errors in recognizing /f/ and /s/ in CCV syllables than in CVC and CV syllables, which were indistinguishable. Taken together, the results of the two studies suggest that syllable-initial consonant clusters, or onsets, are psychologically more complex than singleton consonants. Both children and adults have difficulty analyzing such clusters into their constituent phonemes.

A possible limitation of the present study stems from the use of a 10 KHz sampling rate in digitizing the stimuli. Much of the acoustic-phonetic information for the perception of fricatives is located at high frequencies (e.g., Stevens, 1960), and this information was filtered out by the digital sampling process used to prepare the stimulus materials. Also, since the syllables were pronounced in citation form they may have had abnormally long durations (e.g., Shoup & Pfeifer, 1976). This may have contributed to the long response times that were observed. To overcome these limitations, we carried out a second study. Experiment 2 differed from Experiment 1 in using syllables beginning with stop consonants. The critical acoustic-phonetic information for the perception of stops is located at lower frequencies than the information for fricatives. In addition, the stimulus syllables for Experiment 2 were recorded in a fixed sentence context, thus eliminating the problems associated with syllables produced in citation form.

Experiment 2

Method

Subjects. The subjects were 18 new students drawn from the same pool as for Experiment 1.

Stimuli. The six English stop consonants (/p/, /t/, /k/, /b/, /d/, /g/) served as target phonemes. For each target phoneme, there were nine experimental syllables organized into three groups. Each group contained a CV, CVC, and CCV syllable, as in Experiment 1. For each target, nine additional syllables served as filler items, yielding a stimulus set of 18 items. All stimuli were phonologically legal in English. Those that were real words were evenly divided among the CV, CVC, and CCV categories.

A set of practice syllables using the target phoneme /s/ was also constructed. It contained three experimental syllables and three filler syllables.

A male speaker produced the syllables in the carrier sentence "Peter saw a lovely ----- today." The sentences were recorded, low-pass filtered, and digitized using the same procedures as in Experiment 1. All stimulus syllables were excised from the carrier sentence using a digital waveform editor that permitted fine control over the onsets and offsets of the signal of each spoken syllable.

Procedure. The 18 stimuli in each target set were repeated three times each to produce a block of 54 trials. An experimental session consisted of one practice block of 12 trials (two repetitions of each of the six practice syllables), followed by the six test blocks. All subjects heard a total of 324 test trials. In all other respects, the procedure was identical to that of Experiment 1.

Results and Discussion

For each subject, correct response latencies and error rates were collapsed across the three repetitions of each stimulus item. Repeated measures analyses of variance were performed with syllable structure and syllable group as fixed variables.

Insert Figure 3 about here

Figure 3 shows the response latency data for correctly detected stop consonant targets. A main effect of syllable structure for was found ($F(2,34) = 20.01$, $p < .0001$). The mean response times in milliseconds were 559 for CV

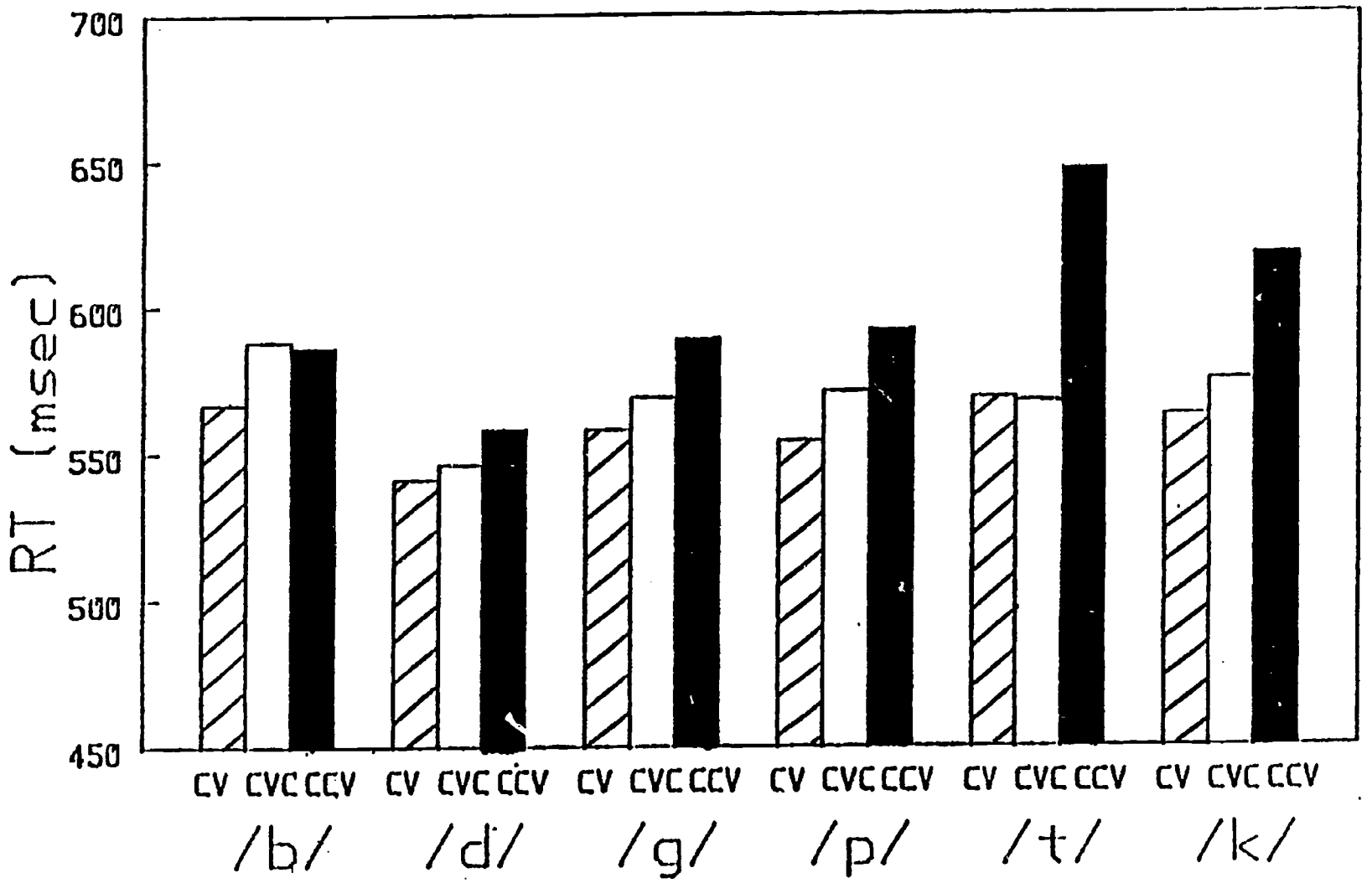


Figure 3. Latencies for correct detections of syllable-initial stop consonant targets in CV, CVC, and CCV syllables.

syllables, 570 for CVC syllables, and 598 for CCV syllables. Planned comparisons showed that response times to CV and CVC syllables did not differ significantly ($F(1,34) = 2.66$, $p > .1$). However, response times to CCV syllables exceeded the mean response times to CV and CVC syllables ($F(1,34) = 32.89$, $p < .0001$). Thus, as for the fricatives in Experiment 1, subjects were slower to detect a stop consonant target when it was part of a cluster than when it was not. A main effect of syllable group ($F(17,289) = 5.06$, $p < .0001$) and an interaction between syllable group and syllable structure ($F(34,578) = 2.82$, $p < .0001$) also emerged.

The interaction between syllable group and syllable structure observed here reflects a tendency for response times to cluster stimuli to be longer for unvoiced target phonemes (/p/, /t/, /k/) than for voiced target phonemes (/b/, /d/, /g/). The interaction between voicing and syllable structure was significant for response latencies ($F(2,32) = 3.42$, $p < .05$), although there was no main effect due to voicing ($F(1,17) < 1.0$). As shown in Figure 4, voicing only affected the speed of subjects' responses to CCV syllables: Response times to unvoiced CCV syllables were slower than response times to voiced CCV syllables. This result was not anticipated on the basis of previous work. Earlier studies comparing response times to voiced and unvoiced stop consonant targets in phoneme monitoring tasks (e.g., Cutler, 1976; Martin, 1977) have not reported differences. However, these previous studies did not report experimental control of the structure of the onsets in which the target phonemes occurred.

Insert Figure 4 about here

The results of Experiment 2 differed from those of the previous experiment in that they failed to show an effect of syllable structure on errors ($F(2,34) = 1.38$, $p > .25$). This difference may have arisen because the error rate on target items was 6.4% in Experiment 2, as compared to 7.5% in Experiment 1. Since subjects were closer to ceiling performance in Experiment 2, there was less room for syllable structure to influence the error rate.

Main syllable structure effects were found for response latencies and error rates for the filler items ($F(2,34) = 3.11$, $p < .06$, and $F(2,34) = 8.65$, $p < .001$, respectively). As in Experiment 1, however, these effects are not interpretable due to phoneme-specific differences among the filler items.

Tests for the effects of phonetic similarity between initial consonants of filler items and target phonemes were carried out using the two indices described in Experiment 1. The number of different distinctive features (Chomsky & Halle, 1968), correlated negatively with both response times ($r = -.422$, $p < .01$) and error rates ($r = -.306$, $p < .02$), as predicted. Correlations with ratings of phonetic similarity (Singh et al., 1971) failed to reach significance ($r = -.049$, $p > .5$, for response times, and $r = -.156$, $p > .2$, for error rates). The significant correlations based on the linguistic index of phonetic similarity suggest that subjects made more errors on and were slower to respond to filler items whose initial consonants were phonetically similar to the target phonemes.

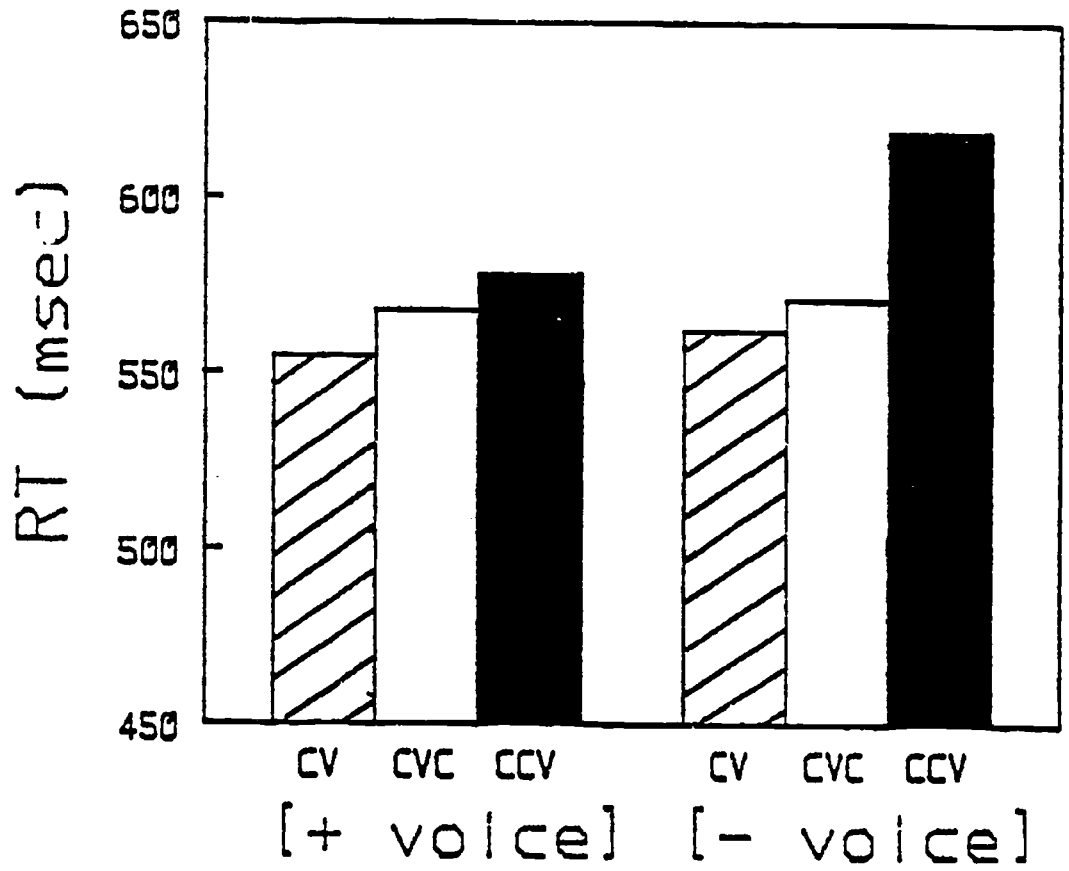


Figure 4. Latencies for correct detections of voiced (/b/, /d/, /g/) and unvoiced (/p/, /t/, /k/) stop consonant targets in CV, CVC, and CCV syllables.

In summary, adult listeners were slower to correctly detect syllable-initial stop consonants when they occurred in clusters than when they occurred as singleton consonants. This result interacted with the voicing feature of the target phoneme.

General Discussion

Experiments 1 and 2 showed that adults take longer to detect syllable-initial consonants when they occur in clusters than when they occur singly in CV and CVC syllables. This pattern was observed for the two fricatives tested in Experiment 1, as well as for the stops tested in Experiment 2. The results suggest that sequences of syllable-initial consonants form a coherent perceptual unit, the syllable onset. In order to detect the first consonant of an onset, subjects must analyze this unit into its constituent phonemes. The need for such analysis leads to lengthened response times and, in Experiment 1, to increased errors.

Although similar effects of syllable structure were found in the two experiments, response times were faster in Experiment 2 than in Experiment 1. Several factors differ between the two experiments, but the difference in phoneme category--stop versus fricative--may be important. Previous phoneme monitoring studies (e.g., Foss & Swinney, 1973; Morton & Long, 1976; Rubin, Turvey, & van Gelder, 1976; Savin & Bever, 1970) have found faster responses to stop targets than for fricative targets. The difference may arise, in part, because the acoustic cues to fricatives are longer in duration than those for stops (Cutler & Norris, 1978).

While adults have more difficulty detecting syllable-initial consonants in clusters than singly, this difficulty emerges primarily in increased response times for correct responses. Error rates on target phonemes were less than 8% in both experiments. For young children, the difficulty of analyzing onsets into their component phonemes appears to be more severe. Treiman (1980) found that 5-year-olds missed the target phoneme in CCV syllables in over 25% of cases. The error rates for CVC and CV syllables were less than 15%. For both children and adults, however, onsets appear to be cohesive units. Additional effort is required to segment them into their constituents.

As noted earlier, evidence from several different sources has been used recently to support the proposal that syllables are organized into onset and rime units. The present results provide an additional source of converging evidence to the linguistic and behavioral evidence cited earlier. While several previous studies suggested that onsets function as units in speech production (e.g. MacKay, 1972) and in the learning of word games (Treiman, submitted), the present results indicate that onsets also play a role in speeded tasks requiring the perception and overt identification of speech sounds.

The present results are also relevant to the long-standing controversy on the units of speech perception, a controversy that has been reviewed recently by Mehler, Dommergues, Frauenfelder, and Segui (1981) and Pisoni (Note 2). Investigators have debated over the units into which the continuous speech stream is segmented, with some theorists (e.g., Cutting, 1975; Foss, Harwood, & Blank,

1980; Studdert-Kennedy, 1976) arguing for the primacy of phonemes as the basic perceptual units and others (e.g., Massaro, 1972; Savin & Bever, 1970) arguing for the primacy of syllables. Still other investigators (e.g., Healy & Cutting, 1976) have suggested that both phonemes and syllables may play important roles in speech perception. The present results support the view that phonemes and syllables should not be considered to be mutually exclusive possibilities. Rather, a hierarchy of units appears to be involved in the perception of speech. This hierarchy includes syllables, phonemes, and intra-syllabic units such as onsets and rimes.

Reference Notes

1. Vergnaud, J.-R., & Halle, M. Metrical phonology. Unpublished manuscript, Department of Linguistics, MIT, 1979.

2. Pisoni, D. B. In defense of segmental representations in speech processing. Paper presented at the 101st meeting of the Acoustical Society of America, Ottawa, May, 1981.

References

Barton, D., Miller, R., & Macken, M. A. Do children treat clusters as one unit or two? Papers and Reports on Child Language Development, 1980, 18, 93-137.

Chomsky, N., & Halle, M. The sound pattern of English. New York: Harper & Row, 1968.

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

Cutler, A., & Norris, D. Phoneme monitoring in sentences. In W. Cooper & E. Walker (Eds.), Sentence processing: Psycholinguistic studies presented to Merrill Garrett. New York: Halsted, 1979.

Cutting, J. E. Aspects of phonological confusion. Journal of Experimental Psychology: Human Perception and Performance, 1975, 1, 105-120.

Foss, D. J., Harwood, D. A., & Blank, M. A. Deciphering decoding decisions: data and devices. In R. A. Cole (Ed.), Perception and production of fluent speech. Hillsdale, NJ: Erlbaum, 1980.

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.

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

Fudge, E. C. Syllables. Journal of Linguistics, 1969, 5, 253-286.

Halle, M., & Vergnaud, J.-R. Three dimensional phonology. Journal of Linguistic Research, 1980, 1, 83-105.

Healy, A., & Cutting, J. E. Units of speech perception: Phoneme and syllable. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 73-84.

Hockett, C. F. Where the tongue slips, there slip I. In M. Halle (Ed.), To honor Roman Jakobson. The Hague: Mouton, 1967.

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

MacKay, D. G. The structure of words and syllables: Evidence from errors in speech. Cognitive Psychology, 1972, 3, 210-227.

Martin, M. Reading while listening: A linear model of selective attention. Journal of Verbal Learning and Verbal Behavior, 1977, 16, 453-463.

Massaro, D. W. Preperceptual images, processing time, and perceptual units in auditory perception. Psychological Review, 1972, 79, 124-145.

Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. The syllable's role in speech segmentation. Journal of Verbal Learning and Verbal Behavior, 1981, 20, 298-305.

Morton, J., & Long, J. Effect of word transitional probability on phoneme identification. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 43-52.

Newman, J. E., & Dell, G. S. The phonological nature of phoneme monitoring: A critique of some ambiguity studies. Journal of Verbal Learning and Verbal Behavior, 1978, 17, 359-374.

Rubin, P., Turvey, M. T., & Van Gelder, P. Initial phonemes are detected faster in spoken words than in spoken nonwords. Perception and Psychophysics, 1976, 19, 394-398.

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

Shoup, J. E., & Pfeifer, L. L. Acoustic characteristics of speech sounds. In N. J. Lass (Ed.), Contemporary issues in experimental phonetics. New York: Academic Press, 1976.

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

Stevens, P. Spectra of fricative noise in human speech. Language and Speech, 1960, 3, 32-49.

Studdert-Kennedy, M. Speech perception. In N. J. Lass (Ed.), Contemporary issues in experimental phonetics. New York: Academic Press, 1976.

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

Treiman, R. The structure of spoken syllables: Evidence from novel word games. Submitted for publication.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

Controlled Perceptual Strategies in Phonemic Restoration*

Howard C. Nusbaum, Amanda C. Walley, Thomas D. Carrell,
and William Ressler

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

*This research was supported by NIMH research grant MH-24027, NINCDS research grant NS-12179, and NIH training grant NS-07134 to Indiana University. William Ressler is now at the Department of Psychology, Yale University, New Haven.

Abstract

When a portion of a spoken word is replaced by noise, listeners tend to perceptually restore the missing information. Recently, it has been shown that this phonemic restoration illusion is facilitated by the presentation of an intact version of a word prior to the same word with noise replacement (Samuel, 1981). The effects of this intact word prime on phonemic restoration have been attributed to an increase in expectations for each phoneme in the test word as a result of lexical access for the prime. If access to the phonological structure of a word is automatic upon word recognition, an intact word prime that is identical to the test word should always facilitate restoration. To test this hypothesis in the first experiment, we confined the restoration manipulation (noise-replacement/noise-addition) to the final syllable of the test words. We found that an identical intact word prime reduced phonemic restoration instead of enhancing restoration. In a second experiment, subjects were given extensive practice with a small set of test words. The increased familiarity with the test words should have increased phonological expectations for each test word, thus enhancing the restoration illusion. Instead, phonemic restoration was reduced with practice. Together, these experiments suggest that the generation of phonological expectations by lexical access is not an automatic consequence of word recognition. Thus, for auditory word perception, the recognition of word patterns seems to be mediated by a mechanism that is different from the lexical access system.

Controlled Perceptual Strategies in Phonemic Restoration

Speech perception entails a complex interaction between two sources of information. One source is the phonetic (or segmental) structure of an utterance, derived from a bottom-up analysis of the speech waveform. The second source of information is the top-down flow of knowledge following lexical access. In the process of understanding fluent speech, top-down and bottom-up mechanisms must cooperate (and sometimes compete) to provide a coherent interpretation of an utterance (e.g., Foss & Blank, 1980; Pisoni & Sawusch, 1975). Perhaps the best demonstration of this interplay between top-down and bottom-up processes is the phonemic restoration illusion.

To produce phonemic restoration, a phoneme in a word is replaced with some other sound such as a cough or white noise. This is done by locating all of the waveform in the original word that perceptually corresponds to a target phoneme like the /m/ in "democrat." This acoustic segment is then removed from the waveform and replaced by noise of the same duration as the excised segment. When noise has replaced a phoneme in a sentence, subjects have a difficult time determining which phoneme is replaced (Warren, 1970; Warren & Obusek, 1971). Listeners tend to hear this type of utterance as an intact sentence with a bit of extraneous noise added to it.

Recently, Samuel (1981) developed a new methodology for investigating phonemic restoration that utilizes the subjective experience of this illusion. Samuel presented subjects with two versions of each test word. In one version, noise replaced a phoneme in the test word, while in the second version, the noise was added to the same phoneme. Therefore, in the first version, a phoneme was missing from the word and in the second version, the word was intact -- all phonemes were present. When a missing phoneme is perceptually restored in a word, the resulting percept should be similar to the word with noise added to the phoneme. On each trial, subjects heard one of the two different versions of a word. The subjects were asked to decide whether noise had replaced a phoneme or was added to a phoneme in the test word. The proportion of noise-replaced words identified as "noise-added" is closest to the measure of phonemic restoration used in previous research (cf. Warren, 1970). This proportion indicates the extent to which words missing a phoneme were perceived as intact. Unfortunately, this measure of restoration is confounded with any bias the subjects might develop for a particular response. In order to dissociate response bias from perceptual sensitivity, Samuel used a signal detection analysis (see Green & Swets, 1966). He computed d' as a measure of the ability to discriminate between noise-added and noise-replaced versions of the same word. A d' of zero is assumed to indicate that noise-replaced words cannot be discriminated from noise-added words. This implies that the listener would have completely restored the missing phonemes in noise-replaced words.[1] Any increase in d' is interpreted as a reduction in the strength of the restoration illusion. When fewer phonemes are perceptually restored, the discriminability of noise-added and noise-replaced words improves. Thus, d' provides a direct measure of the perceptual salience of phonemic restoration.

One of the most interesting experiments reported by Samuel (1981) involved a comparison of phonemic restoration in words and nonwords. This experiment used a priming paradigm. In the control conditions, on each trial, a single noise-added or noise-replaced test item was presented. In the primed condition, the test

word or nonword was preceded by an intact version of itself without noise. For example, the test word "funeral" with noise added to or replacing the /n/ was preceded by the intact prime "funeral" with no noise in it. The effects of the prime on phonemic restoration were determined by comparing d' in the control and primed conditions for words and nonwords.

Insert Figure 1 about here

Figure 1 shows the change in d' produced by priming, relative to the control conditions in Samuel's experiment. Clearly, priming had very different effects on phonemic restoration in words and nonwords. For nonwords (shown by the solid bar on the right), priming significantly increased d' compared to the control condition. The presentation of a prime before a nonword test item facilitated the subjects' ability to determine whether or not a phoneme was missing from the test nonword. Subjects could perceptually compare the nonword prime and test items to locate the missing information.

However, when the test items were words, priming significantly reduced d' compared to the control condition (shown by the striped bar on the left). That is, the presentation of a prime word before a test word facilitated the phonemic restoration illusion. Priming made it harder to determine whether noise was added to a phoneme or replaced a phoneme in the test word. Samuel's (1981) interpretation of this effect was that the prime word activated its representation in the lexicon, increasing the expectation of each phoneme in the test word, thus enhancing restoration.

The most apparent difference in the perceptual processing of words and nonwords is the presence of an entry in the lexicon for words. It is this stored lexical representation that must generate the top-down flow of information responsible for perceptual restorations. According to associative theories of perception (e.g., Shiffrin & Schneider, 1977), lexical access is an automatic consequence of word recognition. In this type of theory, the lexicon is viewed as a content-addressable system in which an input word pattern directly accesses the appropriate lexical knowledge. In other words, access to the meaning and phonological structure of words occurs as a direct and mandatory consequence of word pattern processing. As a result, phonemic restoration should always be facilitated by priming with an intact version of the test word.

In contrast, it might be possible for the listener to actively control word recognition and lexical access. This would imply that under appropriate task constraints, listeners should be able to adopt different strategies of word perception. For example, if a listener directed attention to the sound structure of a word instead of to its meaning, the prime might be effectively used to determine if a phoneme is missing in the test word. This would produce a decrement in phonemic restoration (increasing d') similar to the nonword priming results obtained by Samuel (1981). On the other hand, if lexical access is an automatic consequence of word recognition, phonemic restoration should be enhanced by priming regardless of the task constraints.

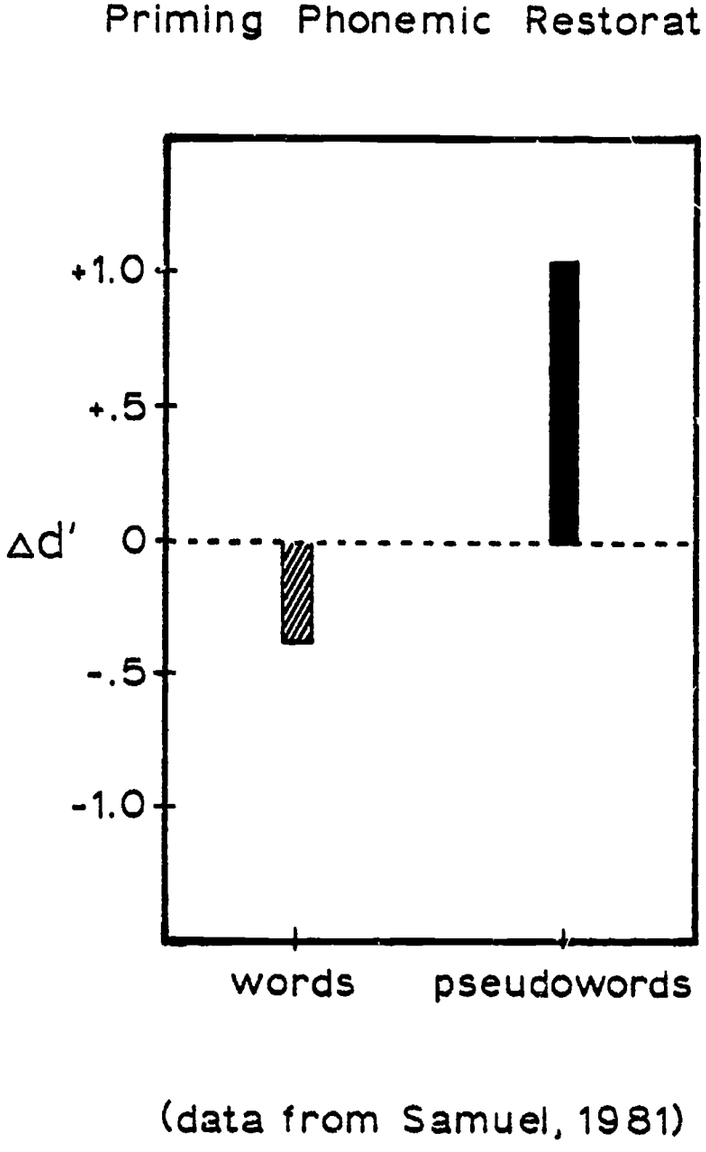


Figure 1. The effects of priming with words and nonwords on phonemic restoration for words and nonwords, respectively.

Experiment 1

In this experiment, the relationship between the prime and test word was varied to manipulate the type of expectations that might develop from processing the prime. In one condition, no prime was presented before the test word to provide a baseline measure of phonemic restoration. A second baseline condition involved presenting prime words and test words that were different in meaning and phonological structure. In this condition, subjects might hear the word "attachment" as a prime followed by the test word "civilized" with noise added to a phoneme or replacing a phoneme in the test word. Presumably, processing a different word as a prime should not generate any phonological or semantic expectations relevant to processing the test word. In a third condition, the prime and test words were the same word. In the test word, noise either replaced a phoneme or noise was added to the same phoneme, while the prime was an intact word with no noise at all. In this condition, subjects might hear the test word "communion" preceded by the prime "communion." According to Samuel (1981), in this condition the prime word should increase the expectation of each phoneme in the test word producing better phonemic restoration than the baseline conditions. However, in our test words, the noise replacement/addition manipulation was always confined to the final syllable in the test word and the number of syllables was constant for all the stimuli. As a result, uncertainty about the location of the manipulated segment was minimized relative to Samuel's experiment where both location and word length varied. If lexical access is a mandatory consequence of word recognition, the uncertainty of the location of the manipulated segment should not matter; the top-down flow of phonological expectations generated by lexical access for the prime should increase phonemic restoration. This would be reflected by a lower d' when the test word is primed by an intact version of itself compared to the baseline conditions. However, if word recognition does not automatically entail lexical access, the subjects should be able to focus attention on the perceptual analysis of the final syllable. That is, they should be able to perceptually compare the pattern structure of the prime and test words to better determine if a phoneme was replaced by noise in the test word.

In the final priming condition, the prime was a nonword that was constructed from the test word. These nonword primes were matched to the test words in all but the initial phoneme. For example, the prime for the test word "democrat" was "lemocrat." This condition allowed us to assess the extent to which priming effects depend on the phonological relationship between the prime and test words.

Method

The test items in this experiment were derived from 90 three-syllable words. Two versions of each test word were created by replacing a phoneme in the third syllable with white noise or by adding noise to the same phoneme. An additional 40 words and 20 nonwords (all three syllables long) were used as primes. The test words and primes were all natural tokens produced in isolation by a single male talker. These stimuli were digitized at 10 kHz and were stored on disc. The noise replacement/addition process was accomplished using a software digital waveform editor with 200 microsec resolution. All the stimuli (the primes and the test words) were presented in real time under computer control and were low pass filtered at 4.8 kHz.

Insert Figure 2 about here

Seventeen subjects identified each of the test words as either "noise-added" or "noise-replaced" in eight conditions. These conditions are shown in Figure 2. After a practice condition, subjects received one control condition without priming. This condition provided one baseline measure of phonemic restoration. The subjects also participated in three different priming conditions at each of two different interstimulus intervals. The interval between the prime and the test word was either 350 msec or 1500 msec. In one of the three priming conditions, the prime was an intact version of the test word without noise. In a second priming condition, the prime and test items were different words. In this condition, the prime words and test words differed in both meaning and phonological structure. In the final priming condition, the primes were nonwords differing from the test words only in the initial phoneme. The order of participation in the different conditions following practice was randomly determined. The subjects were instructed to listen to the prime (when a prime was presented) and then determine whether noise replaced or was added to a phoneme in the subsequent test word. The subjects responded by pressing an appropriately labeled button on a computer-controlled response box.

Results and Discussion

The effects of a different-word prime on phonemic restoration are shown in Figure 3. The dashed line represents d' for discriminating noise-added and noise-replaced test words when no prime was presented. The solid line indicates the effect of a different-word prime preceding the test word by 350 msec or 1500 msec. The different-word primes did not significantly affect phonemic restoration compared to the control condition at the 350 msec ISI ($t(16) = .82$, n.s.) or at the 1500 msec ISI ($t(16) = .37$, n.s.). Thus, performance was not significantly different in the two control conditions. This lack of a priming effect for different prime and test words indicates that the effects reported by Samuel (1981) for words and nonwords were due to the relationship between the primes and test items; the results were not simply due to the presence of some arbitrary utterance before the test item.

Insert Figure 3 about here

A two-way analysis of variance was performed on the d' values with the type of prime (different-word, same-word, or nonword) serving as one factor and interstimulus interval as the second factor. The results showed that only the type of prime produced a significant change in d' ($F(2,32) = 5.68$, $p < .01$). The duration of the interstimulus interval did not significantly affect d' ($F(1,16) =$

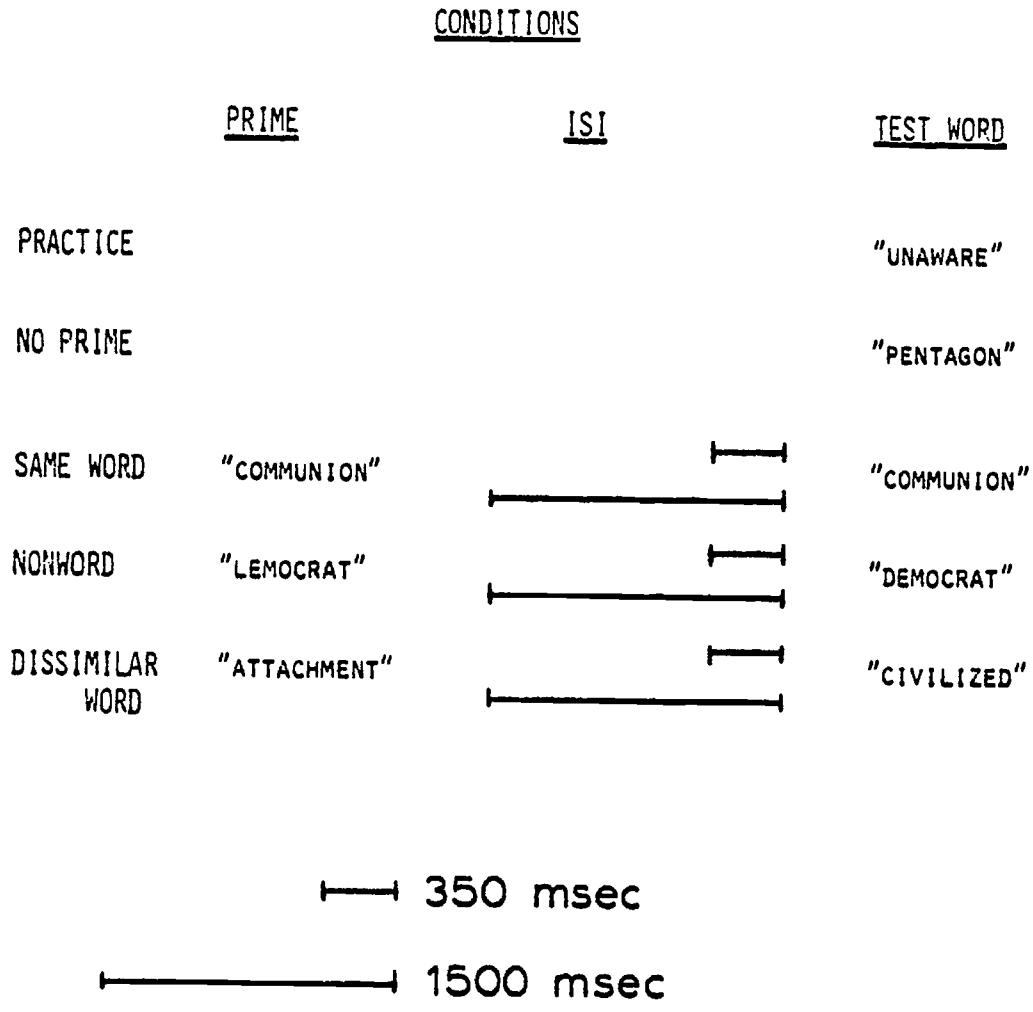


Figure 2. Trial structure for the practice, no-prime and various priming conditions of Experiment 1.

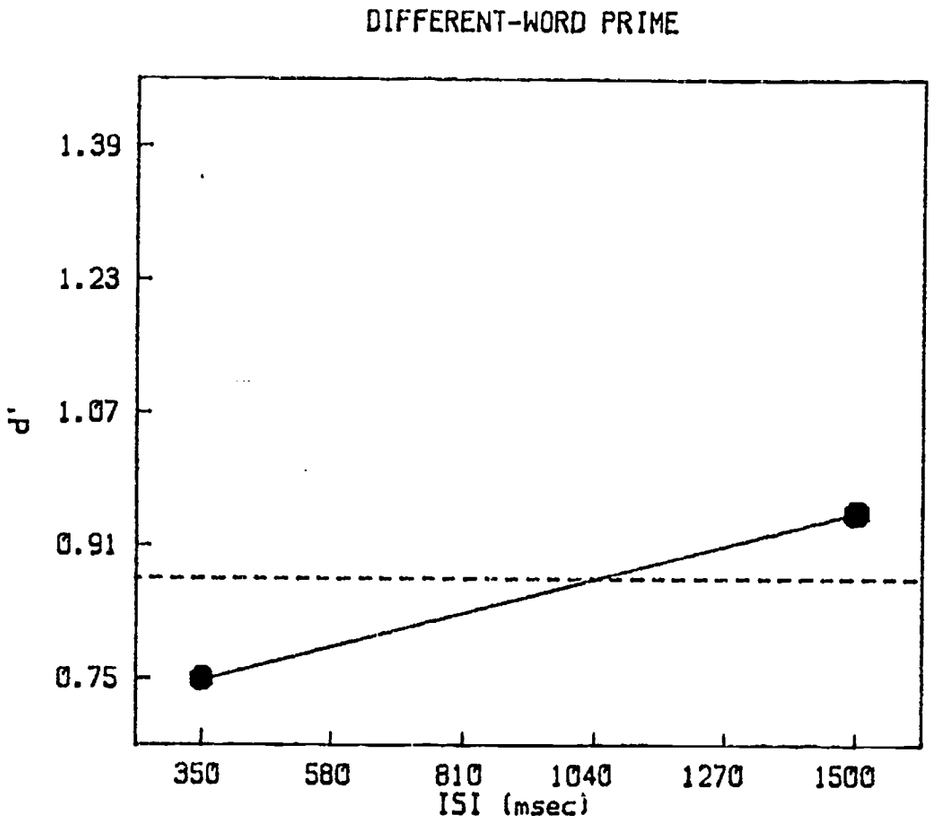


Figure 3. The effects of different-word priming on phonemic restoration relative to the no-prime condition.

.06, n.s.). Moreover, there was no significant interaction between the type of prime and ISI ($F(2,32) = 1.27$, n.s.).

Insert Figure 4 about here

Figure 4 shows the effects of a same-word prime on phonemic restoration compared to the different-word control. The dashed line represents baseline phonemic restoration measured by d' in the different-word prime control condition. A post-hoc Newman-Keuls test revealed that the same-word primes produced a significant increase in d' compared to the different-word prime control. This indicates that phonemic restoration was significantly reduced by the presence of the same-word prime. These results are opposite the priming results obtained by Samuel (1981) with words, but are similar to the results he obtained for nonword prime and test items. Samuel found that the same-word prime facilitated phonemic restoration, while in our same-word priming condition, restoration was inhibited. The primary difference between the procedures was that, in Samuel's experiment, the length of the words and the location of the manipulated phoneme in the words varied to a greater extent. In our experiment, the number of syllables was constant and the manipulated segment was always in the third syllable. This may have prompted subjects to focus attention on the final syllable. The same-word prime would then have allowed listeners to directly compare the third syllable of the prime word and the third syllable of the test word to determine if a phoneme was missing.

Insert Figure 5 about here

Finally, Figure 5 shows the data from the nonword priming condition compared to the different-word prime baseline. If the same-word priming advantage resulted from a perceptual comparison of the prime and test words at the phonemic level, the nonword primes should have been as effective as the same-word primes in increasing d'. However, the Newman-Keuls analysis showed that d' was not significantly different in the nonword prime condition and the different-word control condition. Further, d' was significantly greater for same-word priming compared to the effects of the nonword primes. These results suggest that the reduction in phonemic restoration conferred by the same-word primes may occur at a level of processing responsible for word recognition. Even though the phonological structures of the nonword prime and the test word were matched for the critical syllable, apparently subjects could not use this information for reducing the salience of phonemic restoration. These results could be explained if the effect of the same-word primes is at the word recognition stage of perception. Since nonwords cannot be successfully matched to the stored representations of word patterns, nonword primes should have no effect on this

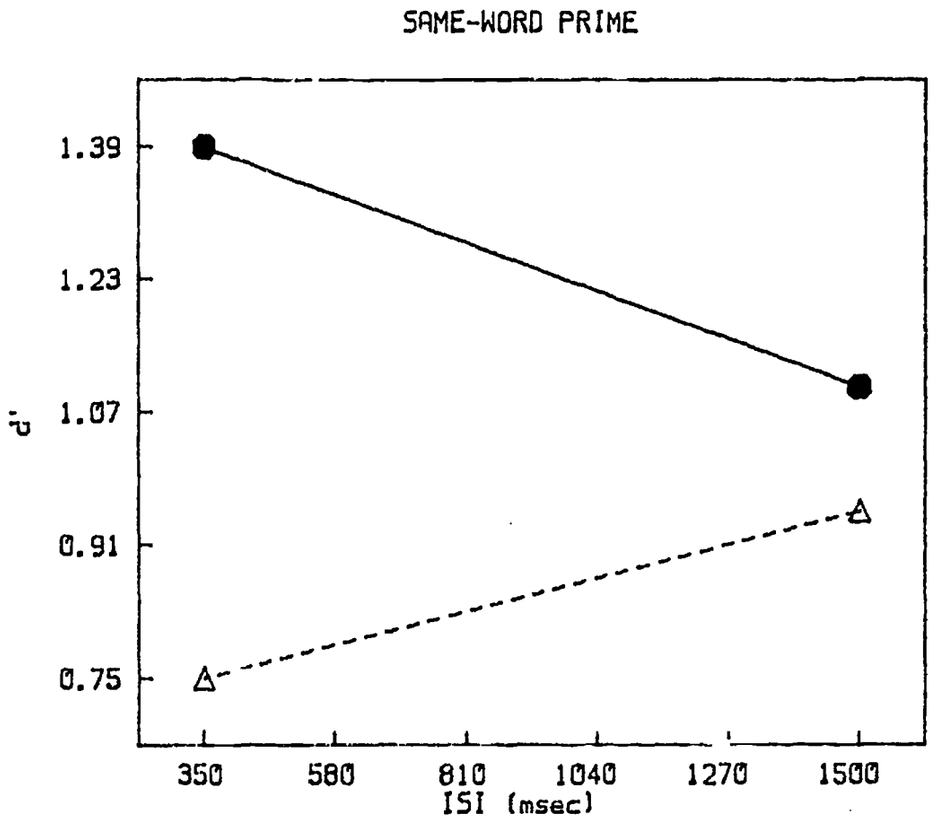


Figure 4. The effects of same-word priming on phonemic restoration relative to the different-word prime condition.

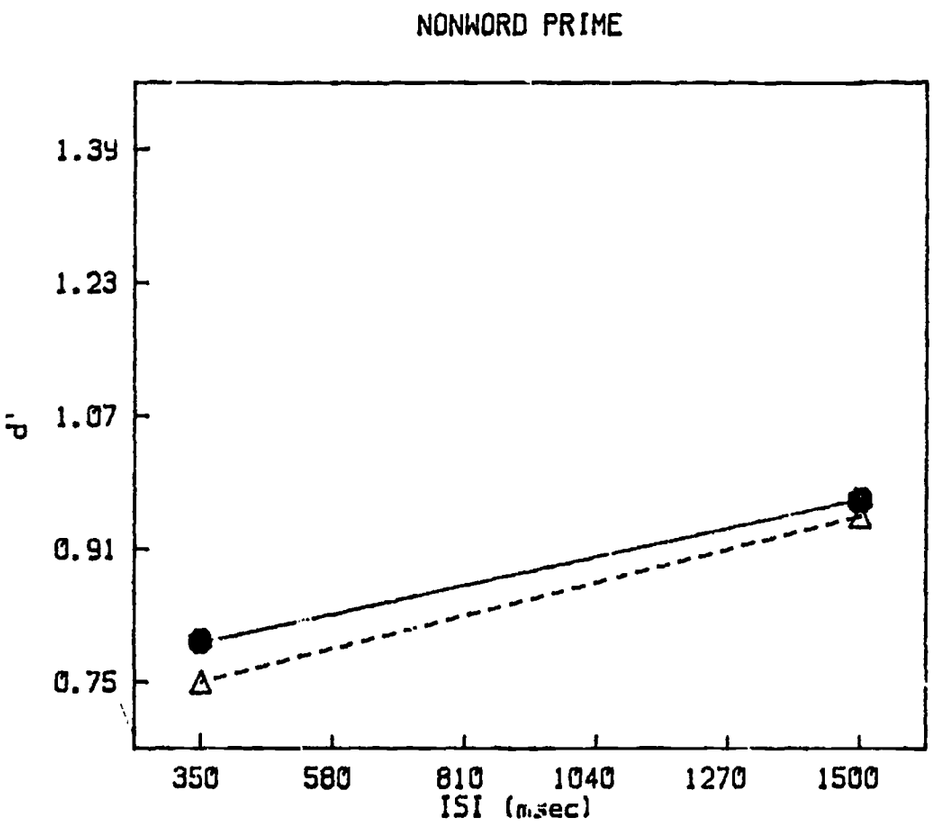


Figure 5. The effects of nonword priming on phonemic restoration relative to the different-word prime condition.

process. This implies that in the same-word prime condition, subjects did not simply compare each phoneme in the prime with the corresponding phoneme in the test word to detect a discrepancy. If that was the case, the nonword primes should have been equally effective in reducing phonemic restoration. Instead, it appears that to determine that a phoneme was missing in the test word, subjects matched the input pattern of the test word against its stored pattern representation that had been previously activated by the same-word prime. Since the nonword prime could not activate this stored representation of the test word, the nonword prime could not serve to locate a missing segment in the test word.

If lexical access is an automatic process, the presentation of the same-word prime should have facilitated phonemic restoration. Instead, the effect of the same-word prime was to inhibit phonemic restoration. These results obtained with prime and test words are similar to the results Samuel found with identical prime and test nonwords. Thus, it appears that our subjects were able to process the words without accessing lexical knowledge for the test items. This indicates that listeners can actively control lexical access and word perception to some extent.

Experiment 2

In the first experiment, we found that priming with the same word used as a test item does not always enhance phonemic restoration. Indeed, it appears that when the location of the manipulated segment in a word is constrained, subjects are able to use the same-word prime as an aid in reducing phonemic restoration. If the presentation of a word automatically generates a top-down flow of phonological information about the presented word, listeners should not have been able to reduce restoration. Assuming that lexical access is the process that produces phonemic restoration by filling in missing phonemes, this process was apparently not triggered by the same-word primes in our experiment. This suggests that subjects can hear a word and yet avoid or inhibit lexical access. It appears that listeners are able to process the acoustic-phonetic pattern of a word without accessing more detailed information about the word in the lexicon.

To test the generality of this conclusion, we conducted a second experiment using a completely different paradigm. Previous research on auditory word perception has indicated that the initial portions of words are critical for lexical access (e.g., Cole & Jakimik, 1980; Grosjean, 1980; Marslen-Wilson & Welsh, 1978). According to this research, the initial syllable (approximately) provides sufficient information for identifying many words. Samuel (1981) provided some support for this claim by demonstrating that there was less phonemic restoration for the initial syllable of a word compared to subsequent syllables in the word. This indicates that identification of the first syllable in a word is primarily the result of bottom-up pattern analysis, while subsequent syllables are generated by lexical access and then simply "confirmed" by bottom-up processes (see Foss & Blank, 1980). Presumably, these top-down expectations are responsible for the enhanced phonemic restoration in the remainder of the word. Accessing the lexicon with the initial portion of a word generates expectations about the phonological structure of the remainder of the word. If lexical access and the concomitant flow of phonological expectations are automatic consequences of the pattern analysis of word-initial syllables, enhancing the listener's expectations from this word-initial information should increase phonemic restoration.

Giving subjects extensive practice with a small set of words would provide a test of this hypothesis. As subjects become familiar with a small set of words, they should find it easier to predict the rest of a word from its initial syllable. (Assuming, of course, that the beginnings of the words in the set are sufficiently distinctive.) Thus, we might expect that extended perceptual experience with a small set of words would increase the phonological expectations produced by lexical access, thereby enhancing phonemic restoration. If lexical access automatically results from word-initial pattern recognition, a listener should be unable to adopt a strategy that prevents or inhibits phonemic restoration. Moreover, if subjects are presented with a novel set of words following practice, we would not expect any transfer of the enhanced phonological expectations.

To test these predictions, we presented feedback about the identity of a test word ("noise-added" or "noise-replaced") after each trial, in addition to providing extensive practice with the test words. This feedback should not affect performance if lexical access is automatic; phonemic restoration should increase with experience with the word set. However, if phonemic restoration decreases with practice (indicated by increases in d'), we would have evidence that subjects were able to use the feedback to focus attention on their pattern analysis of the test words. Furthermore, if the effects of practice transferred to a novel set of stimuli, it would indicate that subjects had adopted a more general strategy of word pattern analysis than just learning the specific differences between noise-added and noise-replaced versions of the training words. This would further support the hypothesis that listeners can exercise some control over lexical access and argue for the separation of word recognition and lexical access into different stages of processing.

Method

The stimuli for this experiment were derived from 20 three-syllable words. As in the first experiment, two versions of each word were created. In one version, a phoneme in the third syllable was replaced by white noise. In the second version, white noise was added to the same phoneme. To avoid the possibility that any changes in d' might be masked by ceiling or floor effects for the small set of words, we used two different classes of phoneme as the targets for the noise replacement/addition manipulation. For half the words, noise replaced or was added to a stop consonant. For the other half, white noise replaced or was added to a nasal consonant. Samuel (1981) found that d' for discriminating noise replacement from noise addition was significantly higher for nasals than for stops. Thus, we were assured of a fairly wide range of d' values for our restricted stimulus set. All the test words were read in isolation by a single male talker. The stimuli were digitized, prepared, and presented using the procedures described in the first experiment.

Twenty-one subjects participated in four blocks of trials in a single 1-hour session. In the first three blocks (the training blocks), the subjects were presented with half of the words with a nasal phoneme manipulated and half of the words with a stop consonant manipulated. In the fourth and final block (the transfer block), the subjects heard the remaining ten words (five words with a manipulated nasal and five with a manipulated stop). The assignment of words to the training and transfer blocks was counterbalanced across subjects.

In each block of trials, the subjects were presented with four repetitions of each of the two versions of the test words. The subjects identified each test word as either "noise-added" or "noise-replaced" by pressing the appropriately marked button on a computer-controlled response box. The computer then indicated which version of the test word had been presented by turning on a light over the correct label. This feedback was provided on every trial in the training blocks and in the transfer block.

Results and Discussion

Insert Figure 6 about here

The effects of training on phonemic restoration are shown in Figure 6. The dashed line indicates d' for the words in which noise replaced or was added to a nasal consonant. The solid line represents the effects of training on d' when a stop consonant was the manipulated phoneme. In this figure, it can be seen that d' for the nasals is significantly higher than d' for stop consonants ($F(1,20)=33.28$, $p<.001$). This replicates the difference in discriminability previously obtained by Samuel (1981) of noise-added and noise-replaced words for stops and nasals.

There was also a significant effect of training on d' ($F(3,60)=5.02$, $p<.005$). Post-hoc Newman-Keuls tests revealed that d' in the second and third blocks of training was significantly higher than d' in the first block. Also, d' in the transfer block was significantly higher than d' in the first training block. Moreover, the second and third training blocks and the transfer block did not differ significantly in d'. Furthermore, there was no significant interaction between the type of manipulated phoneme (stop vs. nasal) and trial block ($F(3,60)= 1.27$, n.s.).

Two conclusions can be drawn from these results. First, increasing the amount of practice did not increase phonemic restoration. Since this practice should have strengthened lexical expectations, a decrease in d' should have been observed. Instead, subjects were able to use the training to improve performance, decreasing the salience of phonemic restoration. If lexical access is an automatic consequence of pattern recognition processes operating on the initial portion of a word, subjects should not have been able to inhibit the top-down flow of phonological expectations. The improvement in d' resulting from training indicates that subjects were able to focus attention on processing the pattern structure of the test words. These results clearly parallel the results of our first experiment indicating that subjects can dissociate the analysis of auditory word patterns from lexical access.

The second conclusion is that the strategies adopted by subjects were not item-specific. Rather, the strategies that subjects learned by the second block of training were equally effective with the training words and the transfer (novel) words. Instead of learning particular differences between the

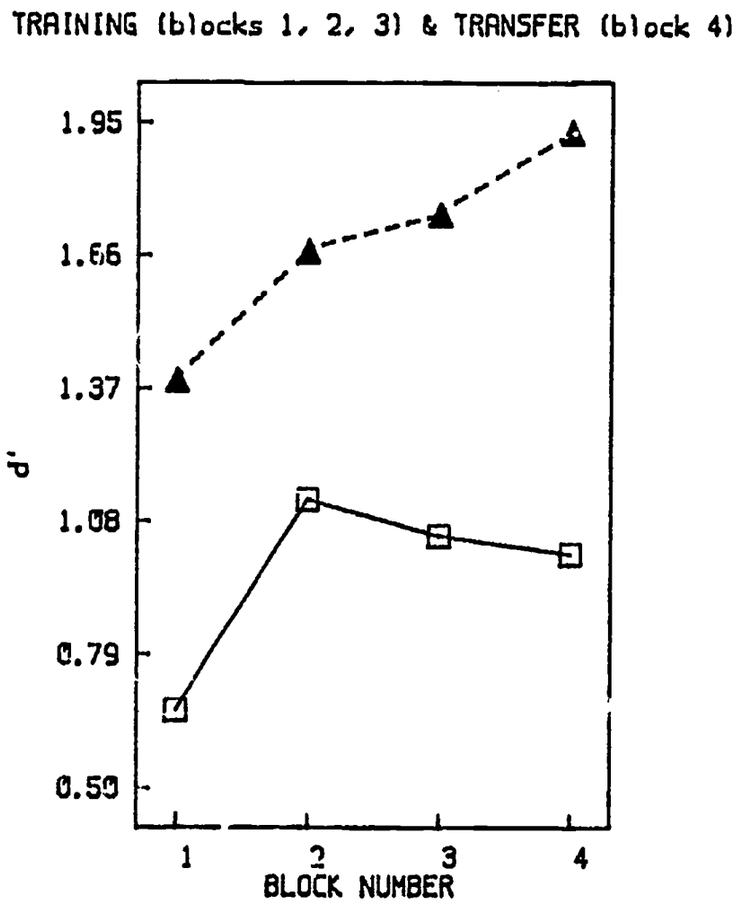


Figure 6. The effects of training on phonemic restoration for stop and nasal consonants in Experiment 2.

noise-added and noise-replaced tokens that were idiosyncratic to the training words, subjects seem to have adopted a more general strategy of attending to the phonetic pattern structure of words. This provides more evidence for the separation of word recognition and lexical access.

General Discussion

In our first experiment, we found that phonemic restoration could be inhibited by the presentation of a prime that was the same word as the test word. This result was opposite the result obtained by Samuel (1981) for same-word priming. In fact, our results were much more similar to the results Samuel found for priming test nonwords with identical nonwords. The major methodological difference between the two studies that might account for this was that in Samuel's experiment there was a great deal of uncertainty about the location of the manipulated phoneme, while in our experiment, the noise-added/noise-replaced manipulation was confined to the last syllable of the test words. This may have allowed subjects to use the same-word prime as an aid in detecting the absence of a phoneme. Moreover, it would seem that the comparison of the prime and test words occurred at the stage of processing responsible for word recognition. This hypothesis is supported by the lack of a significant priming effect for nonwords that were phonologically matched to the test words. If the effects of the same-word prime were due to a phoneme-by-phoneme comparison of the prime and test words, the nonword primes should have been as effective as the same-word primes.

These results suggest that word recognition and lexical access are not intrinsically bound together either as a single process or as an automatic associative system. If lexical access is a mandatory consequence of word recognition (pattern analysis), phonemic restoration should always be facilitated by a same-word prime. Since in our experiment the same-word prime inhibited restoration, it seems reasonable to conclude that word pattern processing can occur without lexical access. This conclusion is bolstered by the results of our second experiment. If phonemic restoration is produced by phonological expectations generated by lexical access, these expectations should be strengthened by practice with a small set of words. However, we found that subjects were able to use this training to reduce the salience of phonemic restoration. Furthermore, since the effects of training transferred to a novel set of words, it appears that subjects adopted a general perceptual strategy to inhibit restoration rather than learning word-specific differences. Taken together, these experiments argue that word recognition does not automatically entail the top-down generation of phonological expectations from lexical access. Subjects can adopt strategies that focus attention on either recognition or lexical access.

At first glance, it may be hard to understand how word recognition and lexical access can be truly separate processes. However, it is possible to conceive of a word recognition system that is independent of lexical access (see Thibadeau, Just, & Carpenter, 1982). For example, Klatt (1980) has proposed a model of word recognition that does not involve accessing the meaning of words.[2] The LAFS system is a network of acoustic templates that represent stored patterns of words. Traversing the network from start to finish only means that a particular sequence of acoustic features -- a word -- has been recognized; this recognition process does not involve nor does it yield any lexical knowledge beyond the stored pattern. Moreover, the LAFS system has no provision for

storing or retrieving semantic, syntactic, pragmatic, or phonological information pertinent to a recognized word pattern. Access to this detailed knowledge must be accomplished by subsequent processes.

Clearly then, it is not necessary to envision word recognition and lexical access as inextricably bound into a single mechanism (see Thibadeau et al., 1982). A model of word recognition need only be a system for the analysis and recognition of word patterns and lexical access is not a requisite component of this type of system. In addition, it would seem necessary to have one such mechanism for auditory word recognition and a different mechanism for visual word recognition, since the pattern structures of words in these modalities are quite different. Of course, once a word is recognized, its internal representation may be different (abstract) from its original pattern structure. As a result, only one mechanism should be needed for lexical access regardless of a word's original modality (see Forster, 1978). The purpose of this lexical access mechanism would be to retrieve the lexical knowledge appropriate to a particular recognized word.

Insert Figure 7 about here

Figure 7 shows the outline for one possible model of auditory word perception. The earliest stages of processing are devoted to acoustic-phonetic recognition. The output of this phonetic recognition mechanism is used for word pattern recognition. Note that in this model, word recognition and access have been separated into different stages of processing (see Pisoni, 1981). To some extent, this separation is dictated by the present results. Our results, in comparison with those obtained by Samuel, indicate that listeners can modify the strategies used in word perception to take advantage of different task constraints.

In our model, word recognition refers to the process that matches input phonetic patterns against stored phonetic representations of words. We assume that phonemic restoration occurs at this stage of processing, produced by feedback derived from lexical access. In this model, word recognition and lexical access together form an interactive system that mediates word perception. The effects of the same-word primes and training may be at the earlier stage of word pattern processing. Since nonwords cannot be accessed in the lexicon and differ from the test words in internal acoustic structure, these primes cannot affect phonemic restoration at either stage of lexical processing. Whether or not our processing model is correct in its detail, we have found evidence that lexical access is not completely automatic upon the presentation of a word. Therefore, it appears that word pattern recognition should be considered as distinct from lexical access.

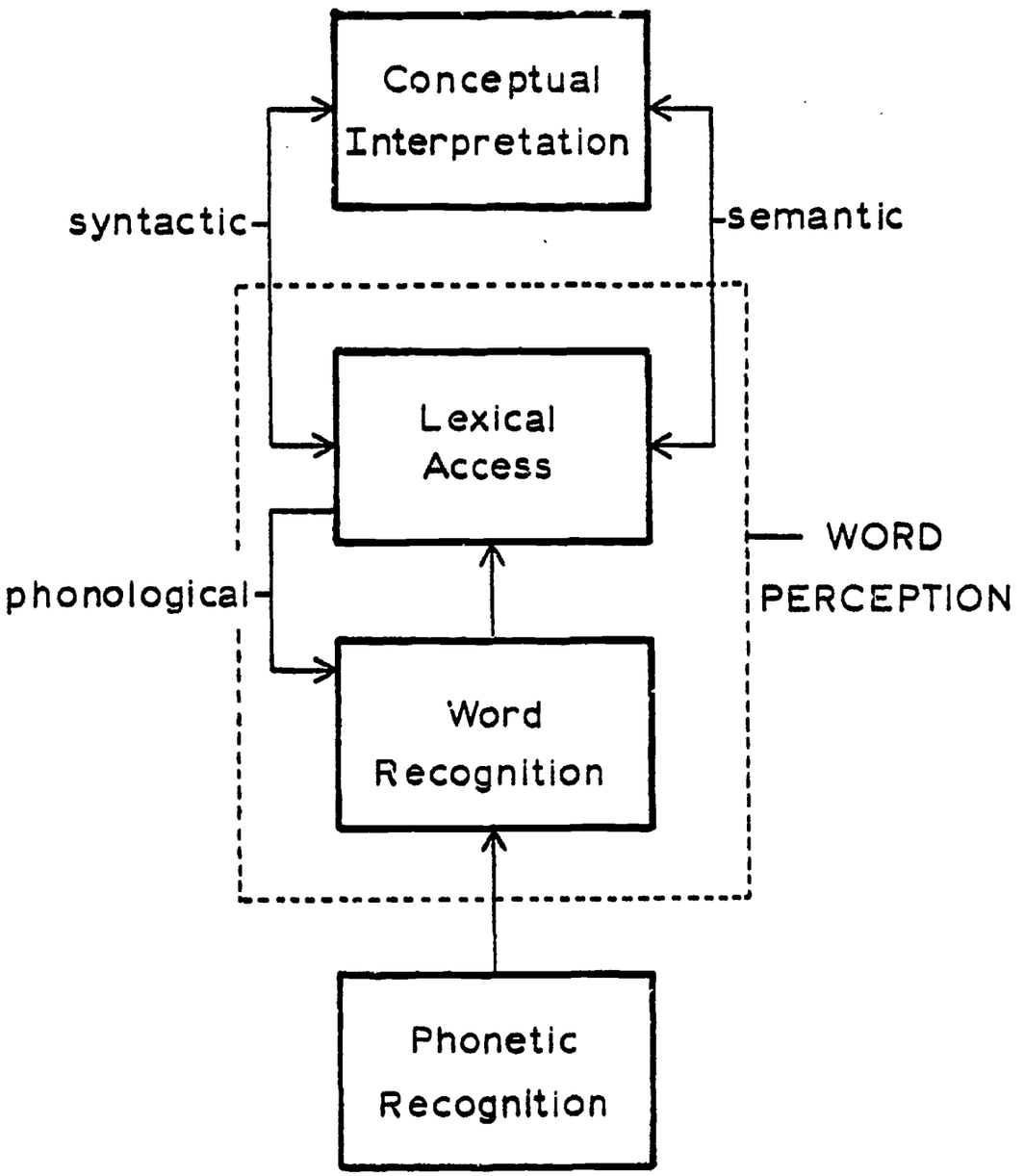


Figure 7. A model of auditory word perception in which word recognition and lexical access are assumed to constitute separate stages of processing.

References

Cole, R. A., & Jakimik, J. A model of speech perception. In R. A. Cole (Ed.), Perception and production of fluent speech. Hillsdale: LEA, 1980.

Forster, K. I. Accessing the mental lexicon. In E. Walker (Ed.), Explorations in the biology of language. Montgomery, Vermont: Bradford Books, 1978.

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

Green, D. M., & Swets, J. A. Signal detection theory and psychophysics. New York: John Wiley & Sons, 1966.

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

Klatt, D. H. Speech recognition: A model of acoustic-phonetic analysis and lexical access. In R. A. Cole (Ed.), Perception and production of fluent speech. Hillsdale: LEA, 1980.

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

Pisoni, D. B. Some current theoretical issues in speech perception. Cognition, 1981, 10, 249-259.

Pisoni, D. B., & Sawusch, J. R. Some stages of processing in speech perception. In A. Cohen & S. G. Nooteboom (Eds.), Structure and process in speech perception. Berlin: Springer-Verlag, 1975.

Samuel, A. G. Phonemic restoration: Insights from a new methodology. Journal of Experimental Psychology: General, 1981, 110, 474-494.

Shiffrin, R. M., & Schneider, W. Controlled and automatic information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological Review, 1977, 84, 127-190.

Thibadeau, R., Just, M. A., & Carpenter, P. A. A model of the time course and content of reading. Cognitive Science, 1982, 6, 157-203.

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

Warren, R. M., & Obusek, C. Speech perception and phonemic restorations. Perception & Psychophysics, 1971, 9, 358-363.

Footnotes

[1]Of course, a small d' does not, by itself, indicate that subjects are perceptually restoring phonemes to the noise-replaced version of test words. A small d' could be produced if phonemic restoration for the noise-replaced items seldom occurred and the white noise in the noise-added test items effectively masked the target phoneme. However, in this situation subjects would most often respond "noise-replaced" rather than "noise-added" indicating that in most cases the test words seemed to be missing a phoneme. This could be determined easily from the pattern of subjects' responses.

[2]We have used the term "word recognition" to refer to the processes that analyze the pattern of a word and match this pattern against stored lexical representations. However, Klatt (1980) uses "lexical access" to refer to this process. Thus, in our terminology, Klatt has proposed a model of pattern recognition, which is not lexical access. We reserve the term "lexical access" to refer to those processes that follow the selection of a lexical candidate. The lexical access system is responsible for determining the meaning, use, and phonological structure of an identified word pattern. This same distinction has been made recently by Thibadeau et al. (1982), using similar terminology for visual word perception.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

Sources of Knowledge in Spoken Word Identification*

Aita Salasoo and David B. Pisoni

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

*We wish to thank Michael Studdert-Kennedy for insightful discussions and Nancy Cooney for help in collecting the data. This research was supported, in part, by NIMH grant MH-24027 and NINCDS grant NS-12179 to Indiana University in Bloomington.

Abstract

A technique was developed to study word recognition in sentence contexts. Subjects listened to increasing durations of content words in either meaningful or semantically anomalous sentences. Initially, each target word was replaced by envelope-shaped noise. In consecutive presentations of a test sentence, 50-msec increments of the original speech waveform replaced either the initial or final segment of the noise until the entire word was presented on the last trial. Subjects identified the words in the sentences after each presentation. Recognition points were collected and incorrect word responses were analyzed in detail. In normal sentences, recognition occurred with less acoustic-phonetic information when the signal increased from the beginning of each target word than from the end of each target: the mean difference was 50 msec. The difference between forward and backward gating did not occur in the anomalous sentences. Error analyses indicated that both top-down and bottom-up sources of knowledge interact to generate a set of potential word candidates. The results 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 of the sentence context.

Sources of Knowledge in Spoken Word Identification

The last thirty years of research in the field of speech perception has focused almost exclusively on the processing of phonemes, syllables and isolated words. The results of studies on phenomena such as categorical perception, dichotic listening, the lag effect, selective adaptation, cue trading, feature sharing and duplex perception, have provided an important source of knowledge concerning the perception of the acoustic-phonetic properties of spoken language, particularly English. These studies have also provided the empirical data base for the development of high-quality synthesis-by-rule systems (e.g. Allen, 1981).

Despite the fact that a great amount of literature currently exists on the perception of phonemes, nonsense syllables and isolated words, relatively little is actually known at this time about the perception of fluent speech. In this paper we are concerned with the perceptual and cognitive processes employed in understanding fluent, continuous speech. In perceiving speech, we assume that the human listener computes a variety of perceptual codes at different levels of abstraction. In order to extract the linguistic message from the physical speech signal, the listener is assumed to use all available sources of knowledge. We will assume that there are several additional sources of knowledge, possibly of different types, that are used in perceiving continuous speech compared to the perception of isolated speech tokens. Furthermore, we assume that listeners are extremely flexible in their reliance on these knowledge sources. When one source of knowledge is impoverished, degraded, or possibly obliterated, listeners are capable of adjusting their "normal" perceptual strategies quite rapidly to reallocate attention and processing resources to other sources of knowledge, thus ensuring that the linguistic message is understood.

The problems of studying the perception and comprehension of fluent speech are enormous when compared to those encountered in studying single tokens (Cole, 1980). In addition to difficulties surrounding the linguistic specification of the early acoustic-phonetic sensory input, another, perhaps more difficult, problem confronts investigators: Namely, the effects of top-down context brought about by the listener's knowledge of morphology, syntax and semantics in the perception and comprehension of fluent speech. A multitude of linguistic, pragmatic and situational variables contribute to top-down knowledge. It seems likely to us that the mechanisms or processes by which this knowledge affects the perception and comprehension of fluent speech will be complex and fairly general in nature.

Evidence that listeners' perception of speech suffers when semantic and syntactic constraints are removed from the linguistic message was demonstrated by Miller, Heise and Lichten (1951) and Miller and Isard (1963). With extremely impoverished speech signals presented against high levels of noise, listeners were able to extract the linguistic content of the message so long as they had access to the semantic and syntactic information. When these top-down knowledge sources were experimentally removed or modified in some way (Miller & Isard, 1963), listeners' perceptual performance suffered substantially as they tried to

make use of the extremely limited and often unreliable information in the speech signal that was the only resource at their disposal. Findings of "trade-offs" in knowledge sources used to recognize words in studies with degraded stimuli may be generalized to the normal fluent speech processing situation, if the potential disruption of the speech signal at any moment is acknowledged. The problem of how top-down context is used to support perception and understanding of fluent speech is perhaps the most important question in the field of speech perception today; it is clearly the central problem that many investigators have concerned themselves with at the present time.

In this paper we are concerned with sources of knowledge available during sentence processing that are employed specifically in the processes leading to spoken word identification. While smaller linguistic units such as phonemes and syllables may not enter into the conscious perceptual and production processes of speaker-listeners of a language, words are undeniably units that are brought to conscious awareness. We will assume, then, that at some stage of perceiving and processing the speech input, words are identified or recognized by listeners. We will adopt the term word identification to stand for the correct belief (and a response contingent on that belief) that a particular word has just occurred. We will reserve the term word recognition for the results of the low-level sensory pattern-matching process that is assumed to occur upon hearing a spoken word. Thus, the word identification process involves a number of component stages including word recognition, lexical access and retrieval, and response execution. By lexical access we mean contact of some consequence of the speech input with a lexical representation (i.e., a word) in memory and retrieval or activation of that item in working memory (Pisoni, 1981). The candidate mechanisms proposed for achieving this contact have been search (Forster, 1976) and direct access (Marslen-Wilson & Welsh, 1978). For the present, decision mechanisms may be characterized as controlled processes (cf. Schneider & Shiffrin, 1977) that impose both response biases and criteria on the output of the other word identification processes.

While it is apparent that additional knowledge sources in fluent speech, lumped together often as sentence context, facilitate word identification processes in comparison to identification in isolation, the locus and mechanism of these context effects are poorly understood. By controlling access to a number of knowledge sources, both "bottom-up" sensory-derived information and "top-down" linguistic contextual constraints, we examined several assumptions about spoken word identification that pervade current theories.

Questions about the operation and precise mechanisms of context effects in speech processing have focused on the issue of autonomous vs. interactive processing (Marslen-Wilson & Tyler, 1980; Norris, 1982; Swinney, 1982; Tyler & Marslen-Wilson, 1982a; 1982b; Cairns, Note 1) and to the special status given to word-initial phonetic segments in lexical access processes (Garrett, 1978; Cairns, Note 1). The last decade has witnessed increasing interest toward the plausibility (or implausibility) of an autonomous, specialized modular linguistic processor system in the human cognitive system (e.g. Fodor, 1979; Forster, 1976; 1979; Marslen-Wilson, 1981; Norris, 1982; Swinney, 1982; Tyler & Marslen-Wilson, 1982a, 1982b; Cairns, Note 1). According to the autonomy principle (e.g. Swinney, 1982, p.164), lexical processing consists of "a set of isolable, autonomous substages, where these substages constitute domain specific processing modules". Lexical access, for example, has been assumed to be autonomous or

context-independent (Forster, 1979; Marslen-Wilson & Tyler, 1980; Swinney, 1982). Support for this assertion comes from lexical ambiguity studies, in which both meanings of an ambiguous word are shown to be briefly sensitive to semantic priming in a lexical decision task, even when sentence context strongly predicts the unrelated meaning of the word (e.g. Seidenberg, Tanenhaus, Leiman & Bienkowski, 1982; Swinney, 1979; Tanenhaus, Leiman & Seidenberg, 1979). Garrett (1978), Forster (1979) and more recently, Swinney (1982) have all suggested that lexical access processes are not sensitive to semantic biases in the sentence context. According to their account of context effects at the lexical level, contextual semantic constraints influence post-access decision stages of word identification and earlier stages of processing are fully determined by bottom-up acoustic information, with the possible supplementation from syntactic sources in Garrett's formulation (1978).

The argument is somewhat more complex than just automatic, obligatory access processes (Tyler & Marslen-Wilson, 1982b) contrasted with controlled, decision processes. The failure to distinguish clearly between automatic (cf. Shiffrin & Schneider, 1977) and obligatory processes by Tyler and Marslen-Wilson (1982a,b) makes the derivation of experimental predictions from their position difficult. In addition, the level at which responses in tasks seen as relevant in this argument are output from the processing system become critical (Forster, 1979). The necessary assumption for evaluating any data in this debate is that the experimental task indeed taps the level of processing under investigation, whether lexical access or post-access decision stages. This is not a trivial assumption in many cases, given the controlled nature of lexical decisions and detection tasks used to support autonomy positions, e.g. mispronunciation and rhyme detection (Cole, 1973; Tyler & Marslen-Wilson, 1977).

As Cairns (Note 1) has pointed out, to date the ambiguity research has had the greatest impact on the issue of autonomy. Two problems pervade this field of research: First, the work addresses meaning interpretations assigned to words, and thereby augments the influences of higher-level post-access processes. Second, homophone processing may not be typical of spoken word identification processes. Therefore, this source of evidence requires supplementing from other phenomena. Thus, the theoretical question of the autonomy of lexical access in fluent speech processing remains largely untouched by current investigations of normal word identification. Nevertheless, it is a profitable research strategy to test aspects of the strong constraints of the autonomy hypothesis in order to gain a more accurate understanding of word identification processes and a more powerful processing model of them.

One particular manifestation of the autonomy principle (not always acknowledged as such) is the special status given to the acoustic-phonetic information contained in the speech signal. In Forster's autonomous search model of lexical access (1976), the master file in the lexicon contains all the phonetic, syntactic and semantic information stored with a word token that is used by the decision processor. Entrance to that master file can only proceed, in the case of speech processing, via the peripheral phonetic file. Thus, one direction for testing Forster's specific implementation of the autonomy principle is to determine whether, in certain conditions, words can be accessed from knowledge sources other than the initial acoustic-phonetic information.

A stronger claim about the role of word-initial acoustic-phonetic information is Marslen-Wilson's "principle of bottom-up priority" (1981; Marslen-Wilson & Tyler, 1980; Tyler & Marslen-Wilson, 1982b), central to the cohort theory of spoken word identification. Unlike the selection of a single word token in Forster's search model, according to the cohort theory, an entire set of lexical candidates, the word's acoustic-phonetic cohort, is directly activated during lexical access. These word candidates overlap in their word-initial phonetic representation with the speech input. Despite the differences between the two accounts, the proposed initial processing stages of both are autonomous, as Swinney (1982) has pointed out.

The initial content of word cohorts is fully determined by acoustic-phonetic input (Marslen-Wilson & Welsh, 1978; Oden & Spira, 1978; Swinney, 1979; Tyler & Marslen-Wilson, 1982a, 1982b). However, once the cohort is activated, it is proposed that both continuing bottom-up acoustic-phonetic information and all other sources of information (including syntactic and semantic constraints) are available to interactively deactivate lexical candidates that are incompatible with any relevant source of information. A word is identified "optimally", at the point where it becomes "uniquely distinguishable from all of the other words in the language beginning with the same sound sequence" (Tyler & Marslen-Wilson, 1982b, p.175). Thus, in the nominally "interactive" cohort theory, the cohort-establishing processes involved in lexical access are still viewed as acoustic-phonetically determined and context-independent, i.e., autonomous.

Given the dependence of Marslen-Wilson's "principle of bottom-up priority" on word-initial acoustic-phonetic information, the proponents of the cohort theory are committed to demonstrating how word beginnings are identified in a continuous speech waveform. Words are not physically discrete units and their physical boundaries can only be located after their identification. To accommodate this state of affairs, Cole and Jakimik (1980) have proposed a sequential speech recognition account proceeding from left to right, with immediate word identification decisions for one word allowing identification of the beginning of the next word. The "word initial sounds" are responsible for determining the product of lexical access, in Cole and Jakimik's view (1980). One problem with this approach is the observation that human listeners recover from identification errors in the middle of sentences remarkably well and that failure to identify (or misidentify) a word in midsentence is not typically problematic for identification of the following words or comprehension of the intended message.

Tyler and Marslen-Wilson (1982b) have, by default, located the effects of context at post-access decision stages of processing. The phenomena to be accounted for here include the well replicated result that words in normal, meaningful sentences can be identified before their physical duration has been processed (e.g. Grosjean, 1980; Marslen-Wilson & Tyler, 1980), and a handful of studies in both visual and auditory domains that demonstrate better perception, production and memory for words when their initial fragments, as opposed to their final fragments, are presented (Bruner & O'Dowd, 1958; Nooteboom, 1981).

Analysis of listeners' responses in an identification task in which the signal duration of target words was gated or strictly controlled (Grosjean, 1980), constituted the first attempt to operationalize Marslen-Wilson's theoretical cohort concept. Grosjean (1980) interpreted his response data as

evidence against the claim that an entirely acoustic-phonetically controlled set of lexical candidates is accessed before a word is identified. Grosjean (1980) suggested that both acoustic and nonacoustic sources of knowledge can interact to select potential word cardidates in lexical access. Based on Grosjean's study using the gating technique, we examined the contextual and sensory knowledge sources used to identify words in sentences.

Experiment 1

The present study investigated the knowledge sources employed in the identification of words in spoken sentences using a sentence gating paradigm. With regard to bottom-up information, we were concerned with the differential informativeness of the acoustic-phonetic properties of the beginnings and endings of words. Our interest in top-down knowledge focused primarily on semantic and syntactic cues. Specifically, the three questions we addressed were: First, is word-initial acoustic-phonetic information obligatory for successful continuous word identification? Second, how does the reliance on acoustic-phonetic information change among normal, meaningful sentences, and syntactically normal, but semantically anomalous sentences? And, third, how are "word-initial" (or "word-final") incorrect response distributions related to the amount of signal duration required for identifying spoken words?

The content words in spoken sentences served as target items to be identified by subjects after each presentation of the sentence. On the first trial, the waveform of each target word was replaced completely by envelope-shaped noise. This noise removed all segmental acoustic-phonetic cues, while at the same time preserving prosodic and duration information. On each consecutive trial, 50-msec increments of the original waveform replaced selected parts of the noise mask. The 50-msec increments accumulated on successive repetitions of the sentence until, on the final trial, the entire waveform of the original word was presented. Since this sentence gating procedure involved repeated presentations of spoken sentences, in which the physical waveform varied in degrees according to the experimental manipulation, it may be thought of as an extension of the psychophysical method of limits. Our use of the term "gate" therefore refers to the incremented duration of the intact speech signal of the target words.[1]

Method

Subjects

The subjects were 194 introductory psychology students, who received course credit for their participation. All subjects were native speakers of English with no known hearing loss. None had been subjects in previous experiments using speech or speech-like materials.

Materials

Two sets of experimental sentence materials were used. Eight Harvard Psychoacoustic sentences (Egan, 1948) were chosen for the meaningful context condition. These sentences covered a wide range of syntactic structures and were balanced according to word frequency and phonological density counts in English usage, e.g. "The stray cat gave birth to kittens". The Harvard sentences

contained semantic (interpretive) and syntactic (structural) contextual cues typical of active declarative English sentences. The second context condition consisted of eight sentences selected from a set originally developed for use in the evaluation of synthesized speech (Nye & Gaitenby, 1974; Pisoni, 1982). These materials, known as the Haskins sentences, were syntactically normal and contained high frequency words. Unlike the Harvard sentences, however, they were semantically anomalous, e.g. "The end home held the press". As such, the Haskins sentences represented a class of impoverished contexts, in which rules of intrasentential semantic relations had been deliberately violated. The two sets of materials will be referred to as meaningful and syntactic sentences, respectively.

All the content words from the meaningful and syntactic sentences served as targets for this study.[2] The target words were also excised from the sentences and presented in isolation. This condition served as a control for the contribution of any sentence context per se in word identification processes (Miller, Heise & Lichten, 1951). Thus, there were three major context conditions: words in meaningful sentences, words in syntactic sentences and words in isolation.

Two properties of the target words were varied orthogonally: first, the amount of acoustic-phonetic information in the waveform, defined by gate duration; and second, the location of that information within the word, defined by gating-direction. The stimulus duration varied in 50-msec increments between successive trials. The two levels of gating-direction were forward, with signal increasing left-to-right from the word beginning, and backward, with increasing amounts of signal, right-to-left, from the end of the word.

Audio tapes of the original sentences, read by a male speaker, were low-pass filtered at 4.8 kHz and stored digitally on a PDP-11/34 computer. Beginnings and endings of target words were located with a digital waveform editor. The gated conditions of the target words in each sentence were produced by simply replacing the appropriate number of consecutive 50 msec intervals with envelope-shaped noise (Horii, House & Hughes, 1971), using a waveform time-domain processing program. For each digital sample of the waveform, the direction of the amplitude was reversed while the absolute value of the amplitude and the RMS energy were preserved. This procedure maintained the prosodic and durational cues of the speech signal, while at the same time obliterating the spectral information (i.e. formant structure) used to identify segmental phonemes.

For each original sentence, two sequences of experimental sentences were produced, one each for the forward and backward gating conditions. In both sets of materials, the first and last trials were identical: On the first trial, all target words were replaced by noise masks, while the last trial was the original, intact, spoken sentence. In the forward- and backward-gated sequences, the second to penultimate trials contained acoustic-phonetic information increasing in 50-msec increments from the beginning and ending of each target word respectively. Figure 1 shows speech spectrograms of the first, third, fifth and last trials (from top to bottom) of both the forward-gated and backward-gated sequences of a meaningful sentence used in the experiment. The isolated word sequences were created by simply excising words from their parent sentences (Pollack & Pickett, 1963). Forward-gated and backward-gated sequences were created separately for each target word using identical procedures. Each gated word presentation was treated as a trial similarly to each parent test sentence.

Insert Figure 1 about here

Sixteen experimental tapes were created from the digitally stored stimuli using a 12-bit D/A converter and a Crown 800 Series tape recorder. For each of the two sets of materials, eight blocks of experimental trials were generated, so that for each gating direction there were two counterbalanced random orders for each context condition.

Procedure

Groups of six or fewer subjects were tested simultaneously in adjacent booths in a sound-treated experimental room. Each group heard one experimental tape at 77 dB SPL peak levels over TDH-39 matched and calibrated earphones. Thus, between 20 and 26 subjects heard each context type by gating direction condition.

Subjects were told they would hear a number of sentences (or words); each one would be repeated so that it would become clearer on each successive trial. Subjects were instructed to write down after each presentation of a test sentence (or word), the word or words they heard. Subjects were encouraged to guess if they were not certain. For both the sentences and isolated word controls, the inter-sequence interval was four seconds. The experimenter stopped the tape recorder manually after each sentence presentation and continued only when all subjects had finished writing their responses in prepared answer sheets. In the isolated word condition, the tape ran without interruption; the intertrial interval was three seconds long; cue tones indicated the start of a new stimulus sequence.

In the sentence context conditions, response sheets contained the syntactic frame for each experimental sentence. The function words and separate lines for each target content word in the sentence were marked on the answer sheet. (The function words remained acoustically intact during every spoken sentence presentation.) Thus, subjects presumably had some access to top-down knowledge provided by the information on the answer sheets, e.g. the possible form class of words following function words, the number of words in the sentences, etc. Subjects were required to respond to each word after each stimulus presentation with either a word or an 'X' if they could not identify a word.

Results

Two types of dependent measures were obtained. Firstly, we computed the "identification point" for all target words. This was defined as the duration of the signal present on the trial during which the word was first correctly identified and maintained thereafter by a subject.[3] Secondly, we collected and carefully categorized subjects' incorrect word responses in order to examine the response distribution of word candidates that was generated in a given gating condition before a target word was correctly identified. We considered these response distributions to be empirical word cohorts (Grosjean, 1980). That is, we assumed that the incorrect word responses before a word was correctly

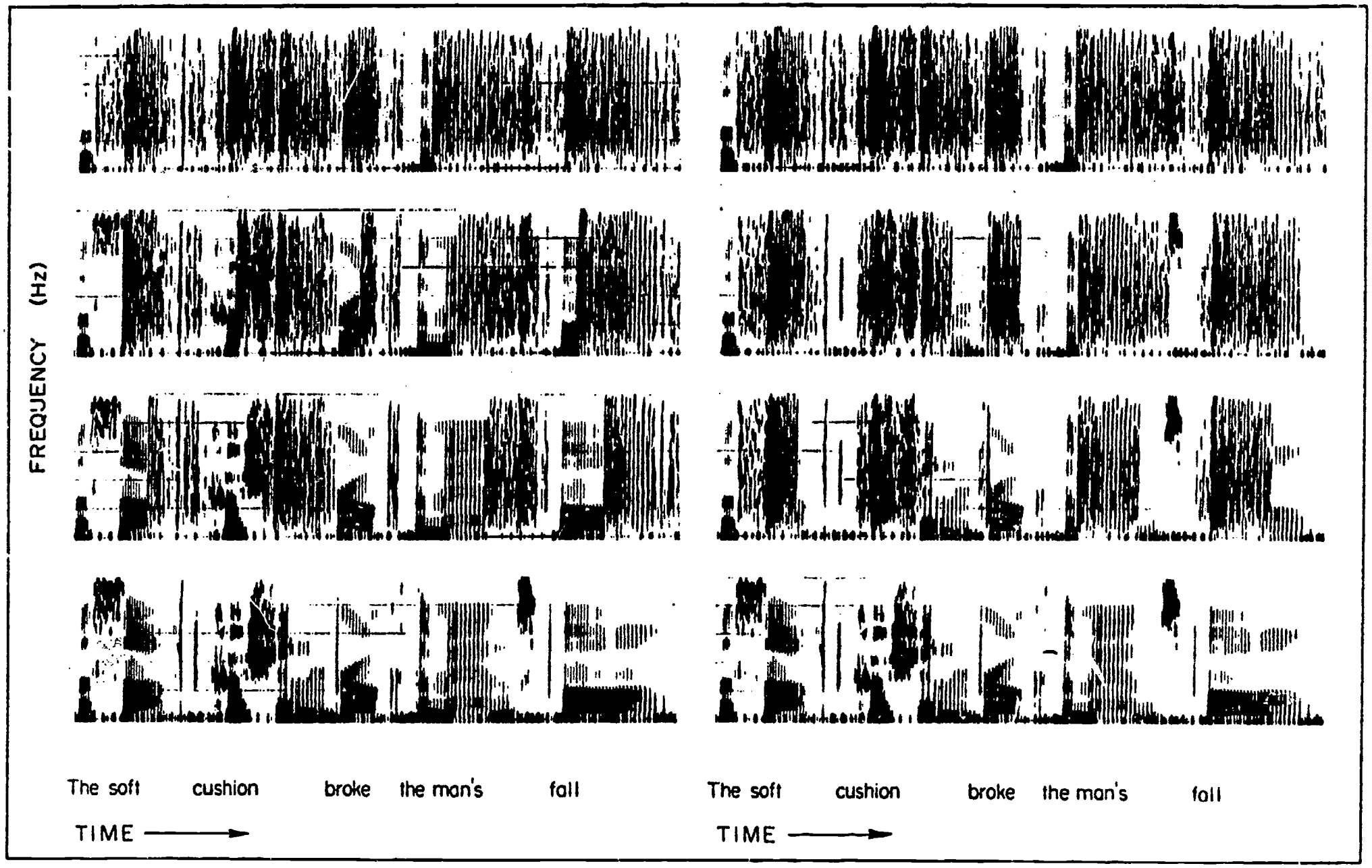


Figure 1. Sample speech spectrograms of forward-gated and backward-gated sentences. Increasing signal duration is shown left-to-right and right-to-left respectively for the target words of one test sentence.

identified would reflect individual word candidates generated in lexical access. In our analyses, we were interested in two properties of these response distributions: first, their overall size; and second, their distribution and structure in terms of the various knowledge sources that were used in spoken word identification.

Initially, the data from the two material sets were analyzed separately for gating-direction effects and then planned comparisons between the material sets were carried out. The data for the words in each sentence position were pooled across the eight test sentences. This was done to test for serial order effects, e.g. that recognizing words early in the sentence might influence the identification of words occurring later in the same sentence. The results for the identification point data and our analyses of the incorrect word response data will be examined separately below.

Identification Point Results

First, we computed the identification points, which were based on the mean performance of each subject. The data for words in the meaningful and syntactic sentences (solid lines) and their isolated controls (broken lines), are shown in the left and right panels of Figure 2, respectively. The actual measured, physical duration of each word at each sentence position is also included as a baseline for comparison (dotted lines). Forward-gated identification points are shown as triangles and backward-gated ones as squares. Also shown in each panel are the mean identification points averaged over all sentence positions.

Insert Figure 2 about here

The raw identification points were converted into proportions of the mean total word duration in each sentence position, to compensate for differences in duration as a function of syntactic structure.[4] Statistical analyses were then carried out on arcsin transformations of the proportions. Analyses of variance by subject and treatment were performed. Gating direction and sentence position (or subjects) were treated as fixed and random factors respectively and F' statistics were calculated (Clark, 1973). Unless otherwise stated, all significance levels are less than $p=.01$.

Examining the data for words in the meaningful sentence context shown in the "mean" column of the left-hand panel of Figure 2, a main effect due to gating-direction was observed: The backward-gated condition required 40 msec greater signal duration than the forward-gated condition, $F'(1,78) = 6.56$. The same words presented in isolation needed 30 msec more signal duration for identification in the backward-gated condition than in the forward-gated condition, $F'(1,90) = 7.30$. Thus, either in the presence or absence of meaningful sentence context, an advantage for word-initial acoustic-phonetic information was observed. Combining over gating-direction conditions, word identification in isolation (open symbols) required 96 msec more signal duration than in the meaningful sentence context (filled symbols), $t(78) = 6.22$.

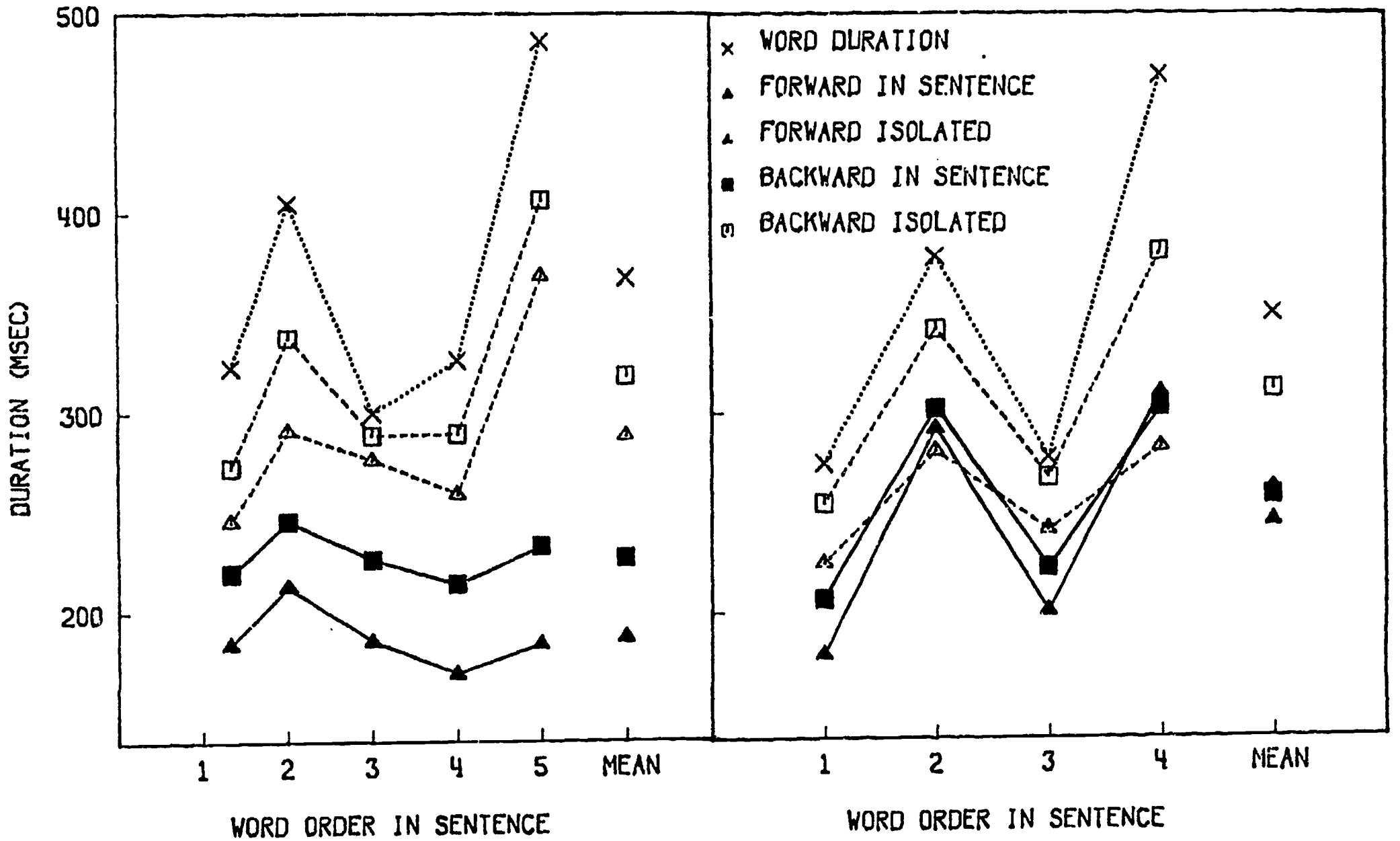


Figure 2. Identification points for words in meaningful and syntactic sentences expressed as msec of signal duration in each sentence position in Experiment 1. Forward-gated and backward-gated words (triangles and squares) are shown in both sentence context and in isolation (filled and open symbols). The measured durations of each target word at each sentence position are marked with X's.

No effect of serial position of words in meaningful sentences was observed, $F(4,44) = 1.67$, $p > .1$. Later-occurring words were not recognized with less signal duration than words occurring prior to them in the sentence. This result indicates that words earlier in a sentence conveyed no predictive information that could be used to facilitate identification of following words. In the present task, subjects were not successful in using the information available from an already identified word to predict the following word. More importantly, in meaningful sentences, subjects did not appear to be hampered by failures to recognize early-positioned words in their identification performance of later-occurring words.

The raw identification points for words from the syntactic materials are shown for each sentence position in the right panel of Figure 2. Unexpectedly, in the syntactic sentence context, there was no significant gating direction difference ($F'(1,56) = 1.26$, $p > .1$), indicating the absence of an advantage for the use of word-initial acoustic-phonetic information. The mean signal duration required for identification was 252 msec. As for the words in the meaningful context, sentence position had no main effect on the identification points in the syntactic context ($F_s(3,129) < 1.0$, $p > .44$ and $F_t(3,29) = 1.43$, $p > .25$).[5] While there was no gating-direction effect in the syntactic context (filled symbols), there was a 54-msec advantage for forward-gated conditions for the isolated words excised from the syntactic sentence contexts (open symbols), $F'(1,80) = 13.84$.

The amount of signal required for word identification in the syntactic sentences was directly proportional to the measured word duration, $r(31) = .92$. This can be seen as the parallel identification point and word duration functions in the right panel of Figure 2. This correlation was observed also for the isolated words from both materials sets ($r(39) = .89$ and $r(31) = .76$, for the Harvard and Haskins isolated controls, respectively). No such relationship was observed for the identification points and durations of words in the meaningful sentences, as evidenced by the identification curves that are not parallel with the duration curve in the left panel of Figure 2, $r(39) = .24$, $p > .10$.

Interestingly, the presence of the syntactic sentence context interacted with the manipulation of gating-direction. Less word-initial acoustic-phonetic information was sometimes required to recognize words in isolation than in the presence of semantically anomalous, but syntactically normal, sentence context. It appears that misleading context inhibited the normal reliance on acoustic-phonetic information for word indentification.

The identification point data for both material sets after transformation into proportions of signal duration are shown in Figure 3.

Insert Figure 3 about here

When the identification point data were compared across the two sentence context types, no differences were observed in the isolated control conditions: Words excised from the meaningful Harvard and syntactic Haskins material sets

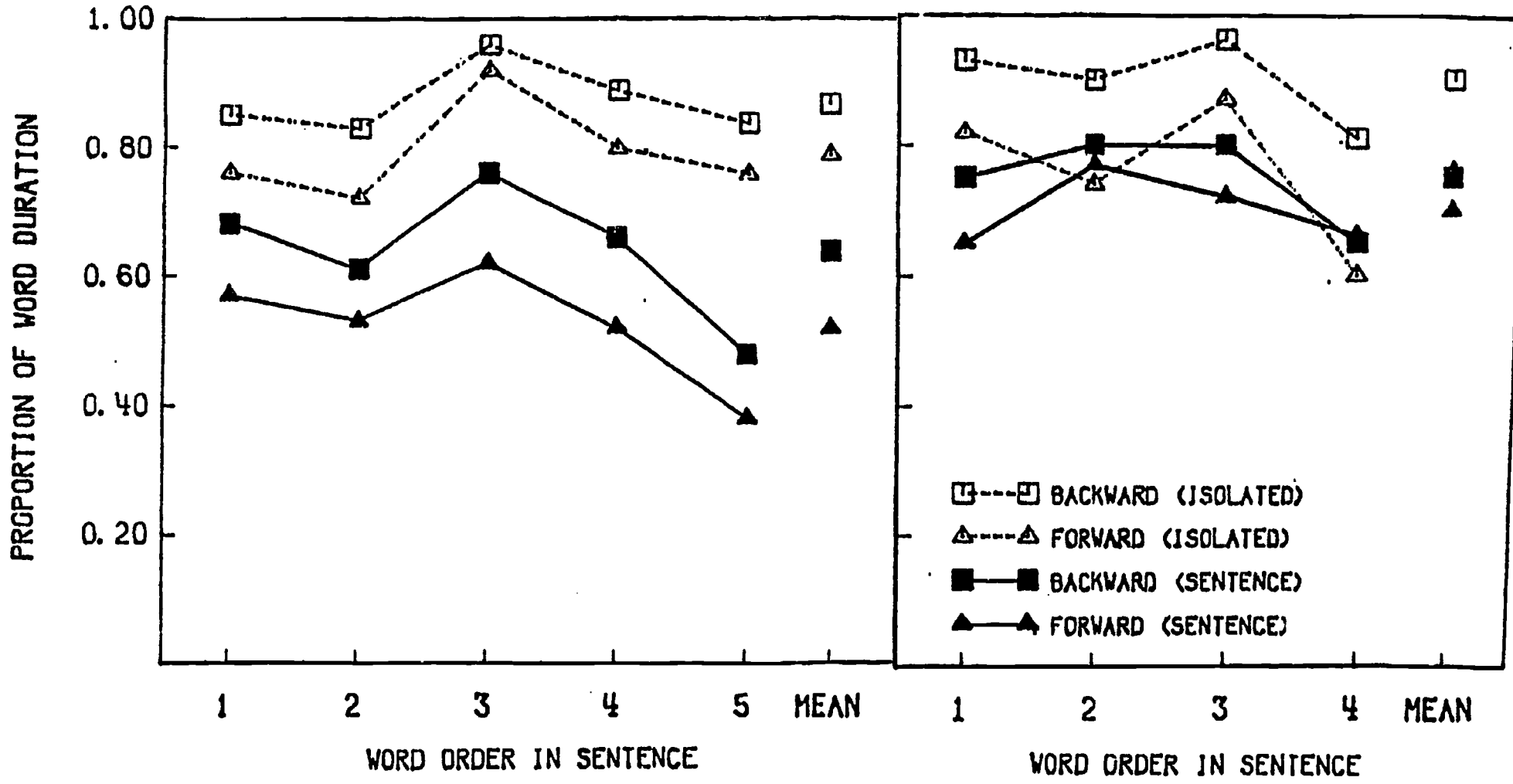


Figure 3. Identification points for words in meaningful and syntactic sentences expressed as proportions of the measured word duration for each sentence position in Experiment 1.

were identified with .83 and .81 of the mean word duration, respectively. However, in the syntactic sentence context, .72 of the mean word duration was needed for identification, while only .56 was necessary to identify the words from the meaningful sentences. This difference between the two context types was significant, $t(82) = 3.31$. Of the four context conditions, only in the syntactic sentence context was there no observed advantage for word-initial acoustic-phonetic information over word-final acoustic-phonetic information.

Analysis of the Response Distributions

We were also interested in the structural organization of the incorrect word candidates generated by listeners before they correctly identified the target words. The number of different incorrect word responses proposed by at least one subject were examined as a measure of response output. Analyses of variance with gating-direction and sentence position as factors and sentences as repeated measures were performed on these output measures.

The mean number of different word responses in each sentence position in the meaningful and syntactic sentences (excluding correct identification responses) are shown in Figure 4. A marginally significant gating-direction effect was found in the meaningful sentences (left panel), $F(1,14) = 4.94$, $p < .05$. The presence of only word-final acoustic-phonetic information yielded more word candidate responses than word-initial acoustic-phonetic information. In addition, an effect of the serial position of a word in the sentence was found: Fewer incorrect responses were proposed by subjects for words that occurred later in a meaningful sentence than for words that occurred earlier in the sentence, $F(4,56) = 5.75$. This is shown in the decreasing slope of the two curves in the left panel of Figure 4.

Insert Figure 4 about here

For the response distributions in the syntactic context (right panel of Figure 4), no gating-direction effects were observed in the incorrect response data ($F(1,14) < 1.0$, $p > .56$). However, sentence position effects, $F(3,42) = 8.68$, can be seen in this figure: A larger number of word candidates was generated for words in the second and fourth sentence positions than in the first and third positions. In this way, the data reflect the variations in the corresponding identification point data for the syntactic context condition (see Figure 1 and Footnote 3). The correlation between the identification points and the number of incorrect word responses for the syntactic context was significant, $r(31) = .96$, suggesting that both of our dependent measures, identification points and number of incorrect word candidates, are indeed related to the same underlying processes involved in continuous spoken word identification.

To compare the incorrect word response data of the two sentence contexts, each with a different number of contributiong subjects, the data were normalized by dividing the raw number of word responses by the total number of responses for each word in a condition. This comparison confirmed that more lexical candidates were proposed in the syntactic sentences than in the meaningful sentences,

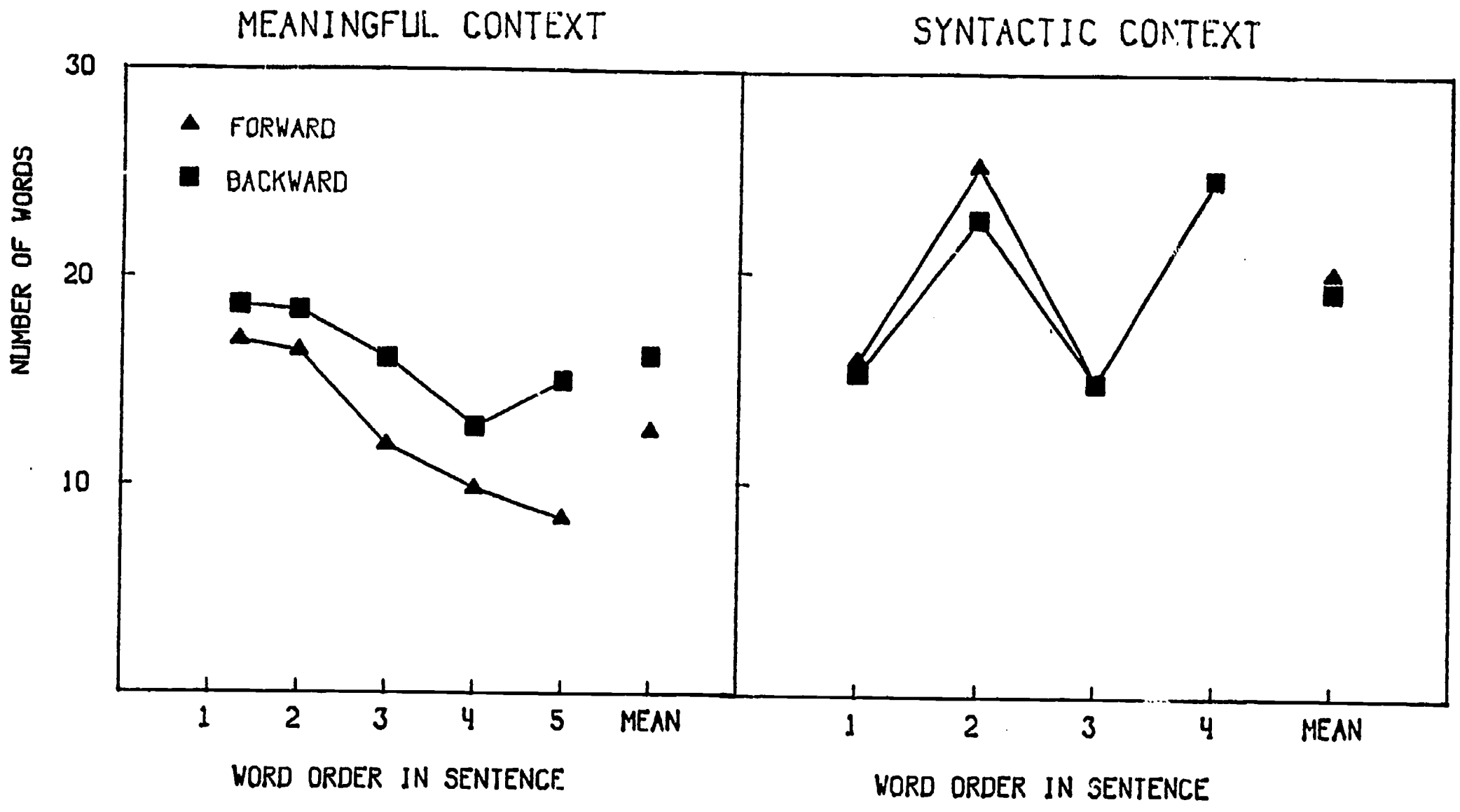


Figure 4. Number of incorrect word responses generated for forward-gated and backward-gated words in each sentence position (triangles and squares) in the meaningful and syntactic sentences in Experiment 1.

$t(70) = 5.41$, despite the fact that the sentence frame was fixed in the syntactic sentences.

To examine the structure of the word candidate responses in greater detail, analyses of the sources of knowledge underlying each proposed lexical candidate were carried out for each subject's response protocol. Each incorrect word response was categorized as originating from one of three possible sources: (1) acoustic-phonetic analysis of the signal; (2) syntactic contextual information; and (3) "other" sources (nonwords, words from an inappropriate form class, or intrusions). Every word candidate from each gating direction condition was classified as belonging to only one of the three categories. In this scoring procedure, preference was given to acoustic-phonetic information as a knowledge source, so that the remaining two categories contained no candidates that were phonetically similar to the target word.[6] Thus, we chose a conservative measure of the nonsensory contributions to the set of words hypothesized before a word was actually identified correctly.

As might be anticipated, the two sentence contexts differed substantially in the extent to which semantic and syntactic cues controlled the response distributions. We assumed that in the meaningful sentences, normal pragmatic, semantic, and syntactic constraints were operative. In contrast, we assumed that no normal pragmatic or semantic relations could be derived from the syntactic sentences. In fact, whatever semantic cues might be generated were incompatible with the normal syntactic cues in this condition. The criterion we adopted for scoring membership in the syntactic knowledge category was based solely on appropriate form class (in the absence of correct acoustic-phonetic information) for both meaningful and syntactic sentence context conditions. Although this knowledge source was not present in the isolated word identification task, response distributions in those conditions were nevertheless scored for this category also, simply as a control measure.

Finally, the "other" category contained primarily response intrusions from other sentences or from other serial positions in the test sentence. Also, in this category were phonemically dissimilar nonwords and words from inappropriate form classes. Since the "other" word responses were not based on the knowledge sources with which we were concerned, they were omitted from analyses of variance performed on the response distributions.

Figure 5 shows the results of analyses of the source of lexical candidates generated for the words in the meaningful sentences (upper left panel) and in the syntactic sentences (upper right panel) and their isolated controls in the lower panels, respectively. The data are shown as proportions of all responses for each sentence position, thus enabling comparisons to be made across different context types. Correct identification responses, "other" responses and null responses (i.e. X's) constituted the remainder of responses not shown in the figure for each gating-direction. Triangles represent responses based on acoustic-phonetic information; squares represent responses based on syntactic information. Filled symbols stand for forward-gated conditions and open symbols stand for backward-gated conditions.

In the meaningful sentence context shown in the upper left panel, three main effects were found: first, a knowledge source effect, $F(1,28) = 9.83$; second, a gating-direction effect, $F(1,28) = 9.83$; and third, a word position effect,

$F(4,168) = 5.61$. Overall, more incorrect word responses were based on correct acoustic-phonetic information than on correct syntactic, but incorrect acoustic-phonetic information. That is, subjects displayed a clear preference for incorrect responses (i.e., potential word candidates) to be controlled by the acoustic-phonetic input. The main effect of word position, in the absence of an interaction between knowledge sources and word position ($F(4,112) = 2.14$, $p > .08$), simply reflected the serial order effect of the overall number of incorrect word responses as already observed in Figure 4. Thus, when word candidates were generated, they were controlled, in large part, by the acoustic-phonetic information in the signal, conforming to the principle of bottom-up priority in spoken language understanding.

In the analyses of the lexical candidates in the isolated condition for words from the meaningful sentences, the knowledge sources underlying the response distribution was found to have a highly significant main effect, $F(1,28) = 286.68$. This result was expected, since in isolation no syntactic or semantic cues are available. In addition, a marginally significant gating direction effect was observed, $F(1,28) = 4.74$, $p < .04$. A marginal interaction ($F(1,28) = 4.87$, $p < .04$) located this gating-direction difference solely to the acoustically-based lexical candidates: When only word-final acoustic-phonetic information was available, the number of acoustic-phonetically based word responses increased. No directional effect was observed for the number of syntactically based word responses. Nonacoustic, i.e. syntactic, sources played a stable, though minimal, role in supporting word responses in both gating-direction conditions.

When the data for the same words in meaningful sentences and in isolation are compared, the effects of the meaningful sentence context are apparent. The number of incorrect lexical responses based on acoustic-phonetic information, represented by triangles, is significantly smaller in the meaningful sentence context than in the isolated words ($F(1,28) = 160.94$). At the same time, the contribution of the nonacoustic syntactic knowledge sources in isolation or in context, represented by squares, remains fairly stable overall ($F(1,28) = 1.36$, $p > .25$). Nevertheless, an interaction between gating-direction and sentence context, $F(1,28) = 7.77$, was observed for the syntactically based responses (squares): When only word-final acoustic-phonetic information was present in the meaningful sentences, the number of responses based on correct syntactic knowledge (open squares) increased compared to the forward-gated condition (filled squares). Thus, the presence of meaningful context decreased the contribution of acoustic-phonetic information to the set of hypothesized lexical candidates. Moreover, the presence of meaningful context co-occurring with the absence of word-initial acoustic-phonetic information, increased the reliance on syntactic contextual knowledge to hypothesize word responses.

Insert Figure 5 about here

Next, the incorrect response distributions for words in the syntactic sentences were examined. The right panels of Figure 5 show the results for the words in the syntactic context, and for their isolated controls. In the sentence

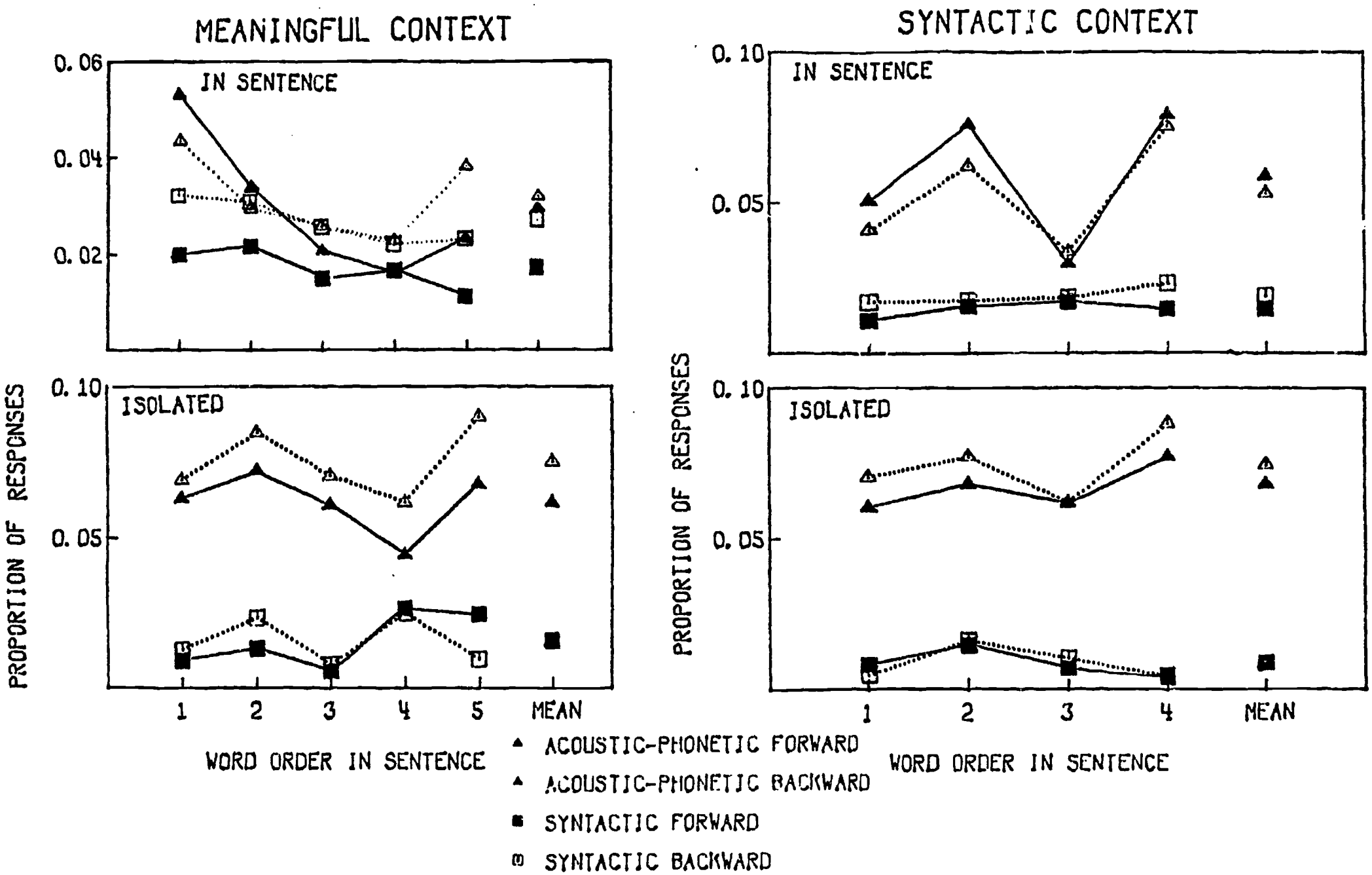


Figure 5. Distribution of incorrect word responses based on acoustic-phonetic and purely syntactic sources (triangles and squares) for words in Experiment 1. Data for meaningful and syntactic contexts are shown in the left and right panels, respectively. Words in sentence context are shown in the top panels; their isolated controls are in the bottom panels. Forward-gated conditions are shown as filled symbols; backward-gated conditions as open symbols connected by dotted lines.

context (upper right panel), no gating-direction differences were observed, ($F(1,28) < 1.0$, $p > .35$). Unlike the data for the meaningful context, the availability of word-initial acoustic-phonetic information in the syntactic sentences did not result in fewer acoustically based incorrect word candidates to be generated when only word-final acoustic-phonetic information was present in the signal. Like the isolated response distributions, there was significantly greater reliance on acoustic-phonetic information than on syntactic knowledge, $F(1,28) = 225.59$. A main effect for word position, $F(3,84) = 9.71$, and an interaction between the source of word candidates and the sentence word position, $F(3,84) = 8.47$, were also observed: In the third serial position, fewer incorrect lexical candidates based on acoustic-phonetic information were proposed than in the other word positions. This result reflects the confounding of word duration, form class and sentence position in the syntactic sentences. (See Footnote 5). No other interactions were observed.

In the isolated word condition (lower right panel), no gating-direction main effect was found, $F(1,28) = 1.82$, $p > .17$, while a highly significant source effect was observed, $F(1,28) = 428.11$. As for the isolated words from the meaningful sentences, most incorrect word responses were acoustic-phonetically based.

We then compared the incorrect response distribution of the isolated words to that of the words in semantically anomalous, syntactic sentence context. More word candidates based on correct acoustic-phonetic information were generated for isolated words, than for the same words embedded in the syntactic context, $F(1,28) = 37.46$. The observed sentence position effect ($F(3,84) = 10.38$) was due to the words occurring in the third sentence position. For only those words, all verbs, the number of word responses based on acoustic-phonetic information decreased in the presence of the anomalous sentence context. Unexpectedly, the role of the nonphonetic, top-down knowledge sources changed in the response distribution for the syntactic sentence context and isolated word conditions ($F(1,28) = 13.53$): More incorrect lexical candidates were based on compatible syntactic knowledge (but incompatible acoustic-phonetic information). Thus, even in the absence of helpful semantic information (in the syntactic context), subjects relied more on syntactic knowledge than in the isolated control condition, where there was little, if any syntactic knowledge. This result is expected, of course, since the syntactic sentences all had the same surface syntactic structure.

Summary of Results

Subjects were able to identify content words in spoken sentences when only word-initial or only word-final acoustic-phonetic information was present. When only word-initial acoustic-phonetic information was present, subjects required less signal duration to identify words, and generated fewer lexical candidates before correctly identifying target words than when only word-final acoustic-phonetic information was present. This advantage of word-initial acoustic-phonetic informativeness was present in meaningful sentences. However, this advaintage was substantially attenuated in the semantically anomalous, syntactic sentence contexts.

The reliance on the acoustic-phonetic knowledge source for generating word candidates as measured by our analyses of incorrect response candidates, was

significantly greater for words in the syntactic context than in the meaningful context. The contextual constraints in meaningful sentences did not appear to facilitate the amount of signal duration needed to identify words that occurred at the ends of sentences, compared to words occurring earlier in the sentences. On the other hand, a serial position effect was found for the overall number of incorrect word responses: Fewer incorrect candidates were proposed for later-occurring words in meaningful sentence contexts.

Closer scrutiny of the knowledge sources used to generate response candidates revealed three main findings: First, when normal semantic and syntactic sentence cues coexisted as in the meaningful sentences, subjects used the available acoustic-phonetic information more effectively (as measured by the number of acoustic-phonetically appropriate but incorrect word responses) than when only syntactic cues occurred in semantically anomalous sentences. Second, in the meaningful sentences, more syntactically controlled responses were made when only word-final acoustic-phonetic information was available. Third, even in the syntactic sentence context, subjects generated more lexical candidates using only syntactic knowledge than in the isolated control conditions.

Discussion

The present results demonstrate clearly the differential informativeness of the acoustic-phonetic information in the beginnings of words compared to the ends of words. We have also uncovered a lawful relationship between the set of incorrect word responses and the final product of the word identification process. We propose, as others have (e.g. Marslen-Wilson & Tyler, 1980), that spoken word identification in sentences is driven primarily by bottom-up processes. We extend previous proposals by suggesting that these processes can use word-initial acoustic-phonetic information more efficiently than word-final acoustic-phonetic information in lexical access. In addition, we have found reliable evidence to support previous suggestions (e.g. Garrett, 1978; Grosjean, 1980), that local semantic constraints, in conjunction with syntactic knowledge normally support the use of acoustic-phonetic information in generating lexical candidates in the processes of spoken word identification. When the acoustic signal is degraded or uninformative (e.g. normally in word endings) and the semantic and syntactic constraints of English are maintained, listeners compensate for the impoverished bottom-up sensory input by using higher-level constraints in the word identification processes. However, when normal semantic constraints are altered or deliberately removed from sentences, subjects do not compensate for the degraded acoustic signal in an analogous manner. Indeed, the entire process of word identification in syntactic context appears to be markedly altered.

We believe these results support the principle of bottom-up priority in spoken word identification, as articulated by Marslen-Wilson (1981). Our analyses of the distributions of lexical candidates reveal that subjects typically rely on acoustic-phonetic information in the stimulus to generate word candidates even from impoverished or unreliable input. The presence of both normal semantic and syntactic sentence context had substantial effects on the distribution of potential word candidates that listeners hypothesized based on various knowledge sources. Meaningful context allowed more accurate and efficient use of available acoustic-phonetic information in generating word candidates:

Subjects made fewer incorrect word responses based on correct acoustic-phonetic information than for both words in contexts with conflicting semantic and syntactic cues, or for the same words presented in isolation.

Experiment 2

The aim of this second study was threefold: to replicate the effects observed in the first study, to study the growth functions of word candidates over increasing amounts of signal duration and to investigate the effects of the repeated-presentation procedure used in Experiment 1. We made two assumptions: first, that all possible knowledge sources, including nonacoustic semantic and syntactic sources, were used in generating word candidates for lexical access; and second, that these multiple sources would be differentially informative at various points in the time-course of the word identification process. If these assumptions hold, we reasoned, then some variation in the balance of currently available information would be expected over time. Further, these changes would be reflected in the set of word candidates generated with different durations of acoustic-phonetic information available to identify a word. This line of argument yields predictions that test the autonomous character of lexical access processes, as they specifically occur in the cohort theory of spoken word identification (Marslen-Wilson & Welsh, 1978; Tyler & Marslen-Wilson, 1982b).

Let us briefly review the relevant claims of cohort theory. According to the principle of bottom-up priority, the acoustic-phonetic information contained in the first 175 msec or so of the signal directly activates a set of lexical candidates, the word cohort, that overlap phonetically in their initial segments with the target word. This set of potential words is developed from both the sensory input and the top-down syntactic and semantic knowledge available from the context. A word is identified or recognized when all but one lexical candidate is deactivated by the interaction of these two knowledge sources. However, the original set of word candidates, according to the cohort theory, is activated solely by the acoustic-phonetic information in the speech signal.

Specifically, we predicted that at short gate durations, when minimal (or no) phonetic segmental information is available in the speech signal, more word candidates based on other knowledge sources should be generated than at longer gate durations. This, in essence, proposes that interactive processes can provide input to the set of lexical candidates generated before a word is identified. If, on the other hand, semantic and syntactic knowledge can only be used to eliminate incorrect (acoustic-phonetically based) word candidates, as suggested by Marslen-Wilson, any nonacoustic syntactic word candidates occurring in the response distribution should represent random noise. This prediction adheres to the notion of autonomous lexical access.

In Experiment 1 the data were examined in terms of sentence position. This provided information about the presence or absence of serial position effects and potential interdependencies among the multiple content word targets in each sentence. However, this method of analyzing the data did not permit an examination of the time-course or growth of word candidates or the changes in the distribution of word candidates over successive gate durations. This was undertaken in the present experiment.

An additional motivating factor for the present study was the sequential nature of the gating paradigm previously used by Grosjean (1980) and employed in our first experiment. The procedure of repeated sentence trials, each with the presentation of greater acoustic-phonetic word signal duration, may have influenced subjects' word identification responses artifactually in several ways. First, repeated presentations of the same signal on early gates of a sentence may have led to facilitation in terms of the amount of signal duration required for word identification. Thus, repetition may have allowed more accurate encoding of the word-initial bottom-up information in the forward-gated conditions and word-final signal information in the backward-gated conditions respectively. This strategy would predict enhanced encoding of word beginnings and endings in their respective gating-direction conditions. The identification response required in Experiment 1 however included other processes in addition to these encoding stages. In these later processing stages, word-initial information may have produced greater facilitation than word-final acoustic-phonetic information. However, subjects may have developed a specialized response strategy during successive presentations of a test sentence. Seeing their responses from earlier trials on their answer sheets, may have influenced their responses to later presentations of the same test sequence. This may have caused reluctance to change some word candidates, even when additional acoustic-phonetic information was present. To determine the validity and generalization of the procedure used in Experiment 1, each subject in this experiment heard each test sentence only once. Whereas separate groups of subjects were presented with the meaningful and syntactic contexts in Experiment 1, in the present study, every subject heard both meaningful and syntactic sentence contexts in both forward and backward gating conditions.

Another procedural aspect that we studied was the role of the syntactic information embodied in the printed sentence frames on subject answer sheets in Experiment 1. This question relates to the generality of top-down knowledge used in this word identification task. The data from the sentence context and isolated word conditions in Experiment 1 suggest that nonsensory syntactic knowledge plays only a minimal role in spoken word identification processes. Subjects may have used very general linguistic knowledge, if any, as opposed to specific knowledge, gained from a bottom-up parsing analysis of the stimulus input. If this was indeed the case, then the function word sentence frames, e.g. "The ----- ----- ----- in the -----.", were not instrumental in providing subjects with syntactic information specific to each test sentence in Experiment 1. In the present study, therefore, subjects had no visual information about the semantic or syntactic cues of the sentence stimuli: They simply wrote down whatever words they heard and were encouraged to guess whenever they were unsure.

Method

Subjects

The subjects were 64 different students drawn from the same pool as those for Experiment 1.

Materials

A subset of the materials used in Experiment 1 were chosen, so that there were 16 gating conditions for each of 8 meaningful and 8 syntactic sentences. In the zero msec gate duration, every content word was entirely replaced with envelope-shaped noise. The "Full" condition comprised the intact spoken sentence. The 50, 100, 150, 200, 250, 300 and 350 msec gates of word signal duration were employed for both forward-gated and backward-gated conditions. Each test sentence was presented only once during a trial, in contrast to the sequence of test sentences that were presented on each trial in Experiment 1.

Design and Procedure

Using a latin square design, 16 groups of four subjects each listened to each of the 16 original sentences in a different gating condition. Each group was given two practice sentences at the beginning of the experimental session. Subjects were instructed to listen and try to identify each sentence as completely as possible on blank answer sheets. Subjects were not informed of the number of words in each sentence or the manipulated variable of meaningful and syntactic sentence contexts. A PDP-11/34 computer controlled the order and presentation of the stimuli. Every group heard a different random order of their particular gating conditions of the 8 meaningful sentences, followed by 8 syntactic sentences. A trial began with a 500-msec cue light, followed by a 1-second pause. Then subjects heard one gated sentence stimulus, presented at 77 dB SPL through their TDH-39 earphones. After writing their responses down, subjects pressed "Ready" buttons. When all subjects in a group had done so, the next trial was automatically initiated. Every subject heard each of the 16 sentences once, in a different gating duration-by-direction condition. Experimental sessions lasted approximately 20 minutes.

The four variables of concern in this study were gating direction (forward vs. backward), context type (meaningful vs. syntactic sentences), word position in a sentence and presentation type (single vs. repeated presentations). The effects of the presentation type involved comparisons with the results of Experiment 1. For each of these variables, we were interested in both the amount of signal duration required for target word identification and the nature of the distribution of potential word candidates, as measured by our analyses of incorrect responses.

Results

For each gating condition, the proportion of subjects who correctly identified the target word was scored. "Identification threshold points" were defined as the amount of signal duration required for 50% of subjects to recognize a spoken word. The identification threshold data and the analysis of incorrect word candidates will be discussed in separate sections.

Identification Thresholds

Figure 6 shows the probability of correct identification of words in the meaningful and syntactic sentences. Forward-gated and backward-gated data points are shown as the letters F and B, respectively. Best-fitting logistic curves have

been plotted through the data points, using the method of least squares. The identification threshold point, representing a .50 probability of correct identification, occurs where each curve is intersected by the broken line. Chi-square tests were used to determine the differences due to gating direction and context type. For each test, two rows (forward and backward gating-direction) and nine columns of signal durations were used.

The major finding was that in both context types, the identification threshold points for backward-gated words were greater than for forward-gated words ($X^2(8) = 24.84$ and $X^2(8) = 24.78$ for meaningful and syntactic sentence materials respectively). The difference in the threshold of identification was 46 msec for words in meaningful sentences and 39 msec for words in syntactic sentences. The advantage for word-initial signal gates was greatest for signal durations between 100 and 300 msec. Thus, with a single presentation of a gated sentence, the advantage of word-initial acoustic-phonetic information over word-final acoustic-phonetic information was observed in both the meaningful and syntactic sentence contexts.

To locate contextual differences, we compared the estimated thresholds for word identification (in terms of signal duration) for single presentations of meaningful and syntactic sentences. The threshold curves for the syntactic materials (lower panel) are shifted to the right compared to the corresponding curves for the meaningful sentence materials (top panel). Forward-gated words in the syntactic sentence context required 31 msec more signal for identification than forward-gated words in the meaningful sentences. The difference between the backward-gated conditions was 24 msec. Thus, when the speech signal of both a meaningful sentence and a semantically anomalous but syntactically normal sentence were equally impoverished in terms of acoustic-phonetic information, identification of the constituent words was more accurate for the meaningful context.

Insert Figure 6 about here

The identification threshold data for words in each sentence position separately are shown for the meaningful and syntactic sentences in the top and bottom panels of Figure 7. In all panels, the plotted numbers represent data points for each word position. The left-hand panels contain the probability of correct identification for each gate of signal duration in forward-gated conditions and the right-hand panels contain the corresponding results for the backward-gated conditions.

Insert Figure 7 about here

The identification threshold data yielded serial order results that replicated the findings observed in Experiment 1. No serial order effects were

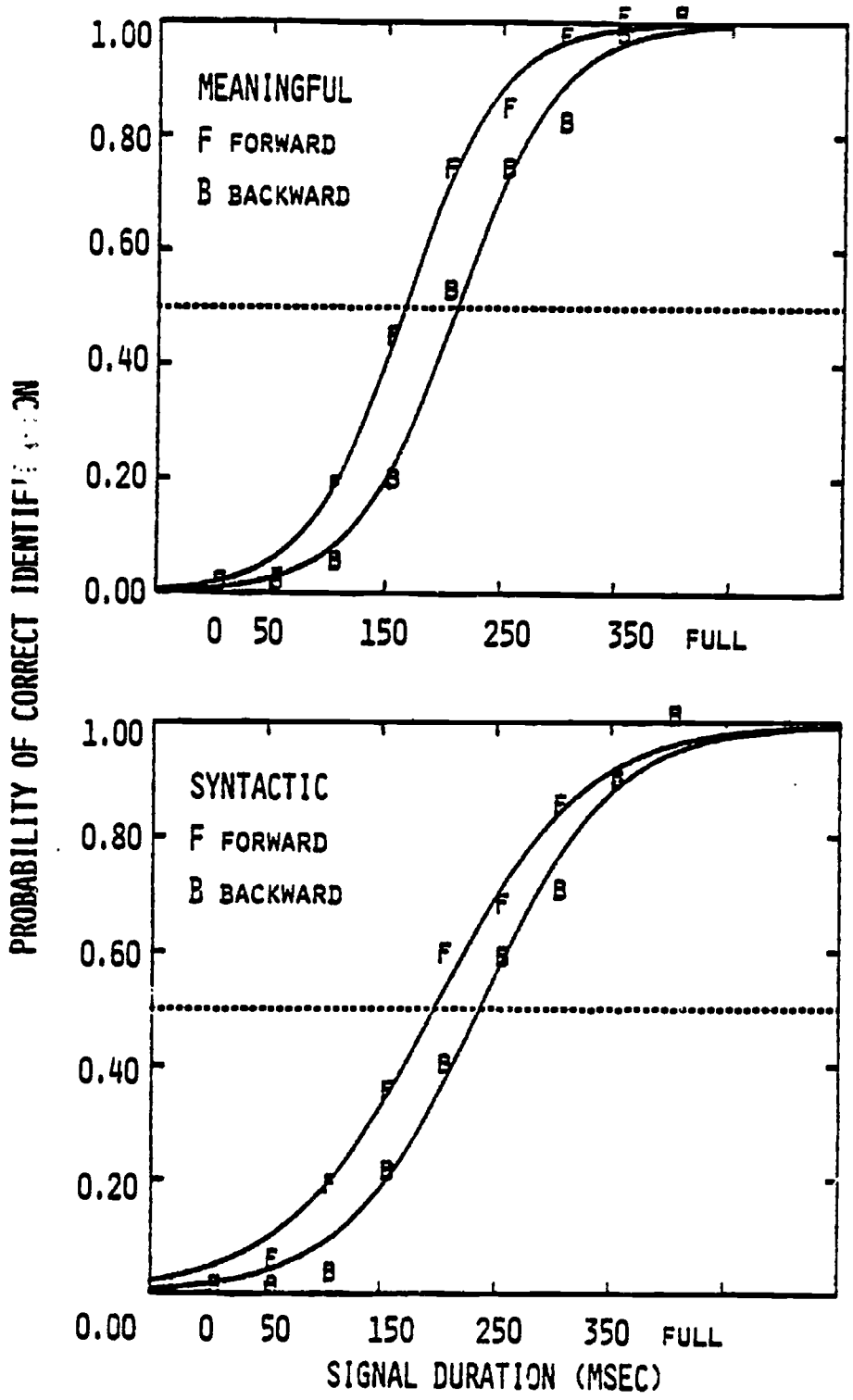


Figure 6. Best-fitting identification threshold curves for words in meaningful sentences and syntactic sentences in Experiment 2, top and bottom panels respectively. The functions were fitted through data points indicating the probability of correctly identifying the target words at each gate of signal duration. The letters F and B represent forward-gated and backward-gated conditions, respectively. The intercepts with the broken lines show the identification thresholds, defined as the .5 probability of correct identification of the target words.

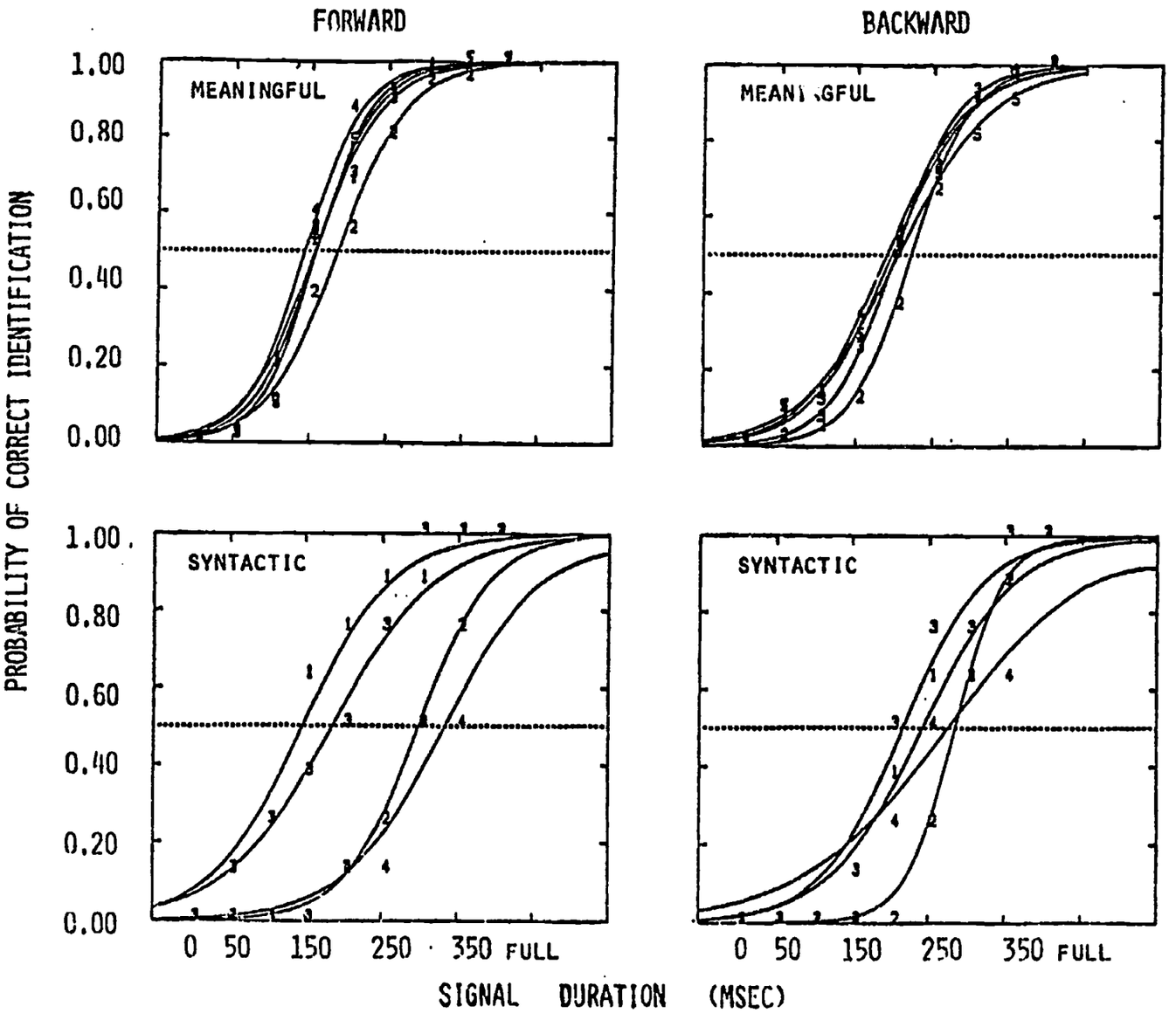


Figure 7. Best-fitting identification threshold curves for forward-gated and backward-gated conditions (in the left and right panels) for words in each sentence position in both the meaningful and syntactic sentences in Experiment 2 (in the top and bottom panels respectively). The numerals represent data points for words at the corresponding sentence position. Overlapping curves indicate no sentence position differences.

observed for the words in the meaningful sentences in either gating direction (See top panels). In contrast, the words in the syntactic sentences (bottom panels) required widely varying signal durations for identification of the individual words, depending on their spoken duration, form class and sentence position. The words in the second and fourth sentence positions, already noted as being both the longest words in the syntactic materials (i.e. nouns), consistently required longer signal durations for identification than words in other sentence positions. This serial position pattern in the syntactic sentences was observed for both forward-gated and backward-gated conditions. No predictive effects of serial position in a gated sentence were observed to facilitate identification of words occurring at the ends of the meaningful sentences (top panels). That is, threshold curves for words in the third, fourth and fifth sentence positions are not shifted to the right in Figure 7, compared to words in the first and second sentence positions. In addition, for the words in the syntactic sentences, identification threshold points in the second and fourth positions were longer than those in the first and third sentence positions, just as in Experiment 1 (see Figure 1 and Footnote 4 for an explanation).

A major methodological concern of this study was the comparison of the observed gating direction effect for the present single-presentation procedure with the results obtained using the sequential presentation method in the first experiment. In order to compare data on the amount of signal duration required with single sentence presentation to the amount required with repeated sentence sequences in Experiment 1, the identification point data from the previous study were converted to identification thresholds for each direction-by-context type group of subjects. To compute these values, the data were rescored so that each identification point in Experiment 1 contributed to the threshold curve at every gate duration shorter than itself. When calculated in this way, each subject in Experiment 1 contributed to many gate durations, corresponding to successful identification of a word on consecutive sentence presentations, but different groups of subjects contributed to each direction and context condition curve. In contrast, the points on each identification threshold curve for the present experiment reflect independent data points, but the same subjects contributed to all four curves. The word identification threshold data at each successive gate duration for the meaningful and syntactic sentences in Experiment 1 are shown in the top and bottom panels of Figure 8. Again, the letters F and B locate actual data points for forward-gated and backward-gated conditions, respectively.

Insert Figure 8 about here

First, we compared the identification threshold curves for both experimental procedures for the meaningful sentences in the top panels of Figures 6 and 8. No presentation effect was observed for either the forward-gated words ($X^2(8) = 7.64$, $p > .45$) or the backward-gated words ($X^2(8) = 14.80$, $p > .08$). Surprisingly, 50% of subjects in the two experiments required almost identical signal durations of 161 and 166 msec to recognize forward-gated words in meaningful sentence contexts. A small interaction was observed, such that identification thresholds were slightly higher for the single sentence presentation method. For both presentation methods, nevertheless, forward-gated

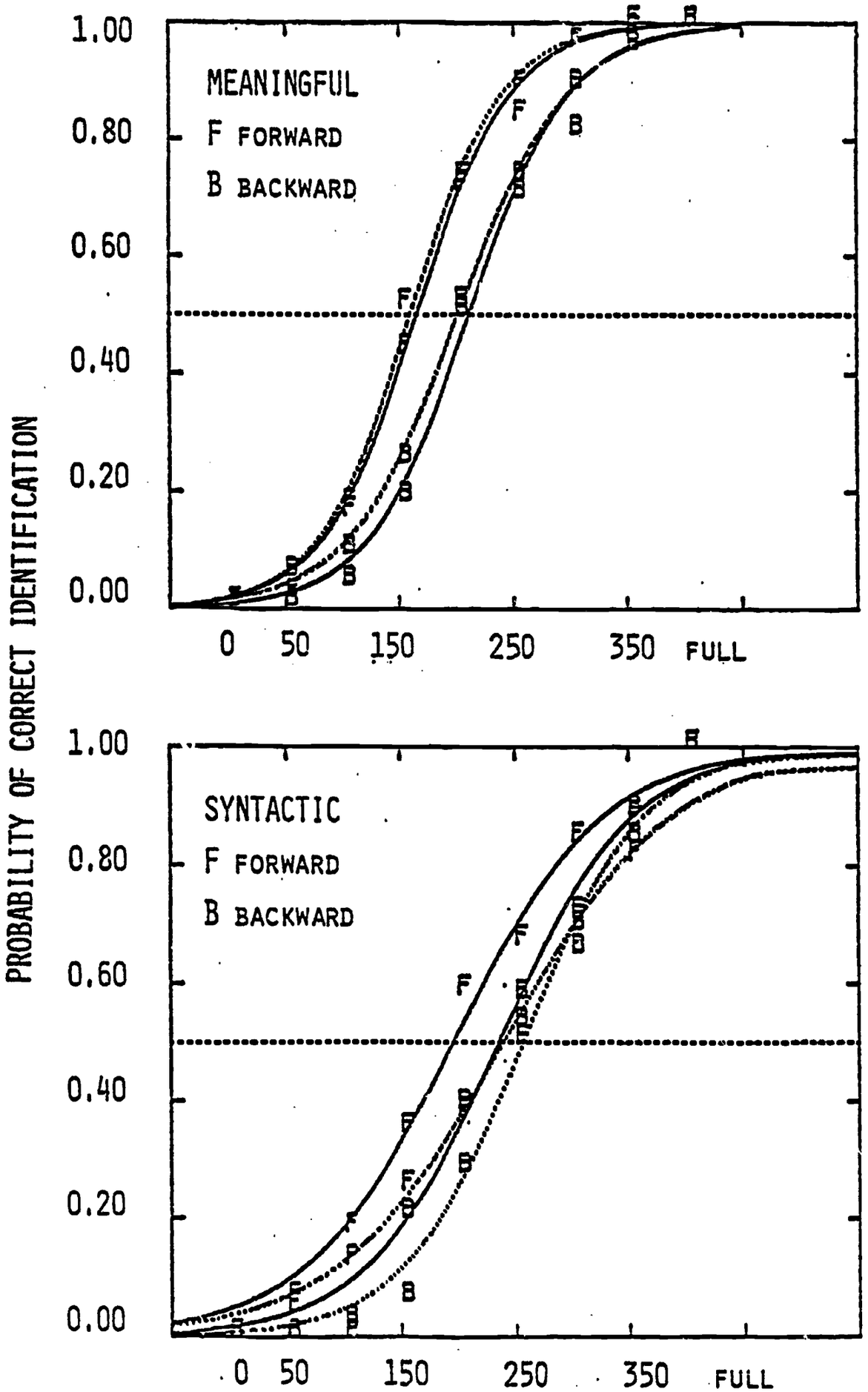


Figure 8. Comparisons of the best-fitting identification threshold curves for words in the meaningful and syntactic sentences (top and bottom panels, respectively) in Experiment 1 (dotted curves) and Experiment 2 (solid curves).

words in meaningful sentences were identified with less signal duration than backward-gated words. Thus, the gating direction effect in meaningful context was robust across procedural differences of single vs. repeated sentence presentations and across different experimental designs (within- vs. between-subjects). We therefore conclude that the use of repeated sentence presentation method in Experiment 1 did not artifactually facilitate the identification of words in the meaningful sentences.

In contrast, for the words in the syntactic sentence context in the bottom panels of Figures 6 and 8, a dramatic reversal was observed: Presentation format did affect the amount of signal duration required for word identification. With single presentations, less signal duration was required for subjects to correctly identify words in syntactic sentences. This effect was accompanied by an interaction with gating direction, such that identification of forward-gated words was actually inhibited when the words were presented repeatedly in syntactic sentence contexts ($X^2(6) = 38.96$). The accumulation of conflicting top-down semantic and syntactic information with successive repetitions appears to constrain the use of word-initial acoustic-phonetic information. As a result, the forward-gated identification threshold points were increased relative to the single presentation condition.

In short, the effects of presenting repeated sentence trials were minimal in the forward-gated meaningful sentence conditions that simulate normal speech processing situations most closely. However, repeated presentations of conflicting semantic and syntactic cues in the syntactic sentences caused subjects in Experiment 1 to require more signal in order to recognize words than in the single-presentation procedure of the present experiment. The identification threshold obtained in the present study replicated the major findings of Experiment 1. On a single presentation, word-initial acoustic-phonetic information was more informative than word-final acoustic-phonetic information. This advantage of word-initial signal occurred for both meaningful and syntactic context types in the present single presentation procedure although it was obscured in the repeated presentation condition for the syntactic sentence context. The failure to observe facilitation due to forward gating in the syntactic sentence conditions of Experiment 1 suggests the operation of a slow accumulation of the conflicting top-down constraints in identification of words in the anomalous, syntactic sentences. This issue will be taken up in greater detail in the discussion section.

Analysis of the Response Distributions

Figure 9 shows our analysis of the response distributions in the single-presentation conditions. These are displayed as proportions of total responses for each sentence position for meaningful and syntactic contexts in the left and right panels, respectively. The complex pattern of word response distribution results obtained for Experiment 1 was replicated in the present study. In both experiments, more incorrect word candidates were generated for words in syntactic contexts than for words in meaningful contexts. Subjects in both experiments also generated more incorrect word responses based on acoustic-phonetic information in the syntactic context conditions. Moreover, in meaningful sentence conditions that contained normal semantic cues, more incorrect word candidates were based only on appropriate syntactic constraints than in the syntactic context conditions. When only word-final acoustic-phonetic

information was available for listeners, the number of syntactically-based word responses was greater than when word-initial accustic-phonetic information was available for the words in the meaningful sentences. Thus, the present findings demonstrate that the observed distributions of word candidates found in Experiment 1 were not artifacts of the repeated-presentations procedure. More importantly, however, a similar trade-off between acoustic-phonetic information and syntactic knowledge as sources for word candidate responses was observed for the meaningful sentences in both experimental procedures.

Insert Figure 9 about here

One focus of the present study was on the analysis of the word candidate "growth functions" with increasing signal durations. This analysis helped to identify the relative temporal course of top-down and bottom-up contributions to specification of the set of potential lexical candidates. In Figure 10, growth functions of the response distributions are shown for the meaningful and syntactic sentence conditions (in the left and right panels), broken down by correct responses, acoustic-phonetic and syntactically based incorrect word responses (represented by X's, triangles and squares, respectively). Forward-gated and backward-gated conditions are shown in the top and bottom panels, respectively.

Three findings are apparent in this figure: First, in each condition the number of correct responses increases with longer signal durations. Second, the proportion of incorrect responses was generally small when compared to the proportion of correct responses. Nevertheless, the number of incorrect responses peaked at short signal durations, well before the identification threshold point was reached, and attenuated at longer gate durations. Third, as reflected over all gate durations, a larger proportion of incorrect word responses was based on acoustic-phonetic information in the syntactic contexts than in the meaningful contexts.

Two particular features of the candidate growth functions are noteworthy. They relate, first, to the relative peaks in the growth functions for the two categories of incorrect responses, and second, to the gate duration at which the proportion of correct responses overtakes the proportion of acoustic-phonetic and syntactically appropriate, but still incorrect, candidates.

For the meaningful contexts, the greatest number of word candidates was based on incorrect acoustic-phonetic information, but compatible word form class (squares), occurred at the 100-msec gate duration. In contrast, for both contexts, the acoustic-phonetically controlled response distribution (triangles) reached its peak at a longer gate duration and, unlike the syntactically controlled responses, maintained its proportion of the response distribution over approximately the next 100 msec of signal duration.

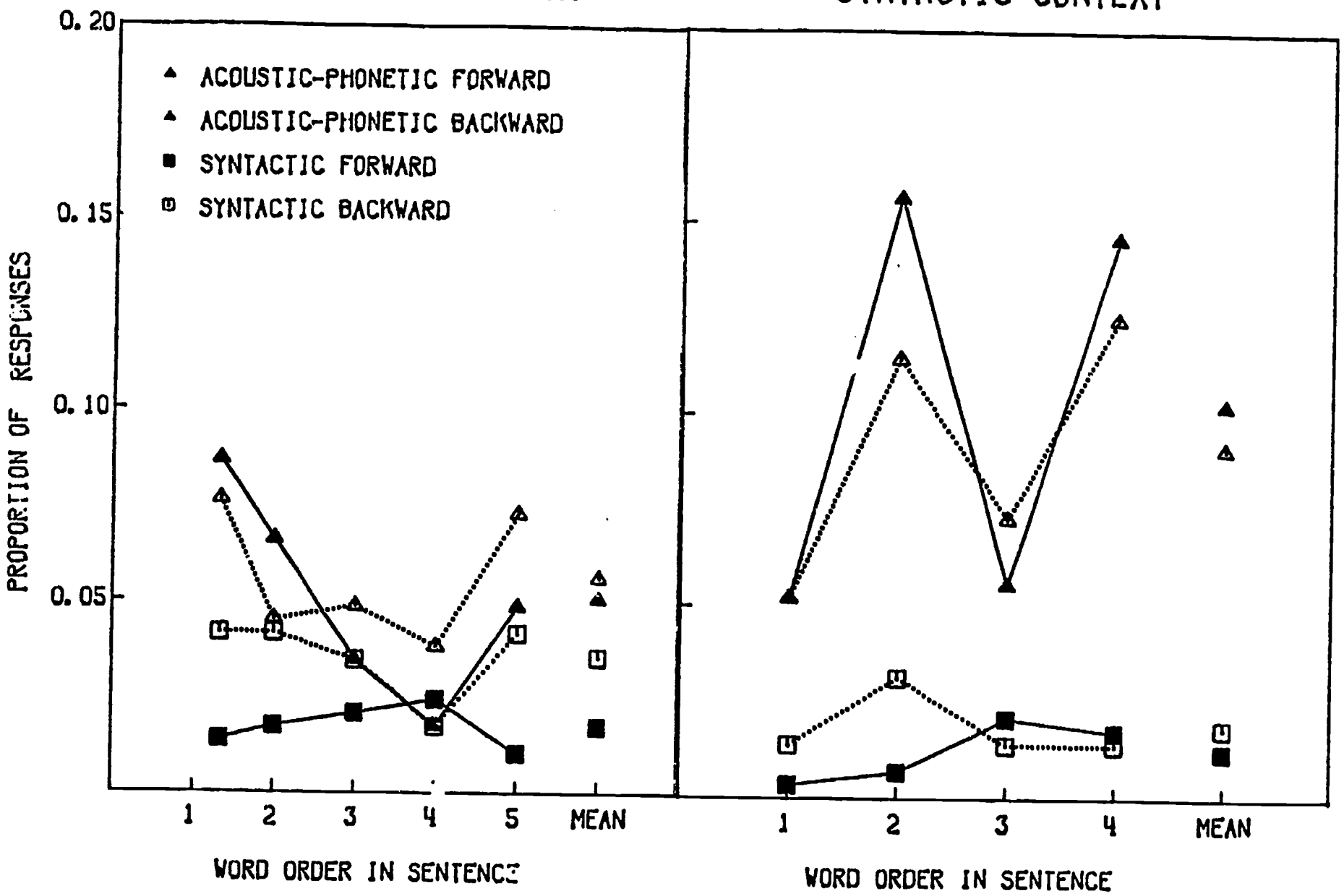


Figure 9. Number of incorrect word responses generated for forward-gated and backward-gated words in meaningful and syntactic sentences in Experiment 2, expressed as proportions of total responses for each sentence position. Responses based on acoustic-phonetic and syntactic information are shown as triangles and squares; forward-gated and backward-gated conditions are shown as filled and open symbols.

Insert Figure 10 about here

The bottom left panel of Figure 10 displays the response distributions for the backward-gated meaningful contexts. The word candidates generated on the basis of syntactic knowledge dominate the distribution at the 0, 50 and 100-msec gate durations. When 150 msec of signal duration was presented, both correct responses and acoustic-phonetic responses surpassed the contribution of purely syntactically based lexical candidates. Thus, when only word-final acoustic-phonetic information was available, the syntactic cues in the meaningful sentences allowed for a fairly early syntactic contribution to the set of potential word candidates.

This pattern of results for the meaningful context conditions suggests that top-down syntactic knowledge may be used to generate word responses early in the identification process, even if they are incompatible with the actual acoustic-phonetic input. Syntactic knowledge therefore appears to contribute to the set of word candidates hypothesized before enough acoustic-phonetic information is available to initiate lexical access. However, when at least 150 msec of acoustic-phonetic information is present in the speech signal, only candidates that are also acoustic-phonetically compatible with the input are maintained as candidates in the response distribution. At this point in processing, syntactically controlled, but acoustically inappropriate lexical candidates appear to be deactivated and their contribution to the response distribution decreases accordingly.

The top and bottom right panels of Figure 10 show the growth functions of the word response distributions for the forward-gated and backward-gated words in the syntactic sentence context, respectively. Again, the number of incorrect word responses based on acoustic-phonetic information begins to decline at shorter signal durations for forward-gated words than for backward-gated words. This relation between the peaks of the growth functions of the acoustic-phonetic component corresponds to the gating direction difference observed in the identification thresholds in Figure 6. As in Experiment 1, the syntactically-based word candidates constitute only a negligible proportion of responses in the syntactic context conditions (7.4%). Nevertheless, more word responses based on correct syntactic knowledge occurred on gating trials in which between 50 and 150 msec of signal duration was presented to listeners. Thus, a small number of syntactically appropriate word candidates were generated, even in the impoverished syntactic contexts, when minimal acoustic-phonetic information was available at the short signal durations.

Taken together, our analyses of the response distribution growth functions suggest that subjects used general syntactic knowledge to generate word responses in gating conditions in which the signal duration was less than 150 msec. With longer word signal durations, the number of purely syntactically based responses decreased to an insignificant proportion of the response candidates hypothesized in both meaningful and syntactic contexts. When compatible semantic information accompanied the syntactic cues, as in the meaningful sentences, the proportion of syntactically controlled responses increased. Thus, in meaningful sentences,

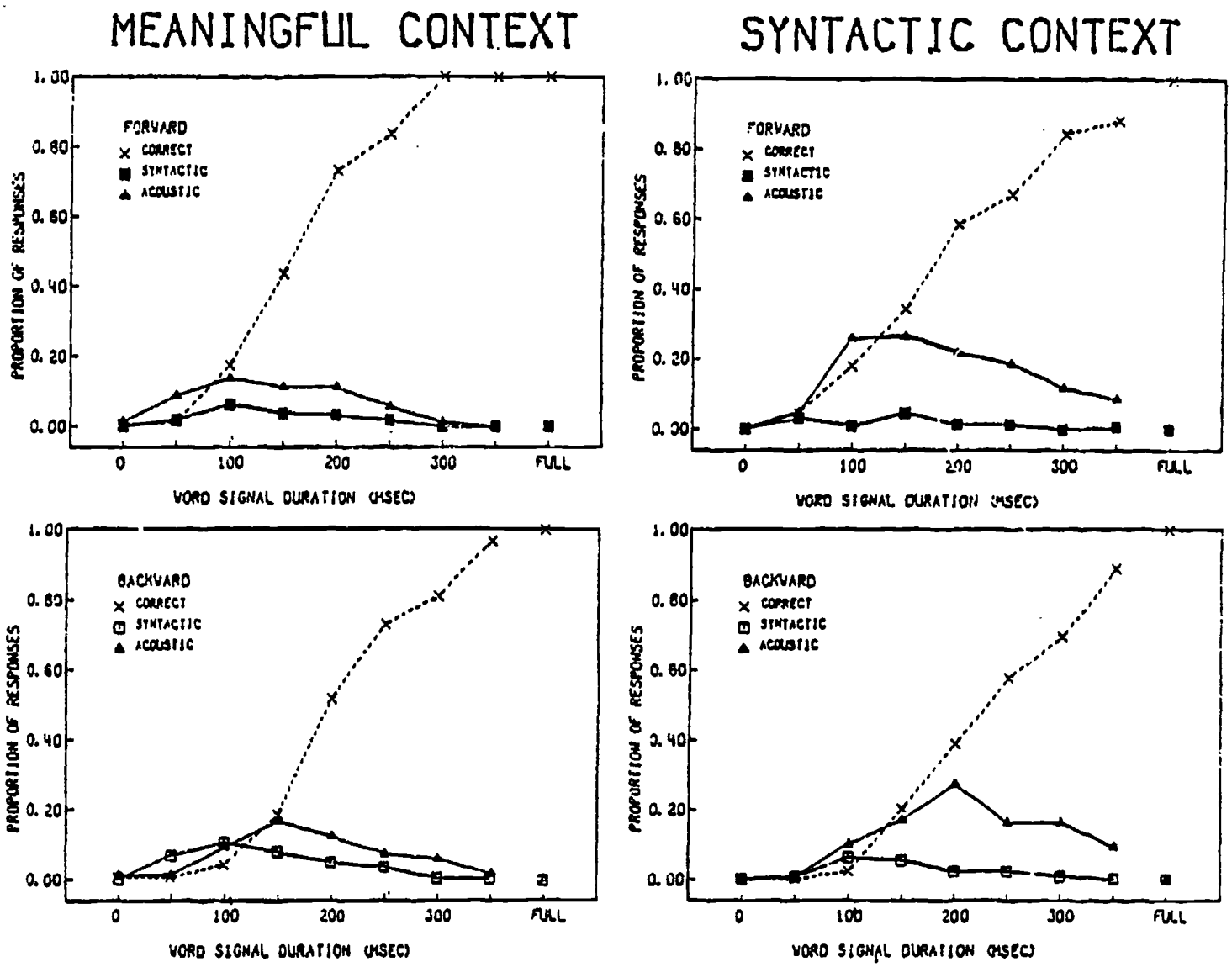


Figure 10. Growth functions of word response distributions for increasing amounts of signal duration in the meaningful and syntactic contexts in Experiment 2. expressed as proportions of the total responses for each gate duration. X's, triangles and squares indicate correct identification responses, incorrect responses based on acoustic-phonetic and syntactic sources, respectively; filled and open symbols indicate forward-gated and backward-gated conditions.

subjects deploy both bottom-up and top-down information to hypothesize word candidates when less than 150 msec of the signal was available. When more complete acoustic-phonetic specification of the word was provided, the contribution of syntactic and semantic knowledge was less prominent: Subjects reduced the number of word candidates based on compatible acoustic-phonetic information.

Discussion

The results of our second experiment, using a different presentation procedure and a different experimental design, replicated the major findings of the first experiment. In addition, the present study identified some of the temporal characteristics of the interaction between bottom-up and top-down knowledge sources in word identification. In both presentation methods, we found an advantage of word-initial acoustic-phonetic information over word-final acoustic-phonetic information in meaningful sentences. The amount of signal duration required for correct word identification and an analysis of the number of incorrect responses supported this conclusion. Further analysis of word candidates generated at increased signal durations revealed that the distributions of hypothesized words were quite sensitive to differences in the use of the various knowledge sources available in both the meaningful sentences and the semantically anomalous sentences.

General Discussion

The results from our two experiments demonstrate that words can be identified in sentences without word-initial acoustic-phonetic information. However, the hypothesized word candidates are sensitive to the presence of normal sentential semantic and syntactic constraints and generally follow the principle of bottom-up priority. Our data suggest that in normal, fluent speech processing, the acoustic-phonetic information contained in the 150 msec at the beginning of words is a major source of information used for lexical access and the ensuing word indentification process. While the processing system is also sensitive to nonsensory sources of knowledge, when normal continuous speech processing conditions are simulated, e.g. in the meaningful sentence contexts in Experiments 1 and 2, word-initial acoustic-phonetic information appears to control the distribution and selection of word candidates with as little as 150 msec of the beginning of a word.

Our analysis of the errors indicated that nonsensory based word candidates were frequently hypothesized when compatible acoustic-phonetic input was not present in the speech signal. We believe that this finding has three important theoretical implications. First, it suggests that interactive processes that employ both acoustic and nonsensory information occur either before or at the level of lexical access. We failed to find evidence for a strictly autonomous level of lexical access in word identification that was unaffected by higher sources of knowledge.

The present results also have several implications for understanding context effects in speech perception and word recognition. In both experiments we found that word identification in semantically anomalous sentences did not resemble the

corresponding processes in meaningful sentences. The similarities between isolated words and words in anomalous sentences in Experiment 1 suggest that in these impoverished contexts, words may have been processed as though they occurred in randomized lists without any internal structural organization. It seems plausible to us that the inhibition observed in the syntactic contexts points to a nonautonomous syntactic processor and/or integration stage of comprehension (Cairns, Note 1). Our data suggest parallel syntactic and lexical processing of words in spoken sentences where semantic constraints constitute a critical source of information for the operation of normal lexical access processes.

Finally, the present findings provide substantial support for the principle of bottom-up priority in word identification. However, there are several qualifications. While acoustic-phonetic information in the speech signal appears to be the primary source of information used by subjects to generate lexical candidates accessed from long term memory in the first stage of word identification, semantic and syntactic information present in sentences also enable nonsensory, syntactically compatible word candidates to be activated and entered into the pool of hypothesized word candidates. As the phonetic determination of a word begins to emerge, fewer and fewer word candidates are entertained by the listener. Thus, listeners use all the available information in both stages of spoken word identification, weighting, if only momentarily, the most reliable knowledge source most. Before acoustic information has accumulated to chunks of approximately 150 msec, syntactic knowledge does play a role in constraining the set of potential lexical candidates. At this point in time, acoustic-phonetic information gains prominence in the lexical access process, while both top-down and bottom-up sources continue to eliminate incorrect word candidates from the hypothesized set. The presence of compatible semantic and syntactic information is therefore an obligatory component of normal word identification in meaningful sentences. The balance among these various sources of information appears to provide an extremely efficient system to support spoken language comprehension over a wide range of listening conditions.

Footnotes

1. Our terminology and our procedure differ somewhat from previous studies using gated stimuli. Gated stimuli were used as early as 1963 (Pickett & Pollack, Pollack & Pickett). In their word identification task, Pollack and Pickett (1963) presented subjects with single presentations of words that had been excised out of spoken sentences; they called these "gated stimuli". Correct identification of such stimuli was often impossible for their subjects. Ohman (1966) has also used isolated gated nonsense stimuli and more recently, Grosjean (1980) and Cotton & Grosjean (Note 3) have employed isolated word stimuli. One difference between our stimulus conditions and Grosjean's "no context" stimuli lies in the treatment of the nonpresented part of the word signal. While Grosjean's gated stimuli were followed by durations of silence, we used noise masks in order to preserve the relative timing and original speech rhythm of the sentences.

Further, in previous studies using sentences (Grosjean, 1980; Cotton & Grosjean, Note 3), the final word has been the sole target for identification. In the present study we used multiple target words in sentences to simulate the demands of normal, continuous word identification in speech processing.

2. Behavioral evidence suggests that function words may be identified with different processes and knowledge sources than content words (Garrett, 1978; Marslen-Wilson & Tyler, 1980; Salasoo & Pisoni, Note 4). Therefore, function words in the experimental sentences were not treated as identification targets, but, instead, remained intact in every condition.

3. One deviant answer after at least two consecutive correct identification trials that was corrected on following trials was allowed.

Our identification points were operationally defined and differed from Grosjean's "isolation points" (1980), as well as from Ottevanger's identification points (1980, 1981). The former is also an empirical term used within the Cohort theoretical framework. It differentiates between the reduction of the cohort set to a single word and subjects' confidence of their conscious identification responses. Since confidence ratings were not collected in this study, the issue about the level of consciousness of words accessed from the mental lexicon will not be addressed. The latter term, "recognition point", adheres to the theoretical definition of the Cohort framework, i.e. "the phoneme going from left to right in the word that distinguishes that word from all others beginning with the same sound sequence." (Ottevanger, 1980, p.85; See also Tyler & Marslen-Wilson, 1982b) We believe our identification point analysis is more directly related to the amount of signal required to correctly identify spoken words in various conditions than these distinct usages.

4. Sentence-final words are generally longer and more stressed than words in other sentence positions. The computation of proportions enabled the comparison of identification points across both sentence positions and context types.

5. The variation in identification points for words in the syntactic context according to sentence position seen in Figure 1 is not observed when the data from the syntactic sentences are viewed as proportions in Figure 3. The shape of

the curve in Figure 1 reflects the fact that the form class and length of the target words in the Haskins syntactic sentences were confounded with sentence position. All the sentences had the same surface structure (i.e. Determiner adjective noun verb determiner noun). The words in the second and fourth positions in the sentences were always nouns and were longer than words in the other sentence positions.

6. In the absence of any standardized criteria for degree of phonetic overlap, the following guidelines were adopted by the first experimenter and a research assistant (N.C.) to determine membership to the acoustic-phonetic knowledge source. Similarity between the initial (or final) phoneme of the response and the target word received greatest weighting. Words whose initial phoneme only differed in its voicing feature from that of the target word, according to the Chomsky and Halle feature system (1968), were included in the category. Finally, word responses, which retained the vowel and at least one other phoneme from the target word in the correct sequence, were also considered to be based primarily on acoustic-phonetic information contained in the signal.

Reference Notes

1. Cairns, H.S. Autonomous theories of the language processor: Evidence from the effects of context on sentence comprehension. Unpublished manuscript, 1982.

2. Norris, D. Word recognition: Context effects without priming. Unpublished manuscript, 1982.

3. Cotton, S. & Grosjean, F. The gating paradigm: successive or individual presentations? Unpublished manuscript, 1982.

4. Salasoo, A. & Pisoni, D.B. Detecting masked content and function words in fluent speech. Paper presented at the meetings of the Midwestern Psychological Association, Minneapolis, May, 1982.

References

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

Bruner, J.S. & O'Dowd, D. A note on the informativeness of parts of words. Language and Speech, 1958, 1, 98-101.

Clark, H.H. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 335-359.

Cole, R.A. Listening for mispronunciations: A measure of what we hear during speech. Perception and Psychophysics, 1973, 13, 153-156.

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

Cole, R.A. & Jakimik, J.A. A model of speech perception. In Cole, R.A., (Ed.), Perception and Production of Fluent Speech. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

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

Fodor, J. A. The Language of Thought. Cambridge: Harvard University Press, 1979.

Forster, K.I. Accessing the mental lexicon. In Walker, E. & Wales, R., (Eds.), New Approaches to Language Mechanisms. Amsterdam: North Holland, 1976.

Forster, K.I. Levels of processing and the structure of language processor. In W.E. Cooper & E.C.T. Walker (Eds.), Sentence processing: Psycholinguistic studies presented to Merrill Garrett. Hillsdale, NJ: Erbaum, 1979.

Garrett, M.F. Word and sentence perception. In R. Held, H.W. Leibowitz, & H-L. Teuber (Eds.), Handbook of sensory physiology, Vol. VIII, Perception. Berlin, Springer Verlag, 1978.

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

Horii, Y., House, A.S. & Hughes, G.W. A masking noise with speech envelope characteristics for studying intelligibility. Journal of the Acoustical Society of America, 1971, 49, 1849-1856.

Marslen-Wilson, W.D. Optimal efficiency in human speech processing. Unpublished paper, 1981.

Marslen-Wilson, W.D. Speech understanding as a psychological process. In J.C. Simon (Ed.), Spoken Language Generation and Understanding. Dordrecht: Reidel, 1980.

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.

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, 329-335.

Miller, G. A. & Isard, S. Some perceptual consequences of linguistic rules. Journal of Verbal Learning and Verbal Behavior, 1963, 2, 217-228.

Nooteboom, S.D. Lexical retrieval from fragments of spoken words: beginnings versus endings. Journal of Phonetics, 1981, 9, 407-424.

Norris, D. Autonomous processes in comprehension: A reply to Marslen-Wilson and Tyler. Cognition, 1982, 11, 97-101.

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

Oden, G.C. & Spira, J.L. Influence of context on the activation and selection of ambiguous word senses. Wisconsin Progress Report No. 6, Dept. of Psychology, University of Wisconsin, Madison, 1978.

Ohman, S. Perception of segments of VCCV utterances. Journal of the Acoustical Society of America, 1966, 40, 979-988.

Ottevanger, I.B. Detection of mispronunciations in relation to the point where a word can be identified. Progress Report Institute of Phonetics University of Utrecht, 1980, 5, 84-93.

Ottevanger, I.B. The function of the recognition point in the perception of isolated mispronounced words. Progress Report Institute of Phonetics University of Utrecht, 1981, 6.

Pickett, J.M. & Pollack, I. Intelligibility of excerpts from fluent speech: effects of rate of utterance and duration of excerpt. Language and Speech, 1963, 6, 151-164.

Pisoni, D.B. Some current theoretical issues in speech perception. Cognition, 1981, 10, 249-259.

Pisoni, D. B. The perception of speech: The human listener as a cognitive interface. Speech Technology, 1982, 1, 2, 10-23.

Pollack, I. & Pickett, J.M. Intelligibility of excerpts from conversational speech. Language and Speech, 1963, 6, 165-171.

Schneider, W. & Shiffrin, R.M. Controlled and automatic information processing: I. Detection, search and attention. Psychological Review, 1977, 84, 1-66.

Seidenberg, M.S., Tanenhaus, M.K., Leiman, J.M. & Bienkowski, Automatic access of the meanings of ambiguous words in context: some limitations of knowledge based processing. Cognitive Psychology, 1982, 14, 489-537.

Shiffrin, R.M. & Schneider, W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological Review, 1977, 84, 127-190.

Swinney, D. Lexical access during sentence comprehension: (Re)consideration of context effects. Journal of Verbal Learning and Verbal Behavior, 1979, 18, 645-659.

Swinney, D.A. The structure and time-course of information interaction during speech comprehension: Lexical segmentation, access, and interpretation. In J. Mehler, E.C.T. Walker and M. Garrett (Eds.), Perspectives on Mental Representation: Experimental and Theoretical Studies of Cognitive Processes and Capacities. Hillsdale, NJ: Lawrence Erlbaum, 1982.

Tanenhaus, M.K., Leiman, J.M. & Seidenberg, M.S. Evidence for multiple stages in the processing of ambiguous words on syntactic contexts. Journal of Verbal Learning and Verbal Behavior, 1979, 18, 427-440.

Tyler, L.K. & Marslen-Wilson, W.D. The on-line effects of semantic context on syntactic processing. Journal of Verbal Learning and Verbal Behavior, 1977, 16, 683-692.

Tyler, L.K. & Marslen-Wilson, W.D. Conjectures and refutations: A reply to Norris. Cognition, 1982, 11, 103-107 (a).

Tyler, L.K. & Marslen-Wilson, W.D. Speech comprehension processes. In In J. Mehler, E.C.T. Walker and M. Garrett (Eds.), Perspectives on Mental Representation: Experimental and Theoretical Studies of Cognitive Processes and Capacities. Hillsdale, NJ: Lawrence Erlbaum, 1982(b).

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

Effects of Perceptual Load in Spoken Comprehension:

Some Interactions with Comprehension Goals*

Hans Brunner, Peter C. Mimmack, Alford R. White and David B. Pisoni

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

*This research was supported, in part, by NIMH Research Grant MH-24027, NIH Research Grant NS-12179, and NIH Training Grant NS-07134 to Indiana University. We are indebted to Jan Luce, Paul Luce, Laurie Walker, Robin Abrams, and Nina Sayer for their assistance in the propositional analyses of the pilot free recall study.

Abstract

This paper is a continuation of work reported by Brunner and Pisoni (Journal of Verbal Learning and Verbal Behavior, 1982,21,186-195) on the effects of subsidiary task demands in spoken text comprehension. Subjects listened to texts either without secondary task demands or then with simultaneous word- or phoneme-monitoring. Comprehension was evaluated through responses to verification statements and with measures of text recall. The effects of comprehension goals were ascertained by requiring only one method of retrieval from each subject and administering the task after instructions priming either that or the opposing method of retrieval (e.g., asking for text recall after instructions for verification statements). In conditions with question instructions and word monitoring we replicated the facilitation in verification latencies reported by Brunner and Pisoni (1982) and, also, found a corresponding enhancement of text memory for high-level propositions. In the conditions with recall instructions, however, the imposition of subsidiary demands only lowered recall and lengthened verification latencies relative to unconstrained comprehension. The implications of this work for research in text comprehension is also discussed.

Recently, Brunner and Pisoni (1982) published an investigation into the artifactual effects of subsidiary task paradigms, such as word- and phoneme-monitoring (q.v., Levelt, 1978), on otherwise-normal, unconstrained spoken text comprehension. The authors' contention was that procedures such as these, requiring subjects to focus specialized attention on some aspect of a text's speech sound structure, transformed normal comprehension into a divided attention task. As a consequence, it was not clear whether the style of comprehension induced by these procedures - or the results obtained from them - were analagous to the kind of comprehension engaged in outside of the laboratory.

Brunner and Pisoni (1982) examined this issue by having subjects listen to texts either without subsidiary task demands (i.e., "normal" or "unconstrained" comprehension) or then with simultaneous word- or phoneme-monitoring (e.g., Bergfeld-Mills, 1980; Foss & Swinney, 1973). Comprehension was evaluated through performance on a sequence of verification statements displayed immediately after the presentation of each text. Contrary to all expectations, Brunner and Pisoni found a facilitation, or decrease in the retrieval times for macro- and high-level propositions after comprehension with simultaneous word monitoring relative to the other two listening conditions. Since the magnitude of this effect was contingent upon the propositional level of the material being verified (low level propositions: -189 msecs; high level propositions: -548 msecs; macropropositions: -682 msecs), it was assumed to reflect some enhancement in the encoding of high level propositions during the course of text integration. In terms of Kintsch and van Dijk's model (1978), selective attention to whole words would increase the number of reinstatement cycles - and hence, the number of elaborative rehearsals - for high- and macro-level propositions.

The present research was prompted by our initial failure to find support for this post hoc interpretation. We reasoned that if the facilitation for high level material were due to some enhancement of encoding, then we should also observe superior recall for high level propositions after comprehension with word monitoring. A followup study was therefore conducted, requiring subjects to free recall each text immediately after presentation rather than answer a sequence of questions about it. In all other respects the design, materials, monitoring conditions and procedures of this experiment were identical to those employed by Brunner and Pisoni (1982).

The results of this study are shown in Figure 1. Text recall

Insert Figure 1 about here.

---

protocols were scored according to the propositional method of analysis devised by Kintsch and his associates (e.g., Kintsch, 1974; Kintsch, Kozminsky, Streby, McKoon & Keenan, 1975; Turner & Greene, Note 1). Protocol sheets from the different monitoring conditions were randomly shuffled so that the scorers were

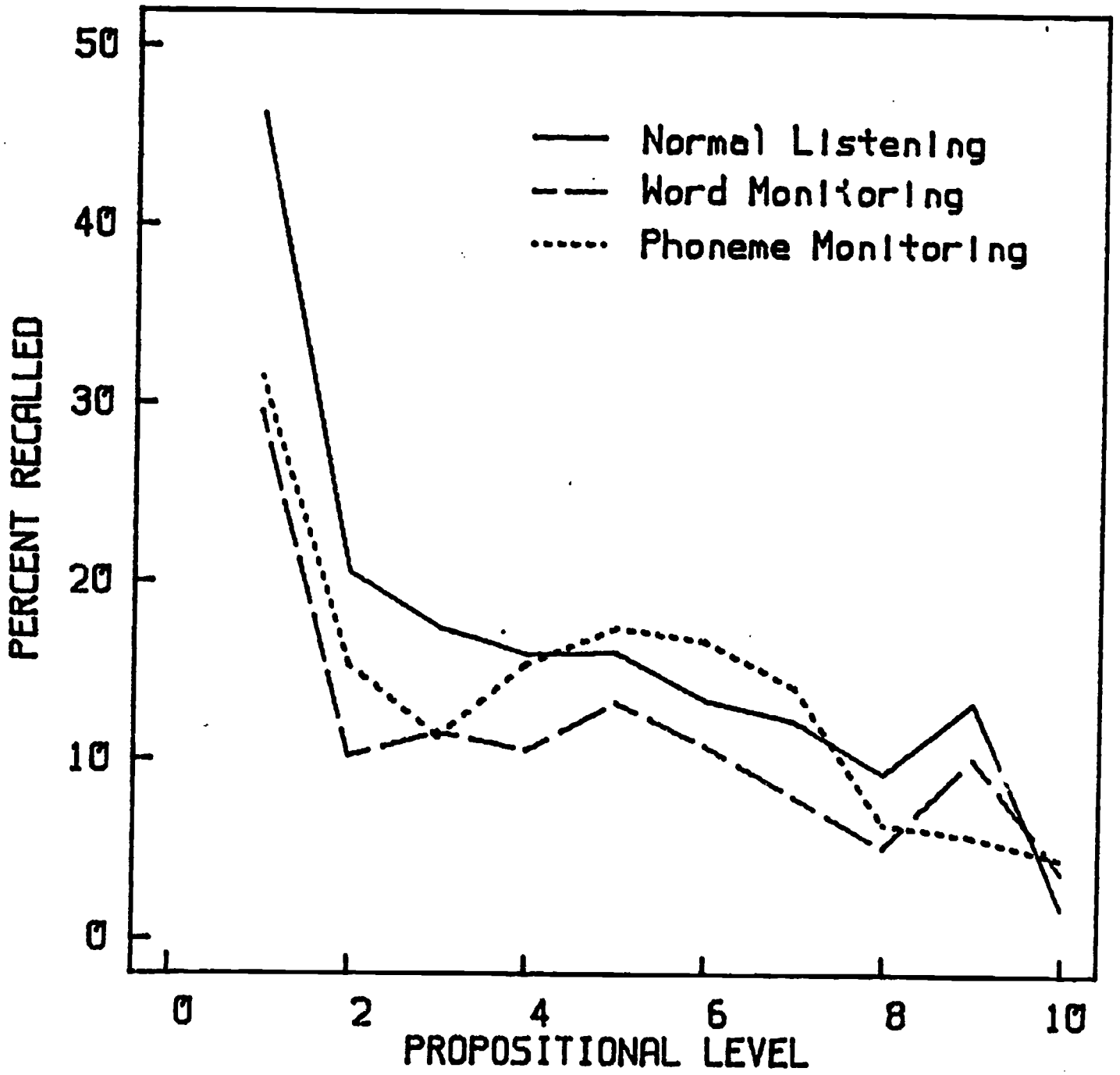


Figure 1. Immediate text recall across propositional levels and secondary listening conditions.

always blind with respect to the monitoring condition corresponding to each protocol. As can be seen, we failed to observe the kind of superior memory for high level propositions after comprehension with word monitoring that an interpretation based on variations in the strength of memory would require. Analyses of variance revealed reliable main effects due to propositional levels (min $F'(9,35)$[1]$=7.12$, $p<.01$) and monitoring conditions ($F_s(2,21)=4.01$, $MS_{err}=.014$, $p<.03$; $F_t(2,8)=11.76$, $MS_{err}=.003$, $p<.01$) and a significant, levels by monitoring conditions interaction ($F_s(18,189)=2.09$, $MS_{err}=.004$, $p<.01$; $F_t(12,48)=2.01$, $MS_{err}=.002$, $p<.04$). Thus, in addition to replicating typical levels effects under conditions of normal comprehension(e.g., Kintsch et al., 1975), we also obtained overall differences between the three monitoring conditions (normal comprehension: 16.6%; word monitoring: 11.3%; phoneme monitoring: 13.8%) which, unlike the results of Brunner & Pisoni (1982), at least reflected the added pressure on finite processing resources that one might expect from performance with simultaneous task demands. However, the enhancement of memory for low level propositions in the two monitoring conditions was quite counterintuitive. Pairwise comparisons of the three listening conditions produced significant, levels by monitoring condition interactions if normal comprehension was contrasted against either word- ($F_s(9,126)=2.2$, $MS_{err}=.004$, $p<.02$) or phoneme-monitoring ($F_s(9,126)=3.16$, $MS_{err}=.004$, $p<.001$), but not when the two monitoring conditions were compared to each other ($F<1.0$). It seemed clear, then, that (1) the encoding of low level propositions was facilitated by both forms of target monitoring and (2) the magnitude of this effect was greater for comprehension with phoneme monitoring, which is the more difficult of the two secondary tasks.

These findings, while adding to the evidence that subsidiary task demands distort comprehension, seemed irreconcilable both with the results of Brunner & Pisoni (1982) and any other extant theories of attention or text processing that we were aware of. Our only explanations for the obvious discrepancies between the two sets of results were either (1) that we had erred somehow in the selection and construction of materials or (2) that we were seeing unique interactions due to divergent comprehension goals.

We addressed both of these possibilities by constructing an entirely new set of materials and then repeating the procedures of both experiments with orthogonal manipulations of subjects' retrieval expectations. Half of our subjects were given instructions and practice trials emphasizing text recall, the other half were given the same amount of training for visually displayed, verification statements. Within each of these two groups, half of the subjects actually performed in the retrieval mode that they were instructed for. The other half, however, were abruptly switched to the opposing retrieval mode after hearing all of the experimental texts. By examining recall and question answering performance on an entirely new set of materials after consistent and inconsistent comprehension expectations, we hoped to dissociate the role of these two factors in producing the discrepant results described above.

There were no real precedents for the effects of comprehension goals on spoken text comprehension. Substantial effects, however, have been found in measures of word-by-word (e.g., Aaronson & Scarborough, 1976) and sentence-by-sentence (e.g., Graesser, Hoffman & Clark, 1980; Cirilo, 1981)

reading latencies. The general finding of this research is that subjects expecting subsequent recall both read more slowly and, also, show more sensitivity to variations in syntax and surface form than subjects reading for the purpose of subsequent question answering. As a consequence, recall protocols obtained from subjects reading for recall have been found to contain fewer inferences and generalizations and bear a closer resemblance to the text's underlying propositional structure than protocols obtained from subjects expecting questions or problem solving and then asked for final free recall (Fredericksen, 1972).

In order to induce realistic expectations and still obtain a reasonable amount of data from each subject, we had to delay recall and question answering procedures until after the presentation of four experimental texts. Thus, it would be unreasonable to expect a precise replication of either the Brunner & Pisoni (1982) or preceding pilot results, which were all collected immediately after the presentation of each text. However, if the differences between these results are due to divergent comprehension goals, then four primary effects should obtain: First, we should replicate the decrease in verification latencies for high level propositions after comprehension with question instructions and simultaneous word monitoring. Second, for subjects expecting questions but given unexpected final text recall we should also find some corresponding enhancement of recall for high level propositions. Third, for subjects both expecting and subsequently providing text recall, a replication of the results in Figure 1, showing worse overall performance in the two monitoring conditions than after normal comprehension, should obtain. Finally, for subjects expecting recall but given final question answering, we should observe uniformly longer verification latencies after comprehension with either form of monitoring than after unconstrained comprehension.

Method

Materials

Twelve expository texts were constructed on the basis of articles chosen from local and national periodicals (see Table 1 for the general characteristics of these texts).

Insert Table 1 about here.

A male speaker (H.B.) recorded all 12 experimental texts and 3 practice stories onto one track of an audiotape with a professional quality microphone (Unidyne III, Model #545) and tape recorder (Ampex AG-500) in a sound-attenuated IAC booth. Each of the 15 stories on this master tape was preceded by the word, "Ready". The 12 experimental texts were then subdivided into 3 materials sets,

Table 1

Experimental Text Characteristics

Materials Set	Topic	Number of Words	Number of Propositions	Number of Levels	Target Phoneme	Target Word
1	Toxic Wastes	243	112	6	/d/	dispose
1	Real Estate	257	115	6	/h/	home
1	Robots	188	78	6	/r/	robots
1	Taxes	224	99	8	/t/	tax
2	Stock Brokers	218	106	6	/b/	broker
2	Florida	210	89	8	/g/	ground
2	Heart Disease	249	120	8	/p/	program
2	Subways	198	82	7	/t/	tunnel
3	Locomotion	319	157	8	/r/	run
3	Archeology	228	103	9	/k/	clay
3	hysterectomies	275	106	6	/d/	doctor
3	Dormitories	195	89	7	/d/	dormitory

each consisting of 4 experimental texts preceded by the 3 practice stories (always played in the same, fixed order). The experimental texts of each set were then permuted into 3 random orders and copied (with practice stories still at the beginning) onto 3 separate audiotapes. The resulting 9 experimental tapes comprised the materials for this experiment.

Phoneme targets for the phoneme monitoring condition consisted of the word-initial phonemes of the word targets in the word monitoring condition (cf., Blank, Pisoni & McClaskey, 1981). Marking tones, inaudible to the subjects, were placed on the second track of the audiotapes at points corresponding to the onset of each word/phoneme target. Each tone initiated an interrupt to the computer which started a time which, in turn, was stopped when subjects pressed a response button.

Verification statements for the question answering conditions consisted of true/false and "Remember" questions, as in Brunner & Pisoni (1982). Examples of these are shown, with one of the experimental texts, in Table 2.

Insert Table 2 about here.

Five questions were constructed for each text. Two of these were always Remember questions, probing memory for surface form. Remember questions were always presented in a standard sentence frame ("Did the word "XXXXX" occur in this story?") with either a target (e.g., "avoid"), a rhyming distractor (e.g., "annoyed") or a synonymous distractor (e.g., "escape") in the test position. The present experiment differed from the work of Brunner and Pisoni (1982) by having both high- and low-level Remember questions. High level Remember questions contained target words serving as arguments in high level propositions of the text's underlying propositional structure; low level Remember questions were similarly derived from low level propositions of the underlying meaning structure.

Propositional content was evaluated with verification statements pertaining either to high level propositions, low level propositions, or inferences, as in Brunner and Pisoni (1982). The high- and low-level probes tested for information explicitly conveyed in the text. Inferences bore no direct relationship to explicit propositional content but could be answered only through the development of macropropositions from emergent relationships between the explicit propositions.

Table 2

Prototypical Text with Comprehension Probes

Boston researchers have recently shown that monkeys in a modest exercise program can avoid many of the ill effects of a high-fat diet known to cause hardening of the arteries. Because of the difficulty of doing studies on humans, the evidence that exercise can prevent heart disease has all been indirect. Monkeys, however, are much like humans in their basic physiology. And in a research laboratory the effects of exercise can be measured with a relatively small number of animals.

The Boston team began with 27 monkeys divided into three groups, each with its own program: one non-exercising group on standard monkey chow; one non-exercising group on a high-cholesterol, high-fat diet; and a third group of exercising monkeys on the fatty diet.

None of the animals on the standard, low-fat diet developed signs of heart disease, but most of the sedentary monkeys on the high-fat diet did. Despite their diet, the monkeys on a program of exercise showed few signs of hardening of the arteries. Their heart rates dropped and their hearts grew larger. The size of critically important coronary arteries also increased in the exercising animals.

The exercise program was deliberately chosen to approximate a moderate jogging regimen in humans. After being worked into good physical condition, the animals were required to exercise three times a week - just enough to maintain their fitness. That such a modest program should have had such dramatic results speaks strongly to the importance of physical fitness for health in today's society.

High-level Remember Question:

Did the word "escape"(S) / "avoid"(T) / "annoyed"(R) occur in this story?

Low-level Remember Question:

Did the word "extremely"(S) / "critically"(T) / "politically"(R) occur in this story?

High-level Proposition:

A high-fat diet without exercise produces fairly substantial hardening of the arteries.

Low-level Proposition:

A research program without monkeys provides only indirect evidence on heart disease.

Inference:

A lowfat diet without exercise produces no significant change in heart size.

Note: "T" indicates the target word, "S" the synonym distractors, and "R", the rhyming distractors.

Procedure and Apparatus

Experimental sessions were conducted in groups of one to five subjects. Each subject was seated in an experimental cubicle equipped with a Ball Brothers standard CRT display monitor (Model TV-120), a seven-button response box, and a pair of TDH-39 headphones connected to an Ampex AG-500 tape recorder. Each booth was interfaced to a PDP-11/34 computer, which presented instructions, monitoring targets, and test questions on the CRT's and recorded all responses and latencies from the response boxes.

Instructions were read to the subject by the experimenter, who told them that they would hear a number of short stories on various topics, which they were to listen to for content. One half of the subjects were then told that they would subsequently have to recall the stories; the other half were given instructions and examples leading them to expect subsequent question answering. All subjects were then given a single practice trial, requiring either free recall or question answering following presentation of the first practice story. Subjects in the conditions with normal comprehension were then given the remaining two practice texts and, depending on their instructional condition, were either asked for recall of the two stories or were required to respond to a sequence of 10 verification statements concerning them.

Subjects in the two monitoring conditions, however, were told that they would also have to listen for the presence of a prespecified, word-initial target phoneme or target word during the presentation of each story. Monitoring targets would be visually displayed to subjects prior to the start of each text. During the text's presentation subjects were instructed to keep their index finger resting lightly on a "READY" button (mounted in the center of each response box) and press it as quickly as possible whenever the prespecified target was detected. Subjects in the monitoring conditions were then given two more practice stories (the same two heard by the normal comprehension group) with monitoring and retrieval instructions appropriate to their particular condition.

Each subject then heard and responded to the 4 experimental texts of only one materials set. The presentation of each text was visually prompted ("ATTENTION! New Story Coming Up. Please press READY to begin.") and delayed until all subjects had pressed their READY buttons. In the two monitoring conditions, new story prompts were augmented with a display of the trial's target word or phoneme (e.g., "Listen for the sound /d/, as in "dancing"). After hearing all four experimental texts through headphones, the subjects were either asked to free recall the four texts or respond to a randomized sequence of 20 verification statements. Each verification statement was subject-initiated, with the response to it being followed by a 7-point confidence rating on the answer just provided and, then, feedback (via lights mounted in each response box) indicating the correct answer for that question.

Subjects and Design

The three listening conditions (i.e., unconstrained comprehension, word monitoring, and phoneme monitoring) and four instructional conditions (i.e., question answering/final question answering; question answering/final free recall; recall/final question answering; recall/final free recall) were all manipulated between subjects. Question types (i.e., high- and low-level Remember questions, high level propositions, low level propositions, and inferences) were manipulated within subjects. For the sake of counterbalancing, three materials sets (manipulated between subjects) were also included in the design.

The data from 216 subjects - 6 for each set of materials within each combination of instructions, listening conditions, and final retrieval mode - were required for this design. Subjects were all Indiana University undergraduates whose participation in the experiment served as partial fulfillment of a course requirement. All subjects were native speakers of English with no prior history of hearing or speech disorders.

Results

Monitoring Data

Mean latencies and detection rates from the two monitoring conditions were computed, separately, for each subject and text. The overall subjects-random means are shown in Table 3.

Insert Table 3 about here.

As can be seen, words were detected more quickly and accurately than phonemes. The 144 millisecond difference (min $F'(1,23)=13.22^2$) in monitoring latencies and the 15.6% spread (min $F'(1,27)=21.25$) in detection rates were both statistically reliable. There were no differences in these measures as a function of prior instructions or subsequent retrieval conditions.

Unlike the findings reported by Brunner and Pisoni (1982), the current data were collected without any rejection of subjects due to insufficient monitoring, recall, or question answering performance. This may account, in part, for the somewhat larger than usual difference between word and phoneme detection rates. Overall levels of performance, however, were comparable to those reported by Brunner and Pisoni (1982) and, therefore, were also somewhat lower than detection rates reported elsewhere in the literature (cf., Foss & Swinney, 1973; Bergfeld-Mills, 1980).

Table 3

Mean Latencies and Detection Rates for Word- and Phoneme-Monitoring Conditions.

	Word Monitoring	Phoneme Monitoring
Latencies (msecs)	575	719
Detection Rates	.899	.743

We have data from two studies, then, showing reductions in monitoring performance as a function of both the imposition of real comprehension standards with text recall and verification probes and, also, the use of whole texts rather than isolated sentences for comprehension materials. As in our previous work, this kind of tradeoff between primary and subsidiary task demands only underscores our concern with the assumptions about simultaneous comprehension processing which have been made in the past. In all other respects, however, the current pattern of monitoring results is entirely consistent with the results of numerous studies already conducted using these measures (e.g., Savin & Bever, 1970; Foss & Swinney, 1973; Blank, Pisoni & McClaskey, 1981).

In order to assess the relative effects of word- and phoneme-monitoring on the time course of ongoing comprehension, it is important to be able to assume roughly equal levels of performance in the two tasks. Brunner and Pisoni (1982) had tried, and failed, to achieve this through the application of a priori subject rejection procedures. Given tneir failure and, also, the difficulty of devising appropriate rejection criteria for all of the different measures of this work, we instead decided to rely entirely on multivariate correction. Thus, recall and verification data were twice analyzed; once in raw form and once after removal of the covariance due to these differences in secondary detection rates (q.v., Kerlinger & Pedhazur, 1973).

Questions: Representation of Surface Form

Overall, our manipulations of instructions, propositional height, and listening conditions had no significant effects on the proportion of correct Remember target identifications (p(correct response) = .71). Levels of confidence, however, were reliably influenced. After removal of the covariance due to monitoring detection performance, there was a reliable listening conditions main effect ($F_s(2,190) = 5.9$, $MS_{err} = 1.779$; $F_t(2,88) = 5.139$, $MS_{err} = 1.31$; cell means -- normal comprehension: 5.47; word monitoring: 4.9; phoneme-monitoring: 4.33) due to higher levels of confidence expressed by subjects in the conditions with less demanding levels of perceptual load. Thus, despite the null effect on proportions of correct responses, it seems clear that the imposition of increasingly pre-lexical secondary demands made subsequent access to surface form more and more difficult.

Analysis of the raw Remember target response latencies produced a significant instructions main effect (min $F'(1,107) = 5.67$, $p < .05$) and a significant propositional height by instructions interaction (min $F'(1,109) = 5.12$, $p < .05$). These were unaffected by the removal of monitoring detection rate covariance (instructions main effect: min $F'(1,192) = 5.867$; instructions by propositional height interaction: min $F'(1,225) = 3.526$). The instructions main effect was due to the extra retrieval time required by subjects expecting final text recall (cell means: recall instructions: 4444 msecs; question instructions: 3885 msecs). However, from the instructions - by - propositional level interaction, shown in Table 4, it can be seen that most of this variance was restricted to the recognition of words associated with low levels of propositional text structure.

Insert Table 4 about here.

For subjects expecting final free recall, target words serving as arguments in low-level propositions took about 500 msecs longer to retrieve and verify than targets associated with high-level propositions. If question-answering was expected then, for precisely the same questions and texts, there was a reversal in this effect, with the lower levels of text organization now requiring roughly 400 msecs less time to recognize than the corresponding high-level words.

Removal of the extraneous covariance also produced an additional, significant instructions by listening conditions interaction ($F_s(2,183) = 2.91$, $MS_{err} = 253000$; $p < .05$; $F_t(2,131) = 69.59$, $MS_{err} = 66119$), which is shown in Table 5.

Insert Table 5 about here.

As can be seen from the superscripts, this is due to a facilitation in response latencies after comprehension with question instructions and simultaneous word monitoring, which did not obtain in the corresponding condition after recall instructions. Target recognition latencies after comprehension with question instructions and phoneme monitoring were also faster, by 575 msecs, than those from unconstrained comprehension. This, however, amounted to no more than an intermediate degree of facilitation in the relevant context of variance. For subjects expecting final text recall, the imposition of subsidiary task demands produced consistent, but statistically unreliable, increases in the target recognition latencies.

Analysis of the probabilities of correct rejection for the two forms of (rhyming vs. synonym) distractors produced, first, a significant main effect due to distractor types (raw data: min $F'(1,14)=6.13$, $p<.05$; covariance-adjusted: min $F'(1,385)=11.27$), indicating that subjects were almost twice as accurate at rejecting rhyming (p(correct rejection) = .67) as opposed to synonymous (p(correct rejection) = .33) distractors. There was also a main effect due to listening conditions (raw data: $F_s(2,102)=2.76$, $MS_{err}=.259$, $p<.06$; $F_t(2,22)=3.27$, $MS_{err}=.099$, $p<.05$; covariance-adjusted: $F_s(2,407)=4.95$, $MS_{err}=.212$; $F_t(2,263)=2.5$, $MS_{err}=.13$, $p<.08$) caused by higher probabilities of correct rejection after unconstrained comprehension (p(correct rejection) = .58) than after comprehension with either form of subsidiary task demand (word monitoring: p(correct rejection) = .44; phoneme monitoring: (p(correct rejection) = .49).

Table 4

Response Latencies (msecs) for the "Remember" Targets;
Instructions by Propositional Level Interaction

	Recall Instructions	Question Instructions
High-level Propositions	4197[b]	4099[b]
Low-level Propositions	4672[bc]	3659[a]

Note: Cell means with the same superscript, do not differ significantly ($p < .05$).

Table 5

Covariance-Adjusted Response Latencies (msecs)
for "Remember" Targets;
Instructions by Listening Condition Interaction

	Recall Instructions	Question Instructions
Normal	4199[b]	4372[b]
Comprehension		
Word Monitoring	4568[b]	3486[a]
Phoneme Monitoring	4560[b]	3761[ab]

Note: Cell means with the same superscript, do not differ significantly ($p < .05$).

Orthogonal to this, analyses of the distractor rejection latencies produced a significant instructions main effect (min F'(1,99) = 12.68) and a significant propositional levels by distractor type interaction (min F'(1,36) = 7.47). The effect of instructions was unchanged by the removal of covariance with monitoring performance (min F' (1,252) = 7.76) and reflected an overall increase in distractor rejection latencies after instructions for final text recall (recall instructions: 5161 msecs; question instructions: 4303 msecs). The propositional levels by distractor type interaction remained only marginally significant after removal of the detection rate covariance ($F_s(1,203) = 3.14$, $MS_{err} = 4600000$, $p < .07$; $F_t(1,183) = 3.09$, $MS_{err} = 394479$, $p < .08$) and is shown in Table 6.

Insert Table 6 about here.

Whereas the rhymes were unaffected by propositional level, it can be seen that synonyms serving as arguments in low-level propositions took longer to reject than either the rhymes or high-level synonyms. When considered in conjunction with the higher probabilities of correct rejection for rhymes, one might conclude that, whereas rhyming distractors were rejected on the basis of obvious, macropropositional inconsistencies with the story, synonyms were most likely rejected on the basis of micropropositional evaluation.

Questions: Representation of Meaning

The effects of listening conditions and comprehension goals on the underlying representation of meaning were examined, first, through analyses of the confidence ratings and proportions of correct responses. Inferences were completely unaffected in either of these domains (mean confidence rating: 5.24; p(correct response): .72). Confidence ratings for the high- and low-level propositions were reliably influenced by the three listening conditions (min F'(2,114) = 5.26). This effect (mean confidence ratings - normal comprehension: 6.19; word monitoring: 5.53; phoneme monitoring: 5.25), however, was eliminated by the removal of monitoring detection covariance. Removal of the extraneous covariance also added to the reliability of differences in proportions of correct responses across propositional levels (min F'(1,205) = 4.93, $p<.05$), with responses to the high-level propositions being 17% more accurate than those to their low-level counterparts (cell means - high-level: .94; low-level: .77).

Propositional levels effects were also obtained from the analysis of response latencies (raw data: min F'(2,127) = 11.24; covariance-adjusted: min F'(2,419) = 7.65). As in Brunner and Pisoni (1982), inferences (cell mean: 5664 msecs) required more verification time than low-level propositions (cell mean: 5089 msecs), which in turn took longer to respond to than high-level propositions (cell mean: 4667 msecs). In addition, there was also a significant main effect due to instructions (raw data: min F'(1,111) = 11.35; covariance-adjusted: min

Table 6

Response latencies (msec) for the "Remember" Distractors;
Propositional Level by Distractor Type Interaction

	Rhyming Distractors	Synonym Distractors
High-level Propositions	4753[a]	4698[a]
Low-level Propositions	4373[a]	5328[b]

Note: Cell means with the same superscript, do not differ significantly ($p < .05$).

F'(1,421) = 12.52), reflecting, as before, an overall increase in response latencies due to the expectation of final text recall (cell means - recall instructions: 5481 msecs; question instructions: 4799 msecs).

More germane to the replication of Brunner and Pisoni (1982), however, was a significant listening conditions by instructions interaction (raw data: min F'(2,118) = 3.47; covariance-adjusted: min F'(2,362) = 4.87), which is shown in Table 7.

Insert Table 7 about here.

As is evident from the superscripts, comprehension with either form of subsidiary task and the expectation of final recall produced substantial increases in the verification latencies, relative to unconstrained comprehension. This was especially pronounced after comprehension with word monitoring, where the difference from normal comprehension was over 1100 milliseconds in magnitude. For subjects expecting question answering, however, the Brunner and Pisoni (1982) facilitation due to word monitoring was nicely replicated.

Although the three-way, instructions by levels by listening conditions interaction failed to reach significance across all forms of analysis, it can be seen from the cell means in Tables 8 and 9 that the magnitude of the instructions by listening condition interaction differed across propositional levels. Performance after instructions for recall is shown in Table 8.

Insert Table 8 about here.

Here it can be seen that, relative to unconstrained comprehension, the increase in response latencies due to word monitoring was much greater for low-level propositions (2322 msecs) than for high-level propositions (785 msecs), which, in turn, underwent a larger increase due to word monitoring than the inferences (428 msecs). There was a decrease in the magnitude of effect at the higher propositional levels, then. An opposing trend is evident from the cell means of Table 9, showing performance after question-answering instructions.

Table 7

Covariance-Adjusted Verification Latencies (msecs) for Inferences, High- and Low-level Propositions; Instructions by Listening Condition Interaction

	Recall Instructions	Question Instructions
Normal Comprehension	5690^{abc}	5523^{ab}
Word Monitoring	6867^{c}	4752^{a}
Phoneme Monitoring	6159^{bc}	5273^{ab}

Note: Cell means with the same superscript, do not differ significantly ($p < .05$).

Table 8

Covariance-Adjusted Verification Latencies (msecs) for Inferences, High- and Low-level Propositions; Recall Instructions

	High Level	Low Level	Inferences
Normal Comprehension	5151^{a}	5356^{ab}	6563^{c}
Word Monitoring	5934^{abc}	7678^{d}	6991^{cd}
Phoneme Monitoring	5267^{ab}	6340^{bc}	6869^{cd}

Note: Cell means with the same superscript, do not differ significantly ($p < .05$).

Insert Table 9 about here.

Here, as in Brunner and Pisoni (1982), there is an increase in the magnitude of effect at higher propositional levels. For low-level propositions, the facilitation due to word monitoring (17 msecs) is completely negligible. This becomes much higher, however, as the comparisons proceed, first to high-level propositions (920 msecs), and then to the inferences (1378 msecs). Thus, selective attention to whole words interacts in opposing ways with intentions for recall vs. question-answering: In the former case, there is an increase in latencies which is smaller at the higher levels of meaning; in the latter, there is a decrease in latencies which gets larger at the high- and macro-propositional levels of meaning.

Looking only at the normal comprehension of high- vs. low-level propositions, Tables 8 and 9 also reveal an instructionally induced reversal in propositional levels effects ($F_s(1,102) = 5.84$, $MS_{err} = 2166446$, $p<.02$; $F_t(1,11) = 9.44$, $MS_{err} = 245182$, $p<.01$). Previous research by McKoon (1977) has already demonstrated longer verification latencies (by about 300 msecs) for low-level propositions, a finding which follows nicely from Kintsch's (1974) theory on the representation of text in permanent memory, and also from Cirilo and Foss' (1980) more recent finding of longer reading times for high-level propositions. In the present results, for subjects expecting final recall, we replicated this effect: low-level propositions required roughly 200 milliseconds longer for verification than high-level propositions. For subjects expecting final questions, however, it can be seen that low-level propositions were verified in roughly 500 milliseconds less time than their high-level counterparts. Recall that this is precisely the same, instructions by levels interaction found for Remember target latencies, shown in Table 4.

Text Recall

Taken together, the preceding pattern of verification latencies both replicates the findings of Brunner and Pisoni (1982) and, also, provides strong support for the importance of comprehension goals as a determinant of the form that the interaction with perceptual load will take. The minimal criterion for an explanation based on differential encoding, however, lies in the convergence of these latencies with performance in text recall. In general, for each decrease in response latencies there should be a corresponding increase in the analogous measure of text recall, and vice versa. Thus, where we had observed a facilitation in the response latencies for high-level propositions due to word monitoring, there should be a corresponding increase in recall of high-level propositions. Similarly, where we had observed longer verification latencies following comprehension with either form of subsidiary task and prior instructions for recall, there should be corresponding decreases in levels of correct recall for these conditions.

Table 9

Covariance-Adjusted Verification Latencies (msecs) for Inferences, High- and Low-level Propositions; Question Answering Instructions

	High Level	Low Level	Inferences
Normal Comprehension	5080^{abc}	4509^{ab}	6981^{d}
Word Monitoring	4160^{a}	4492^{ab}	5603^{bc}
Phoneme Monitoring	4692^{ab}	5045^{abc}	6082^{c}

Note: Cell means with the same superscript, do not differ significantly ($p < .05$).

From the trends[3] shown in Figure 2, it can be seen that this is exactly what we found. Unlike our pilot recall data,

---

Insert Figure 2 about here.

---

which were collected immediately after the presentation of each text, there was a significant (min F'(7,1047) = 14.49), monotonic decrease in performance across propositional levels (cf., Kintsch, Kozminsky, Streby, McKoon, and Keenan, 1975) for all three of our listening conditions. There were also significant (min F'(2,934) = 12.89), overall differences among the three listening conditions, with levels of recall after normal comprehension (p(correct recall) = .26) and comprehension with word monitoring (p(correct recall) = .25) collectively being about 10% higher than recall after comprehension with simultaneous phoneme monitoring (p(correct recall) = .16). Finally, there was a significant effect due to instructions (min F'(1,925) = 3.86, $p<.05$): Subjects expecting questions (p(correct recall) = .24) generally recalled about 3% more than the subjects expecting final text recall (p(correct recall) = .21).

Although the two-way, instructions by listening conditions interaction failed to reach significance ($F_s(2,816) = 3.28$, $p<.03$; $F_t(2,396) = .267$), the magnitude and direction of these effects was not equivalent for all three listening conditions, under both forms of instruction. The breakdown of overall cell means is shown in Table 10:

Insert Table 10 about here.

Here it can be seen that, whereas recall after normal comprehension was unaffected by differences in prior instructions[4], performance after word monitoring and question instructions was 6% higher than that found for subjects given word monitoring and recall instructions ($F_s(1,272) = 8.4$, $MS_{err} = .032$; $F_t(1,132) = 4.32$, $MS_{err} = .025$). There was also an analogous, 4% increase in performance following comprehension with phoneme monitoring, which failed to reach significance. It seems clear, then, that most of the instructions main effect variance was actually an artifact of interactions with the two differing forms of secondary task.

There is strong convergence between these results and the preceding verification latencies. For subjects expecting recall, the imposition of increasingly stringent subsidiary demands produced corresponding decreases in

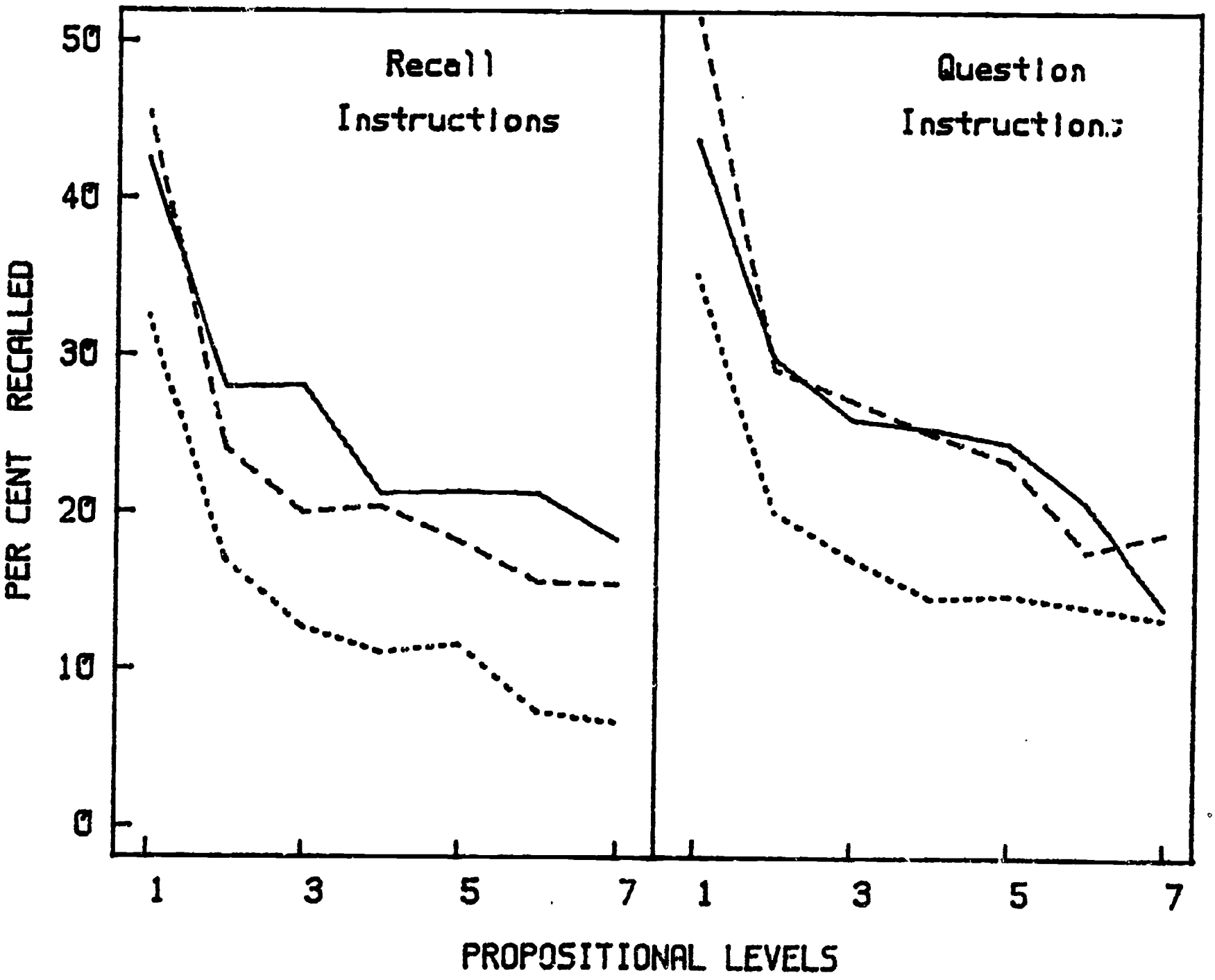


Figure 2. Delayed text recall across propositional levels and secondary listening conditions (---------- normal comprehension; - - - - word monitoring; . . . . phoneme monitoring).

Table 10

Proportions of Correct Recall
Following All Combinations of Instructions
and Listening Conditions

	Recall Instructions	Question Instructions
Normal Comprehension	.26^{cd}	.25^{cd}
Word Monitoring	.22^{bc}	.28^{d}
Phoneme Monitoring	.14^{a}	.18^{ab}

Note: Cell means with the same superscript, do not differ significantly ($p < .05$).

proportions of correct recall. For subjects expecting questions, however, there was a facilitation in recall closely corresponding to that obtained for the verification latencies, above. Comprehension with word monitoring produced a recall function, in Figure 2, virtually superimposed over that for normal comprehension over most propositional levels. The exception to this was the highest propositional level, where the proportion of correct recall was some 9% higher than that resulting from unconstrained comprehension. Thus, there is good agreement between these proportions of correct recall and the preceding latency data: Where the subsidiary tasks produced generally longer response latencies we find lower levels of text recall; where they produced some facilitation in response latencies, we find a corresponding increase in recall for high-level propositions.

Discussion

Taken as a whole, these results are in strong agreement with the earlier findings of Brunner and Pisoni (1982). With respect to the retention and use of surface form, we replicated the facilitation in target recognition latencies after comprehension with word monitoring and prior instructions for question-answering. For subjects listening to the texts with the aim of subsequent recall, the imposition of simultaneous word monitoring had no such effect, producing instead a small and statistically unreliable increment in response latencies. With respect to the two forms of distraction, we also replicated the almost two-to-one ratio in probabilities of correct rejection reported by Brunner and Pisoni (1982). Selective attention to the speech sound structure might have produced higher false alarm rates to rhyming distractors after comprehension with phoneme monitoring but the evidence for this hypothesis simply failed to materialize. Considering all of these factors together, then, we must conclude that the encoding of surface form was rapid, semantic in nature, and completely unaffected by the qualities of subsidiary task demands. This is consistent with recent work on the deficiencies of synthetic speech (e.g., Luce, Feustal & Pisoni, in press; Luce, Note 2) showing impairments in the encoding of meaning, rather than interference in the perceptual processing, per se. It is possible, then, that the processing of surface form is automatic in nature (q.v., Shiffrin & Schneider, 1977; Schneider & Shiffrin, 1977), the incidental byproducts of perturbations in the perceptual stream being passed on to later, more controlled stages of text processing.

The representation of surface form was influenced, however, by the propositional level associated with each to-be-retrieved target or distractor word. For "Remember" target words recognized after recall instructions, placement in the lower propositional levels produced longer retrieval latencies. However, if the same materials were presented after question-answering instructions, then the opposing pattern was obtained, with low-level target words now requiring less time for recognition than their high propositional level counterparts. The presence of these levels effects indicates that the representation of surface form is both integrated with, and accessed through, some hierarchical representation of text meaning in memory.

A similar reversal in the propositional levels effect was also obtained, after normal comprehension, for the verification of high- and low-level propositions. As noted above, the finding of longer verification latencies for low level propositions both replicates (i.e., McKoon, 1977; Cirilo & Foss, 1980), and is consistent with, the general tenor of research into the propositional representation of text in memory (e.g., Kintsch, 1974; Kintsch et al., 1975). However, the restriction of this effect - for both semantic verification statements and the recognition of individual target words - to conditions emphasizing verbatim memory raises the serious possibility that previous theorizing on the basis of materials comprehended in this manner (e.g., Spiro, 1980; Hayes-Roth & Thorndyke, 1979; McKoon & Ratcliff, 1980; Fletcher, 1981; Manelis & Yekovich, 1976) could be an artifact of subjects' encoding expectations, rather than being a reflection of "normal" comprehension processing, per se. The finding of similar, instructionally induced differences in verification latencies by Green (1975) further legitimizes this possibility. In any event, the finding is noteworthy, and the possibility of such an artifact is clearly in need of further investigation.

Our most important findings, however, lay in the correspondence between semantic verification latencies and the probabilities of correct text recall. Here, we found support for all four objectives outlined in the Introduction: First, with different materials, we replicated the facilitation due to word monitoring reported by Brunner and Pisoni (1982). Second, for subjects expecting questions but asked for recall, we found a corresponding enhancement of recall for high level propositions. Third, for subjects both expecting and providing final recall, we replicated that aspect of the pilot recall data indicating worse overall performance in the two monitoring conditions relative to unconstrained comprehension. Finally, for subjects expecting recall but given questions, we found uniformly longer verification latencies after the imposition of either form of subsidiary task. We conclude, then, that the facilitation in verification times reported by Brunner and Pisoni (1982) was in fact the result of some task-induced enhancement in the encoding of high- and macro-level propositions (e.g., Kintsch & van Dijk, 1978; van Dijk, 1980).

There are some notable exceptions to this picture of perfect convergence. First, we failed to replicate the enhancement of recall for lower level propositions obtained in our pilot work (Figure 1). However, both the extreme robustness and internal consistency of these fidings persuade us that they are not artifactual and that the failure to replicate the low level enhancement is a result of retrieval delay. If secondary attentional demands served to inhibit the integration of micropropositions into larger, macro level units, then some relative "enhancement" in the immediate recall of unintegrated, low level propositions from temporary memory would be a natural result. This interpretation would predict greater enhancement with increasing secondary task difficulty, which is also indicated in the pilot data. Moreover, according to any theory of reconstructive memory (e.g., Bartlett, 1932; Loftus, 1979), nonintegrated, low level details should be the first to be forgotten with increasing delay (cf., also, McKoon, 1977); this too is consistent with the present findings. Thus, we do have a reasonable account for the pilot recall data, and we prefer them as yet another important attentional interaction in the natural time course of comprehension and encoding.

We invoke a distinction between the accessibility and availability of long term memory (e.g., Tulving & Thomson, 1973) to account for the disproportionate increase in verification latencies following comprehension with word monitoring and prior recall instructions (Table 8). If texts were represented in memory as a hierarchy of lexically nondecomposed word concepts (q.v., Kintsch, 1974; Anderson, 1976), then selective attention to whole words might be expected to produce more interference with, or facilitation of, the encoding of whole word concepts than selective attention to phonemes. However, levels of recall after word monitoring were not lower than recall after phoneme monitoring, as an interpretation based on the longer verification latencies due to word monitoring might suggest. Measures of recall, of course, reflect the prevalence of facts both retrieved and reconstructed from permanent memory (e.g., Bartlett, 1932; Hasher & Griffin, 1978; Anderson & Pichert, 1978). Nonetheless, we account for this discrepancy by assuming that the frequency of reconstructions was probably constant across listening conditions and, therefore, that measures of recall would be the more accurate index of what is available in long term memory. With these assumptions, we can then attribute these truly excessive verification latencies following recall instructions and either form of monitoring to some task-induced inhibition in the development of associative retrieval paths (cf., Anderson, 1976); i.e., a task-induced reduction in the accessibility of that information.

Comprehension for the purpose of answering questions appears to carry with it a relatively low cognitive overhead. The present results indicate that, when coupled with some form of subsidiary monitoring task, there seems to be enough remaining, unassigned processing capacity for subjects to productively integrate the secondary load into their ongoing text and memory processing. With the imposition of subsequent recall, however, these dual task demands become a triple task load, the subject now having to (1) comprehend the text, (2) perform the subsidiary task and then (3) impose intentional encoding procedures on the already comprehended material. Under these circumstances, the imposition of secondary demands exceeds the available processing capacity, producing generally worse memory and verification performance. Our current interpretation is that subjects, in their timesharing between these three demands, are probably giving higher priority to intentional encoding than comprehension, thereby interfering with the assimilation of incoming propositions. One implication of Brunner and Pisoni's (1982) results was that the assumption of a fixed-capacity decision mechanism underlying the interpretation of secondary detection latencies (e.g., Foss & Swinney, 1973; Bergfeld-Mills, 1980) in phoneme monitoring research might be a false one. The present results indicate that, for comprehension with the intent of subsequent recall, the presupposition of strict capacity limitations is in fact valid. Under less stringent conditions, however, capactity limitations have not yet been established, and any interpretation of secondary monitoring data on the basis of that assumption is still premature.

At the outset of this work we had expected to find, at most, only some moderate decrements in comprehension as a result of secondary task demands. Quite clearly, what began as a minor paradigmatic investigation has produced a number of unexpected, and potentially quite important, attentional and perceptual interactions. By now, it is quite evident that subsidiary task paradigms have their own artifactual effects on otherwise normal, unconstrained comprehension

and that generalization from these kinds of results to natural language understanding must take these interactions into account.

Footnotes

1. All results presented in this paper were analyzed twice, alternately treating subjects and texts as the random variable (q.v., Clark, 1973). Throughout, Fs will represent subjects-random F ratios and Ft, text-random F ratios. Where significant, however, it is min F' which will be reported.

2. Unless otherwise stated, all p's reported in this section are less than .01.

3. Removal of the monitoring, detection rate covariance had no effect on the pattern of results obtained from analyses of the raw text recall data. Therefore, analyses of the covariance-adjusted data have been excluded from discussion in this section.

4. The propositional levels effect is significant within each individual listening condition: normal comprehension (min F'(7,354) = 4.07); word monitoring (min F'(7,359) = 5.76); phoneme monitoring (min F'(7,377) = 4.86).

Reference Notes

1. Turner, A., and Greene, E. The construction and use of a propositional text base (Tech. Rep. #63). Boulder, CO: U of Colorado, Institute for the Study of Intellectual Behavior, 1977.

2. Luce, P.A. Comprehension of fluent speech produced by rule (Progress Report #7). Bloomington, IN: Department of Psychology, Indiana University, Speech Research Laboratory, 1982.

References

Aaronson, D, & Scarborough, H. Performance theories for sentence coding: Some quantitative evidence. Journal of Experimental Psychology: Human Perception and Performance, 1976, &2, 1, 56-70.

Anderson, J.R. Language, Memory, and Thought. Hillsdale, N.J.: Erlbaum, 1976.

Anderson, R.C. & Pichert, J.W. Recall of previously unrecallable information following a shift in perspective. Journal of Verbal Learning and Verbal Behavior, 1978, 17, 1-12.

Bartlett, F.C. Remembering. London: Cambridge University Press, 1932.

Bergfeld-Mills, C. Effects of context on reaction time. Journal of Verbal Learning and Verbal Behavior, 1980, 19, 75-83.

Blank, M.A., Pisoni, D. & McClaskey, C.L. Effects of target monitoring on on understanding fluent speech. Perception & Psychophysics, 1981, 29,383-388.

Brunner, H., & Pisoni, D.B. Some effects of perceptual load on spcken text comprehension. Journal of Verbal Learning and Verbal Behavior, 1982, 21, 186-195.

Cirilo, R.K. Referential coherence and text structure in story comprehension. Journal of Verbal Learning and Verbal Behavior, 1981, 20, 358-367.

Cirilo, R.K. & Foss, D.J. Text structure and reading time for sentences. Journal of Verbal Learning and Verbal Behavior, 1980, 19, 96-109.

Clark, H.H. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 335-339.

van Dijk, T.A. Macrostructures. Hillsdale, N.J.: Erlbaum Associates, 1980.

Fletcher, C.R. Short-term memory processes in text comprehension. Journal of Verbal Learning and Verbal Behavior, 1981, 20, 564-574.

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.

Fredericksen, C. Effects of task-induced cognitive operations on comprehension and memory processes. In R. Freedle and J. Carroll, (Eds.), Language comprehension and the acquisition of knowledge, N.Y.: Wiley, 1972.

Graesser, A., Hoffman, N., & Clark, L. Structural components in reading time. Journal of Verbal Learning and Verbal Behavior, 1980, 19, 135-151.

Green, D.W. The effects of task on the representation of sentences. Journal of Verbal Learning and Verbal Behavior, 1975, 14, 275-283.

Hasher, L. & Griffin, M. Reconstructive and reproductive processes in memory. Journal of Experimental Psychology: Human Learning and Memory, 1978, 4, 318-330.

Hayes-Roth, B. & Thorndyke, P.W. The integration of knowledge from text. Journal of Verbal Learning and Verbal Behavior, 1979, 18, 91-108.

Kerlinger, F.N. & Pedhazur, E.J. Multiple Regression in Behavioral Research. New York: Holt, Rinehart & Winston, 1973.

Kintsch, W. The Representation of Meaning in Memory. Hillsdale, N.J.: Erlbaum, 1974.

Kintsch, W., Kozminsky, E., Streby, W., McKoon, G. & Keenan, J.M. Comprehension and recall of a text as a function of content variables. Journal of Verbal Learning and Verbal Behavior, 1975, 14, 196-214.

Kintsch, W. & van Dijk, T.A. Toward a model of text comprehension and production. Psychological Review, 1978, 85, 363-394.

Levelt, W.J.M. A survey of studies in sentence perception: 1970-1976. In Levelt, W.J.M. & Flores-d'Arcais, G.B. (Eds.), Studies in the Perception of Language. Wiley & Sons, New York: 1978.

Loftus, E.F. The malleability of human memory. American Scientist, 1979, 67, 3, 312-320.

Luce, P.A., Feustal, T.C. & Pisoni, D.B. Capactiy demands in short term memory for synthetic and natural speech. Human Factors, in press.

McKoon, G. Organization of information in text memory. Journal of Verbal Learning and Verbal Behavior, 1977, 16, 247-260.

McKoon, G. & Ratcliff, R. Priming in item recognition: The organization of propositions in memory for text. Journal of Verbal Learning and Verbal Behavior, 1980, 19, 369-386.

Manelis, L. & Yekovich, F.R. Repetitions of propositional arguments in sentences. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 301-312.

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

Schneider, W. & Shiffrin, R.M. Controlled and automatic human information processing: I. Detection, search, and attention. Psychological Review, 1977, 1, 1-66.

Shiffrin, R.M. & Schneider, W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological Review, 1977, 2, 127-190.

Spiro, R.J. Accomodative reconstruction in prose recall. Journal of Verbal Learning and Verbal Behavior, 1980, 19, 84-95.

Tulving, E. & Thomson, T. Encoding specificity and retrieval processes in episodic memory. Psychological Review, 1973, 80, 5, 352-373.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

Cognitive Processes and Comprehension Measures in Silent and Oral Reading*

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

*This research was supported by NIMH research grant MH-24027 to Indiana University in Bloomington. I am grateful to David B. Pisoni for insightful comments on an earlier draft, to Jerome C. Harste for helpful discussions, and to Hans Brunner for access to the experimental materials and programs.

Abstract

On-line reading rates and post-reading comprehension data in the form of error rates, response latencies and confidence ratings of questions probing recognition of various levels of text structure were collected for passages read orally and silently by 16 college students. Oral reading rates were slower than silent reading rates. The four types of comprehension items showed a levels effect for response latencies: More encompassing or higher-level representations were verified more slowly than surface structure or low-level text representations. Differences due to reading mode were found only for low- and high-level proposition verification latencies: Prior silent reading of the text led to slower verification responses. The error data and response confidence ratings failed to show reading mode or levels effects. The present results suggest the critical differences in processing between oral and silent reading involve the relative rates of the temporal course of ongoing higher-level comprehension processes. Slower comprehension microstructure construction processes in oral reading create memory traces that are accessed faster during memory-based comprehension tasks than traces established by similar but faster processes that occur during silent reading.

Introduction

To both the observer and the reader, the acts of oral and silent reading have obvious differences. The former includes an immediate vocalization response and the latter does not (necessarily). In order to appreciate the similarities between these ways of reading it must be assumed the the reader's primary goal in both reading modes is to understand the written text. Logically, perceptual and cognitive processes mediate between the printed page and the reader's resultant state of knowledge abstracted from the printed page, i.e. comprehension. Accepting this fundamental assumption about the relation of reading and comprehension, then, our interest turns to the level at which processing differences between oral and silent reading might occur.

The present study addresses the issue of the extent to which common cognitive processes underlie oral and silent reading. Previous research using both on-line measurements of reading behavior and performance on subsequent memory and comprehension tests of the read material have yielded contradictory results and implications for resolving this issue. The locus of observable differences in eye movements (e.g. Anderson & Swanson, 1937; Fairbanks, 1937; Wanat, 1976) and reading rates (e.g. Juel & Holmes, 1981; Mead, 1915, 1917; Rogers, 1937) has not been agreed upon. Oral reading errors, also known as "miscues", (Danks & Hill, 1981; Goodman, 1969; Levy, 1981; Weber, 1968) are another online source of information about oral reading processes. One problem with error evidence, however, is that no analogous measurement of silent reading processes is available because of the covert nature of normal reading in mature subjects. Errors of omission or incorrect substitutions in silent reading can only indirectly be inferred from regressive eye movements (Rayner & McConkie, 1976). Oral reading miscue analysis (Goodman & Burke, 1978) is the foundation of one widely accepted theory of reading (Goodman, 1967, 1970, 1979). This particular account of reading assumes that oral reading miscue data may be used to make inferences about silent reading processes, implicitly suggesting a unitary cognitive base for processes involved in both reading modes. However, Goodman (1970, p.482) also writes that "When silent reading becomes proficient, it becomes a very different process from oral reading." Goodman suggests that silent reading involves more sampling and prediction of print than oral reading. Thus, in order to evaluate this ambiguous position on an empirical basis, clear-cut sources of evidence for the assumed common cognitive core of oral and silent reading processes must be obtained. More realistically, such commonality may be characterized as a complex interactive system with multiple specific component processes that contribute to reading comprehension (e.g. Just & Carpenter, 1980; Levy, 1981; Rumelhart, 1977; and others). As part of this endeavor currently in progress (see e.g. Perfetti & Lesgold, 1981), component cognitive processes are proposed, suitable measurement techniques are identified and process contributions and interactions have been tested in a variety of conditions.

Historically, two kinds of evidence have been held to support assumptions of predominantly shared cognitive processes in both oral and silent reading modes. First, a sizeable number of studies have failed to find comprehension differences after oral and silent reading (e.g. Anderson & Dearborn, 1951; Anderson & Swanson, 1937; Gray, 1958; Jones, 1932; Juel & Holmes, 1981; Poulton & Brown, 1967; Rogers, 1937; and others). Second, much attention has been given to phonological recoding, one process implicated in both modes of reading. Most

studies manipulating phonological recoding factors, however, have been limited to reading of isolated words (e.g. Baron, 1979; Glushko, 1981; Kleiman, 1975; Rubenstein, Lewis & Rubenstein, 1971; Spoehr & Smith, 1975) and have not directly addressed comparisons of silent and oral reading processes that involve measures of comprehension.

In addition, these two areas of research are still surrounded by controversy, methodological problems and multiple intrerpretations in terms of reading mode differences. The body of literature on silent and oral reading comprehension is inconclusive. As many of the early (and more recent) classroom studies of oral and silent reading comprehension have reported advantages to oral reading (Collins, 1961; Elgart, 1975; Swalm, 1973) or to silent reading (e.g. Mead, 1915, 1917), as were cited for claims of null findings above. With regard to phonological recoding, vocalization suppression studies of reading (Hardyck & Petrinovich, 1970; Levy, 1975, 1977, 1981) have shown interactions between ongoing articulating responses and properties of the text structure itself. Readability interactions with reading mode (e.g. Coke, 1974) and other linguistic structural measures (e.g. Wanat, 1976) reveal similar complexities due to context in interpreting the effects of oral and silent reading on comprehension. Thus, neither extremist view of the relation between oral and silent reading, i.e. completely identical cognitive comprehension processes or entirely disjunctive processes, is feasible. To date, the reading mode literature has not weighed greatly on the research directions necessary to discover the subtleties of the human psycholinguistic processing system as it behaves in silent and oral reading.

One methodological problem with previous studies is the often loose, global definition of comprehension (if one is attempted at all). Inevitably, inappropriate or insensitive measurement techniques follow from this type of conceptual shortcoming. Another problem with earlier comprehension measurement tools has been a wide range of production components in the required response, varying from multiple-choice forced-choice questions of low written demands [Collins, 1961; Elgart's use (1975) of Gates-MacGinitie tests (1964)], to overt spoken responses in oral Cloze procedures used by Swalm (1973). Such problems in measuring comprehension of read texts have been affected by developments in the field of text and discourse processing.

Recently, a number of detailed models of knowledge structure and comprehension have been proposed (e.g. ACT, 1976; de Beaugrande, 1980; Kintsch & van Dijk, 1978; Minsky, 1975, 1979; Schank, 1975, 1981). These advances have allowed for precise characterization of the internal structure of texts used for reading passages and meaning structures derived from them in terms of hierarchical propositional structures and their interrelations and associatuve links with the central theme of the text (e.g. Kintsch & van Dijk, 1978).

By carrying out a propositional analysis of texts based on such models (e.g. Turner & Greene, 1977), levels of knowledge structures can be identified in relation to the central theme of the text. These models suggest that shallower levels of processing (Craik & Lockhart, 1972) are given to meaning units of lower levels, i.e. closer to the surface structure of the words and sentences in the text than to meaning units at higher or more encompassing levels of text microstructure and text macrostructure (Kintsch, 1974, 1977). Higher structural levels require more processing and have more enduring memory traces (Kintsch,

1977). In support of this notion, robust microstructure levels effects have been found in reading time and recall measures (Cirillo & Foss, 1980; Kintsch & Keenan, 1973; Kintsch, Kozminsky, Streby, McKoon & Keenan, 1975). These studies reveal that more time is spent reading higher level propositions, which are also recalled with greater probability after (silent) reading. From traditional memory phenomena relating recall and recognition (e.g. Tulving, 1975) follow the predictions that higher-level statements from the text would be recognized with less accuracy and longer response latencies as well as with less confidence than lower-level knowledge structures.

Finally, comprehending a text also entails the ability to derive inferences consistent with the meaning macrostructures of the text (Frederiksen, 1981; van Dijk, 1979). In sum, recent advances in text analysis and comprehension modelling have enabled more accurate measurement of various aspects of reading comprehension.1 Thus, it may be possible to resolve some of the differences in the results of studies of comprehension in oral and silent reading.

The present study addresses a question which came to light as a consequence of the acceptance of the techniques of propositional analysis in studying text comprehension. Namely, do oral and silent reading processes differentially affect the ensuing surface level, low- and high-level propositional and inferential representations constructed during reading? If the ongoing articulating response of the oral reader competes with simultaneous comprehension processes for resources (Goodman, 1970; Levy, 1981; Wanat, 1976), higher levels of comprehension may be predicted to be completed less rapidly and somewhat less efficiently in the course of oral reading compared to silent reading. An alternative hypothesis may be phrased in terms of the temporal course of processing: Slower progress in oral reading compared to silent reading may be related to (by being a partial causative factor or a consequence of) compensatory processes in the development of more abstract knowledge structures involved in complete understanding of a text. These questions relate to issues of automaticity of the pronunciation response in oral reading (Danks & Hill, 1981) and the effect of speed of reading on ensuing comprehension processes (Stanovich, 1981), which will be taken up later on in the discussion section below. To test these two hypotheses, a number of different dependent variables were used -- one to confirm the temporal relation of oral and silent reading rates in real time, and the other three being measures of comprehension at each of four different levels of propositional complexity of text meaning.

Method

Subjects. Sixteen Indiana University students were tested individually. They received credit as partial fulfillment of the requirements for an introductory psychology course.

Materials. Twelve test passages and three practice passages between 150 and 300 words in length, previously used in studies of listening comprehension (Blank, Pisoni & McCluskey, 1981; Brunner & Pisoni, 1982), were chosen as expository reading texts. For each text, four verification statements had been constructed to evaluate various levels of comprehension. An example passage and its affirmative questions are shown in Table 1.

Insert Table 1 about here

Remember questions tested surface level memory of the text, by requiring recognition responses to test words as having occurred in the reading passage or not. The questions were of the standard form, "Did the word "XXXXX" occur in this story?" Either correct words or distractor words synonymous to words which had occurred in the test passage occurred in the target position.

The propositional representation constructed during reading and comprehension of a text was examined on two levels: Verification statements consisting of high-level and low-level propositions (Kintsch et al., 1975) were used to test recognition memory for text microstructure and inferential statements were presented to examine macrostructure representations formed during reading and comprehension (Kintsch and van Dijk, 1978). These three question types addressed more global levels of comprehension (than surface level questions) and occurred in two versions, each of which required either a "Yes" or a "No" response from the reader. Low-level propositions presented for recognition were either exact repetitions of one-clause sentences from the test passage or had one substituted word that rendered the statement incongruous with the content of the text. High-level proposition test items similarly repeated or misrepresented a clause more central to the passage theme, but one which did not necessarily occur entirely within a single sentence. True inferences required subjects to synthesize information explicitly conveyed in the passage; false inferences were contradictory to such syntheses.

Design. Reading mode and levels of comprehension questions were within-subjects variables. Reading mode was blocked within each subject, so that each subject read six stories aloud and six silently. The order of silent and oral response blocks was counterbalanced across subjects. Two counterbalanced orders of the twelve test passages were used each with 8 subjects. Within each of these two subject groups, "Yes" responses were made on the left-hand side of the response box by half of the subjects and on the right-hand side by the other half. In addition, correct and incorrect question versions were counterbalanced among each group of four subjects. The order of the four comprehension questions, one at each level, was also randomly determined for every passage presentation.

Procedure. Subjects were tested individually in a sound-attenuated room in the presence of the experimenter. The instructions, prompts during the course of the experiment, the reading passages and the comprehension questions were all presented on a GBC Standard CRT display monitor (Model MV-10A) placed at eye-level about 40 cm in front of the subject. Stimulus presentation and response collection was carried out by a PDP-11/34 computer. The subject interacted with the experimental visual prompts by pressing appropriate buttons on a seven-button response box connected to the computer.

The beginning of each reading passage was announced on the center of the CRT screen by a prompt, "Attention! New Story Coming Up. Please press READY button to begin." When the subject had indicated his/her readiness, the experimenter

Table 1

An Example Reading Passage

In ancient Rome, Julius Caesar banned chariot driving at night. It seems the thundering chariot wheels made too much noise. Now -- over 2,000 years later -- people are starting to realize that noise isn't good for them. It affects their hearing, their peace of mind, their ability to work efficiently, and, as some doctors point out, their general health.

Most people still accept noise as a routine part of their daily lives: sirens, horns, airplanes, household appliances, power mowers, jackhammers. Some even seek out noise in the form of loud rock music. People can see a smog-filled sky or a filthy lake and they recognize pollution, but noise is not usually regarded with equal concern.

Noise 'is' a form of pollution and, like other forms, it's getting worse. A U.S. government study says that noise pollution is doubling every ten years. Says Dr. Vern O. Knudsen, a noise expert at the University of California: "If noise continues to increase for the next 30 years as it has for the past 30, it could become lethal."

Questions

Surface structure: Did the word "doctors" occur in this story?

Low-level proposition: Noise pollution affects one's work.

High-level proposition: People are starting to realize the harmful effects of noise.

Inference: Modern technology has contributed to the rise in noise pollution.

announced the reading mode (oral or silent) for the passage. The passage appeared on the CRT screen one sentence at a time and was advanced by the reader's button-press control.

Subjects were instructed to read once through and press the "Ready" button in order to continue reading as fluently as possible. The presentation technique prevented regressive eye movements to previous sentences and also allowed for the collection of sentence-by-sentence reading latencies for each test passage. At the end of each passage, the question phase was announced on the screen by the centered prompt, "Attention! Questions. Please press READY button to begin." Subjects initiated each question presentation themselves and were instructed to respond as quickly and accurately as possible once the question had appeared on the screen. After making their "Yes/True" or "No/False" responses to each question, subjects entered a confidence rating of their response on a scale from 1 to 7. A rating of 7 indicated a highly confident response and a rating of 1 indicated a guessing response whose correctness was highly uncertain. A rating scale reminder on the screen at this time was terminated by the confidence rating button-press response. Then, feedback about the correct answer to the comprehension question was provided in the form of a flashing light immediately above the correct button on the response box. Following the four questions, the next reading passage was announced. Thus, for each passage, sentence-by-sentence reading latencies and comprehension data in the form of the number of errors, question-answering latencies and confidence ratings for each question type were collected.

Three passages served as practice for all subjects, the first being read silently and the second and third passages serving as oral reading practice at the beginning of the experimental session. Instructions emphasized that the subject's task was to read in order to understand the content of the passages. The oral fluency and intonation of reading aloud were not stressed in the instructions and the experimenter attempted to avoid any performance pressure during the oral reading blocks, by reassuring worried subjects that their responses were quite satisfactory.

Results

Unless otherwise noted, all reported results are statistically significant at the $p<.01$ level.

Reading Latencies. A two-way analysis of variance with reading mode and stories as fixed factors was performed on subjects' mean sentence-by-sentence reading times for each passage. Subjects took a mean of 1.76 seconds longer to read passages aloud (mean = 8.71 sec) than to read them silently (mean = 6.95 sec), ($F(1,15)=22.06$). This difference was found for all 16 subjects. There were differences in the mean reading latencies between the passages ($F(5,65)=10.28$) reflecting differences in sentence lengths, but these differences did not interact with the reading mode effect ($F(5,65)<1.0$). Thus, as expected, oral reading latencies were greater than silent reading latencies for all test passages.

Comprehension. For initial overall analyses of variance, the four question types were treated as one factor called Question Level. This factor was related

to increasing representation abstraction from the text presented on the CRT screen. For each subject, the total number of errors for each question level for both oral and silent reading (of total possible of six) were summed together. In addition, the mean latency and confidence rating for each question level over all reading passages for each reading mode condition were computed. Separate analyses of variance with the order of reading mode blocks as a between-subject factor and question level and reading mode as fixed, within-subject factors were also carried out on the number of errors, mean latencies and mean confidence ratings.

Error data. No effect of order of reading mode was found, so this factor was not included in further analyses. An expected overall question level effect indicated differential error patterns for the various comprehension level probes ($F(3,45)=8.18$). However, no differences in the number of errors made in oral and silent reading conditions were found ($F(1,15)=3.54$, $p>.07$). In addition, no interaction was found between the level of comprehension probes and the reading mode ($F(3,45)=2.22$, $p>.09$).

Closer examination of the error data at each question level by test for simple effects revealed no differences in performance after oral or silent reading for questions tapping surface structure ($F(1,15)<1.0$), low- or high-level propositional representations in memory ($F(1,15)=3.20, p>.09$), or inferences drawn from the read texts ($F(1,15)<1.0$). The analysis of propositional microstructure did, nevertheless, yield the expected levels effect ($F(1,15)=8.04, p<.02$) and no interaction between question level and reading mode ($F(1,15)<1.0$). Thus, whether a passage was read aloud or silently did not significantly affect the number of errors made on any level of comprehension as measured by the four question types after each passage. Importantly, the various question levels were not affected differentially by the preceding reading mode.

Question Latency data. Since the initial analysis of variance did not reveal practice or fatigue effects due to the order of reading mode blocks ($F(1,14)=3.04, p>1.0$), this factor was excluded from further consideration. Main effects of both reading mode ($F(1,15)=11.52$) and question level ($F(3,45)=32.13$) were obtained in the subsequent analysis with the four question types as one factor. Furthermore, for question latencies, reading mode and question level interacted significantly ($F(3,45)=4.85$).

No effects due to silent or oral reading mode were found for analyses of the latencies to questions probing the surface structure of the texts and inferences following from those texts ($F(1,15)=2.02, p>.17$ and $F(1,15)<1.0$ respectively). In contrast, oral reading led to faster verification of propositions reflecting the microstructure knowledge representation of the texts ($F(1,15)=27.95$). The failure to obtain a difference between latencies to low- and high-level propositions was unexpected ($F(1,15)=2.96, p>.10$). This result differs from the question error data above and indicates the increased sensitivity of the latency measure compared to error rate measures of comprehension, as well as the speed of processing difference between silent and oral reading. The comprehension question latency and error data are shown in Figure 1.

Insert Figure 1 about here

Confidence Rating data. Contrary to expectations from use in listening comprehension tasks (e.g. Brunner & Pisoni, 1982), in the present reading comprehension study, confidence ratings failed to meaningfully reflect differences in silent and oral reading processes in subsequent comprehension questions. An overall analysis of variance revealed an effect of question level, indicating differences between the four types of comprehension level probes ($F(3,45)=9.20$), but no differences were observed due to reading mode ($F(1,15)<1.0$). The mean confidence rating was 5.70, where a rating of 7 indicated a very confident response and a rating of 1 indicated a guessing response. Subsequent analyses indicated no significant question levels effect between low- and high-level microstructure proposition items. Thus, in the present study, subjects' confidence in their prior recognition or verification responses to information from texts they had just read appeared to be minimally related to underlying cognitive processes used in reading comprehension.

In summary, differences in the component cognitive processes between oral and silent reading modes in a task requiring comprehension were revealed in terms of speed of processing. These effects were observed for both reading rates and latencies to comprehension probes. Microstructure representations of the text base revealed the obtained differences due to reading mode: Oral reading led to somewhat faster comprehension responses, but these responses were as accurate as the slower responses made after silent reading of the texts. Comprehension questions were answered equally confidently after both oral and silent reading. Thus, our findings have implicated temporal processing factors in the construction and retrieval of comprehension structures in memory as the loci of differential cognitive processing in oral and silent reading.

Discussion

The present results identified specific levels of knowledge structures used in text comprehension which were affected by prior reading mode. When the text information had been read aloud, subjects correctly recognized low- and high-level propositional statements as having occurred in the text faster than when they had read the text silently. This effect was not reflected in the data from errors or confidence ratings of subjects' responses to comprehension questions. In contrast, however, the reading mode did affect reading rate. The observed on-line temporal differences were in the opposite direction to the question latency data: Mean oral sentence-by-sentence reading times were 1.76 seconds longer than equivalent silent reading times. Thus, only the two temporal dependent measures in the present study, reading rates and response latencies to comprehension questions, were sensitive to the effects of reading mode. Further, only questions pertaining to text microstructure revealed differences due to oral or silent reading of the preceding text.

In the sections below, we will consider these findings in light of our two experimental hypotheses. Then, their implications for models of silent and oral reading comprehension will be discussed.

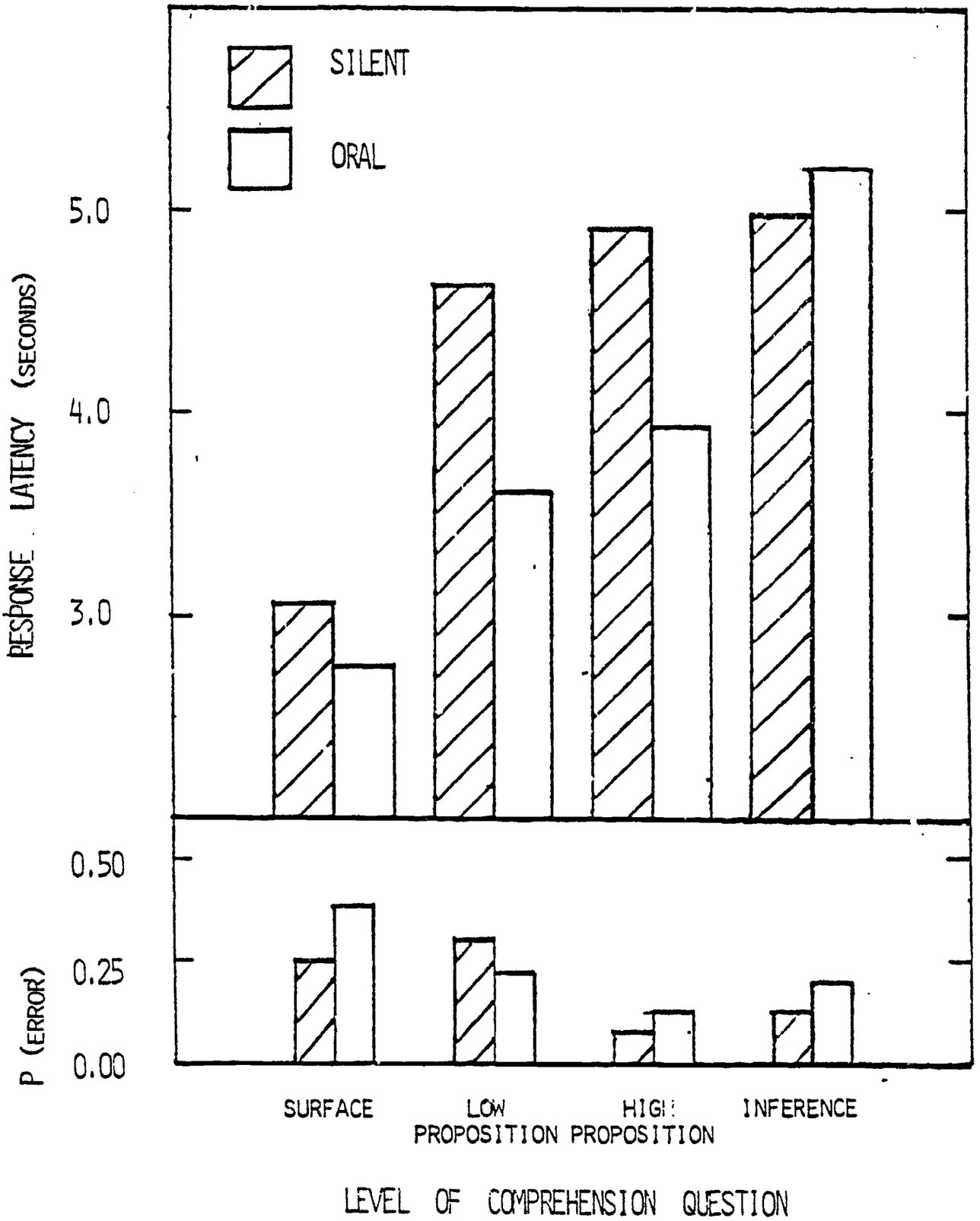


Figure 1. Question latency data (top panel) and error data (bottom panel) after silent reading (striped bars) and oral reading (open bars) at four levels of comprehension.

According to our first hypothesis, the ongoing vocalization response in oral reading requires attentional capacity that is shared with other underlying cognitive processes involved in comprehension (Goodman, 1970; Wanat, 1976). Predictions for our experiment were that slower and less accurate performance was expected in the oral reading condition. Also, it was expected that higher levels of comprehension would suffer more than lower levels in the oral reading mode. In fact, the present results contradict both of these predictions: Faster comprehension latencies occurred after oral reading than after silent reading (for low- and high-level propositions). This finding may be taken to support views that adult readers have automatic pronunciation responses in oral reading (Danks & Hill, 1981).

The second hypothesis we considered was that the slower speed of oral reading would be accompanied by compensatory decrements in comprehension performance. Stanovich (1981), among others, has suggested that slower word recognition in reading cooccurs with increased reliance on contextual cues in the text (presumably used for integrating incoming material with other comprehension structures in working memory). The prediction follows, therefore, for the present study that differences in on-line temporal rates of oral and silent reading would not yield comparable differences in comprehension data. This, again, was not confirmed by our data. We found response latency advantages for verification of low- and high-level propositions from the passage after oral reading.

This finding of faster response latencies following oral reading supports previous reports of spontaneous vocalization during (silent) reading of difficult texts (e.g. Hardyck & Petrinovich, 1970). The suggestion, here, is that when comprehension is difficult, the perceptual and cognitive processes in reading are slowed down (Levy, 1981).2 The additional inference required is that the vocalization of the oral reading response is functional in compensating for otherwise faster-occurring encoding of text microstructure for comprehension. The mean error rate of 21%, combined with subjects' comments about the difficulty of the experimental task, support our claim that the set of passages used were indeed difficult reading materials for our group of subjects to comprehend.

Thus, the present study has shown that the vocalization component of oral reading functions to compensate for faster processing of text microstructure. (In contrast, memory for the occurrence of particular words in the texts was unaffected by reading mode, both in accuracy and in speed of response).

Given these findings, we propose as the locus of this compensation a post-lexical working memory store for parsing and clause integration stages of text comprehension (e.g. Just & Carpenter, 1980). We suggest that phonological recoding for access to the mental lexicon in order to recognize printed words takes place in both reading modes (although we have not shown this to be the case in this study). This proposal will be discussed in relation to two specific accounts of oral and silent reading.

Because Goodman's model (1970, 1979) rests almost entirely on oral reading error data, he has attempted to specify the relation of processes in both reading modes in some detail. We have already alluded to one point of confusion, but at this time Goodman's perspective should be examined with greater depth. For most adult oral readers Goodman (1970) suggests that "... primarily, oral output is produced after meaning has been decoded." (p.483) Thus, oral reading requires

both the decoding of meaning component, proposed to be identical with silent reading by Goodman, and, then, a derivative recoding process, interpretable as phonological recoding, "to produce an oral language equivalent of the graphic input which is the signal in reading" (p.502).

The underlying cognitive processes leading to meaning reconstruction, i.e. comprehension, are presumed to be the same processes in both silent and oral reading, according to Goodman (1967). This appears to be inconsistent with his assertion in other places of active sampling, prediction and processing speed differences between the two reading modes (cf. 1967, p.502). Thus, while Goodman indicates awareness of differences in the relative temporal rates of processing, he does not associate them with reading comprehension processes and therefore makes predictions of no differences in performance between comprehension following oral and silent reading.

A more flexible account of differences in oral and silent reading has been proposed by Danks and Fears (1979). In complete opposition to Goodman's model, the two alternative models of the oral reading processes postulated by Danks and Fears, called the decoding and comprehension hypotheses, both necessarily include phonological recoding or as they name it, decoding (not to be confused with Goodman's similar mechanism referred to as recoding). The hypotheses of Danks and Fears differ in the inclusion of comprehension prior to the spoken oral reading response. Which of the two models is used depends on variables such as the reading skill and motivation of the reader, the specific task and the text difficulty. The essential point for the present study is that phonological recoding, according to these authors, always precedes comprehension in both silent and either model of oral reading. Our findings lead us to concur with this suggestion, but it is necessary to expand on the specific stages involved between phonological recoding and comprehension.

The results of our study, in conjunction with the assumption of common phonological recoding processes in both oral and silent reading, suggest that the temporal differences due to reading mode in the process of reading and in response latencies to comprehension questions to the reading passages occur after words have been recognized, in case role assignment and text unit and clausal integration processes that take place in working memory. This is similar to Just and Carpenter's recent model of silent reading (1980), but our proposal also assumes that the additional time spent in working memory during completion of these higher-level comprehension processes in oral reading (compared to silent reading), results in memory traces that are retrieved faster in later, memory-based comprehension probes of these higher-level units of the text meaning. Word-level differences in memory-based questions were not found in the present study, we suggest, because separate traces of word entries already existing in the mental lexicon were not constructed during text comprehension during initial reading. Similarly, inferences may not be automatically drawn during on-line reading comprehension and may be instead computed from stored information upon demand (Frederiksen, 1981). Thus, what readers actually construct and store as they read for comprehension appears to be propositional structures of the levels specifically probed by our microstructure verification statements (cf. Kintsch & van Dijk, 1978). For this reason, additional working memory storage during the slower clausal integration processes of oral reading, facilitates later verification latencies for only low- and high-level microstructures from the text.

In conclusion, the present data reveal that differences due to reading mode are primarily a function of the speed of higher-level parsing and comprehension processes that occur when subjects read texts for comprehension. In our study, slower reading latencies in oral reading led to faster responses to comprehension items on low- and high-level propositional structures from the text. These results support models of oral and silent reading with common early stages of phonological recoding (e.g. Danks & Fears, 1979). Such models obviously require further specificity in identifying post-lexical larger text unit storage as the critical level on which time differences in oral and silent reading will be reflected in memory-based comprehension tests. The results obtained in the present study also point to the importance of two theoretical endeavors: first, extending unitary processing models of reading comprehension to account for temporal differences due to reading mode; and second, viewing comprehension as a structured hierarchy of component levels of meaning and structure rather than a global, unitary process that somehow reveals itself after a reader encounters a printed text. In addition, it is very likely that the reader's comprehension goals will interact in substantial ways with the time-course of processing of the individual component levels of comprehension. Having found reliable differences in sentence-by-sentence reading speed and response latencies for comprehension questions probing text microstructure components between oral and silent reading, we are encouraged to pursue this problem further in future work that explicitly examines the reader's goals.

Footnotes

[1]In contrast to wide employment of nonspecific or low-level comprehension tests, testing primarily memory for text vocabulary (e.g. Swalm, 1973) or ambiguous levels of comprehended knowledge such as "information in the story" (Collins, 1961), one early researcher, Mead (1915, 1917), adopted a measure reported as "percentage of points reproduced of points read". This is seen as a precursor, albeit nonspecific, of current propositional text base analyses as a unit to measure comprehension.

[2]Other factors determining reading rate and the extent of the use of compensating overt reading responses include visually noisy stimulus quality of the text (Levy, 1981), the word-encoding skill of the reader (Perfetti & Roth, 1981), as well as the text structure and the reader's goals (Frederiksen, 1981).

References

Anderson, I.H. & Dearborn, W.F. The Psychology of Teaching Reading. New York: Ronald Press, 1952.

Anderson, I.H. & Swanson, D.E. Common factors in eye-movements in silent and oral reading. Psychological Monographs, 1937, 48, 3, 61-69.

Anderson, J.R. Language, Memory, and Thought. Hillsdale, NJ: Erlbaum, 1976.

Baron, J. Mechanisms for pronouncing printed words: Use and acquisition. In D. LaBerge & S. Samuels (Eds.), Basic processes in reading: Perception and comprehension. Hillsdale, NJ: Erlbaum, 1977.

Blank, M.A., Pisoni, D.B. & McClaskey, C.L. Effects of target monitoring on understanding fluent speech. Perception & Psychophysics, 1981,29, 383-388.

Brunner, H. & Pisoni, D. B. Some effects of perceptual load on spoken text comprehension. Journal of Verbal Learning and Verbal Behavior, 1982, 21, 186-195.

Cirillo, R.K. & Foss, D.J. Text structure and reading time for sentences. Journal of Verbal Learning and Verbal Behavior, 1980, 19, 96-109.

Coke, E.U. The effects of readability on oral and silent reading rates. Journal of Educational Psychology, 1974, 66, 406-409.

Collins, R. The comprehension of prose materials by college freshmen when read silently and when read aloud. Journal of Educational Research, 1961, 55, 79-82.

Craik, F.I.M. & Lockhart, R.S. Levels of processing: A framework for memory research. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 671-684.

Danks, J.H. & Fears, R. Oral reading: Does it reflect decoding or comprehension? In L. Resnick & P. Weaver (Eds.), Theory and practice of early reading. Vol. 3. Hillsdale, NJ: Erlbaum, 1979.

Danks, J.H. & Hill, G.O. An interactive analysis of Oral Reading. In A.M. Lesgold & C.A. Perfetti (Eds.), Interactive processes in reading. Hillsdale, NJ: Erlbaum, 1981.

de Beaugrande, R. Text, discourse, and process: Toward a multidisciplinary science of texts. Norwood, NJ: Ablex, 1980.

Elgart, D.B. Oral reading, silent reading, and listening comprehension: a comparative study. Journal of Reading Behavior, 1975, 10, 203-207.

Fairbanks, G. The relation betwen eye-movements and voice in the oral reading of good and poor silent readers. Psychological Monographs, 1937, 48, 3, 78-101.

Frederiksen, J.R. Sources of process interactions in reading. In A.M. Lesgold & C.A. Perfetti (Eds.), Interactive processes in reading. Hillsdale, NJ: Erlbaum, 1981.

Glushko, R.J. Principles for pronouncing print: The psychology of phonography. In A.M. Lesgold & C.A. Perfetti (Eds.), Interactive processes in reading. Hillsdale, NJ: Erlbaum, 1981.

Goodman, K.S. Analysis of oral reading miscues: Applied psycholinguistics. Reading Research Quarterly, 1969, 5, 9-30.

Goodman, K.S. Reading: A psycholinguistic guessing game. (1967) In H. Singer & R. Ruddell (Eds.) Theoretical models and processes of reading. Newark, Del: IRA, 1970.

Goodman, K.S. Behind the eye: What happens in reading. In H. Singer & R. Ruddell (Eds.) Theoretical models and processes of reading. Newark, Del: IRA, 1970.

Goodman, K.S. & Burke, C. Theoretically based studies of patterns of miscues in oral reading performance. Washington, D.C.: U.S. Department of Health, Education & Welfare, 1973.

Gray, W.S. The teaching of reading and writing. Chicago: UNESCO, Scott, Foresman, 1958.

Hardyck, C.D. & Petrinovich, L.F. Subvocal speech and comprehension level as a function of the difficulty of reading material. Journal of Verbal Learning and Verbal Behavior, 1970, 9, 647-652.

Jones, E.E. A comparison of comprehension results in oral and silent reading, Peabody Journal of Education, 1932, 9, 292-296.

Juel, C. & Holmes, B. Oral and silent reading of sentences. Reading Research Quarterly, 1981, 4, 545-568.

Just, M.A. & Carpenter, P.A. A theory of reading: From eye fixations to comprehension. Psychological Review, 1980, 87, 329-354.

Kintsch, W. The representation of meaning in memory. Hillsdale, NJ: Erlbaum, 1974.

Kintsch, W. Memory and Cognition. Hillsdale, NJ: Erlbaum, 1977.

Kintsch, W. & Keenan, J.M. Reading rate as a function of the number of propositions in the base structure of sentences. Cognitive Psychology, 1973, 5, 257-274.

Kintsch, W., Kozminsky, E., Streby, W., McKoon, G. & Keenan, J.M. Comprehension and recall of a text as a function of content variables. Journal of Verbal Learning and Verbal Behavior, 1975, 14, 196-214.

Kintsch, W. & van Dijk, T.A. Toward a model of text comprehension and production. Psychological Review,1978, 85, 363-394.

Kleiman, G.M. Speech recoding in reading. Journal of Verbal Learning and Verbal Behavior, 1975, 14, 323-339.

Lesgold, A.M. & Perfetti, C.A. (Eds), Interactive processes in reading. Hillsdale, NJ: Erlbaum, 1981.

Levin, H. Reading silently and aloud. In A. Pick (Ed.), Perception and its development. Hillsdale, NJ: Erlbaum, 1979.

Levy, B.A. Interactive processing during reading. In A.M. Lesgold & C.A. Perfetti (Eds), Interactive processes in reading. Hillsdale, NJ: Erlbaum, 1981.

Mayer, R.E. & Cook, L.K. Effects of shadowing on prose comprehension and problem solving. Memory & Cognition, 1981, 9, 101-109.

Mead, C.D. Silent versus oral reading with one hundred sixth-grade children. Journal of Educational Psychology, 1915, 6, 345-348.

Mead, C.D. Results in silent versus oral reading. Journal of Educational Psychology, 1917, 8, 367-368.

Minsky, M. A framework for representing knowledge. In P. Winston (Ed.), The psychology of computer vision. New York: McGraw-Hill, 1975.

Pintner, R. Oral and silent reading of fourth-grade pupils. Journal of Educational Psychology, 1913, 4, 333-337.

Poulton, E.C. & Brown, C.H. Memory after reading aloud and reading silently. British Journal of Psychology, 1967, 58, 219-222.

Rayner, K. & McConkie, G.W. What guides a reader's eye movements? Vision Research, 1976, 16, 829-837.

Rogers, M.V. Comprehension in oral and silent reading. Journal of General Psychology, 1937, 17, 394-397.

Rowell, E.H. Do elementary students read better orally or silently? Reading Teacher, 1976, 367-370.

Rubenstein, H., Lewis, S. & Rubenstein, M. Evidence for phonemic recoding in visual word recognition. Journal of Verbal Learning and Verbal Behavior, 1971, 10, 635-647.

Rumelhart, D.E. Toward and interactive model of reading. In S.Dornic (Ed.), Attention and performance. Vol. VI. Hillsdale, NJ: Erlbaum, 1977.

Schank, R. Conceptual dependency theory. In R. Schank, N. Goldman, C. Rieger & C. Riesbeck (Eds.), Theoretical issues in natural language processing: An interdisciplinary workshop. Cambridge: Bolt, Beranek & Newman, 1975.

Spoehr, K. & Smith, E. The role of orthographic and phonotactic rules in perceiving letter patterns. Journal of Experimental Psychology: Human Perception and Performance, 1975, 1, 21-34.

Stanovich, K. Attentional and automatic context effects in reading. In A.M. Lesgold & C.A. Perfetti (Eds.), Interactive processes in reading. Hillsdale, NJ: Erlbaum, 1981.

Swalm, J.E. A comparison of oral reading, silent reading, and listening comprehension. Education, 1973, 92, 111-115.

Swanson, D.E. Common elements in silent and oral reading. Psychological Monographs, 1937, 92, 111-115.

Tulving, E. Ecphoric processes in recall and recognition. In J. Brown (ed.), Recall and recognition. London: Wiley, 1975.

Turner, A. & Greene, E. The construction and use of a propositional text base. Technical Report No. 63, Institute for the Study of Intellectual Behavior, University of Colorado, Boulder, Colorado, 1977.

van Dijk, T.A. Macro-structures. Hillsdale, NJ: Erlbaum, 1979.

Weber, R.M. The study of oral reading errors: A survey of the literature. Reading Research Quarterly, 1968, 4, 96-119.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

Perceptual and Cognitive Constraints on the Use of Voice Response Systems*

Howard C. Nusbaum and David B. Pisoni

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

*This work was supported by NIH research grant NS-12179 to Indiana University in Bloomington. We thank Tom Carrell, Paul Luce and Eileen Schwab for interest and advice.

Abstract

With the increasing proliferation of voice response and voice data entry systems, it has become extremely important to understand how humans interact with these devices. Speech perception is a complex process involving a number of stages of analysis beginning with the extraction of phonetic information and ending with a conceptual representation of the input message. Each level of perceptual analysis may be subject to limitations of the human information processing system. Moreover, these perceptual processes may impose further restrictions on the performance of other cognitive tasks. Perceptual encoding, maintenance of items in short-term memory, transfer of information to long-term storage, decision making, and response production may compete for mechanisms or resources needed for visual or auditory information processing, vocal responding, or manual responding. Thus, when an observer must make complex decisions and responses to synthetic speech messages, performance may depend on the interaction of perceptual clarity of the input, nature of the responses, and other task parameters. This paper will discuss how the relationship of such task parameters and cognitive limitations may affect task performance in situations employing voice I/O using synthetic speech.

INTRODUCTION

The era of speech technology has begun. We are just now starting to see the introduction of practical, commercially available speech synthesis and speech recognition devices. Within the next few years, these systems will be utilized for a variety of applications to facilitate human-machine communication and as sensory aids for the handicapped. Soon we will be conversing with vending machines, cash registers, elevators, cars, clocks, and computers. Pilots will be able to request and receive information by talking and listening to flight instruments. In short, speech technology will provide the ability to interact rapidly with machines through our most efficient communications channel -- speech.

However, while there has been a great deal of attention paid to the development of the hardware and systems, there has been almost no effort made to understand how humans will utilize this technology. To date, there has been very little research concerned with the impact of speech technology on the human user. The prevailing assumption seems to be that simply providing automated voice response and voice data entry will solve most of the human factors problems inherent in the user-system interface. But at present, this assumption is untested. In some cases, the introduction of voice response and voice data entry systems may create a new set of human factors problems. To understand how the user will interact with these new speech processing devices, it is necessary to understand much more about the human observer. In other words, we must understand how the human processes information. More specifically, we must know how the human perceives, encodes, stores, and retrieves speech and how these operations interact with the specific tasks the observer must perform.

In the Speech Research Laboratory at Indiana University, we have been carrying out a number of research projects investigating various aspects of human speech perception (see Pisoni, 1982, for a review). Strictly speaking, this work is not human factors research; that is, it is not designed to answer specific questions regarding the development and use of specific products. Rather, the goal of this research is to provide more general and basic knowledge about the perception of synthetic speech. This basic research can then serve as a foundation for subsequent human factors studies that may be motivated be specific problems.

Several areas of research are currently under investigation in our laboratory. In general, this research is concerned with the ability of human listeners to perceive synthetic speech under various task demands and conditions. Typically, two aspects of the listener's performance in these tasks are measured; we measure the speed and accuracy of subjects' responses. These performance variables allow us to make inferences about the complex cognitive processes that mediate human speech perception. For example, the speed of a subject's response (called reaction time or response latency) indicates to some degree the complexity and extent of perceptual and cognitive processing required to make the response. Thus, if subjects are slowest to respond in one particular condition of an experiment, then that condition may require more cognitive computation or processing capacity than other conditions.

CONSTRAINTS ON HUMAN PERFORMANCE

To interpret these results, it is necessary to consider the three basic factors that interact to affect an observer's performance: (1) the inherent limitations of human information processing, (2) the constraints on the structure and content of the speech signal, and (3) the specific task requirements. The nervous system cannot maintain all aspects of sensory stimulation. Moreover, there are severe processing limitations in the capacity for storing raw sensory data. To circumvent these capacity limitations, sensory information must be recoded or transformed into more abstract forms for stable storage and subsequent cognitive operations (Lindsay & Norman, 1977). By using long-term knowledge about the structure of language and real-world events, it is possible to make inferences that supplement and elaborate upon sensory input. Of course, this process is computationally intensive. One of the cognitive resources that is utilized during this type of processing is short-term memory. Short-term memory is the "working memory" of the human cognitive processor and is extremely limited in capacity (Shiffrin, 1976). As a result, the capacity limitations of short-term memory impose severe constraints on almost all perceptual and cognitive processes (Shiffrin & Schneider, 1977). In addition, these memory limitations may be exceedingly critical in speech perception since speech is a dynamic and transient signal. Unlike the printed word, once a word is spoken it must be processed immediately in real time or it is lost. It is impossible for the listener to "glance back" and hear a previous word over again because the physical waveform dissipates so quickly. To avoid this possible loss of information the input speech is buffered through an auditory memory store prior to short-term memory (Crowder, 1978; Pisoni, 1973). This auditory memory holds speech in a relatively unencoded form acting as sort of a sensory tape recorder. However, this sensory memory is even more limited in capacity than short-term memory. Thus, speech perception is a very time-critical process. If it is interrupted for too long a time, important sensory data could be lost and subsequent processing severely affected.

To handle the interruptions of noise and signal distortion, the perceptual process must take advantage of the constraints on the structure of spoken language. Speech production uses a complex and hierarchically organized system of linguistic rules to encode meanings into sounds. At the lowest level of the system, the distinctive properties of the signal are constrained by vocal tract acoustics and articulation. Consonants and vowels are coarticulated onto each other to form a complex coded stream of acoustic events (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Even at this level of representation, there is a great deal of redundancy, so that a single acoustic segment may transmit information relevant to several phonemes (Liberman, 1970). By utilizing this redundancy, in the acoustic speech code, the perceptual system is less sensitive to unwanted distortion of any single acoustic segment.

At the next level of linguistic complexity, the selection and ordering of consonants and vowels in words is constrained by the phonological rules of language. This provides another degree of predictability in the perceptual processing of speech. For example, in English, a initial liquid such as /l/ is never followed by a stop consonant such as /b/ within a single word. This means that information about some of the phonemes in a word may limit the possibilities for other phonemes in that word. Listeners have access to this knowledge and use it extensively in perceiving speech.

Similarly, the arrangement of words in sentences is constrained by the syntactic rules of language. Nouns, verbs, and other syntactic categories are not produced in random or arbitrary sequences. Instead, the allowable combinations of words are ordered by syntax. Research has shown that listeners use syntax to aid word recognition, even when the sentences presented are not meaningful (Miller & Isard, 1963).

Of course the meaning of words and sentences also provides constraints on the selection of words. Moreover, the context in which a sentence occurs constrains the meaning of a sentence. The word-by-word interpretation of a sentence in context will limit the possible alternatives for subsequent words. In this way, as a sentence is perceived, there become fewer and fewer plausibly alternative words available for finishing the sentence. Thus, a listener can sometimes understand a sentence even before the end of the sentence is spoken. Listeners may even be able to infer the identity of words that are obscured or obliterated by noise or distortion.

Clearly, there is a great deal of redundancy in speech from the acoustic structure to the semantic content. This redundancy provides some noise immunity and can help guide the listener's perceptual processing. However, the extent to which this redundancy can be utilized depends on the specific demands of the task confronting the listener. Humans are unusually flexible in their ability to develop specialized perceptual and cognitive strategies to maximize performance for different tasks (cf. Moray, 1975). For example, if a listener is required to respond to single isolated command words, it would be impossible to use syntactic or semantic predictability as an aid for word recognition. However, if the listener has a priori knowledge of the message set, that knowledge can help in recognition by delimiting the possible message just as syntax does.

Human performance in tasks is affected by a variety of factors. Some of these factors are closely related to the structure of the message ensemble (the stimuli), instructions and expectations about the task and stimuli, perceptual set, the nature of the responses, the mapping of the messages onto the responses, and past experience or practice. These factors can be manipulated in various ways to induce the observer to adopt different performance strategies. Observers are also capable of varying the "depth of processing" (Craik & Lockhart, 1972) that a message receives, depending on the requirements of the task. If an observer is required to remember every word in a spoken passage of text, the listener's strategy is very different from the strategy used in a task that only requires general comprehension of the spoken passage. The implication of this is that it is important to evaluate voice response systems under the same task demands that will be imposed on the ultimate users of the systems. The results of an evaluation test in the laboratory that places minimal demands on the listener may not generalize to the actual use of the voice response system in a more demanding application. For this reason, it is important to study the perception of synthetic speech using different tasks such as comprehension, recall, and speeded classification which require several levels of message complexity: passages, sentences, and words.

PERCEPTUAL EVALUATION OF SYNTHETIC SPEECH

When presented with a spoken passage of text, a listener is able to utilize the full range of psycholinguistic knowledge available in order to aid perception. To assess the contribution of this knowledge to listening performance, Pisoni and Hunnicutt (1980) conducted a series of experiments using synthetic speech generated by the MITalk text-to-speech system (see Allen, 1981). In one experiment, subjects were presented with fifteen narrative passages and an appropriate set of multiple-choice comprehension questions. The questions were designed to test general comprehension of the passages and were drawn, along with the passages, from several standardized adult reading comprehension tests. The passages covered a variety of topics, writing styles, and vocabulary. Three groups of subjects were tested. One group read printed versions of the passages, while the remaining two groups listened to continuous spoken versions. One of these two groups heard natural speech passages, while the other group listened to synthetic speech.

In this experiment, two results were extremely interesting. First, the group that heard the MITalk speech improved in comprehension performance from 64.1% correct in the first half of the experiment to 74.8% correct in the second half of the experiment. This 10.7% improvement in comprehension performance indicates that the synthetic speech group may have been learning the dialectal idiosyncrasies of the MITalk synthetic speech. Neither the group that read the passages nor the natural speech group showed any similar change in performance over the course of the experiment.

The second result of interest was that the average comprehension performance of the MITalk group (70.3% correct) was not statistically different from either the natural speech group (67.8% correct) or the reading group (77.2% correct). This suggests that even after a small amount of experience, listeners could understand the main points of the MITalk passages as well as the natural passages. A similar result was reported by Jenkins and Franklin (1981) in a comparison of natural speech with synthetic speech produced by the FOVE synthesis-by-rule system (Ingemann, 1978). Jenkins and Franklin found that the gist or general points of grade-school level passages could be remembered equally well for natural and synthetic speech.

At first glance, it seems somewhat surprising that comprehension performance for natural and synthetic speech should be the same. It is tempting to conclude that there are no problems in understanding the content of synthetic passages. However, it is important to note that in these experiments, listeners only had to understand the general ideas in the spoken passages. 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. 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. With this issue in mind, Luce (1981) conducted a more detailed examination of the comprehension of fluent synthetic speech. In this study, Luce used more specific questions that 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 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 the spoken passages and subjects responded by pressing the appropriately marked button on a response box that was interfaced to a minicomputer.

One group of subjects heard synthetic MITalk-produced passages and another group listened to natural speech 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 the previous studies, this indicates that comprehension of the natural passages was better than comprehension of the synthetic speech. However, there was a surprising result. The 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 heard natural speech remembered fewer specific words from the passages than the MITalk listeners.

These results seem to present something of a paradox. After all, if listeners can understand the words in a passage, they should be able to understand the passage at all levels of comprehension. The resolution to this paradox may be that subjects were asked if a particular word occurred in a passage after the text was heard. The group that listened to synthetic speech 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 understand the words may have made the words 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. Thus for these subjects, the specific words (as opposed to the concepts involved) were less salient in memory. Previous research has shown that during sentence comprehension, the surface structure is quickly forgotten while the basic concepts are retained (Sachs, 1967).

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.

Indeed, Pisoni and Hunnicutt (1980) have demonstrated that this psycholinguistic knowledge can have an important effect on the perception of synthetic speech. Using MITalk-generated speech, they compared word recognition in two types of isolated sentences. The first type of sentence was syntactically correct and meaningful. An example is given in (1) below:

(1) Add salt before you fry the egg.

The second type of sentence was also syntactically correct. However, these sentences were semantically anomalous. In other words, these test sentences had the syntactic form of normal sentences, but they were nonsense. An example of this type of nonsense sentence is given in (2) below:

(2) The yellow dog sang the opera.

By comparing word recognition performance for these two classes of sentences, it was possible to determine the influence of sentence meaning on word perception. Correct word recognition was very good in the meaningful sentences with a mean of 93.2% of the words recognized. However, only 78.7% of the words in the anomalous sentences were correctly recognized. Clearly the meaning of a sentence is a significant factor in perceiving synthetic words. Moreover, an analysis of the errors produced for both the meaningful and anomalous sentences indicated that the meaning of a sentence constrained the perceptual selection of words.

Thus, even though psycholinguistic knowledge can guide word recognition to some extent, it may be that word perception is harder for synthetic words than for natural words. Pisoni (1981) used a lexical decision task to compare the perception of natural and synthetic words in isolation. In this experiment, subjects were presented with two types of test items. On each trial, subjects had to decide as quickly as possible whether a test item was a "word" or a "nonword." Reaction time and accuracy were both measured. 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. This indicates that synthetic speech perception requires more cognitive "effort" than natural speech perception. But most important, this result was found for words 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 experience 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 (Norman & Bobrow, 1975); that is, the limitation may be in the structure of the synthetic speech itself.

In another experiment, Pisoni (1981) asked subjects to name synthetic and natural test items. On each trial, a subject repeated as quickly as possible a word or nonword that was heard through the headphones. The time required to make the naming response was measured together with the accuracy of the response. As in the previous studies, it was found that 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 the synthetic nonwords than the natural 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. This reinforces the conclusion that encoding the phonological structure of synthetic speech may require more computation than

encoding natural speech. Taken together, these findings suggest that the intelligibility of the phonological segments of synthetic speech should be worse than the intelligibility of the phonemes in natural speech.

Pisoni and Hunnicutt (1980) used the Modified Rhyme Test (MRT) to investigate the segmental intelligibility of synthetic and natural speech. On each trial the subjects were presented with a single isolated monosyllabic word. The subjects then selected one of six alternative responses. The alternatives were also monosyllabic words differing from each other in a single phoneme. On some trials, the responses differed only in the initial consonant; on other trials, the responses differed only in the final consonant. Therefore, the subjects were essentially choosing one of six different phonemes.

Performance on this task was very good for both the natural and synthetic speech. The error rate for natural speech was only .6% -- 99.4% correct. For the MITalk-produced synthetic speech, the error rate was 6.9% or 93.1% correct. It seems apparent that listeners had very little trouble deciding which phoneme was the correct response for both types of speech. However, in this forced-choice format (the listener is required to pick one response from six alternatives), the MRT constrains responding in much the same way that syntax or semantics in sentences can guide in word selection. As a result, this forced-choice testing procedure could inflate the estimates of segmental intelligibility by artificially constraining response alternatives. Indeed, Pisoni and Koen (1982) found that when these constraints were removed, segmental intelligibility was significantly worse. In their experiment, subjects were not restricted to six alternative responses. Instead, after hearing a test word, the subjects were free to select any response (from all the words they knew) that seemed appropriate; the set of allowable responses was not designated by the experimenter. The results of this change from the traditional forced-choice MRT procedure to the free-response were very striking. The error rate for the natural speech increased only slightly from .6% in the forced-choice procedure to 2.8% in the free-response format. However, for the synthetic speech, the error rate increased dramatically from 6.9% in the forced-choice procedure to 24.6% in the free-response paradigm. While listeners were able to identify 97.2% of the natural speech correctly, only 75.4% of the synthetic speech was correctly classified in the free-response format. Thus, even though segmental intelligibility appeared to be quite good in the traditional MRT forced-choice test, the ability to select <u>any</u> word as a response significantly impaired intelligibility in the free-response format. Apparently, subjects were able to use the constraints imposed by the restricted set of responses provided in the standard MRT. It is very important to emphasize the implications of these results. Data obtained in the standard forced-choice MRT cannot be generalized to predict performance in the free-response format. This is because there is an interaction between the type of testing procedure (forced-choice vs. free-response) and the type of speech presented (natural vs. synthetic); it is not possible to simply add or subtract a constant to performance in the forced-choice MRT to predict performance in the free-response paradigm. This clearly illustrates why it is important not to assume that a standardized testing procedure (like the MRT) will predict performance in applications where the task demands may be quite different.

Up to this point, the research we have summarized has been concerned with the intelligibility of speech. In general, it appears that even synthetic speech produced by a system as sophisticated as MITalk is less intelligible than natural

speech. However, it is clear from these studies that the intelligibility of synthetic speech will depend on the structure of the message set and the demand of the task. Moreover, it also appears that the segmental intelligibility of synthetic speech will be a major factor in word perception. For low cost speech synthesis systems where the quality of segmental synthesis may be poor, the best performance will be achieved when the set of possible messages is small and the user is highly familiar with the message set. It may also be important for the different messages in the set to be maximally distinctive like the military alphabet (alpha, bravo, etc.). In this regard, the human user should be regarded in somewhat the same way as an isolated-word speech recognition system.

Of course, this consideration becomes less important if the spoken messages are accompanied by a visual display of the same information. When the user can see a copy of the spoken message, any voice response system will seem, at first glance, to be quite intelligible. While providing visual feedback may reduce the utility of a voice response device, a low cost text-to-speech system could be used in this way to provide adequate spoken confirmation of data-base entries. Where visual feedback cannot be provided and the messages are not restricted to a small predetermined set, a more sophisticated text-to-speech system would be advisable.

Assessing the intelligibility of a voice response unit is an important part of evaluating any system for applications. But it is equally important to understand how the use of synthetic speech may interact with other cognitive operations carried out by the human observer. If the use of speech I/O interferes with other cognitive processes, performance of other tasks might be impaired if carried out concurrently with other speech processing activities. For example, a pilot who is listening to talking flight instruments might miss a warning light, forget important flight information, or misunderstand the flight controller. Therefore, it is important to understand the capacity limitations imposed on human information processing by synthetic speech.

LIMITATIONS ON SYNTHETIC SPEECH PERCEPTION

Recent work on human selective attention has suggested that cognitive processes are limited by the capacity of short-term (working) memory (Shiffrin & Schneider, 1977). Thus, any perceptual process that imposes a load on short-term memory may interfere with decision making, perceptual processing, and other cognitive operations. If perception of synthetic speech imposes a greater demand on the capacity of short-term memory than perception of natural speech, then the use of synthetic speech in applications where other cognitive operations are critical might produce serious problems.

Recently, Luce, Feustel, and Pisoni (1982) conducted several experiments to determine the effects of processing synthetic speech on short-term memory capacity. 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 visually presented on a CRT screen. On some trials no digits were presented and on other trials there were either three or six digits in the visual 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 write down all the digits in the order of presentation. After the digits were recalled, the subjects then wrote 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 revealed by the number of subjects who could recall all the digits presented in correct order. As the size of the digit lists increased, there were significantly fewer subjects recalling 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 indicate 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, Luce et al. (1982) 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 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 demonstrated that, in the synthetic lists, the words heard later in each list interfered with active maintenance of the words heard earlier in the list. 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.

The data on serial ordered recall support the conclusion from the lexical decision research that the processing of synthetic words and nonwords seems to require more computation than perception of natural speech. Thus, the perceptual encoding of synthetic speech requires more cognitive capacity and may in turn restrict other cognitive processing. Previous research on capacity limitations in speech perception demonstrated that paying attention to one spoken message seriously impairs the listener's ability to detect specific words in other spoken messages (e.g., Bcokbinder & Osman, 1979; Treisman & Riley, 1969). Moreover, several recent experiments have shown that attending to one message significantly impairs phoneme recognition in a second stream of speech (Nusbaum, 1981). Taken together, these studies indicate that speech perception requires active attention and cognitive capacity, even at the level of encoding phonemes. As a result, increased processing demands for the encoding of synthetic speech may place important limitations on the use of voice response systems in high information load conditions. This is especially true in cases where a listener may be expected to pay attention to several different sources of information.

SUMMARY AND CONCLUSIONS

Evaluating the use of voice response systems is not just a matter of conducting standardized intelligibility tests. Different applications will impose different demands and constraints on observers. Thus, it is necessary to take into account the three factors that interactively combine to affect human performance. First, the intrinsic limitations of human information processing must be considered. Perceptual and cognitive processes are primarily limited by the capacity of short-term memory. Since synthetic speech perception imposes a severe load on short-term memory, it is reasonable to assume that in highly demanding tasks, synthetic speech perception may impair the performance of other concurrent cognitive operations. Of course the converse situation may occur also; that is, performing a demanding task may interfere with synthetic speech perception. The human observer is not an interrupt-driven computer that can respond immediately to the presentation of an input signal. During complex cognitive processing, an observer may not be able to make the appropriate response to a speech signal; even worse, the presentation of a synthetic message might not be detected. Therefore, in highly demanding tasks, it is important to provide messages that maximize redundancy and distinctiveness. This is where the second factor, the structure and content of the message set, becomes critical. As the message set becomes simpler (e.g., isolated command words), the perceptual distinctiveness of the messages should be increased accordingly. For isolated words, the listener is unable to rely on the psycholinguistic constraints provided by syntax and meaning. Moreover, the discriminability of the messages is most important when the quality of phoneme synthesis is poor. In this type of synthetic speech, redundancy in the acoustic structure of the signal is minimized. As a result, more effort may be required to encode the speech. This implies that low cost synthetic speech should only be used when the task demands (the third factor) are not severe. It would be more advisable to use a low cost synthesizer to provide spoken confirmation of data-base entries than as a voice response system in the cockpit of a jet fighter.

Furthermore, it should be recognized that the ability to respond to synthetic speech in very demanding applications cannot be predicted from the results of the traditional forced-choice MRT. In the forced-choice MRT, the listener can utilize the constraints inherent in the task, provided by the restricted set of alternative responses. However, outside the laboratory, the observer is seldom provided with these constraints. There is no simple or direct method of estimating performance in less constrained situations from the results of the forced-choice MRT. Instead, evaluation of voice response systems should be carried out under the same task requirements that are imposed in the intended application.

From our research on synthetic speech perception, we can specify some of the constraints on the use of voice response systems. However, there is still a great deal of research to be done. Basic research is needed to understand the effects of noise and distortion on synthetic speech processing, how perception is influenced by practice and prior experience, and how naturalness interacts with intelligibility. Now that the technology has been developed, research on these problems and other related issues will allow us to realize both the potential and the limitations of voice response systems.

References

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

Bookbinder, J., & Osman, E. Attentional strategies in dichotic listening. Memory & Cognition, 1979, 7, 511-520.

Craik, F. I. M., & Lockhart, R. S. Levels of processing: A framework for memory research. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 671-684.

Crowder, R. G. Audition and speech coding in short-term memory: A tutorial review. In J. Requin (Ed.), Attention and performance VII. New York: Academic Press, 1978.

Ingemann, F. Speech synthesis by rule using the FOVE program. Haskins Laboratories Status Report on Speech Research, 1978, SR-54, 165-173.

Jenkins, J. J., & Franklin, L. D. Recall of passages of synthetic speech. Paper presented at the 22nd meeting of the Psychonomic Society, Philadephia, November, 1981.

Liberman, A. M. The grammars of speech and language. Cognitive Psychology, 1970, 1, 301-323.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. Psychological Review, 1967, 74, 431-461.

Lindsay, P. H., & Norman, D. A. Human information processing (2nd ed.). New York: Academic Press, 1977.

Luce, P. A. Comprehension of synthetic speech produced by rule. Journal of the Acoustical Society of America, 1982, 71, S96.

Luce, P. A., Feustel, T. C., & Pisoni, D. B. Capacity demands in short term memory for synthetic and natural speech. Submitted to Human Factors.

Miller, G. A., & Isard, S. Some perceptual consequences of linguistic rules. Journal of Verbal Learning and Verbal Behavior, 1963, 2, 217-228.

Moray, N. A data base for theories of selective attention. In P. M. A. Rabbitt & S. Dornic (Eds.), Attention and performance V. New York: Academic Press, 1975.

Norman, D. A., & Bobrow, D. G. On data-limited and resource-limited processes. Cognitive Psychology, 1975, 7, 44-64.

Nusbaum, H. C. Capacity limitations in phoneme perception. Unpublished doctoral dissertation, S.U.N.Y. at Buffalo, September, 1981.

Pisoni, D. B. Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception & Psychophysics, 1973, 13, 253-260.

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(2), 10-23.

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

Pisoni, D. B., & Koen, E. Intelligibility of natural and synthetic speech at several different signal-to-noise levels. Paper presented at the 103rd meeting of the Acoustical Society of America, Chicago, April, 1982..

Sachs, J. S. Recognition memory for syntactic and semantic aspects of connected discourse. Perception & Psychophysics, 1967, 2, 437-442.

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: LEA, 1976.

Shiffrin, R. M., & Schneider, W. Controlled and automatic information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological Review, 1977, 84, 127-190.

Slowiaczek, L. M., & Pisoni, D. B. Effects of practice on speeded classification of natural and synthetic speech. Journal of the Acoustical Society of America, 1982, 71, S95-S96.

Treisman, A. M., & Riley, J. G. A. Is selective attention selective perception or selective response? A further test. Journal of Experimental Psychology, 1969, 79, 27-34.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

Perceptual Anchoring of a Speech-Nonspeech Continuum*

Howard C. Nusbaum

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

*This research was supported by NIMH grant MH-31468 to SUNY/Buffalo. Preparation of this manuscript was supported, in part, by NIH training grant NS-07134 to Indiana University. The author would like to thank David B. Pisoni, Eileen C. Schwab, James R. Sawusch, Amanda C. Walley, Thomas D. Carrell, and Steven Greenspan for a number of helpful comments and suggestions.

Abstract

Previous research has demonstrated that selective adaptation effects can be found for speech-nonspeech continua which vary on a phonetically irrelevant acoustic dimension (Remez, 1979, 1980). The conclusion of this research was that, by the logic behind selective adaptation research, the set of feature detectors involved in speech perception does not appear to be constrained by any acoustic-phonetic principles. The present experiment used an anchoring procedure with a similar speech-nonspeech stimulus series to test this claim. The stimuli in the series varied in formant bandwidth from a /bae/ at one endpoint to a nonspeech "buzz" at the other end of the series. When one of the endpoint stimuli was presented (and identified) more often than the rest of the series, the category boundary for the test series was displaced relative to the baseline rating function. The contrast effects produced by anchoring with endpoint stimuli in the present study are quite similar to the adaptation results obtained by Remez. These results suggest that the adaptation effects with speech-nonspeech continua probably reflect judgmental anchoring rather than the feature detector desensitization that would otherwise be suggested by adaptation-induced contrast effects. Thus, contrary to the claims made by Remez, previous adaptation experiments with speech-nonspeech continua do not constitute evidence against feature detector theories.

Perceptual Anchoring of a Speech-Nonspeech Continuum

Recently, Studdert-Kennedy (1981) has strongly criticized cognitive psychology, claiming that "this information-processing approach to speech perception . . . eventually led to a dead end, as it gradually became apparent that this undertaking was mired in tautology" (p. 302). This indictment of cognitive psychology primarily stems from a growing disaffection with one particular research paradigm -- selective adaptation -- and with the feature detector theories that are supported by this research (see Diehl, 1981; Repp, 1982; Studdert-Kennedy, 1977, 1981, 1982, for criticisms of feature detector theories of speech perception). In the past few years, several studies have questioned the validity of selective adaptation as a procedure for investigating phoneme perception (see Diehl, Elman, & McCusker, 1978; Diehl, Lang & Parker, 1980; Remez, 1979, 1980; Remez, Cutting, & Studdert-Kennedy, 1981; Rosen, 1979; Simon & Studdert-Kennedy, 1978). At present, selective adaptation is the only paradigm used to provide direct evidence for the operation of feature detectors in human speech perception. Thus, while there have been theoretical arguments against feature detector theories (e.g., Studdert-Kennedy, 1977, 1982), discrediting selective adaptation would most effectively eliminate support for these theories of phoneme perception (see Diehl, 1981).

The first selective adaptation experiments using speech attempted to find evidence for phonetic feature detectors (see Cooper, 1975, 1979, for reviews of this early research). Abbs and Sussman (1971) proposed that the perception of phonetic categories could be mediated be feature detectors tuned to respond directly to phonetic information in the speech waveform. Sets of feature detectors would independently register phonetic features such as voicing. The outputs of these phonetic feature detectors would then be combined to identify a single phonetic segment. The existence of phonetic feature detectors would also provide a mechanism capable of explaining categorical perception (see Abbs & Sussman, 1971). If phonetic features are represented by discrete neural units (at least one for each phonetic feature value), intermediate states of perception between feature values should not be computed easily. Thus, the continuous acoustic variation present in the speech signal would be converted to a distribution of neural activity in discrete feature detectors. Two consonants that share the same phonetic feature values but differ acoustically should produce the same pattern of activation throughout the detector population. This would make discrimination of two segments from the same phonetic category more difficult than discrimination of two segments from different phonetic categories. Furthermore, this system would produce perceptual constancy of speech sounds despite acoustic variation.

To test the phonetic feature detector hypothesis, Eimas and Corbit (1973) reasoned that these putative detectors would be adapted if the same units were stimulated often enough. By rapidly and repeatedly presenting the same phonetic information to listeners, it was hoped that the corresponding feature detectors would be fatigued. This would make the detectors less responsive to subsequent stimulation. For example, since /b/ is a voiced stop consonant, repeated exposure to /ba/ should reduce the sensitivity of a voiced feature detector. This would make the voiced detector less able to compete with a counterpart voiceless detector. Therefore a stimulus that had elicited equal responses from both voiced and voiceless detectors prior to adaptation should elicit more of a voiceless response after adaptation. This contrast effect represents the dependent measure of the perceptual impact of selective adaptation.

Eimas and Corbit synthesized test series that varied acoustically between voiced and voiceless endpoints. These stimuli were created by systematically changing the voice onset time (VOT) of one endpoint in small steps until the second endpoint was produced. While the endpoints were perceived as good examples of the voiced and voiceless categories, stimuli near the voiced/voiceless category boundary were much more ambiguous on the voicing dimension. The category boundary represents the point where the hypothetical detectors that mediate endpoint perception respond equally. In essence, the boundary stimulus is phonetically ambiguous on the test dimension. Eimas and Corbit found that adaptation with the voiced endpoint produced fewer voiced responses in subsequent identification testing, while adaptation with the voiceless endpoint reduced the number of voiceless responses. This contrast effect was especially pronounced for stimuli near the category boundary prior to adaptation. After adaptation, the category boundary shifted toward the adapting endpoint of the series. As a result, stimuli that were ambiguous before adaptation became identified with the category opposite the adaptor after adaptation.

Insert Figure 1 about here

This type of shift in the category boundary resulting from selective adaptation is illustrated in Figure 1. This figure shows a schematic representation of the results of an hypothetical adaptation experiment for a seven-element test series. Baseline identification in the control condition, where each stimulus is presented equally often in random order, is shown by the solid line in the middle of the figure. The dashed lines represent the effects of selective adaptation on stimulus identification. The effect of adaptation with Stimulus 1 is indicated by the dashed line to the left of the baseline function. The dashed line on the right shows the result of adaptation with Stimulus 7. In both cases, adaptation produces a contrast effect. After adaptation, fewer stimuli are perceived as belonging to the same phonetic category as the adaptor. As can be seen in Figure 1, these effects are most pronounced at the boundary between phonetic categories. Eimas and Corbit (1973) interpreted this type of shift in the category boundary resulting from adaptation as a change in the point of equal detector response due to desensitization of one detector of a voiced/voiceless pair.

Eimas and Corbit (1973) also found that adaptation could be obtained even when the adaptors were not members of the test series. For example, adapting with the voiceless stop /p/ shifted the category boundary of a /d/-/t/ test series in the direction of the /t/ endpoint. They interpreted this result as an indication that the feature detectors were responding to phonetic features instead of syllable-specific acoustic patterns. Moreover, a number of subsequent studies demonstrated that cross-series adaptation could be obtained when the adaptor and an endpoint of a test series shared a phonetic feature but had little or no overlap in acoustic features (see Cooper, 1979, for a review).

Although cross-series adaptation was initially seen as support for phonetic feature detectors (see Cooper, 1975, 1979; Eimas & Corbit, 1973), other research provided evidence against the phonetic feature detector interpretation (see Ades,

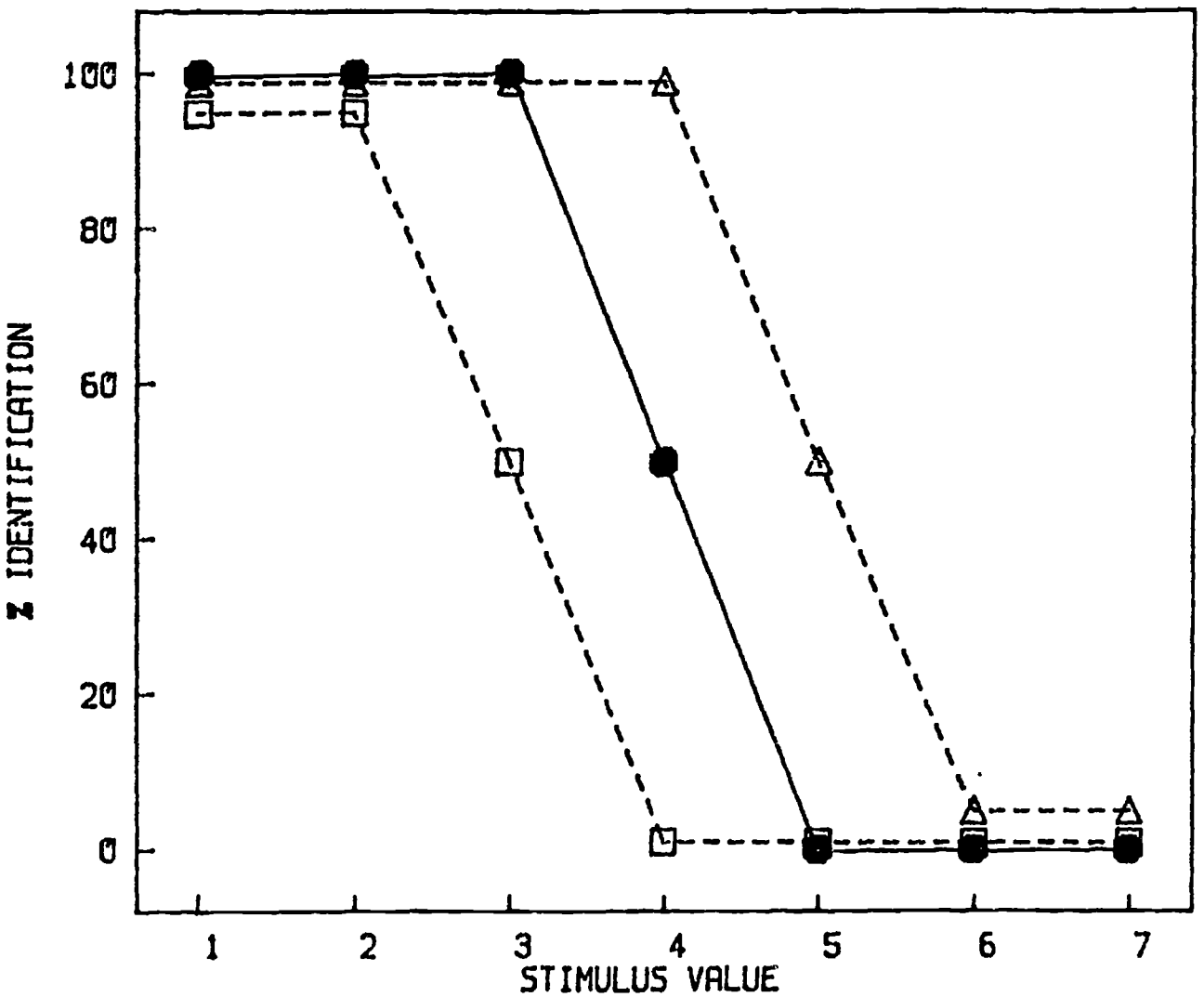


Figure 1. The results of an hypothesized adaptation experiment. Baseline identification of each stimulus in the control condition is shown by the solid. Adaptation with Stimulus 1 is shown with the open squares and adaptation with Stimulus 7 is shown by the open triangles.

1976; Eimas & Miller, 1978, for reviews). One set of experiments showed that cross-series adaptation could be obtained with nonspeech adaptors that shared an acoustic feature or pattern with an endpoint of a speech test series (e.g., Pisoni & Tash, 1975; Samuel & Newport, 1979; Tartter & Eimas, 1975). These studies indicated that the feature detectors involved in adaptation could be tuned to respond to acoustic features instead of phonetic features. However, the strongest argument against phonetic feature detectors came from experiments which found that an adaptor in one position within a syllable would not adapt perception of test items in a different position (Ades, 1974; Sawusch, 1977). Linguistically speaking, a consonant in any position should be the same; phonetic features do not take into account their relative location in an utterance. However, Ades (1974) found that an adaptor in final position (e.g., /aeb/) had no effect on the perception of consonants in initial position (e.g., /bae/-/dae/). These results suggested that adaptation induces fatigue of auditory -- not phonetic -- feature detectors (Ades, 1976; Eimas & Miller, 1978). But regardless of the actual locus of adaptation (i.e., auditory or phonetic), the mechanism of adaptation was generally thought to be feature detector fatigue (Cooper, 1979; Eimas & Miller, 1978).

While there is some neurophysiological evidence for the operation of auditory and call-specific feature detectors in nonhuman animals (see Evans, 1974; Scheich, 1977), similar neurophysiological experiments have not been performed on humans for the obvious reasons. Indeed, the only evidence for the role of feature detectors in human speech perception comes from the contrast effects produced by selective adaptation. But feature detector fatigue is not the only possible explanation of the contrast effects produced by adaptation. In fact, contrast effects have been obtained using other types of psychophysical paradigms without the adaptation methodology.

The concept of feature detector fatigue is closely linked to the procedure of repeatedly stimulating one set of detectors without allowing them to recover. Thus, in the adaptation procedure, adaptors are quickly presented one after another in an uninterrupted sequence, with no intervening stimuli. To the extent that contrast effects can be found using other experimental paradigms which would permit detector recovery, feature detector explanations of selective adaptation become suspect and other explanations of perceptual contrast would be necessary to explain adaptation.

In psychophysics, it is generally acknowledged that perceptual judgments are influenced by the response context (Parducci, 1965) or stimulus context (Helson, 1964) in which perception occurs. This provides an alternative to the feature detector account of selective adaptation in that subjects may treat the sequence of adaptors as a biasing perceptual context (cf. Bryant, 1977; Elman, 1979). In adaptation conditions, subjects perceive one endpoint of the test series more often than any other stimulus in the continuum. Covert identification of the adaptors would result in the disproportionate use of one phonetic category relative to the phonetic category of the other endpoint. This could bias subjects against using the adaptor's category for responding to other test stimuli (cf. Parducci, 1965). On the other hand, perception of the adaptor might bias or modify some internal standard used for perceptual reference (cf. Anderson, Silverstein, Ritz, & Jones, 1977; Helson, 1964). In either case, perception of the adaptor would "anchor" a subjects's judgment of other stimuli instead of desensitizing feature detectors.

However, there is one fundamental difference between this "judgmental anchoring" view of adaptation and the feature detector fatigue interpretation. According to the feature detector fatigue explanation, contrast effects should depend on repetitively stimulating the same feature detectors without relief. If the adaptors were presented in such a way as to allow detector recovery, no contrast effects should be obtained. This means that detector fatigue should be contingent on the arrangement of the adaptor presentations. In comparison, the judgmental anchoring position would predict that the production of contrast effects should depend upon the greater probability of occurrence of an "extreme" or exemplar stimulus relative to the rest of the stimulus ensemble (Anderson et al., 1977; Helson, 1964; Parducci, 1965). Thus, even if other stimuli are presented between instances of the adaptor (now called an anchor), contrast effects should be found. Indeed, just such effects have been obtained for the perception of brightness (Helson, 1964), dots varying in numerosity (Helson & Kozaki, 1968), heaviness of lifted weights (Parducci, 1963, 1965), and sounds varying in frequency or intensity (Cuddy, Pinn, & Simons, 1973; Sawusch & Pisoni, Note 1). In all cases, subjects were presented with one stimulus -- the anchor -- more often than the other test items. The anchoring procedure essentially parallels the selective adaptation paradigm. Subjects identify randomly ordered stimuli in each of two conditions. The control condition is identical to the control condition in an adaptation experiment, where all stimuli occur with equal frequency. In the anchoring condition, one of the endpoint stimuli is presented and identified more often than any other test stimulus. Compared with the adaptation condition, it is as if the adaptors were mixed in and presented along with the other test items (see Simon & Studdert-Kennedy, 1978). The interval between anchors is equal to the intertrial interval and stimuli from other perceptual categories can intervene between anchors.

Anchoring effects with stop consonants have been found, and at first glance, these results seem to closely parallel the effects of selective adaptation on stop consonant perception (Rosen, 1979; Simon & Studdert-Kennedy, 1978). In principle, these contrast effects produced by anchoring cannot be explained by feature detector fatigue. The similarity of adaptation and anchoring of stop consonant perception should imply that these effects are mediated by a common mechanism which is <u>not</u> detector fatigue. In other words, these are exactly the kind of results that judgmental anchoring could explain. However, this is really an oversimplification of the situation. There are, in fact, two alternative explanations of these anchoring results which are consistent with feature detector theories.

First, it is possible that the endpoint stop consonants used in the anchoring experiments were not good phonetic exemplars. Since the speech stimuli used in adaptation and anchoring experiments are typically synthetic in origin, it is possible that these stimuli were not only perceived in a "speech mode" (Liberman, 1970; Liberman & Studdert-Kennedy, 1977) because the acoustic features in those sounds did not accurately reflect naturally produced acoustic-phonetic cues. If these stimuli were perceived using auditory (nonspeech) processes, or a mixture of phonetic and auditory processes due to the lack of veridicality of the acoustic-phonetic structure of the sounds, the obtained anchoring effects could reflect judgmental contrast in the auditory processes alone. Previous research has already demonstrated that perception of nonlinguistic auditory dimensions such as loudness and pitch is extremely susceptible to anchoring influences (Simon & Studdert-Kennedy, 1978; Sawusch & Pisoni, Note 1). Therefore, if the perception of synthetic speech is only partially mediated by nonspeech processes, anchoring effects might be obtained.

For example, in one anchoring experiment reported by Rosen (1979), subjects rated the endpoints of a test series as only "probably" /b/ or /d/ in the control condition. This was not the highest rating available; the subjects did not consider the endpoints to be "definite" (the highest rating) members of the phonetic categories /b/ and /d/. Moreover, there was no evidence that these speech sounds were categorically perceived; no discrimination data were presented. Also, Simon and Studdert-Kennedy (1978) did not report any discrimination data for their stop consonant anchoring experiments. As Studdert-Kennedy, Liberman, Harris, and Cooper (1970) have pointed out, categorically perceived speech should be immune to the contextual influences in anchoring; such speech sounds should be identified absolutely rather than in relation to other test stimuli. Sawusch and Pisoni (Note 1) have also claimed that anchoring effects cannot be obtained with categorically perceived speech sounds.

A second possible explanation of the anchoring effects found for consonant perception is that adaptation (instead of anchoring) actually occurred. Since, by definition, the anchor is the most frequently occurring stimulus, it is not unusual for several anchors to be presented in a contiguous sequence in any particular random order. At high ratios of anchors to test stimuli, these sequences could become quite long. If the length of such sequences is not controlled, adaptation -- rather than anchoring -- could be the result. Sawusch and Pisoni (Note 1) have shown that when anchor sequence length was constrained to prevent adaptation, no anchoring of stop consonant perception was obtained. But when the same number of anchors was presented as an adapting sequence prior to identification of the test series, significant adaptation effects were produced. In other words, when the number of adaptors and anchors was the same in two experiments, and the length of anchor sequences was controlled, only the adaptation paradigm produced contrast effects for stop consonant perception. Thus, it is the arrangement of the adaptors/anchors which appears to be crucial for modifying stop consonant perception. This is exactly the pattern of results predicted by feature detector fatigue and opposite the prediction made by judgmental anchoring explanations. As a result, previous anchoring experiments with stop consonants do not constitute strong evidence against feature detector interpretations of selective adaptation.

Another method for producing contrast effects has been described by Diehl et al. (1978). In this procedure, subjects were presented with pairs of consonant-vowel (CV) syllables. Subjects were asked to identify both members of a pair after the stimuli were heard. When a stop consonant that was ambiguous on one phonetic dimension (e.g., voicing) was paired with a phonetic exemplar on that dimension (e.g., a "good" example of /b/ or /p/), a contrast effect was obtained. The ambiguous test consonant was heard as more /b/-like when paired with a /p/ context and more /p/-like with a /b/ context. Moreover, this contrast effect was also found when the exemplar context followed the ambiguous test item. This effectively ruled out one-trial adaptation by the exemplar context. Diehl et al. (1978) interpreted these results as indicating that adaptation effects are mediated by the type of judgmental anchoring described by Helson (1965). Based on these and other similar results (e.g., Diehl et al., 1980), Diehl (1981) claimed that "adaptation results no longer constitute evidence for feature detectors" (p. 7).

More recently, however, this claim has been refuted by experiments that have dissociated selective adaptation from the successive contrast procedure employed

by Diehl et al. (1978). Sawusch and Jusczyk (1981) compared the effects of these two experimental paradigms on the perception of a /ba/-/pa/ speech series. The /ba/ endpoint produced the same type of contrast effect when presented as an adaptor and when presented as a context stimulus in the successive contrast procedure. The /pa/ endpoint also produced similar contrast effects in the two paradigms. But a /spa/ stimulus produced a very different pattern of results. The /spa/ was constructed by combining /s/ frication noise with a /ba/ syllable. The resulting stimulus contained a stop consonant-vowel syllable with the acoustic structure of /ba/ but the perceptual identity of /pa/. When /spa/ was paired with an ambiguous test syllable in the successive contrast procedure, the effect was similar to the contrast effect obtained with the /pa/ endpoint context -- the test item was labeled as /ba/ more often. In comparison, when /spa/ was used as an adaptor, the effects were identical to the /ba/ endpoint adaptor -- the boundary syllable was labeled as /pa/ more often. In adaptation, the voiced spectro-temporal structure of the stop consonant cues in /spa/ governed the direction of the contrast effect. In the successive contrast paradigm, the obtained contrast effects were determined by the perception of the stop consonant in /spa/ as voiceless. These results clearly demonstrate that the judgmental contrast effects found by Diehl et al. (1978) occur at a locus of processing which is different from the locus of selective adaptation effects. Similar results dissociating adaptation from successive contrast have been found for the perception of place of articulation in stop consonants (Sawusch & Nusbaum, Note 2). In short, it appears that selective adaptation may produce contrast effects at a level of auditory feature processing, while the locus of successive contrast effects may be at a stage of processing responsible for identifying phonemes.

In general then, attempts to produce contrast effects using experimental procedures other than selective adaptation have not really been successful in arguing against feature detector theories of consonant perception. Since judgmental contrast procedures and selective adaptation can produce different patterns of results, it is hard to account for all the data with one mechanism such as response bias (cf. Diehl, 1981). It is apparent that feature detectors alone cannot explain the successive contrast effects reported by Diehl et al. (1978) and Sawusch and Jusczyk (1981). But feature detectors were never invoked to explain this sort of judgmental contrast effect. What is most important is that adaptation of stop consonant perception cannot be dismissed as response contrast or judgmental anchoring. The hypothesis that stop consonant perception is mediated at an early processing stage by auditory feature detectors is still supported by selective adaptation research (Sawusch & Jusczyk, 1981; Sawusch & Pisoni, Note 1; Sawusch & Nusbaum, Note 2).

However, Remez (1979, 1980) has taken a very different approach in testing the hypothesis that feature detectors play a role in phoneme perception. Remez demonstrated that selective adaptation could induce contrast effects for an acoustic dimension that does not directly cue a phonetic distinction. The stimuli for these experiments were synthetic sounds that varied in formant bandwidth from a speech token at one end of a series to a nonspeech "buzz" at the other end. Subjects used the appropriate speech or nonspeech label to categorize the test series in both control and adaptation conditions. Adaptation with either the speech or the nonspeech endpoint produced significant contrast effects. Remez interpreted these results as evidence against both phonetic and auditory feature detectors. The rationale for his conclusion was that the distinction between speech and nonspeech is not a phonetic feature distinction and should not be mediated by phonetic feature detectors. Furthermore, formant

bandwidth is not an acoustic feature relevant to phonetic decisions and should not be processed by auditory feature detectors used in speech perception. By the logic of selective adaptation, however, these results should suggest the existence of either a speech/nonspeech feature detector set or detectors tuned to different formant bandwidths. According to Remez, from the point of view of phonetics, the addition of these detectors to the human perceptual system would not seem warranted.

On the other hand, the contrast effects Remez obtained might not reflect the same type of auditory feature adaptation isolated by Sawusch and Jusczyk (1981). Instead, these effects could have been produced by contrast in higher-level judgmental (decision) mechanisms rather than in lower-level auditory feature detectors. While it has been possible to eliminate judgmental contrast explanations of stop consonant adaptation effects dissociating adaptation effects from judgmental contrast (Sawusch & Jusczyk, 1981; Sawusch & Pisoni, Note 1; Sawusch & Nusbaum, Note 2), this may not be true for adaptation of the perception of speech-nonspeech continua.

In comparison with stop consonants, when vowels were presented in an anchoring experiment, significant contrast effects were obtained (Sawusch & Nusbaum, 1979). This was true even though the length of anchor sequences had been constrained to prevent adaptation. Further, these vowel anchoring effects closely parallel the effects of selective adaptation with vowels (cf. Morse, Kass, & Turkienicz, 1976). Thus, for vowel perception, adaptation can be explained by judgmental contrast rather than by feature detector fatigue. It is possible that this type of judgmental anchoring could mediate the adaptation effects found with speech-nonspeech continua.

In one experiment, Remez (1979) used a series which varied from /a/ to a "buzz", and in a second experiment, he used an /ae/-"buzz" stimulus series. These stimuli were created by increasing formant bandwidths in small steps starting with the vowel endpoint and finishing when a "buzz" was produced. After citing several potential problems with these vowel-based stimuli, Remez (1980) replicated his earlier results with a /ba/-"buzz" test series constructed in the same way. For all these continua, one endpoint contained a vowel and the other endpoint was a nonspeech "buzz" sound. Previous research has established the sensitivity of vowel identification to extra presentations of a vowel endpoint (e.g., Sawusch & Nusbaum, 1979). Moreover, anchoring effects have been found when a change in vowel identity is correlated with a change in consonant identity such as in a /bae/-/dE/ test series (Simon & Studdert-Kennedy, 1978). Recently, Sawusch (Note 3) has repeated this experiment using the entire set of stimuli that would be produced by all possible combinations of a /b/-/d/ consonant series with an /ae/-/E/ series. These stimuli included /bae/, /dae/, /bE/, and /dE/ as the "corners" (i.e., the four endpoints) of the set. This allowed Sawusch to separate anchoring by consonant category (e.g., all /b/ items) from vowel anchoring (e.g., all /ae/ items). Consonant identification was unaffected by anchoring, but significant anchoring effects were produced for vowel identification. These results suggest that when changes in consonant and vowel are correlated (e.g., as in the /ba/-"buzz" series), shifts in identification of the stimuli may be attributed to changes in vowel perception alone. Therefore, anchoring could be predicted for the speech endpoint of all the test series used by Remez.

Similarly, anchoring effects have been found for nonspeech sounds (e.g., Cuddy et al., 1973), so that anchoring for the "buzz" endpoint could also have been predicted. The present experiment was designed to test these predictions. If anchoring can be obtained for perception of a series of sounds ranging from speech to nonspeech, the adaptation effects reported by Remez could then be explained by judgmental contrast. This would eliminate the empirical basis for his argument against selective adaptation and thus would eliminate one argument against feature detectors in phonetic perception. By comparison, if anchoring cannot be produced for a speech-nonspeech continuum, it would have to be conceded that there is some flaw in the logic behind selective adaptation. This result would imply that adaptation does not tap the processes mediating phoneme perception.

Method

Subjects

The subjects in this experiment were 20 graduate and undergraduate students at the State University of New York at Buffalo. None of the subjects had previously participated in a speech experiment. All subjects were right-handed native speakers of English with no reported history of either speech or hearing disorder. The subjects were paid $3/h for their participation.

Stimuli

The stimuli were a set of nine synthetic sounds that ranged perceptually from the consonant-vowel syllable /bae/ (Stimulus 1) to a nonspeech "buzz" (Stimulus 9). These sounds were created using the cascade branch of a software speech synthesizer (Klatt, 1980a) which was modified by Kewley-Port (Note 4). The synthesis parameters for the /bae/ endpoint were derived from measurements of a spectrogram of /bae/ spoken by an adult male talker. The first three formants increased linearly in frequency over the first 40 msec of the syllable: F1 increased from 280 Hz to 670 Hz; F2 rose from 1300 Hz to 1600 Hz; and F3 changed from 2300 Hz to 2700 Hz. After the first 40 msec, these formant frequencies were constant over the remainder of the syllable. The values of F4 and F5 were fixed at 3300 Hz and 3850 Hz respectively. The fundamental frequency (FO) contour was chosen to approximate a natural CV syllable produced in isolation. These parameters were the same for each of the nine stimuli. The stimuli in the test series differed from each other in formant bandwidths only. The bandwidths of the first five formants were increased in 50 Hz steps for each successive token starting with the initial values in Stimulus 1 (/bae/). In Stimulus 1, the bandwidths of these formants were 65 Hz, 95 Hz, 130 Hz, 250 Hz, and 200 Hz for F1 through F5. Figure 2 shows spectrograms of Stimulus 1 (the /bae/ endpoint), Stimulus 5 (the series midpoint), and Stimulus 9 (the "buzz" endpoint). The progression across the series from the prototypical /bae/ formant pattern to a buzz with no clear formant structure can easily be seen in these spectrograms.

Insert Figure 2 about here

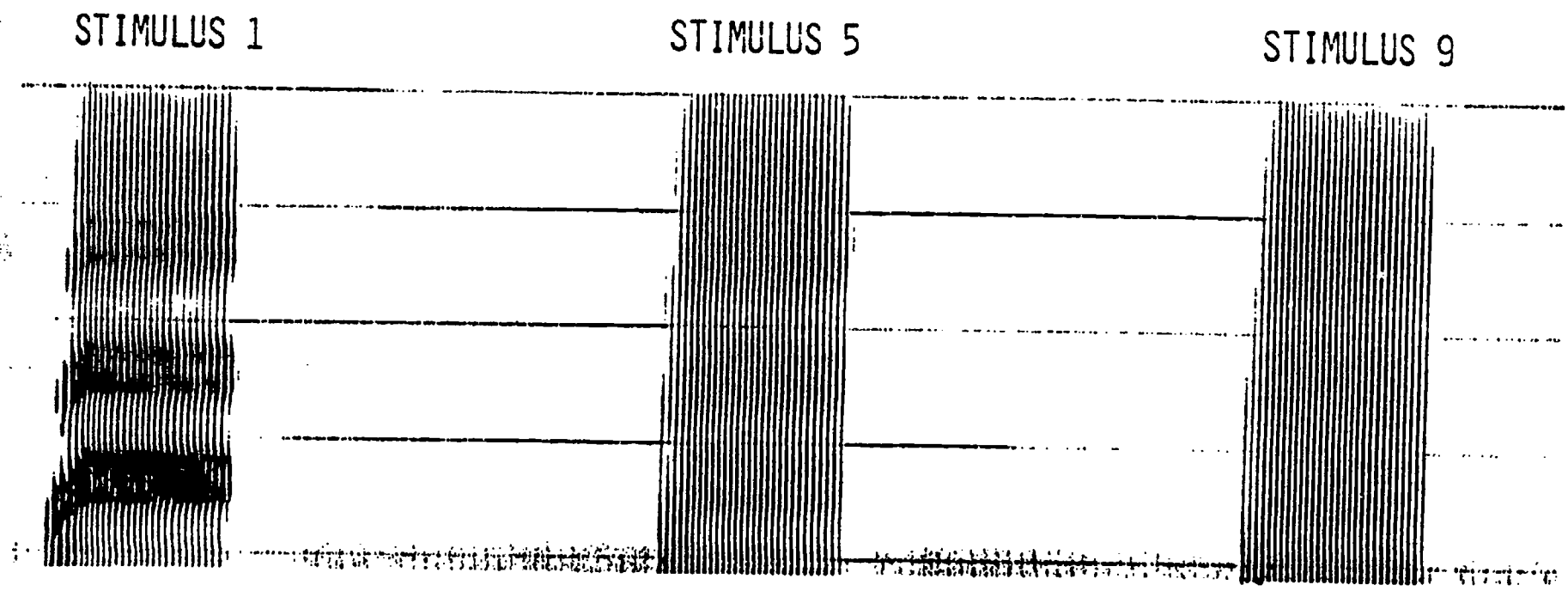


Figure 2. Spectrograms of Stimulus 1 (/bae/), Stimulus 5 (the series midpoint), and Stimulus 9 ("buzz").

Another way of looking at the effects of bandwidth change is to examine smoothed power spectra of the onsets of the endpoint stimuli. The linear prediction power spectrum (see Markel & Gray, 1976) of the first 25.6 msec of the /bae/ endpoint is shown in Figure 3. The spectral peaks corresponding to the first five formants are clearly visible. The tilt of the spectrum corresponds to the diffuse-falling pattern that Blumstein and Stevens (1979) consider indicative of a labial consonant. Thus, both the onset spectrum and transition pattern for Stimulus 1 are appropriate for /b/. The LPC spectrum of 25.6 msec at onset of the "buzz" endpoint is shown in Figure 4. The spectral peaks have been substantially reduced by the bandwidth increase, resulting in a nearly flat onset spectrum. A comparison of these /bae/ and "buzz" onset spectra with the spectra presented by Remez (1980) shows that the effects of formant bandwidth increase are similar for both sets of stimuli.

Insert Figure 3 and Figure 4 about here

The stimuli were converted to analog form by a 12-bit digital-to-analog converter, low-pass filtered at 4.8 kHz, and presented to subjects in real time under computer control. The sounds were presented binaurally through Telephonics TDH-39 matched and calibrated headphones. The intensity of the stimuli was set to 76 dB SPL.

Procedure

The subjects were assigned to two groups of 10 subjects each. One group received the /bae/-anchored condition and the second group received the "buzz"-anchored condition. Each subject participated in a single 1 h session with one to four subjects in any particular session. Experimental sessions were conducted under the control of a PDP-11/34a computer which determined random orders, presented stimuli, and collected responses.

At the beginning of each session, subjects responded to a practice random order of three repetitions of each stimulus. The data from this practice set were discarded. After practice, subjects were presented with two control random orders of 90 trials each. The control orders were followed by two anchored random orders of 140 trials each. In the control random orders, each of the nine stimuli was presented ten times. In the anchor conditions, one of the endpoints was presented 60 times and the rest of the stimuli occurred 10 times each. The anchored random orders were constrained so that no more than four anchors ever occurred in a sequence. One group of subjects received the /bae/-anchored random orders while the other group heard the "buzz"-anchored random orders. All stimuli were separated by a fixed 4 sec intertrial interval. By the end of each session, each subject had provided at least 20 responses to each of the nine stimuli in the control and anchored conditions, excluding practice.

The subjects were told that they would be listening to computer generated stimuli that would sound like either the syllable /bae/ as in "bat" or a nonspeech "buzz." Subjects were asked to rate each sound on a six-point scale by pressing the appropriately labeled button on a response box. The first button was pressed to indicate a good instance of /bae/, the sixth button was pressed

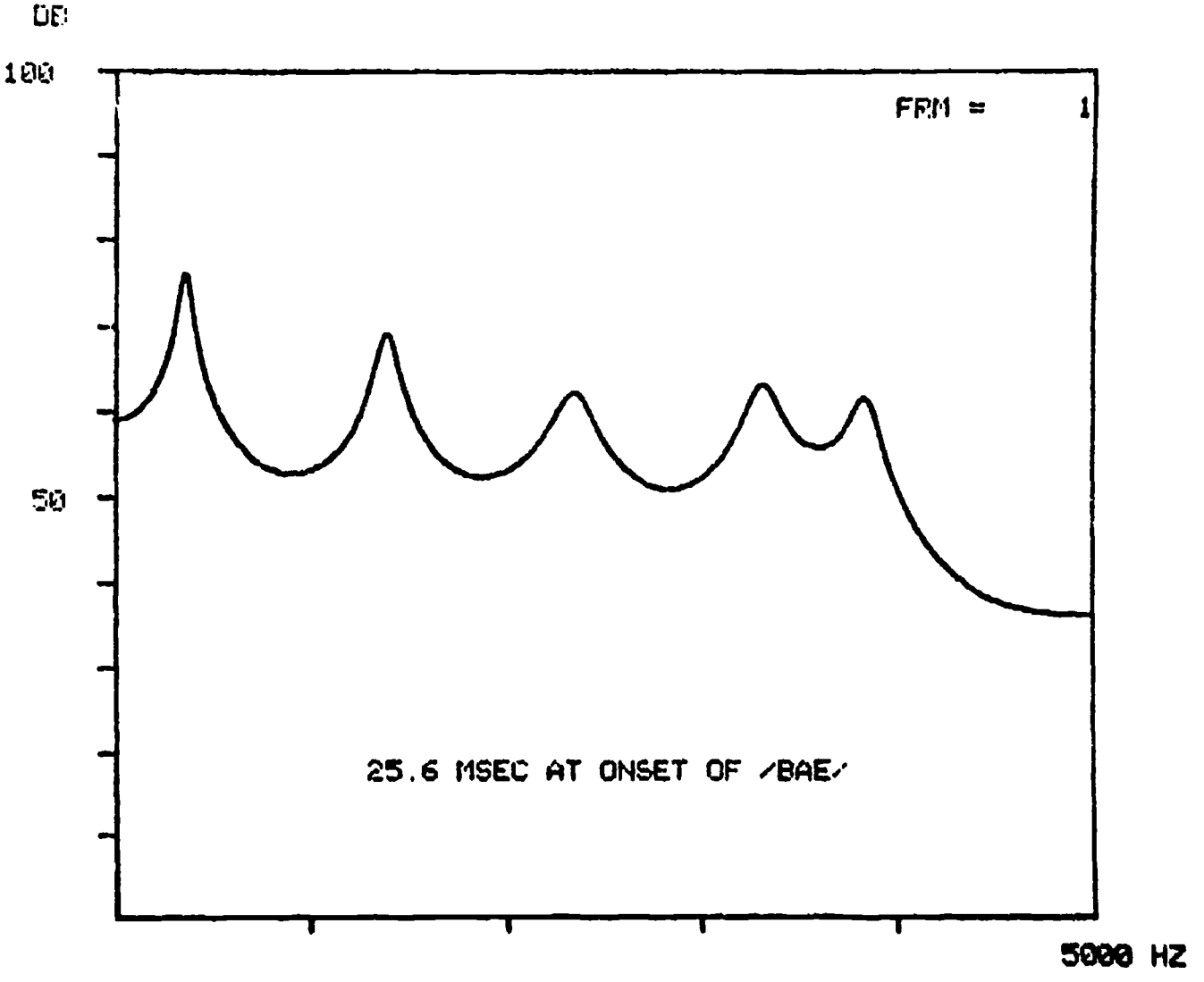


Figure 3. Smoothed spectrum of 25.6 msec at the onset of Stimulus 1 (/bae/).

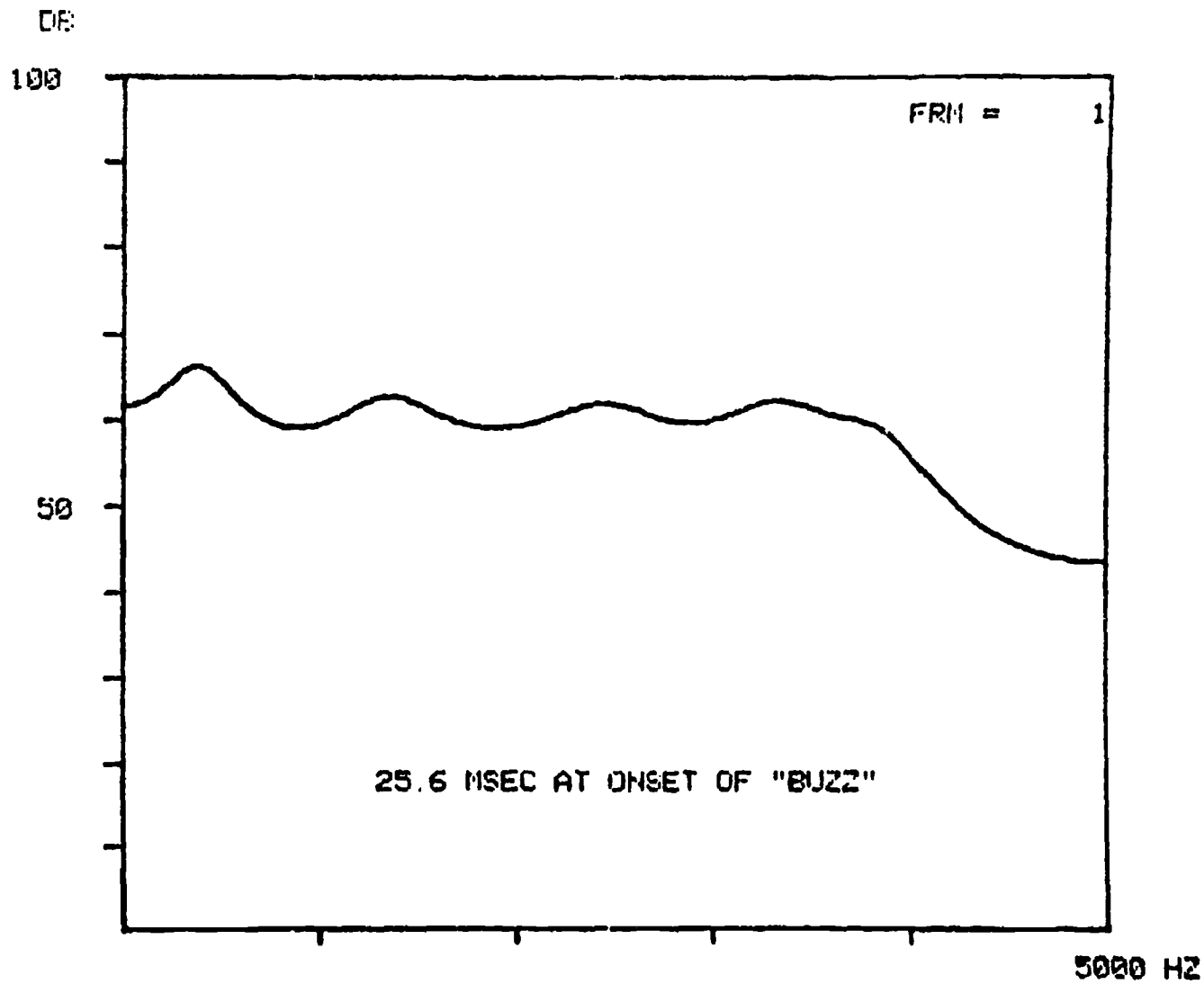


Figure 4. Smoothed spectrum of 25.6 msec at the onset of Stimulus 9 ("buzz").

for a good example of the "buzz", and the intervening buttons were used to identify variations between those endpoints.

Results and Discussion

Mean rating functions were computed for the control and anchor conditions for both the /bae/-anchor and "buzz"-anchor groups. The rating functions for the /bae/-anchor group are shown in Figure 5 and the rating functions for the "buzz"-anchor group are shown in Figure 6. Each point represents the mean of at least 200 judgments. Category boundaries in all conditions were determined by linear interpolation between the points on either side of the boundary.

Insert Figure 5 and Figure 6 about here

Both anchors produced contrast effects in the perception of the speech-nonspeech test series. Anchoring with the /bae/ produced a significant shift of the category boundary by .66 stimulus units (33 Hz of bandwidth) toward the /bae/ endpoint ($t(9) = 2.97$, $p < .02$, for a two-tailed test). Similarly, the "buzz" anchor caused a significant change in the placement of the category boundary by .92 stimulus units (46 Hz of bandwidth) towards the "buzz" end of the series ($t(9) = 3.72$, $p < .01$, for a two-tailed test).

Anchoring the perception of the speech-nonspeech continuum produced contrast effects which closely resemble the adaptation results reported by Remez (1979, 1980). Furthermore, these anchoring results were obtained even though the anchored random orders were constrained to prevent adaptation by long sequences of anchors. Thus, these anchoring results effectively eliminate the empirical basis for the argument against feature detectors made by Remez. Instead of affecting early sensory stages of auditory feature processing, the adaptors used by Remez may have served as a judgmental context that biased perception of the speech-nonspeech test series.

Remez (1979, 1980) has claimed that feature detector theories cannot account for adaptation of the speech/nonspeech distinction. Phonetic feature detector theories would have to add this distinction as a phonetic property. Clearly, this addition would not be reasonable since the putative role for phonetic feature detectors is to differentially classify phonetic segments and not distinguish speech from nonspeech. It is also hard to understand why the set of auditory feature detectors proposed for phonetic processing should include detectors for formant bandwidth, as this dimension does not directly cue any phonetic feature. Indeed, considering the adaptation experiments alone, feature detector theories cannot reasonably account for the speech/nonspeech adaptation results. But when the present anchoring results are considered, there is no need for feature detector theories to explain the adaptation effects reported by Remez. The perceptual plasticity shown by subjects identifying stimuli in the speech-nonspeech continuum is beyond the domain of feature detector theories. It is up to theories of psychophysical judgment (Helson, 1964; Parducci, 1965) or probability learning (e.g., Anderson et al., 1977) to explain these contrast effects. Such theories operate at a higher cognitive level than the auditory feature detectors proposed as the first stage of phonetic processing (see Pisoni

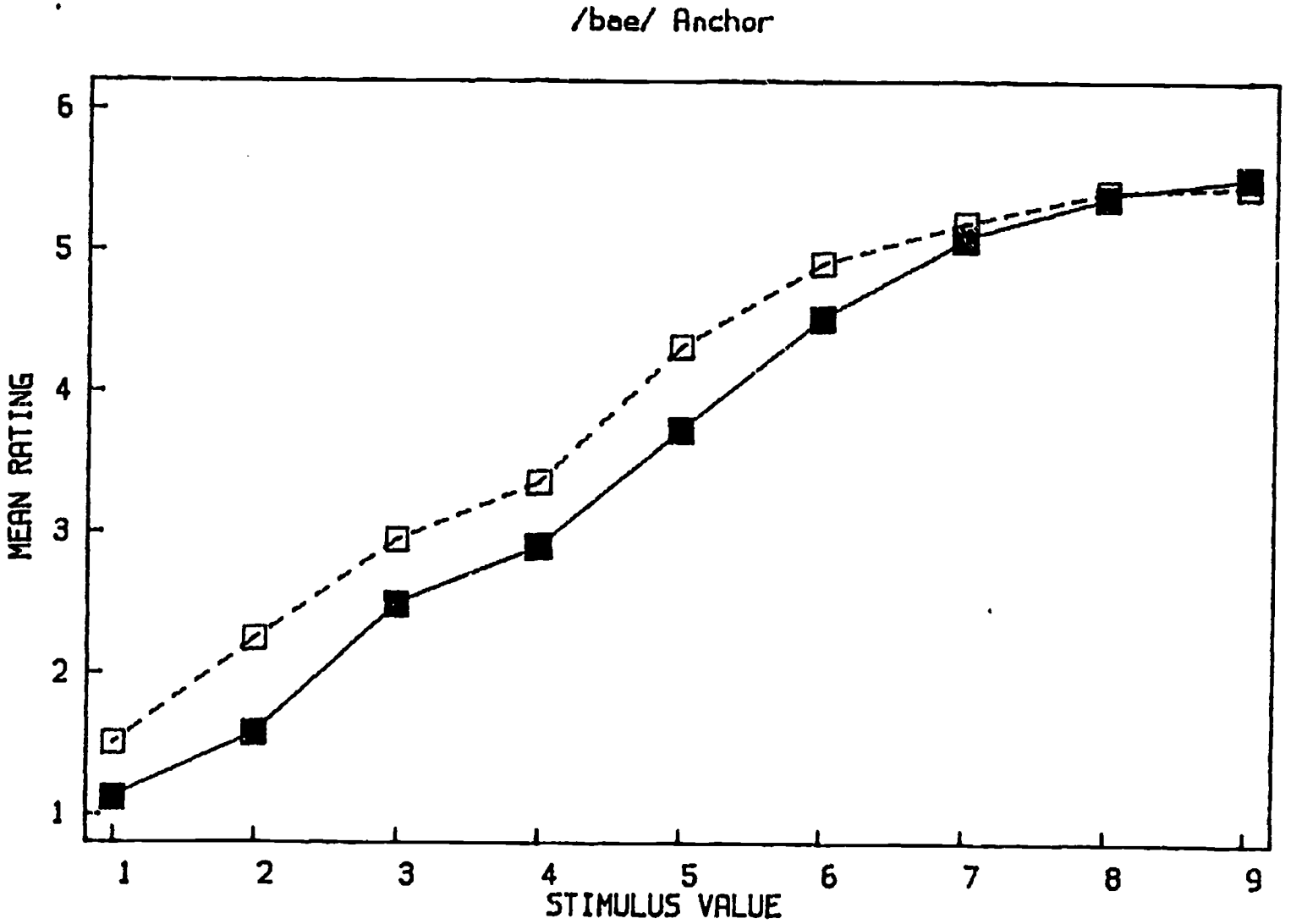


Figure 5. The effects of anchoring with the /bae/ endpoint (Stimulus 1). Mean ratings from the control condition are shown by the solid line and mean ratings from the condition anchored by Stimulus 1 are shown with the dashed line.

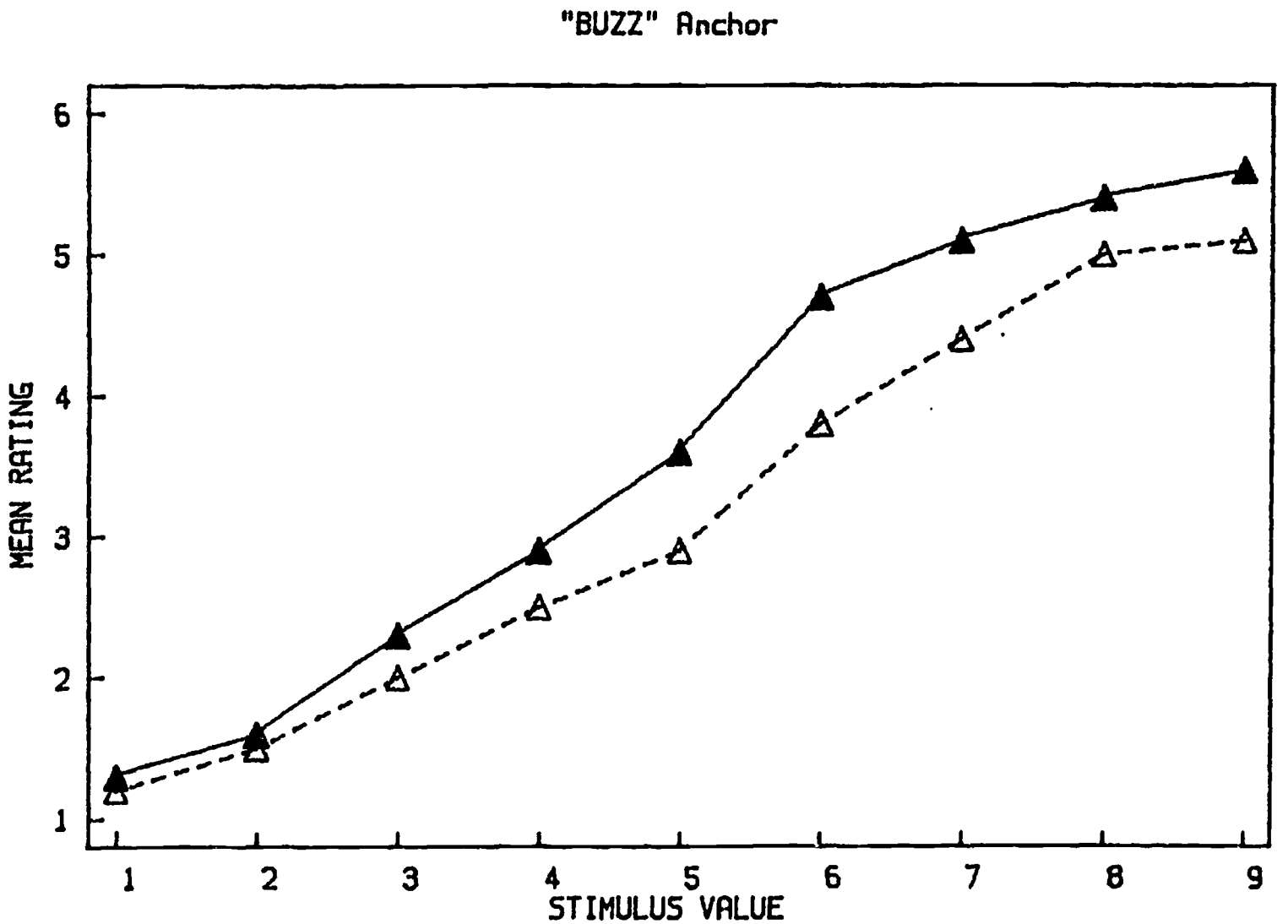


Figure 6. The effects of anchoring with the "buzz" endpoint (Stimulus 9). Mean ratings from the control condition are shown by the solid line and mean ratings from the condition anchored by Stimulus 9 are shown by the dashed line.

& Sawusch, 1975). Moreover, the adaptation results no longer suggest an unbounded proliferation of auditory feature detectors because sensory adaptation effects (as opposed to judgmental contrast effects) have not been unequivocally demonstrated for acoustic cue dimensions that are irrelevant to phonetic processing. For any perceptual dimension not directly involved in phonetic labeling, it may be that contrast effects are mediated by changes in higher-order judgmental (decision) mechanisms instead of fatigue of lower-level auditory feature detectors.

Beyond the specific implications of these anchoring results, there is a more general methodological significance to the present experiment: The conclusions drawn from adaptation experiments alone can be misleading. The extra presentations of an adaptor may have a number of perceptual consequences at different stages of processing. The adapting sequence could desensitize low-level auditory feature detectors and also induce a response bias in higher-level labeling processes. In order to tease apart these different effects, it is necessary to use experimental designs that are more complicated than have typically been employed in selective adaptation research. The effects of anchoring or successive contrast procedures must be compared to selective adaptation results.

Cooper (1979) has made a similar argument for the use of cross-series adaptation experiments using speech. In cross-series adaptation, the effect of an adaptor drawn from the test series is compared to the effects of adaptors not contained in the test series. To the extent that cross-series and within-series effects are the same, it can be claimed that the different adaptors share a common locus of processing. However, cross-series adaptation alone is not sufficient to assure that the shared locus of processing is a common set of feature detectors. The shared locus could instead be some higher-level phonetic decision mechanism.

For instance, nonspeech adaptation of the perception of a test series varying between phonetic endpoints has been used as evidence for complex auditory feature detectors (Samuel & Newport, 1979). Unlikely though it may seem, these results might have been produced by judgmental anchoring due to some dimension of perceptual similarity (e.g., periodicity) between nonspeech adaptors and speech test items. This explanation could be ruled out by a simple anchoring experiment or successive contrast procedure.

The converse argument has also been made; to the extent that cross-series adaptation is not found, the perception of the test stimuli and the cross-series adaptors must be mediated by different sets of detectors. Based on the logic of selective adaptation, Remez et al. (1980) have claimed that the lack of cross-series adaptation between similarly perceived stimuli would suggest a senseless proliferation of feature detectors. Pisoni (1980) conducted both within-series and cross-series adaptation experiments with analogous speech and nonspeech test continua. The speech series varied in voice-onset time (i.e., voicing), while the nonspeech stimuli were a set of tone pairs varying in tone-onset time. Even though both series were perceived categorically, with similar category boundaries (see Pisoni, 1977), no cross-series adaptation was found (Pisoni, 1980); the tone adaptors did not affect perception of the voicing dimension and the speech adaptors did not affect labeling of the tones. Rather than attribute the separate within-series adaptation effects found for these stimuli to different sets of detectors -- one set for VOT and one set for TOT

(cf. Remez et al., 1980) -- it is possible that the adaptation effects for the nonspeech stimuli were produced by judgmental contrast. If this were the case, anchoring effects might be predicted for the tone stimuli but not for the speech stimuli.

General Discussion

In order to determine the locus of adaptation, it is necessary to compare the effects of selective adaptation with results from other contrast-inducing procedures. When the influence of higher-level perceptual processes can be eliminated, selective adaptation seems to depend on the match between the acoustic structures of the adaptor and test stimuli (e.g., Sawusch & Jusczyk, 1981). Minimally this suggests that selective adaptation affects low-level auditory coding mechanisms (Simon & Studdert-Kennedy, 1978). Inferences about the specific nature of these mechanisms depend upon the formulation of a psychophysical linking hypothesis (see Weisstein, 1973). For feature detector theories, the linking hypothesis specifies that the salience of a particular perceptual feature relates to the firing rate of the associated feature detector. Prolonged excitation of one detector reduces its sensitivity to subsequent input, thus decreasing the perceptual salience of the mediated feature. It is by this mechanism of desensitization that repeated exposure to an adaptor modifies perception of test stimuli.

This linking hypothesis forms the theoretical foundation for interpreting selective adaptation research. Predictions derived from this hypothesis concerning the effects of adaptation on perceptual salience of phonetic exemplars have been supported (Miller, 1975; Sawusch, 1976a). Recent comparisons of selective adaptation with other contrast-inducing procedures (Sawusch & Jusczyk, 1981; Sawusch & Pisoni, Note 1; Sawusch & Nusbaum, Note 2), including the present experiment, have also found that auditory feature adaptation seems to occur under the conditions dictated by the psychophysical linking hypothesis. In fact, there do not appear to be any remaining empirical arguments against the use of selective adaptation as a paradigm for investigating phoneme perception.

However, feature detector models of speech perception have also been criticized on theoretical grounds. There are basically three theoretical criticisms of feature detectors that have been raised. First, Remez (1979) has claimed that the tuning of feature detectors must be context-sensitive to properly extract phonetic information from the "encoded" waveform (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). The implication is that a feature detector would need to modify its sensitivity for each different context that is presented in order to properly respond to its target feature. This ability would make feature detectors intelligent little homunculi entailing more complexity than is normally attributed to these neural units. Contrary to this assertion however, it is not necessary for feature detectors to be self-tuning for different contexts. Current conceptions of feature detector theories (Ades, 1976; Cooper, 1979; Eimas & Miller, 1978; Sawusch, 1976b; Searle, Jacobson, & Kimberley, 1980) propose that the first stage of processing extracts context-dependent auditory features (see Sawusch, 1977). (One example of a context dependent feature is a rising frequency transition that occurs within a specific frequency range. This would imply the existence of a different rising transition detector for each different frequency range of sensitivity.) These context-dependent features form an auditory substrate for subsequent

categorization processes. By using many banks of auditory feature detectors tuned a priori to different contexts, the need for self-modifying detectors is eliminated. Context sensitivity in this type of system is achieved by the higher-level perceptual mechanisms which presumably integrate auditory information over frequency and time. (Also, it is conceivable that these categorization processes might even feed back to the early stages of auditory feature encoding to "normalize" the tuning functions for different talkers and contexts.)

A second criticism of feature detector theories, raised by Studdert-Kennedy (1977), concerns the mapping of auditory features onto phonetic categories. His assertion is simply that there are no known auditory principles which can govern the integration of diverse auditory cues into phonetic percepts (cf. Fant, 1967). This is, without a doubt, absolutely true, but it really does not constitute a serious problem. The investigation of the auditory factors mediating speech perception is still a relatively new endeavor (see Pastore, 1981), so these currently unknown principles may be discovered yet. However, it seems more likely that the auditory-phonetic mapping rules have a phonetic -- not auditory -- basis. If the auditory property detectors (specialized for phonetic perception) evolved together with the neuromotor system for speech production (see Lieberman, 1973), many of the mapping rules for deriving phonetic percepts from auditory features may have a partially innate phonetic basis. Further, these auditory-to-phonetic integration principles may be modified developmentally (cf. Jusczyk, 1981) as the listener learns which acoustic cues are associated together in the production of phonetic segements. As a consequence, there is no need to specify auditory principles of feature integration since an auditory feature detector system could produce percepts based on phonetic integration rules that take into account the acoustic consequences of speech production.

Finally, Studdert-Kennedy (1982) has asserted that feature detector theories are "tautology, not explanation" (p. 225). The claim is that feature detector theories make a descriptive property of language into a perceptual mechanism. Of course, the descriptive nature of linguistic features means that such features serve a communicative function. A phonetician can identify a phonetic segment from its phonetic features alone; for example, a voiced labial stop consonant can be identified as [b]. In a similar way, the human perceptual system, if endowed with appropriate feature-to-segment mapping rules, should be able to classify phonetic segments given only phonetic features or even acoustic features (see Cole, Rudnicky, Zue, & Reddy, 1980, for evidence that phonetic segments can be labeled using the acoustic features in spectrograms). It is hard to see the theoretical importance of the distinction Studdert-Kennedy (1982) makes between a descriptive attribute and a constituent property of language. Instead of saying that feature detectors decompose an utterance into constituent features, it could be stated that feature detectors describe speech on various phonetic dimensions. These are simply two different perspectives on the same process, just as Studdert-Kennedy's attribute/constituent distinction represents two views of the same linguistic property. This is not a functional difference. Regardless of whether the perceptual system describes speech using phonetic attributes or analyzes speech into constituent features, the result will be the same -- segmentation and labeling of speech as phonemes.

There is a large body of evidence which indicates that humans perceive and represent the phonetic features of speech (see Pisoni, 1981; Studdert-Kennedy, 1976, for reviews), or at least auditory correlates of those features (see Klatt,

1980b). Thus, the issue does not seem to be the perceptual reality of phonetic features, but rather the perceptual function of those properties. While this may seem to be a moot point, given the evidence against phonetic feature detectors (Ades, 1976; Eimas & Miller, 1978), the basic issue has been extended to encompass auditory feature detectors as well (Repp, 1982). In essence, the explanatory power of all feature detector theories of speech perception has been questioned.

One way to approach this question is to compare feature detector theory with an alternative approach such as motor theory (Liberman, Cooper, Harris, & MacNeilage, 1962). The fine details of feature detector theory have changed considerably from its original conception (Abbs & Sussman, 1971) to its current form (see Cooper, 1979; Eimas & Miller, 1978). But the supporting framework of this position has remained the psychophysical linking hypothesis discussed previously. The details of this hypothesis have been sufficiently explicit to provide a host of testable predictions (see Ades, 1976; Cooper, 1975, 1979; Eimas & Miller, 1978, for reviews). In comparison, both the fine details and the theoretical framework for motor theory have been modified over time. Initially, motor theory asserted that phonetic perception was accomplished by active reference to neuromuscular invariants (Liberman, Cooper, Harris, MacNeilage, & Studdert-Kennedy, 1967). The proposition that neuromuscular invariants exist was the only explicitly testable hypothesis generated by motor theory and such invariants were never found (see MacNeilage, 1970, for a discussion). At present, the premise of motor theory seems to be that phoneme perception must take into account the articulatory origins of speech (see Best, Morrongiello, & Robson, 1981; Repp, 1982). This is not a very distinctive theoretical position; that is, this statement does not distinguish motor theory from other theories of speech perception. Moreover, since motor theory no longer specifies the mechanisms by which speech perception is accomplished, it is general enough to account for speech perceived by eyes or by ears.

For example, an expert spectrogram reader can identify the phonemes in a spectrogram using knowledge about the acoustic consequences of speech production (see Cole et al., 1980). In this way, the spectrogram reader actively takes into account the articulatory origins of speech. But the current version of the motor theory of speech perception does not, despite the obvious differences, distinguish between the processes utilized by an expert spectrogram reader and the human (auditory) speech perceiving system. It is apparent that very different perceptual and cognitive mechanisms are employed in spectrogram reading and speech perception. Thus, a theory of speech perception should somehow differentiate perception in the auditory modality with its attendant perceptual phenomena (e.g., categorical perception) from perception in the visual modality.

Assuming that speech perception and production evolved together and are subject to similar developmental influences, an auditory feature based speech perceiving system would necessarily take into account the articulatory origins of speech at some level of description. This is because the effects of developmental tuning on the auditory system will be linked to the source of stimulation -- the vocal tract. Therefore, an auditory theory of speech perception could take into account the articulatory origins of speech without ever positing any sort of connection between the mechanisms used in production and perception. Thus, the motor theory of speech perception is evidently not theoretically distinct from other theories. In other words, it is motor theory that may be tautological since it fails to provide a distinctive explanation of speech perception.

Moreover, this explains the inability of motor theory to generate testable hypotheses. Support for motor theory comes from evidence that suggests that phoneme perception is mediated by a system specialized for that purpose (Best et al., 1981; Liberman, 1974, 1982; Repp 1982). However, this evidence also supports feature detector theories which propose the existence of auditory detectors that are specialized for phonetic perception. But while evidence in support of motor theory also can support feature detector theories, the converse is not true; there is evidence which favors feature detector theories that does not fit into the framework of motor theory. Feature detector theories have predicted that infants should have certain phonetic-perceptual abilities (Cutting & Eimas, 1975) and nonhuman animals should be able to learn to use the auditory cues in speech to emulate human phonetic perception. In fact, prearticulate infants do seem to use acoustic-phonetic information in a manner similar to adults (see Jusczyk, 1981). Also, nonhuman animals, without the articulatory systems for speech production (e.g., chinchillas), can learn to classify phonetic information despite variations in talkers and context (see Kuhl & Miller, 1978; Miller, 1977). These results can be explained by feature detector theories of speech perception (cf. Cooper, 1979; Eimas & Miller, 1978), while proponents of motor theory just seem to dismiss these data as irrelevant (Repp, 1982; Studdert-Kennedy, 1982).

Finally, there is no evidence that an articulatory-based theory is even sufficient to perform phonetic perception. Without a detailed specification of the mechanisms needed for articulatory reference in perception, there is no way to ascertain whether such a theory could work. How would the signal be segmented and compared with the articulatory system? At what level of analysis would such comparisons be made? Repp (1982) has posed similar questions regarding the need for some specification of the mechanisms of motor theory. These questions must be answered before the sufficiency of motor theory can be tested.

In comparison, feature detector theories have, in some cases, been explicit in answering questions about segmentation and labeling phonemes in speech. While the details may differ from model to model, the point is that the feature detector is a simple enough mechanism to be specifically described. The implementation of particular feature detector models as computer programs has demonstrated that these mechanisms are indeed sufficient to account for phoneme perception in consonant-vowel syllables (Sawusch, 1976b; Searle et al., 1980). Although this is far from the full range of human speech-perceiving abilities, it is a significant beginning. Certainly these simulations have demonstrated that feature detectors are capable of performing the task that has been attributed to them.

By comparing feature detector theory with motor theory it can be seen that feature detectors are far from tautology. Instead, feature detector theory provides a precise and testable explanation of phoneme perception. Further, this account has been shown to be sufficient for explaining phoneme perception; that is, feature detector models can actually take the waveform of a syllable as input and produce phonetic labels as output. In short, the theoretical criticisms of feature detector theory (Remez, 1979; Repp, 1982; Studdert-Kennedy, 1982) do not warrant discarding this theoretical approach. Earlier research has already refuted empirical arguments that selective adaptation could be explained by mechanisms other than feature detectors (see Sawusch & Jusczyk, 1981; Sawusch & Pisoni, Note 1; Sawusch & Nusbaum, Note 2). Moreover, the present experiment has demonstrated that, contrary to the claims made by Remez (1979, 1980), the

hypothetical set of auditory feature detectors is not arbitrarily extensible. Thus, at present, the experimental evidence and theoretical arguments seem to support feature detector models of speech perception.

Reference Notes

1. Sawusch, J. R., & Pisoni, D. B. Anchoring, contrast effects, and the perception of speech. Manuscript in preparation.

2. Sawusch, J. R., & Nusbaum, H. C. Auditory and phonetic processes in speech. Paper presented at the 22nd meeting of the Psychonomic Society, Philadelphia, November, 1981.

3. Sawusch, J. R. Personal communication, March 12, 1982.

4. Kewley-Port, D. KLTEXC: Executive program to implement the KLATT software speech synthesizer. (Research on Speech Perception Progress Report 5, pp. 327-346). Bloomington, Ind.: Department of Psychology, Indiana University, 1978.

References

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

Ades, A. E. How phonetic is selective adaptation? Experiments on syllable position and vowel environment. Perception & Psychophysics, 1974, 16, 61-67.

Ades, A. E. Adapting the property detectors for speech perception. In R. Wales & E. Walker (Eds.), New approaches to language mechanisms. Amsterdam: North-Holland, 1976.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. Distinctive features, categorical perception, and probability learning: Some applications of a neural model. Psychological Review, 1977, 84, 413-451.

Best, C. J., Morrongiello, B., & Robson, R. Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception & Psychophysics, 1981, 29, 191-211.

Blumstein, S. E., & 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.

Bryant, J. S. Feature detection process in speech perception. Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 610-620.

Cole, R. A., Ruddnicky, A. I., Zue, V., & Reddy, D. R. Speech as patterns on paper. In R. A. Cole (Ed.), Perception and production of fluent speech. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1980.

Cooper, W. E. Selective adaptation to speech. In F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman, & D. B. Pisoni (Eds.), Cognitive theory (Vol. 1). Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1975.

Cooper, W. E. Speech perception and production. Norwood, New Jersey: Ablex, 1979.

Cuddy, L. L., Pinn, J., & Simons, E. Anchor effects with biased probability of occurrence in absolute judgment of pitch. Journal of Experimental Psychology, 1973, 100, 218-220.

Cutting, J. E., & Eimas, P. D. Phonetic feature analyzers and the processing of speech in infants. In J. F. Kavanagh & J. E. Cutting (Eds.), The role of speech in language. Cambridge, Mass.: MIT Press, 1975.

Diehl, R. L. Feature detectors for speech: A critical reappraisal. Psychological Bulletin, 1981, 89, 1-18.

Diehl, R. L., Elman, J. L., & McCusker, S. B. Contrast effects on stop consonant identification. Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 599-609.

Diehl, R. L., Lang, M., & Parker, E. M. A further parallel between selective adaptation and contrast. Journal of Experimental Psychology: Human Perception and Performance, 1980, 6, 24-44.

Eimas, P. D., & Corbit, J. D. Selective adaptation of linguistic feature detectors. Cognitive Psychology, 1973,4, 99-109.

Eimas, P. D., & Miller, J. L. Effects of selective adaptation on the perception of speech and visual patterns: Evidence for feature detectors. In R. D. Walk & H. L. Pick (Eds.), Perception and experience. New York: Plenum Press, 1978.

Elman, J. L. Perceptual origins of the phoneme boundary effect and selective adaptation to speech: A signal detection analysis. Journal of the Acoustical Society of America, 1979, 65, 190-207.

Evans, E. F. Neural responses for the detection of acoustic patterns and for sound localization. In F. O. Schmitt & F. G. Worden (Eds.), The neurosciences: Third study program. Cambridge, Mass.: MIT Press, 1974.

Fant, G. Auditory patterns of speech. In W. Walthen-Dunn (Ed.), Models for the perception of speech and visual form. Cambridge, Mass.: MIT Press, 1967.

Helson, H. Adaptation-level theory: An experimental and sytematic approach to behavior. New York: Harper, 1964.

Helson, H., & Kozaki, A. Anchor effects using numerical estimates of simple dot patterns. Perception & Psychophysics, 1968, 4, 163-164.

Jusczyk, P. W. Infant speech perception. In P. D. Eimas & J. L. Miller (Eds.), Perspectives on the study of speech. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1981.

Klatt, D. H. Speech recognition: A model of acoustic-phonetic analysis and lexical access. In R. A. Cole (Ed.), Perception and production of fluent speech. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1980. (a)

Klatt, D. H. Software for a cascade/parallel formant synthesizer, Journal of the Acoustical Society of America, 1980, 67, 971-995. (b)

Kuhl, P. K., & Miller, J. D. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. Journal of the Acoustical Society of America, 1978, 63, 905-917.

Liberman, A. M. The grammars of speech and language. Cognitive Psychology, 1970, 1, 301-323.

Liberman, A. M. The specialization of the language hemisphere. In F. O. Schmitt & F. G. Worden (Eds.), The neurosciences: Third study program. Cambridge, Mass.: MIT Press, 1974

Liberman, A. M. On finding that speech is special. American Psychologist, 1982, 37, 148-167.

Liberman, A. M., Cooper, F. S., Harris, K. S., & MacNeilage, P. F. A motor theory of speech perception. Proceedings of the speech communication seminar, Stockholm, 1962.

Liberman, A. M., Cooper, F. S., Harris, K. S., MacNeilage, P. F., & Studdert-Kennedy, M. Some observations on a model for speech perception. In W. Walthen-Dunn (Ed.), Models for the perception of speech and visual form. Cambridge, Mass.: MIT Press, 1967.

Liberman, A. M., Cooper, F. S., Shankweiler, D. S., & Studdert-Kennedy, M. Perception of the speech code. Psychological Review, 1967, 74, 431-461.

Liberman, A. M., & Studdert-Kennedy, M. Phonetic perception. In R. Held, H. W. Leibowitz, & H. -L. Teuber (Eds.), Handbook of sensory physiology (Vol. 8): Perception. New York: Springer-Verlag, 1978.

Lieberman, P. On the evolution of language: A unified view. Cognition, 1973, 2, 55-94.

MacNeilage, P. F. Motor control of serial ordering of speech. Psychological Review, 1970, 77, 182-196.

Markel, J. D., & Gray, A. H. Linear prediction of speech. New york: Springer-Verlag, 1976.

Miller, J. D. Perception of speech sounds by animals: Evidence for speech processing by mammalian auditory mechanisms. In T. H. Bullock (Ed.), Recognition of complex acoustic signals. Berlin: Dahlem Konferenzen, 1977.

Miller, J. L. Properties of feature detectors for speech: Evidence from the effects of selective adaptation on dichotic listening. Perception & Psychophysics, 1975, 18, 389-397.

Morse, P. A., Kass, J. E., & Turkienicz, R. Selective adaptation of vowels. Perception & Psychophysics, 1976, 19, 137-143.

Parducci, A. Range-frequency compromise in judgment. Psychological Monographs, 1963, 77, 2, 1-50.

Parducci, A. Category judgment: A range-frequency model. Psychological Review, 1965, 72, 407-418.

Pastore, R. E. Possible psychoacoustic factors in speech perception. In P. D. Eimas & J. L. Miller (Eds.), Perspectives on the study of speech. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1981.

Pisoni, D. B. Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. Journal of the Acoustical Society of America, 1977, 61, 1352-1361.

Pisoni, D. B. Adaptation of the relative onset time of two-component tones. Perception & Psychophysics, 1980, 28, 337-346.

Pisoni, D. B. Phonetic representations and lexical access. Journal of the Acoustical Society of America, 1981, 69, S32. (Abstract)

Pisoni, D. B., & Sawusch, J. R. Some stages of processing in speech perception. In A. Cohen & S. G. Nooteboom (Eds.), Structure and process in speech perception. Heidelberg: Springer-Verlag, 1975.

Pisoni, D. B., & Tash, J. Auditory property detectors and processing place features in stop consonants. Perception & Psychophysics, 1975, 18, 401-408.

Remez, R. E. Adaptation of the category boundary between speech and non-speech: A case against feature detectors. Cognitive Psychology, 1979, 11, 38-57.

Remez, R. E. Susceptibility of a stop consonant to adaptation on a speech-nonspeech continuum: Further evidence against feature detectors in speech perception. Perception & Psychophysics, 1980, 27, 17-23.

Remez, R. E., Cutting, J. E., & Studdert-Kennedy, M. Cross-series adaptation using song and string. Perception & Psychophysics, 1980, 27, 524-530.

Repp, B. H. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Review, 1982, 92, 81-110.

Rosen, S. M. Range and frequency effects in consonant categorization. Journal of Phonetics, 1979, 7, 393-402.

Samuel, A. G., & Newport, E. L. Adaptation of speech by nonspeech: Evidence for complex acoustic cue detectors. Journal of Experimental Psychology: Human Perception and Performance, 1979, 5, 563-578.

Sawusch, J. R. Selective adaptation effects on end-point stimuli in a speech series. Perception & Psychophysics, 1976, 20, 61-65. (a)

Sawusch, J. R. The structure and flow of information in speech perception: Evidence from selective adaptation of stop consonants. Unpublished doctoral dissertation, Indiana University, 1976. (b)

Sawusch, J. R. Peripheral and central processes in selective adaptation of place of articulation in stop consonants. Journal of the Acoustical Society of America, 1977, 62, 738-750.

Sawusch, J. R., & Jusczyk, P. W. Adaptation and contrast in the perception of voicing. Journal of Experimental Psychology: Human Perception and Performance, 1981, 7, 408-421.

Sawusch, J. R., & Nusbaum, H. C. contextual effects in vowel perception I: Anchor-induced contrast effects. Perception & Psychophysics, 1979, 25, 292-302.

Scheich, H. Central processing of complex sounds and feature analysis. In T. H. Bullock (Ed.), Recognition of complex acoustic signals. Berlin: Dahlem Konferenzen, 1977.

Searle, C. L., Jacobson, J. Z., & Kimberley, B. P. Speech as patterns in 3-space of time and frequency. In R. A. Cole (Ed.), Perception and production of fluent speech. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1980.

Simon, H. J., & Studdert-Kennedy, M. Selective anchoring and adaptation of phonetic and nonphonetic continua. Journal of the Acoustical Society of America, 1978, 64, 1338-1357.

Studdert-Kennedy, M. Speech perception. In N. J. Lass (Ed.), Contemporary issues in experimental phonetics. New York: Academic Press, 1976.

Studdert-Kennedy, M. Universals in phonetic structure and their role in linguistic communication. In T. H. Bullock (Ed.), Recognition of complex acoustic signals. Berlin: Dahlem Konferenzen, 1977.

Studdert-Kennedy, M. The emergence of phonetic structure. Cognition, 1981, 10, 301-306.

Studdert-Kennedy, M. A note on the biology of speech perception. In J. Mehler, E. C. T. Walker, & M. Garrett (Eds.), Perspectives on mental representation. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1982.

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., & Cooper, F. S. Motor theory of speech perception: A reply to Lane's critical review. Psychological Review, 1970, 77, 234-249.

Tartter, V. C., & Eimas, P. D. The role of auditory feature detectors in the perception of speech. Perception & Psychophysics, 1975, 18, 293-298.

Weisstein, N. Beyond the yellow Volkswagen detector and the grandmother cell: A general strategy for the exploration of operations in human pattern recognition. In R. L. Solso (Ed.), Contemporary issues in cognitive psychology: The Loyola Symposium. Washington, D. C.: V. H. Winston, 1973.

II. SHORT REPORTS AND WORK-IN PROGRESS

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

Perceiving Durations of Silence in a Nonspeech Context*

Howard C. Nusbaum

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

*This research was supported by NIH training grant NS-07134 and NIH research grant NS-12179 to Indiana University. The author would like to thank David B. Pisoni for a number of helpful comments and suggestions.

Abstract

Previous research has demonstrated that the interpretation of stop consonant closure as voiced or voiceless depends both on the duration of the closure and the duration of the preceding syllable. This interaction between acoustic cues could result either from auditory contrast in duration judgment or from a phonetic integration process. The present experiment tested these alternatives by presenting nonspeech analogs of VCV stimuli to subjects in a duration judgment task. Subjects identified two nonspeech (sinewave) test series as containing a "short interval" or a "long interval" of silence. Each series consisted of nine stimuli that varied in the duration of silence between two sinewave pairs. In one series, the silence was preceded by a short tone pair and, in the other series, a long tone pair preceded the silence. In contrast to the perception of medial voicing, fewer "short interval" responses were made in the context of the long tone pair than in the context of the short tone pair. Subjects identifying the duration of silence in nonspeech context do not show auditory contrast effects. Thus, the integration of closure duration and the duration of the preceding syllable in perception of medial voicing appears to have a phonetic basis rather than an auditory basis.

Perceiving Durations of Silence in a Nonspeech Context

One of the fundamental questions in speech research concerns the way spectro-temporal cues in the speech waveform are integrated to form phonetic percepts (see Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). In general, the early studies investigating phonetic perception concluded that perceptual processing of speech is mediated by a specialized mechanism (see Liberman, 1970, 1974; Liberman et al., 1967; Liberman & Studdert-Kennedy, 1978, for reviews). The claim was that coarticulatory effects in speech production encode phonemes into sound, distributing the acoustic representation of a phoneme in time so that the acoustic cues to different phonemes overlap and interact with each other. To perceptually decode or unravel phonemes from this tapestry of sound requires knowledge of the encoding process (Liberman, 1970). Thus, according to this argument, phonetic perception must entail (at some level) knowledge about speech production and, of course, this knowledge is only relevant to perception of speech sounds. The utilization of this production knowledge distinguishes speech perception from perception of other auditory signals.

Unfortunately, the evidence offered in support of a specialized speech processor has never been conclusive (see Lane, 1965; Schouten, 1980). Some investigators have argued that speech perception is carried out by generic auditory processes that map acoustic cues onto phonetic features (cf. Fant, 1967). According to this view, phonetic perception results from psychoacoustic constraints -- that is, purely auditory (nonlinguistic) principles operating on the speech waveform (cf. Pastore, 1981). Alternatively, phonetic perception could represent a direct associative response to the acoustic information in the speech signal (Anderson, Silverstein, Ritz, & Jones, 1977). Neither of these alternatives invokes or requires any specialized knowledge about the articulatory processes that encode phonemes into sound.

More recently, however, new evidence for a specialized speech processor has come from phonetic cue trading experiments (see Liberman, 1982; Repp, 1982). These experiments have investigated the extent to which extremely different acoustic attributes are treated as perceptually equivalent in cueing phonemes. For example, Best, Morrongiello, and Robson (1981) examined the interaction of silence duration and the frequency extent of a formant transition in the distinction between "say" and "stay." Using a discrimination procedure, Best et al. found that in speech, listeners treated longer durations of silence following the initial /s/ as functionally equivalent to a greater frequency extent of the first formant (F1) transition. In another experiment, Best et al. presented nonspeech (sinewave) analogs of the "say"-"stay" stimuli. Subjects who heard these analogs as speech treated silence duration and F1 transition extent as perceptually equivalent in discrimination. However, subjects who heard the analogs as nonspeech discriminated the stimuli using only one of the two available cues; they did not perceptually integrate the cues. Best et al. concluded that the basis for the trading relation between silence duration and F1 transition extent was phonetic and not auditory (also see Liberman, 1982; Repp, 1982). In other words, when perceived as speech, the acoustic cues were integrated into a unitary phonetic percept but when perceived as nonspeech, the cues could be attended to separately. Thus, acoustic cues in speech are perceived according to phonetic principles rather than auditory principles (Studdert-Kennedy, 1977).

Unfortunately, this conclusion may not generalize equally well to all phonetic distinctions. Miller and Liberman (1979) have shown that the interpretation of formant transition duration in the distinction between /b/ and /w/ is affected by the duration of the following vowel in a consonant-vowel (CV) syllable. Their findings indicate that both transition duration and vowel duration contribute to the stop-semivowel distinction. Miller and Liberman (1979) concluded that this contribution reflects a form of normalization for speaking rate that is special to speech (also see Miller, 1981). However, Carrell, Pisoni, and Gans (1980) have shown that exactly the same effects can be obtained with sinewave analogs of the syllables used by Miller and Liberman, even though these analogs are not perceived as speech. Since the interpretation of transition duration and steady-state (vowel or tone) duration produces the same pattern of results for speech and nonspeech, it cannot be claimed that the integration of these cues is somehow special or unique to speech perception. Instead, listeners may simply use the vowel duration as an "auditory ground" (cf. Simon & Studdert-Kennedy, 1976) against which transition duration is compared. Thus, the results obtained by Miller and Liberman can be explained as psychophysical anchoring (cf. Helson, 1964; Parducci, 1965).

Clearly, the integration of transition duration and vowel duration in the stop-semivowel distinction is mediated by a different mechanism than the process that integrates silence duration and F1 transition extent in the perception of "say" and "stay." At first glance, it is difficult to see why there should be two distinct mechanisms of cue integration. However, there are two important differences between the duration cues investigated by Miller and Liberman (1979) and Carrell et al. (1980) and the spectro-temporal cues in the "say"-"stay" distinction investigated by Best et al. (1981).

First, the two cues in the stop-semivowel distinction are both temporal cues -- transition duration and vowel duration. Thus, the listeners could contrastively compare these attributes, using vowel duration as a perceptual standard for judging transition duration. In contrast, for the "say"-"stay" distinction, silence duration is a temporal cue while F1 transition extent is a spectral cue. As a result, neither cue could serve as the basis for judging the other.

The second important difference between the pairs of cues underlying these distinctions is the difference in their respective perceptual functions. For the stop-semivowel distinction, vowel duration may serve as a cue to normalize differences in overall speaking rates (see Miller, 1981). However, rate normalization is not necessarily a perceptual function used only in processing speech. Rate normalization could be a more general auditory function serving to maintain perceptual constancy in the perception of other nonspeech signals such as music. In contrast, the juxtaposition of silence duration and F1 transition extent is a consequence of the production differences between "say" and "stay" (see Best et al., 1981). In other words, these cues actually form a single spectro-temporal unit resulting from constraints on speech production. Therefore, it should not be surprising that these cues are only perceptually equivalent in speech.

In short, cue trading may appear to have a phonetic basis because the two cues have a common origin in production or because the cues do not share a psychophysical dimension (e.g., one cue is spectral and one cue is temporal). To

distinguish these possibilities, it is necessary to examine the interaction of two cues that are produced together and share a common psychophysical attribute. Recently, Port and Dalby (1982) have demonstrated that the duration of an initial syllable interacts with perception of a subsequent stop closure interval as voiced or voiceless. In one experiment, several series of stimuli varying from "digger" to "dicker" were created by systematically increasing the closure interval from the medial /g/ endpoint (35 msec) to the medial /k/ endpoint (155 msec). The different series were produced by manipulating the duration of the stressed syllable preceding the stop closure. Thus, for each series, there was one syllable duration combined with nine closure durations. Port and Dalby found that longer initial syllables caused the closure to be perceived as more voiced (i.e., shorter in duration). Figure 1 shows the identification data for the series with the shortest initial syllable (shown by the solid line) and the series with the longest initial syllable (shown by the dashed line) from Port and Dalby (1982). This figure clearly shows that more /g/ responses were made to the test series with the long initial syllable and more /k/ responses were made to the series with the short initial syllable. Port and Dalby argued that this interaction occurred because listeners directly processed the two duration cues as a single perceptual unit -- the ratio of closure duration to preceding vowel (syllable) duration (C/V ratio). Part of the motivation for this suggestion comes from the claim made by Port (1981) that the duration of stop closure and the duration of the preceding vowel are articulated as a single unit. Thus, for medial voicing, the two cues are articulated together.

Insert Figure 1 about here

However, vowel duration and closure duration are also both temporal cues. It is entirely possible that the interaction between these cues is an auditory contrast effect resulting from judging closure duration by comparison with the preceding vowel. This would provide an auditory explanation of the interaction that does not require any knowledge of the association of the two cues in speech production. Thus, for medial voicing, the interaction could result from either an auditory process or a phonetic integration mechanism. Nonspeech analogs of the medial voicing cue can be constructed to test these alternatives. Perey and Pisoni (1980) have shown that subjects can accurately categorize the duration of a silent interval in sinewave analogs of medial voicing in vowel-consonant-vowel (VCV) syllables. If these nonspeech analogs produce the same type of effect found by Port and Dalby (1982) shown in Figure 1, this would suggest that these cues are integrated by a general auditory process rather than a specialized speech processor. However, if nonspeech analogs do not display the same effects as speech, then it would appear that the integration process is mediated by a mechanism that takes into account the phonetic association of acoustic cues during production (see Summerfield, 1982).

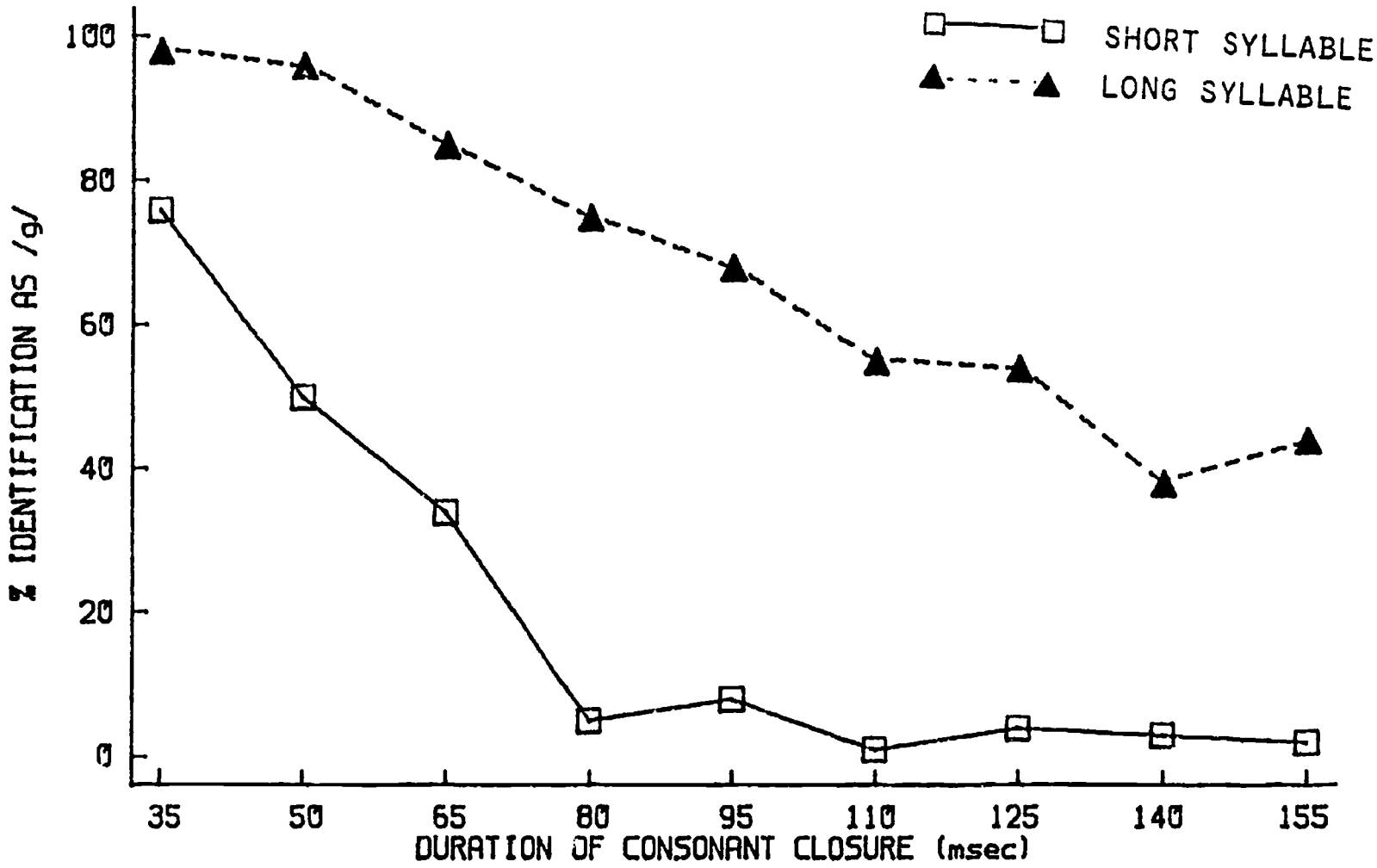


Figure 1. Identification of medial voicing in the "digger"-"dicker" series varying in closure interval (replotted from Port & Dalby, 1982). The solid line shows the percentage of /g/ responses for the series with the shortest initial syllable. The dashed line shows the percentage of /g/ responses for the series with the longest initial syllable.

Method

Subjects

The subjects were 15 undergraduate students at Indiana University. All subjects were right-handed native speakers of English, with no reported history of either speech or hearing disorder. The subjects participated as part of a course requirement.

Stimuli

Two series of nine nonspeech sounds each were used as stimuli. Each stimulus consisted of a pair of tones followed by a short interval of silence (representing the closure interval) followed by another tone pair. The duration of the first pair of tones was constant throughout each series and varied between the test series. For one test series (the short tone series), the duration of the initial pair of tones was 140 msec while for the other series (the long tone series), the duration of the first pair of tones was 260 msec. The duration of the final pair of tones was constant at 115 msec for all the stimuli. Within each series the duration of silence between the tone pairs varied in 15 msec steps from 35 msec for the first stimulus to 155 msec for the ninth stimulus. In the first tone pair of each stimulus, the frequency of one sinewave was held constant at 400 Hz for 100 msec in the short tone series and 220 msec in the long tone series, and then decreased to 380 Hz over the last 40 msec before the silence. The second sinewave component of this tone pair was set to 1800 Hz for 100 msec in the short tone series and 220 msec in the long tone series, and then increased to 2000 Hz over the last 40 msec before the silence. For the tone pair following the silence, one component increased in frequency from 380 Hz to 540 Hz over the first 15 msec of the tone and then was constant at 540 Hz for the final 100 msec. The second component started at 1600 Hz, fell to 1310 Hz over the first 15 msec of the tone and then was constant for the final 100 msec. The frequency transitions were included to simulate the formant transitions into the velar closure in medial position in the stimuli used by Port and Dalby (1982). The frequencies of the sinewaves were chosen to approximate the first two formants of Port and Dalby's "digger"-"dicker" stimuli excluding the initial consonant.

The stimuli were generated under computer control, presented in real-time through a 12 bit digital-to-analog converter, and low-pass filtered at 4.8 kHz. The sounds were presented binaurally through matched and calibrated Telephonics TDH-39 headphones. The intensity of the stimuli was set at 76 dB SPL.

Procedure

Subjects participated in a single 1 h session. Experimental sessions were conducted with small groups of two to six subjects each. Each session consisted of six blocks of trials. The first three blocks provided practice in categorizing the duration of silence between tone pairs as a "short interval" or a "long interval." In one practice block (10 trials), subjects identified five repetitions of the two endpoints (Stimulus 1 and Stimulus 9) of the short tone series; in a second practice block (10 trials), they identified the endpoints of the long tone series. The order of these blocks was determined randomly across groups. In the third practice block (40 trials), ten repetitions of each of the

four endpoints were presented in random order. On each practice trial, subjects were presented with an endpoint stimulus and were instructed to identify it as containing a "short interval" of silence or a "long interval" of silence by pressing the appropriately marked button on a response box that was interfaced to the computer. After all subjects responded on each trial, the computer indicated the correct response by illuminating a light over the appropriate label on the response boxes. This feedback was provided only on practice trials.

Following practice, subjects received three blocks of identification trials. In each of these blocks, five repetitions of each of the nine stimuli in the two test series were presented in random order for a total of 90 trials. Subjects responded to each stimulus by pressing a button labeled "short interval" or "long interval" on the response box. Each subject provided 15 responses to each stimulus in the long tone and short tone test series, excluding practice.

Results and Discussion

The mean percentage of "short interval" responses for each stimulus in both test series is shown in Figure 2. The identification function for the long tone series is shown by the dashed line whereas the identification function for the short tone series is shown by the solid line. Category boundaries between the "short interval" and "long interval" responses were determined by computing the mean of a logistic function fit to the identification data for each series for each subject. The category boundary for the long tone series occurred 1.34 stimulus units (20 msec) earlier (i.e., at a shorter duration of silence) than the boundary for the short tone series. This difference in the location of the category boundaries for the two series is significant ($t(14) = 2.43$, $p < .03$, for a two-tailed test). Thus, in the context of the long tone, silent intervals sounded longer than in the context of the short tone. Comparing these nonspeech identification functions with the speech identification functions obtained by Port and Dalby (1982) shows that the nonspeech functions are reversed relative to the speech functions. For speech stimuli, Port and Dalby found that the long syllable series was identified with more voiced responses (i.e., containing a shorter closure interval) than the short syllable series. However, for nonspeech, the long tone series was identified with fewer "short interval" responses than the short tone series.

Insert Figure 2 about here

Clearly, these results demonstrate that the perception of an interval of silence is not the same in speech and nonspeech contexts. The influence of the duration of a signal preceding a silent interval is different for speech and nonspeech. In speech, a long vowel or syllable preceding closure produces more voiced responses -- the closure sounds shorter. However, this is not due to an auditory contrast effect in judging the duration of the closure, since for nonspeech stimuli, an assimilation effect is obtained -- a long tone pair preceding silence caused the silence to sound longer. Thus, the perception of these acoustic attributes appears to be mediated by different processes in speech and nonspeech. This suggests that the interpretation of stop closure as voiced

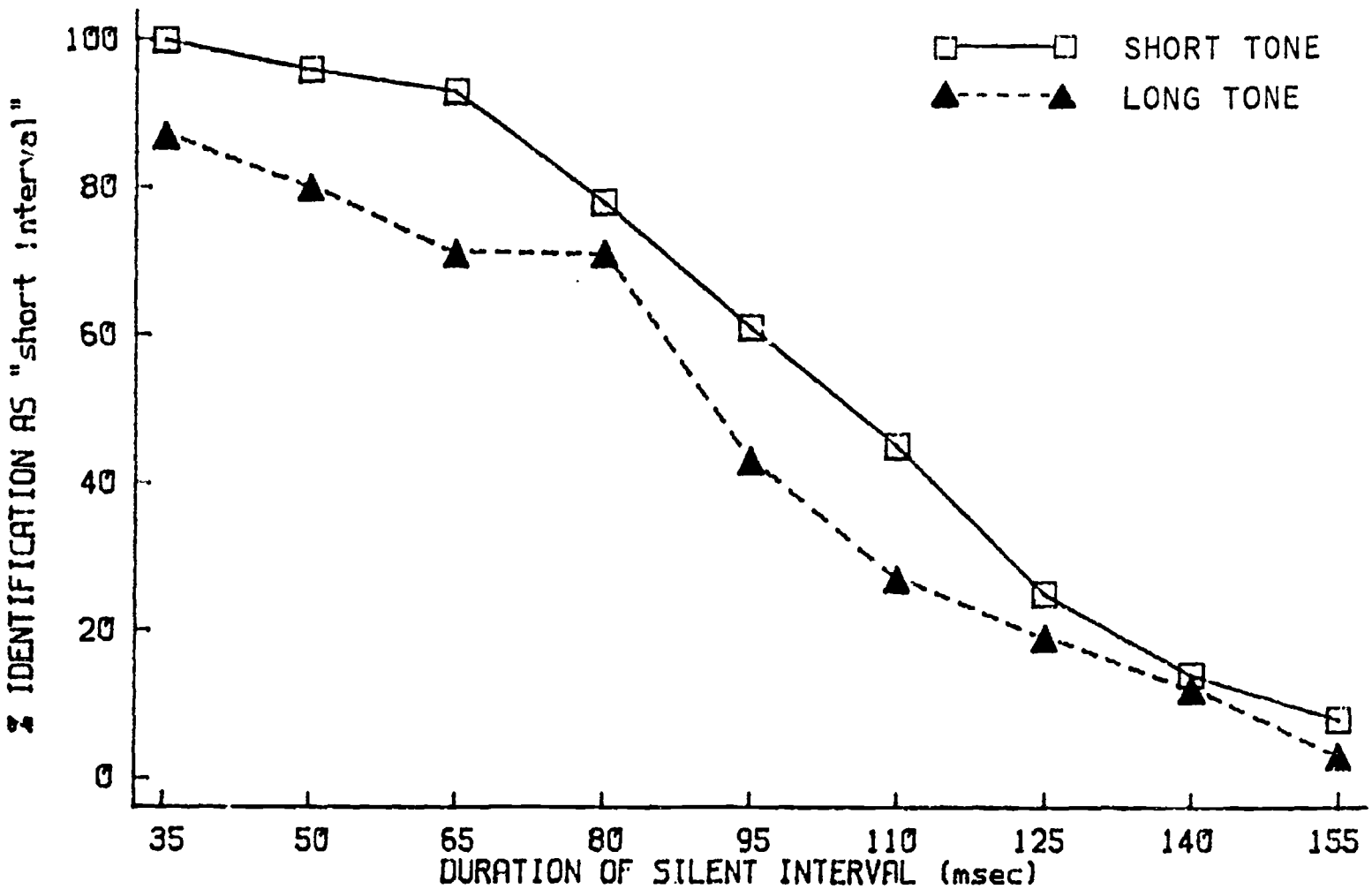


Figure 2. Identification of the duration of silence between two tone pairs. The solid line shows the percentage of "short interval" responses for the series with the short initial tone pair. The dashed line shows the percentage of "short interval" responses for the series with the long initial tone pair.

or voiceless following a syllable is not a generic auditory function. Rather, these temporal cues are apparently integrated in accordance with phonetic principles of perception. Precisely what these principles are remains to be determined in future research. However, for the present time it is clear that the perception of durations of silence is different in the two types of contexts implying that there are substantial differences in the underlying modes of processing speech and nonspeech signals.

References

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. Distinctive features, categorical perception, and probability learning: Some applications of a neural model. Psychological review, 1977, 84, 413-451.

Best, C. J., Morrongiello, B., & Robson, R. Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception & Psychophysics, 1981, 29, 191-211.

Carrell, T. D., Pisoni, D. B., & Gans, S. J. Perception of the duration of rapid spectrum changes: Evidence for context effects with speech and nonspeech signals. Paper presented at the 100th meeting of the Acoustical Society of America, Los Angeles, November, 1980.

Fant, G. Auditory patterns of speech. In W. Walthen-Dunn (Ed.), Models for the perception of speech and visual form. Cambridge, Mass.: MIT Press, 1967.

Helson, H. Adaptation-level theory: An experimental and systematic approach to behavior. New York: Harper, 1964.

Lane, H. L. The motor theory of speech perception: A critical review. Psychological Review, 1965, 72, 275-309.

Liberman, A. M. The grammars of speech and language. Cognitive Psychology, 1970, 1, 301-323.

Liberman, A. M. The specialization of the language hemisphere. In F. O. Schmitt & F. G. Worden (Eds.), The neurosciences: Third study program. Cambridge, Mass.: MIT Press, 1974

Liberman, A. M. On finding that speech is special. American Psychologist, 1982, 37, 148-167.

Liberman, A. M., Cooper, F. S., Shankweiler, D. S., & Studdert-Kennedy, M. Perception of the speech code. Psychological Review, 1967, 74, 431-461.

Liberman, A. M., & Studdert-Kennedy, M. Phonetic perception. In R. Held, H. W. Leibowitz, & H. -L. Teuber (Eds.), Handbook of sensory physiology (Vol. 8): Perception. New York: Springer-Verlag, 1978.

Miller, J. L. The effect of speaking rate on segmental distinctions: Acoustic variation and perceptual compensation. In P. D. Eimas & J. L. Miller (Eds.), Perspectives on the study of speech. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1981.

Miller, J. L., & Liberman, A. M. Some effects of later occurring information on the perception of stop consonant and semivowel. Perception & Psychophysics, 1979, 25, 457-465.

Parducci, A. Category judgment: A range-frequency model. Psychological Review, 1965, 72, 407-418.

Pastore, R. E. Possible psychoacoustic factors in speech perception. In P. D. Eimas & J. L. Miller (Eds.), Perspectives on the study of speech. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1981.

Perey, A. J., & Pisoni, D. B. Identification and discrimination of durations of silence in nonspeech signals. Research on Speech Perception Progress Report, Department of Psychology, Indiana University, 1980, 6, 235-270.

Port, R. F. Linguistic timing factors in combination. Journal of the Acoustical Society of America, 1981, 69, 262-274.

Port, R. F., & Dalby, J. Consonant/vowel ratio as a cue for voicing in English. Perception & Psychophysics, 1982, 32, 141-152.

Repp, B. H. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Review, 1982, 92, 81-110.

Schouten, M. E. H. The case against a speech mode of perception. Acta Psychologica, 1980, 44, 71-98.

Simon, H. J., & Studdert-Kennedy, M. Selective anchoring and adaptation of phonetic and nonphonetic continua. Journal of the Acoustical Society of America, 1978, 64, 1338-1357.

Studdert-Kennedy, M. Universals in phonetic structure and their role in linguistic communication. In T. H. Bullock (Ed.), Recognition of complex acoustic signals. Berlin: Dahlem Konferenzen, 1977.

Summerfield, Q. Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in stops. Journal of the Acoustical Society of America, 1982, 72, 51-61.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

Perception of Synthetic Speech by Children: A First Report*

Beth G. Greene and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47401

*This research was supported by NIMH grant MH-24027, NIH grant NS-12179 and NICHD grant HD-11915 to Indiana University at Bloomington. We appreciate the cooperation of the Monroe County School Corporation, especially the parents, staff and children at Lakeview School. We also thank Amy Ross, Maria Sera and Mary Buuck for their assistance in data collection. Special thanks go to Dr. Jared Bernstein of Telesensory Speech Systems, Inc. for providing the initial synthetic speech stimuli used in this study.

Abstract

Second grade children listened to natural and synthetic speech tokens and were required to identify a test word from among four alternatives in a picture pointing task. Subjects consistently showed higher performance when listening to natural speech. Presentation order significantly affected the results. The natural-synthetic blocked condition led to more errors overall than the synthetic-natural condition. In another study, fourth graders listened to natural and synthetic digit strings and were required to repeat the digit strings aloud to the experimenter exactly as heard. Again, there was a performance decrement when listening to synthetic speech even for highly familiar material such as digits. These results have implications for the design, selection and use of voice-response systems that are to be used in teaching and learning environments with children and adults. Synthetic speech may require additional processing capacity and attentional demands that may produce large decrements in other complex cognitive tasks that children are expected to perform routinely in educational settings.

Perception of Natural and Synthetic Speech by Children

Beth G. Greene and David B. Pisoni

High technology has encompassed all facets of life in the last two decades. A television screen has replaced the pacing professor in many college classrooms. Microcomputers are used to drill school children in spelling and elementary mathematics. Telephone numbers are checked through the telephone company's computerized system which relays information through re-synthesized natural speech. Thus, voice technology has become increasingly important as people of all ages receive information from synthetically generated speech.

As the number of products that incorporate voice output increases in the commercial and educational marketplace, it is becoming clear that there is a greater need for data about the comprehensibility and intelligibility of the synthetic or resynthesized speech employed in the product. Currently, a number of educational toys, such as "Speak and Spell" and "Touch and Talk" produced by Texas Instruments, are widely available. Children are using these toys both for entertainment and learning. And, an increasing number of personal computers are equipped with voice synthesis capabilities. Soon many schools will use computers and voice synthesis intensively for instructional purposes. However, at the present time, there is almost no information in the literature dealing with how young children respond (i.e., perceive and understand) computer generated speech, particularly meaningful connected speech.

Perception of Synthetic Speech

In a recent study, elementary school children were tested using several different speech synthesis systems (cf. Laddaga, Sanders and Suppes, 1981). Laddaga et al. studied the intelligibility of four speech synthesis systems: an MIT system, two Votrax models, and an LPC system of their own design. The speech of a male human talker served as a control condition. The first experiment examined the performance of first graders in recognizing individual letters by name as spoken by the various synthesis systems. The students did quite well: mean scores for each session ranged between 83% and 98% correct. The data were examined to see what letters caused difficulties for the systems. All the systems had difficulty with the letter Z, which was often heard as V. Votrax, MIT and the control system all had difficulty with N, which was heard as M. Both the MIT and LPC systems had a problem with G, which was heard as either B or D in both systems.

A second experiment examined the performance of fifth graders in recognizing initial and final consonants and consonant clusters. The scores on the word recognition experiment were again quite high: mean correct scores for each session ranged between 78% and 100%. The data analysis also focused on the specific consonant errors for each system. The results indicated that the LPC system led to confusions among the /th/, /s/, and /f/ sounds. Most of the problems with the MIT system occurred with consonant clusters rather than with singleton consonants. Votrax had problems with stops which were often dropped from initial position in a consonant cluster. The only serious problem for the human talker (i.e., control) was the /th/ sound which was often heard as /f/.

Laddaga et al. concluded that "the high probability of recognition of sounds shown by the letter and word experiments indicates that some form of synthetic speech is adequate for use in computer-assisted instruction (CAI) in initial reading" (p.395). The scores for the LPC and the MIT systems were generally well above 90% correct in both the letter and word experiments which strongly supported the adequacy of these systems. The scores for Votrax, however, were generally between 80% and 90% in these tests. Except for the Laddaga et al. study, we have been unable to find other research that examined perception of synthetic words by children. Two recent studies, described below, examined perception of synthetic CV syllables by young children.

In a study of the perception of stop consonants by kindergarten and second grade children, Wolf (1973) assessed their ability to identify and discriminate a series of synthetic speech stimuli varying in voice onset time (VOT). Her results showed that the perception of these sounds was found to be nearly categorical. No differences were observed in performance between the two age groups in either the identification or discrimination tests. A comparison of the group identification and discrimination functions with those of adults (Abramson and Lisker, 1970; Lisker and Abramson, 1970) indicated that children identified the stimuli only slightly less consistently than adults; however, the adults' performance on the discrimination task was superior to the performance of the children.

More recently, Elliott, Longinotti, Meyer, Raz and Zucker (1981) studied developmental differences in identifying and discriminating synthesized CV syllables. Their results showed that across ages there were no significant differences in the subjects' ability to label synthesized syllables as compared to natural speech stimuli. However, 6-year old children differed significantly from adults in the location of the category boundaries (i.e., crossover points for /ba/, /da/ and /ga/) while 10-year old children did not differ from the adults. From these results, Elliott et al. concluded that there are important developmental differences in perception tasks such as those employing identification and discrimination paradigms.

Perception of Natural Speech in Noise

Numerous studies have examined the perception of natural speech under a variety of noise conditions. These experiments have generally been conducted using adult subjects and have sought to determine the intelligibility of speech materials used in testing situations.

Elliott, Connors, Kille, Levin, Ball and Katz (1979) obtained precise measures of the sound intensities at which children understand monosyllabic words. They were interested in determining whether there were any developmental effects in children's speech understanding when the Peabody Picture Vocabulary Test (PPVT) was used under different conditions: quiet, open set, closed set, speech presented against a 12-talker "babble" or against filtered noise. Their results showed that no developmental change occurs in "perceptual" masking between the ages of 5 years and adulthood. However, there were prominent developmental changes in speech understanding "thresholds" in quiet across the 5 to 10 year age range. By the age of 10 years, performance of normal children achieves a level that is comparable to adult performance. Furthermore, this

age-related change occurs even though the monosyllabic stimuli are well within the receptive vocabularies of three-year-old children. This study on perception of natural speech in noise indicates that although noise interferes with perception, the speech is still understandable. Both natural speech presented in noise and synthetic speech may interfere with the human observer's perception and subsequent understanding of the linguistic message.

Context Effects in Speech Perception

Another group of recent studies has examined the effects of context on subjects' performance. Over a wide age range, subjects were shown to be sensitive to syntactic, semantic and discourse constraints when discriminating speech sounds. Cole and Perfetti (1979) found that the detection of mispronounced words was affected by syntactic, semantic and discourse constraints. Children as young as four years of age detected mispronunciations more accurately in predictable context (green grass) than unpredictable context (clean grass) while grade school age children and college students detected them more quickly. These results demonstrated that even very young children and children not yet skilled in reading comprehension use contextual information to recognize words from fluent speech.

Schwartz and Goldman (1974) studied several variables that influence performance on speech-sound discrimination tests. Monosyllabic nouns that were common to young children's vocabulary were chosen as stimulus items. Stimulus items were presented to nursery, kindergarten and first grade children in three different contexts (paired comparison, carrier phrase and sentence) and under two different listening conditions (quiet and noise). The results indicated that both the stimulus context and presence of background noise influenced performance of young children on speech-sound discrimination tests. Contexts with the most limited grammatical and phonetic cues led to more errors. Noise adversely affected performance in all conditions across all ages. Three times as many errors were made under the noise conditions than in quiet.

It is well known that children learn to speak and understand their native language rapidly and with relatively little difficulty. Much of the speech input the child receives is less than perfect; it is often garbled, distorted or produced in noisy environments. Even so, children manage to understand the speech presented to them. By studying comprehension of spoken language under a variety of less than ideal conditions, we hope to determine which aspects of the speech signal are most important for understanding and which ones may be particularly distracting to the young child's attention span.

The present investigation was designed to assess school age children's perception of synthetic speech as compared to natural speech. Two tasks, a Picture Vocabulary Task and a Digit Span Task, were designed to assess differences in speech perception. We expected subjects to show a higher level of performance when presented with natural stimuli than with synthetic stimuli. The present study extends the very limited prior research dealing with school age children's perception and understanding of synthetic speech produced by rule.

Method

Subjects

The subjects were second and fourth grade children taken from regular classes at Lakeview Elementary School in Bloomington, Indiana. Parental permission was required for participation. Twenty-seven second grade children participated in the Picture Vocabulary Task. The sample consisted of 12 second grade girls and 15 second grade boys between 7 and 9 years of age. Thirty-four fourth graders participated in the Digit Span Task. This group consisted of 19 fourth grade girls and 15 fourth grade boys between 9 and 11 1/2 years. All children who participated in the study were native speakers of English. In addition, small groups of nursery school and kindergarten children participated in the Picture Vocabulary Task.

Materials and Equipment

Picture Vocabulary Task. The first 63 words from the Peabody Picture Vocabulary Test (PPVT) were chosen as stimulus items. Both natural and synthetic tokens for each item were produced for these tests. All items were pretested to eliminate mispronounced synthetic words and their corresponding natural words. The 46 remaining stimulus items were recorded on audio tape. The items on each tape were arranged sequentially from the easiest to hardest items. For each item, an accompanying picture card was used. There were four different pictures on each card, only one of which was the correct choice. The stimulus item was played via tape recorder and headphones. Each new item was introduced with the prompt "show me _____. There were two practice items. For example, one item was "show me apple." A picture of a sock, pencil, butterfly and apple were on the card and the child's task was to point to the picture of the item heard. The entire set of picture cards was bound in a loose-leaf notebook for presentation along with the spoken items via audio tape.

Digit Span Task. The second task consisted of twenty-four digit sequences selected from the auditory-vocal sequencing subtest of the Illinois Test of Psycholinguistic Abilities. This task was designated the Digit Span Task. The digits ranged from a two digit sequence to an eight digit sequence. As in the Picture Vocabulary Task, there were both natural and synthetic versions for each digit sequence. One practice trial consisted of the child repeating a four digit number, 3-5-2-8, that was spoken by the experimenter. All two and three digit sequences, that is the first 4 stimuli of the experimental task, served as familiarization trials during the experimental sessions. If a subject was unable to perform the repetition task, the session was terminated. No subject was eliminated on this basis.

Equipment. A portable UHER 4000 Report-L reel-to-reel tape recorder, two pairs of Telephonics TDH-39 headphones, a junction box and a portable VTVM were used to present the stimuli in this experiment. Prepared response sheets were used by the experimenter to record all responses.

The natural materials were recorded by a male talker in a sound attenuated IAC booth using a professional quality microphone and tape recorder. The

synthetic materials were produced on the Prose 2000, a digitally controlled text-to-speech synthesizer (Groner, Bernstein, Ingber, Pearlman & Toal, 1982). All items were recorded on audio tape for subsequent use. Both natural and synthetic tokens were processed through a 12-bit analog-to-digital converter in preparation for subsequent editing. Each item, now in digital form, was assigned to a separate stimulus file for subsequent retrieval and audio tape preparation. All test tapes were generated using a specially designed audio tape making program.

Procedure

The equipment was set up in a small quiet room located in the media center at the school. The child's teacher decided when it was a good time for the child to participate in the project. The output volume from the tape recorder was calibrated daily to reflect the equivalent of 80 dB SPL at the headphones. Each subject was tested individually and was seated next to the experimenter at the table.

Picture Vocabulary Task. The second grade subjects participated in this task. The specific instructions for the task were as follows:

> Hi! My name is Beth. I have some pictures to show you. (TURN TO EXAMPLE A) See, there are four pictures on this page. (POINT TO EACH PICTURE) I'll say a word. Then I'd like you to point to the picture that tells the meaning of the word. Let's try one.
>
> Point to the picture that shows 'dog'. (SAY "GOOD" IF CORRECT; IF WRONG REPEAT) (TURN TO EXAMPLE B) Now, "Show me apple." (SAY "THAT'S FINE" IF CORRECT; IF INCORRECT REPEAT)
>
> Instead of me saying all the words, we're going to use this tape recorder here and these earphones. Sometimes you'll hear a man's voice tell you to point to a picture. Other times you'll hear a robot's voice tell you to point to a picture.
>
> Look at all the pictures and choose the one that you think is right. Please listen carefully to each word before you choose a picture.
>
> Ready?
>
> Okay, let's begin!

The subjects were given three practice items to make sure they understood the task. A "show me" prompt introduced each item. Three different audio tapes were used for presentation to subjects. The natural/synthetic (N-S) and synthetic/natural (S-N) tape orders were blocked conditions consisting of twenty-three items in each block. Tape 1 contained a block of 23 natural items (Block A) followed by a block of 23 synthetic items (Block B); Tape 2 contained 23 synthetic items (Block A) followed by 23 natural items (Block B). A third

tape, the mixed natural/synthetic list condition, was blocked as well but the stimulus items were randomly arranged tokens of the natural and synthetic items within blocks A and B. All subjects received the same 46 words in the same serial order from easy to hard within both Block A and Block B.

Digit Span Task. Fourth grade children participated in the Digit Span task. The specific instructions for the task are as follows:

> Hi! I'm Beth. Here's what we're going to do: I'll say a list of numbers and I want you to repeat the numbers. Let's try it. Experimenter: 3, 5, 2, 8 (Prompt if needed - NOW YOU SAY THEM) (Continue until child repeats numbers without prompt.)
>
> Instead of me saying the numbers, we're going to use this tape recorder and these headphones. Sometimes a man will say the numbers, other times a robot will say the numbers.
>
> Please, listen to the entire list before you start to repeat it. OK? Let's do it!

Two random orders of a mixed natural/synthetic stimulus tape were used. All responses were recorded by the experimenter on prepared response sheets. The experimenter turned the tape recorder on and off for each trial and tested subjects in a self-paced format. The subjects were not under any external time pressure and were given verbal encouragement to let them know they were performing the task properly as testing proceeded.

The average length of the experimental sessions was 20 minutes. However, sessions ranged from 12 to 25 minutes depending on the subject. Data collection took about 10 weeks to complete overall.

Results and Discussion

Picture Vocabulary Task. Across all tape orders, the percentage of correct responses for the natural stimuli exceeded the percentage of correct responses for the synthetic stimuli (98.1% vs. 93.8%; $t(52)=3.23$, $p<.002$). All of the subjects consistently made more errors on synthetic items than on natural items. Presentation order significantly affected the results. When subjects heard the natural block of items before the synthetic items (tape order N-S), they made more errors overall than when the stimuli were presented in the synthetic-natural (S-N) order. Figure 1 displays these results.

For the third presentation order, the mixed voice condition, subjects also showed superior performance on natural over synthetic items. Regardless of presentation order, subjects showed equal or better performance on the natural versus the synthetic items. Figure 1 displays the percentage of correct responses for each presentation order. The results in the upper panel show each presentation order separately; the lower panel shows the percentage of correct responses averaged across all presentation orders.

Insert Figure 1 about here

Overall, most of the errors, for both natural and synthetic items, occurred on Block B. Even though the stimuli were selected from a graded series of items used extensively in clinical and educational applications, we found that our division of the items on a binary basis resulted in two lists that were not equivalent. While this imbalance complicated our analyses and results, it did not interfere with testing the underlying hypothesis of this experiment regarding comparisons of natural and synthetic speech.

In the N-S and mixed orders subjects made more errors than in the S-N order, and more of those errors were on Block B. Table 1 displays the overall percentages of correct responses for each block separately.

Insert Table 1 about here

As shown in Table 1, the overall percentage of correct responses is lower for Block B than for Block A. We could conclude that Block B contained more difficult items and therefore subjects performed less well. On the other hand, we might conclude that our subjects, 7 and 8 year old children, simply became tired towards the end of the session and therefore performance decreased. A third possibility is that listening to synthetic speech places increased processing demands on memory and attention and as a result fatigue and/or processing capacity limitations become important factors during sustained listening (Pisoni, 1982). By collapsing the data across all tape orders and across natural and synthetic speech, performance on Block A was superior to performance on Block B (97.6% vs 94.3%; $t(52)=2.3$, $p<.02$).

Digit Span Task. Fourth grade children showed higher levels of performance on natural digit strings than on synthetic strings. When scored for free recall, that is scoring for the correct digit regardless of its position in the series of digits, subjects recalled 82.9% of the natural digits and 78.5% of the synthetic digits. These results are considerably lower percentages of correct responses than were found in the Picture Vocabulary Task. In the Digit Span Task, subjects not only have to perceive the stimulus but they must also be able to encode a series of items. Therefore, we could reasonably predict that subjects would show lower levels of performance as the length of the digit strings increased. Results for free recall as a function of list length are shown in Figure 2.

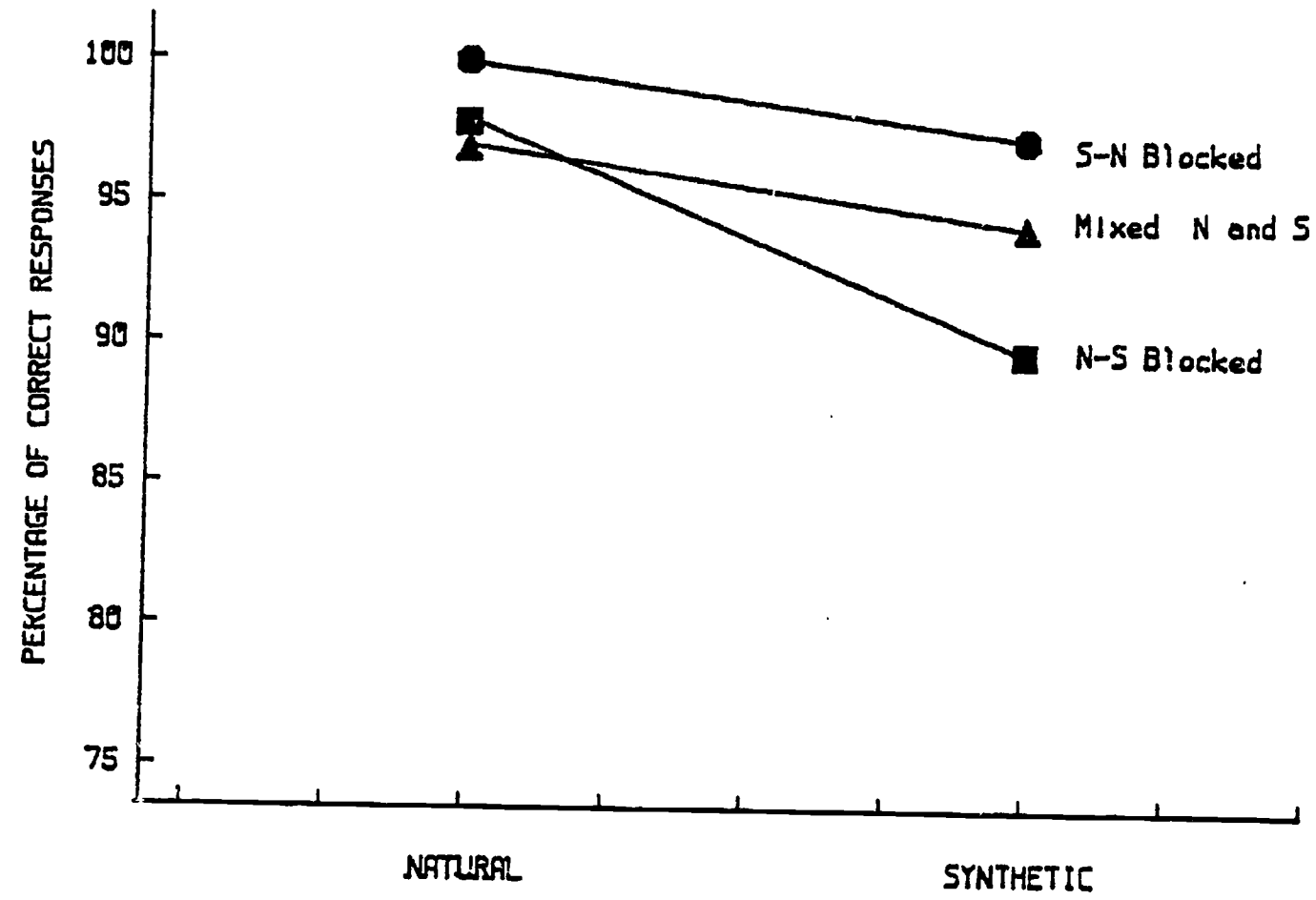


Figure 1. Percentage of correct responses for the Picture Vocabulary Task displayed for natural and synthetic speech. The upper panel displays the results for each presentation order separately. The lower panel displays the results averaged across all orders.

Table 1

Percentage of Correct Responses: Picture Vocabulary Task

	Block A		Block B	
Tape Order	Natural	Synthetic	Natural	Synthetic
N/S	97.8			89.7
S/N		97.3	100	
Mixed	99.2	95.9	94.7	92.6
Mean	97.6		94.3	

Insert Figure 2 about here

As shown in Figure 2, subjects' recall performance decreased as the list length became larger ($F(4,128)=47.03$, $p<.001$). The greater the number of digits in the string, the more information subjects had to perceive, remember and repeat back to the experimenter. The longer digit strings obviously put an increased demand on the subjects' active rehearsal processes in short-term memory and this no doubt affected final recall. Subjects may be able to perceive (i.e., identify, recognize or encode) the input message but during the course of perceptual processing and subsequent transfer of some of the information may be lost and therefore the subject cannot retrieve the correct input items in recall.

Figure 2 also shows a consistent trend for performance on natural strings to be better than synthetic digit strings ($F(1,32)=8.60$, $p<.01$). While subjects had some difficulty recalling the longer digit strings, they recalled natural strings better than synthetic strings overall. Recall of the synthetic digit strings was difficult for at least two reasons: first, as the string gets longer the subject must actively rehearse more items in short term memory; and second, the increased processing load is added to the greater demands required to encode synthetic items. Thus, the combination of the longer string and the synthetic speech produces greater decrements in free recall performance.

Measure of Output. It is possible that children, and perhaps adults as well, tend to perceive the synthetic materials in a manner that is quite different than the natural materials. The results for the free recall measure indicated that subjects showed superior performance overall for recall of natural digit strings (82.9% vs 78.5% correct for natural and synthetic strings respectively). We calculated a "measure of output", that is the absolute number of responses given regardless of whether the response given was correct. This measure thus provided an estimate of the likelihood of responding for each digit string. For example, a subject might hear the natural five digit list 6-4-3-2-1 and respond 6-1-2-3. In this case, the subject's score for free recall would be 4, and, the measure of output would also be 4. If the response were 6-4-3-2-2, the free recall score would still be 4 but the measure of output would be 5 because the subject responded with 5 digits. Results for this analysis also favor responses to natural over synthetic lists (87.3% vs 83.3%).

Free recall scores and a gross measure of output performance showed a similar pattern of responses. Subjects generated more responses when listening to natural speech than when listening to synthetic speech. In this task, the additional processing demands placed on subjects when listening to synthetic digit strings may lead to lower levels of responding both for correct responses and the absolute number of responses generated in the free recall task.

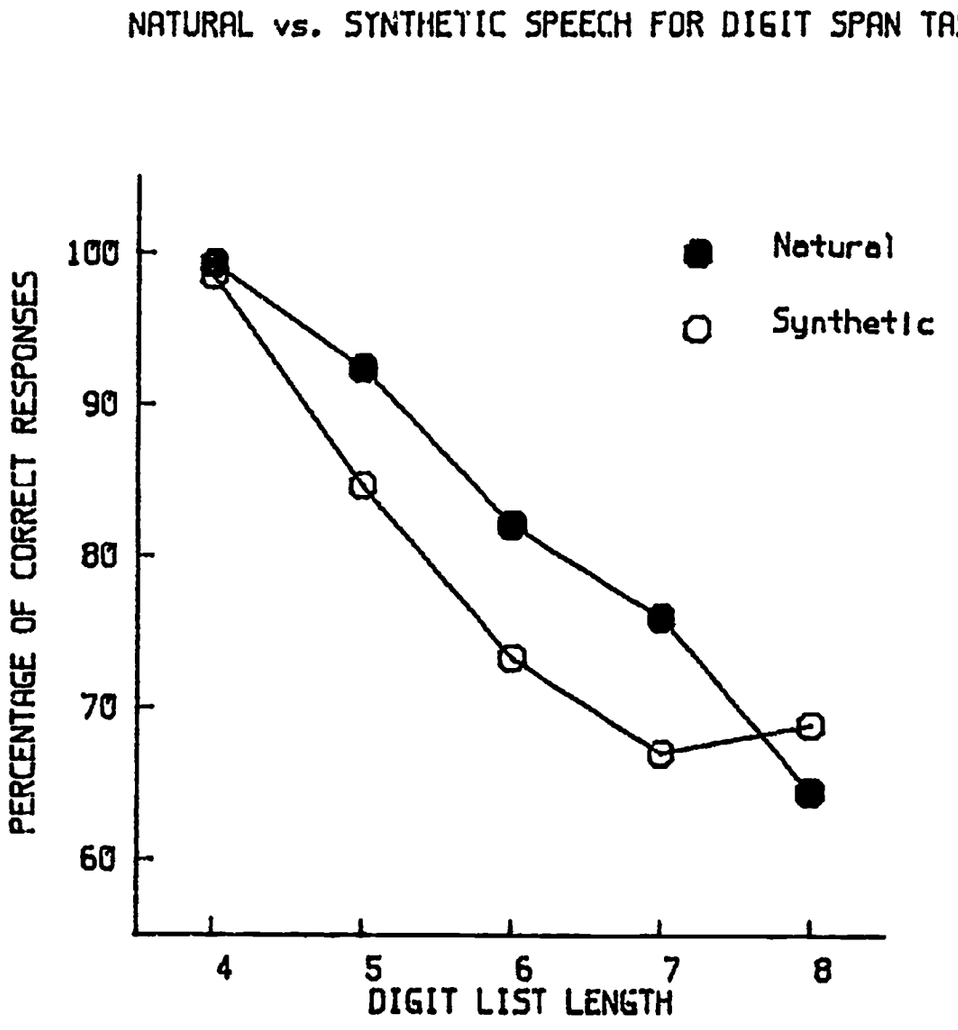


Figure 2. Percentage of correct responses for the Digit Span Task for natural and synthetic speech. The data displayed were scored for free recall and are shown separately for each list length.

Summary and Conclusions

School-age children perform in much the same manner as adults in listening and responding to synthetic speech. Like adults, children show performance decrements when the input stimulus materials are generated using a synthetic voice (cf. Luce, Feustel & Pisoni, 1982; Pisoni, 1982; Nusbaum & Pisoni, 1982). Our results are also consistent with previously reported data which showed the superiority of natural over synthetic speech (Laddaga et al., 1981). The studies conducted by Wolf (1973) and Elliott et al. (1981) demonstrated that children performed in much the same way as adults when presented with synthetic CV syllables. In the present experiments, children's performance did not reach the levels attained by adults. Adults made virtually no errors on the Picture Vocabulary Task and only a few errors on the longer strings in the Digit Span Task. This result is not surprising considering the simple materials used in the Picture Vocabulary Task and the longer memory span available for adults in the Digit Span Task. The findings from these studies always show higher performance levels with natural speech stimuli than with synthetic speech stimuli, regardless of the synthesis systems used to prepare the stimuli. Furthermore, if the synthetic speech task presented to young children was more unstructured, we would expect them to show much poorer performance since the speech quality combined with the less constrained experimental situation would no doubt interact significantly (cf. Schwartz & Goldman, 1974; Nusbaum & Pisoni, 1982).

Our results on the perception of synthetic speech obviously have important implications for the design, selection and use of voice-response systems that are to be used in teaching and learning environments with young children. Further research is needed to assess the perception of synthetic speech under a variety of environmental conditions particularly those involving differential cognitive demands. It is apparent from our preliminary studies that the speech quality and intelligibility of voice-response devices varies substantially from product to product. Children may soon be required to interact routinely in educational settings with machines that incorporate voice output using speech synthesis; some of these devices may produce speech that is difficult to understand for a variety of different reasons. Moreover, the synthetic speech may require additional processing capacity and attentional demands that may well produce large decrements in other complex cognitive tasks that children are expected to perform routinely in educational settings using these learning devices. Additional work is currently underway in our laboratory on these problems and reports of our findings will be forthcoming.

Reference Notes

1. Pisoni, D. B. Perceptual evaluation of voice response systems: Intelligibility, recognition & understanding. Paper presented at the Workshop on Standardization for Speech I/O Technology, National Bureau of Standards, Gaithersburg, Maryland, March, 1982.

2. Nusbaum, H.C. & Pisoni, D. B. Perceptual and cognitive constraints on the use of voice response systems. Paper presented at the Voice Data Entry Systems Applications Conference '82, Lockheed Missiles and Space Company, San Mateo, California, September, 1982.

References

Abramson, A. & Lisker, L. Discriminability along the voicing continuum: Cross-language tests. In Proceedings of the Sixth International Congress of Phonetic Sciences, Prague, 1967. Prague: Academia, 1970.

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.

Elliott, L. L., Connors, S., Kille, E., Levin, S., Ball, K. & Katz, D. Children's understanding of monosyllabic nouns in quiet and in noise. Journal of the Acoustical Society of America, 1979, 66, 12-21.

Elliott, L. L., Longinotti, C., Meyer, D., Raz, I. & Zucker, K. Developmental differences in identifying and discriminating CV syllables. Journal of the Acoustical Society of America, 1981, 70, 669-677.

Groner, G. F., Bernstein, J., Ingber, E., Pearlman, J. and Toal, T. A real-time text-to-speech converter. Speech Technology, 1982, 1, 2, 73-76.

Laddaga, R., Sanders, W. & 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.

Lisker, L. & Abramson, A. The voicing dimension: Some experiments in comparative phonetics. In Proceedings of the Sixth International Congress of Phonetic Sciences, Prague, 1967. Prague: Academia, 1970.

Luce, P. A., Feustel, T. C. & Pisoni, D. B. Capacity demands in short-term memory for synthetic and natural speech. Human Factors, 1983. In press.

Pisoni, D. B. Perception of synthetic speech: Some contributions by the human listener. Speech Technology, 1982, 1, 2, 10-23.

Schwartz, A. H. & Goldman, R. Variables influencing performance on speech-sound discrimination tests. Journal of Speech & Hearing Research, 1974, 17, 25-32.

Wolf, C. G. The perception of stop consonants by children. Journal of Experimental Child Psychology, 1973, 16, 318-331.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

Context Effects in the Perception of English /r/ and /l/ by Japanese*

Patricia Dissosway-Huff
Robert F. Port

Department of Linguistics

and

David B. Pisoni

Department of Psychology
Indiana University
Bloomington, Indiana 47405

*This research was supported, in part, by NIH grant NS-12179 and NICHD grant HD-12511 to Indiana University in Bloomington. We thank Tom Carrell for his help and assistance in preparing the stimuli.

Abstract

Perception of English /r/ and /l/ is a well-known difficulty for Japanese speakers learning English. An identification test for minimal pairs produced by two American speakers was administered to 32 Japanese students of English. This was done just after their arrival in the United States, and then nine weeks later at the completion of an intensive English program emphasizing oral skills. No significant improvement was observed in their overall performance after oral training. The perception of /r/ and/l/ as singletons and in consonant clusters exhibited quite opposite trends. In clusters, /l/ was perceived more accurately than /r/ (66% vs. 52%), while for singletons, /l/ was somewhat worse than /r/ (63% vs. 70%). Singleton consonants in word-final position were more accurately perceived than initial singletons (77% vs. 57%) while for clusters, the finals were slightly worse than initials (56% vs. 62%). Thus, both the /r-l/ effect and word-position effect interact with the singleton-cluster factor but not with each other. It is proposed that the longer duration of /r/ and /l/ in final position relative to all others may account for the better performance in that position.

Introduction

What is the nature of the difficulty that Japanese speakers have in learning English /r/ and /l/? This general problem is faced by any adult who learns a new contrast in a foreign language. Aside from obvious difficulties in speech production and the fine details of articulatory control with these sounds, Japanese students learning English tend to have difficulty in perceiving contrasts in these sounds--as has been shown in a number of studies. The problem stems from the fact that Japanese has only a single "resonant" sound, usually written with the letter r in English transliterations. It is typically pronounced as an apical tap, although, in initial position in Japanese, it is often pronounced rather similarly to an American /r/ in a word like red.

A recent paper by Mochizuki (1981) reported several interesting observations about the perception of /r/ and /l/ by Japanese. First, she found that how well subjects could perform depended on the context in which the contrast occurred. Thus, word-final /r/ and /l/ were identified better than /r/ and /l/ in several other positions. In addition, Mochizuki found some evidence that speakers who could produce English /r/ and /l/ that Americans could accurately identify could nevertheless not reliably perceive /r/ and /l/ spoken by Americans. This seems to reverse the common-sense idea about how the acquisition of a sound contrast ought to proceed. And it raises the possibility that sometimes learning new articulations may be easier than listening for the right new acoustic cues.

Our goals in this experiment were first to examine the context effects on /r-l/ perception by Japanese learners of English in the hope of finding some more systematic account for them. Thus, we looked at English r/l minimal pairs in the same five contexts that Mochizuki used. In addition, since it is clearly the case that Japanese are able to improve their perception over time, the second goal was to examine this ability by speakers across a range of different skills and training at speaking English.

Methods

Table 1 shows examples of the list of real English minimal pairs selected from the five environments. Three singleton environments were used: initial, final and intervocalic. And two cluster environments: initial and final in monosyllables. Altogether 59 minimal pairs, or 118 test words were used. Two phoneticians who are native speakers of American English read two randomizations each of the words taking care to pronounce each pair identically except for the r-l difference. After digitization on a computer, the words were randomized with a 3 sec silent interval inserted between words. The task of the subjects was to circle the correct word containing either /r/ or /l/ on a typed answer sheet.

Insert Table 1 about here

The subjects were 32 engineering students from Nihon University in Japan. They were selected from a group of 17-year-old freshman visiting the United States for a ten-week intensive English program at the University of Tennessee at Martin. All subjects were scheduled to return to Japan at the end of the ten-week program. The subjects had had about six years study of English language from Japanese teachers in Japanese schools. None had studied English with native English-speaking teachers.

The Japanese students were given the Michigan Test of English Language Proficiency (Ann Arbor, MI) and sorted into four levels. We chose half our subjects from the lowest level (Level 1) and half from the highest level (Level 4). Even the more advanced group could not speak English nearly well enough to be admitted to an American university. Thus the overall English ability of these subjects was quite low.

These 32 subjects listened to our tapes once in the first week of the training course and again 9 weeks later at the conclusion of the program.

Results and Discussion

Table 2 shows the basic results of the effect of the ten-week training program and the effect of English-language ability of the subjects measured at the beginning of the training. Looking at the means across the bottom, it can be seen that the beginning level group performed at 57%--only slightly better than chance. The advanced beginners did somewhat better but still quite poorly. Of course, the Michigan Placement Test which was used to evaluate them employed both written and oral test components, so the difference between groups is not a surprise.

Insert Table 2 about here

What may seem a surprise is that neither of the two subject groups improved its score as a result of the 9 weeks of intensive training in conversational English from American teachers. English language courses in Japanese schools are concerned primarily with written English so it is not surprising that the subjects oral performance was quite poor. But why didn't they improve? Experience teaching English to foreign students suggests two possibilities. First, improvement in any one particular area often snows up only several weeks after training in that area is completed. This possibility makes it very difficult to determine the effectiveness of the training without later testing.

Table 1

Word List for Perception Experiment

Initial:	Final:
1. row - low	1. fear - fill
2. rye - lye	2. fire - file
3. rate - late	3. soul - soar
4. reek - leek	4. cool - Coor
5. rude - lewd	5. tear - tell
6. rent - lent	6. pair - pale
7. reap - leap	7. more - mole
8. rip - lip	8. sear - sill
9. rock - lock	9. mire - mile
10. ramp - lamp	10. fair - fail
11. right - light	11. bore - bowl
12. rest - lest	12. poor - pool
13. rim - limb	13. door - Dole
14. rhyme - lime	14. share - shale
15. red - led	15. dire - dial

Initial and Final Clusters:	Intervocalic:
1. fry - fly	1. berry - belly
2. free - flee	2. oreo - oleo
3. pry - ply	3. arrive - alive
4. glass - grass	4. battering - battling
5. cord - cold	5. berated - belated
6. brink - blink	6. array - allay
7. frank - flank	7. hoary - holy
8. pray - play	8. firing - filing
9. fray - flay	9. correct - collect
10. gourd - gold	10. erect - elect
11. hoard - hold	11. far off - fall off
12. Myers - miles	12. wiry - wily
13. bright - blight	13. believe - bereave
14. toward - told	14. tiring - tiling
15. tours - tools	15. pirate - pilot

Table 2

Percent Correct Identification

	Subject Group			
	Beginners	Advanced Beginners	Mean	
Before Training	56%	66%	62%	N.S.
After Training	58%	67%	63%	
Mean	57%	67%		

$p < .01$, Anova

The second possibility is simply that two months, involving training in all aspects of English instruction, is just not enough time to produce any real improvement in this particular aspect of speech. Only subsequent testing could give us clearer sense of why immediate improvement was not observed.

Figure 1 displays the results showing the effect of the context or word position on identification. The scores were all better than chance. Results for the final position in a word--as in pairs like fear-fill--were significantly better than in the other positions.

Insert Figure 1 about here

However, if we look at the results separately for /r/ and /l/ for the same set of contexts, as in Figure 2, a number of interesting interactions can be seen. Overall, /l/ was identified correctly somewhat better than /r/, but this difference obscures the fact that the two segments did better in different positions. The interaction between the two factors is significant by analysis of variance on percent correct. Most of the improvement in the final position in a word is due to the /r/ which does much better here than elsewhere. The identification of /l/ is fairly good everywhere except in initial position. Here performance is at chance.

Insert Figure 2 about here

These results agree quite well with the data that Mochizuki published recently for the same set of contexts but for a much smaller group of subjects. Mochizuki suggests that dark /l/ is better perceived than light /l/. Our results do not fully support that claim since final clusters and intervocalic /l/ score about the same yet differ in "darkness". We would like to suggest a different reason why both /r/ and /l/ are better identified in final position--but one that we can only partially defend at this point.

Measurements from a sample of spectrograms of our test utterances showed that both /r/ and /l/ are quite a bit longer in word-final position than in any other position. This is illustrated in Figure 3. Here we have presented sample spectrograms of words containing /r/ in different positions. Since it is notoriously difficult to measure the beginning and end of /r/ and /l/ from spectrograms, we developed some arbitrary conventions so that comparisons could be made. Because of the gradual onset of these resonants in many contexts, we have measured both the beginning and end of the transitions into (and out of) the /r/ or /l/ (these are marked as A and B in the figure), and then measured from the midpoint of the transition. Thus, the line C-D is considered the resonant duration. Using conventions developed along this line, notice that the duration

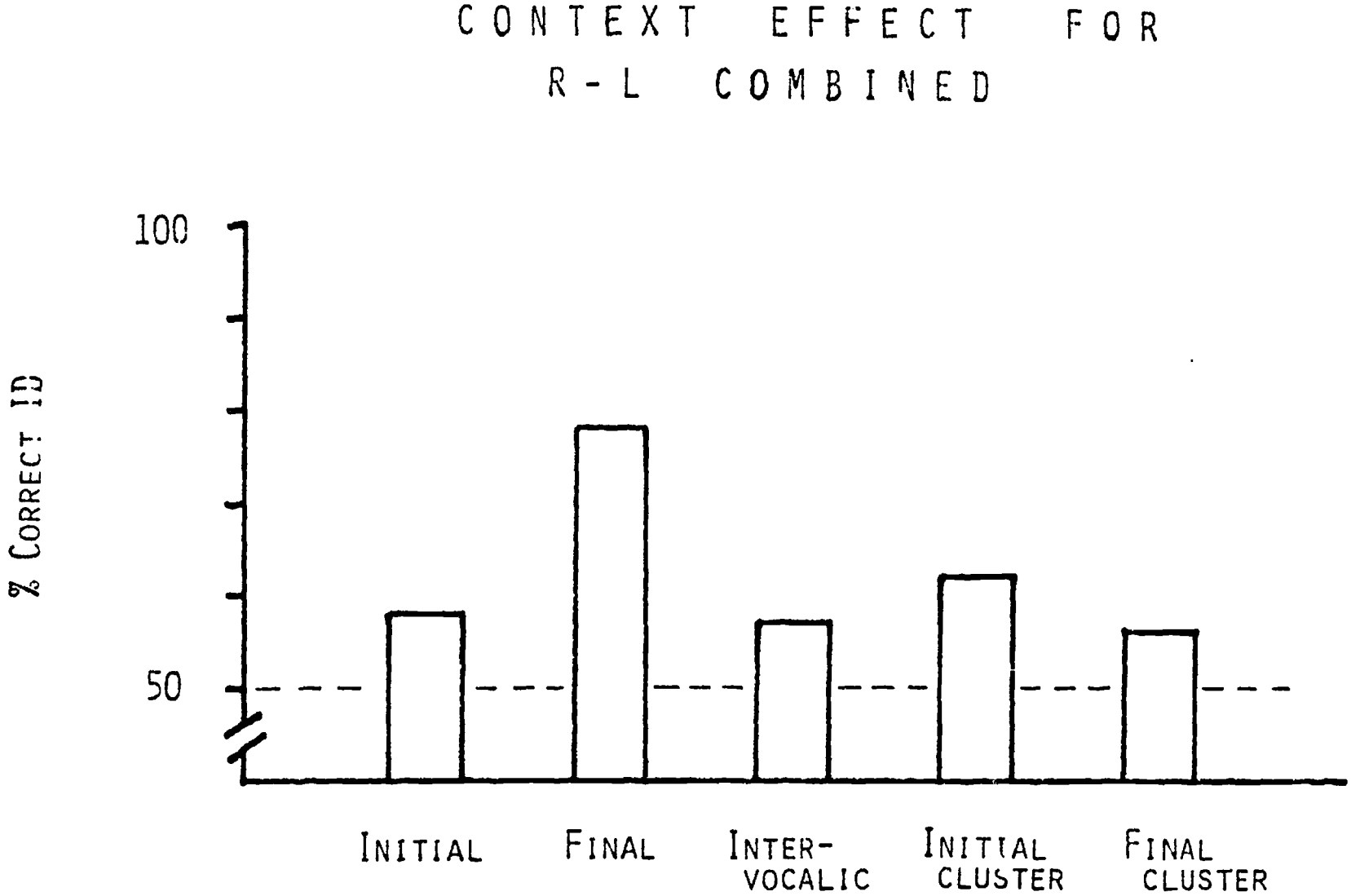


Figure 1. Percent correct identification of /r/ and /l/ responses combined across different environments.

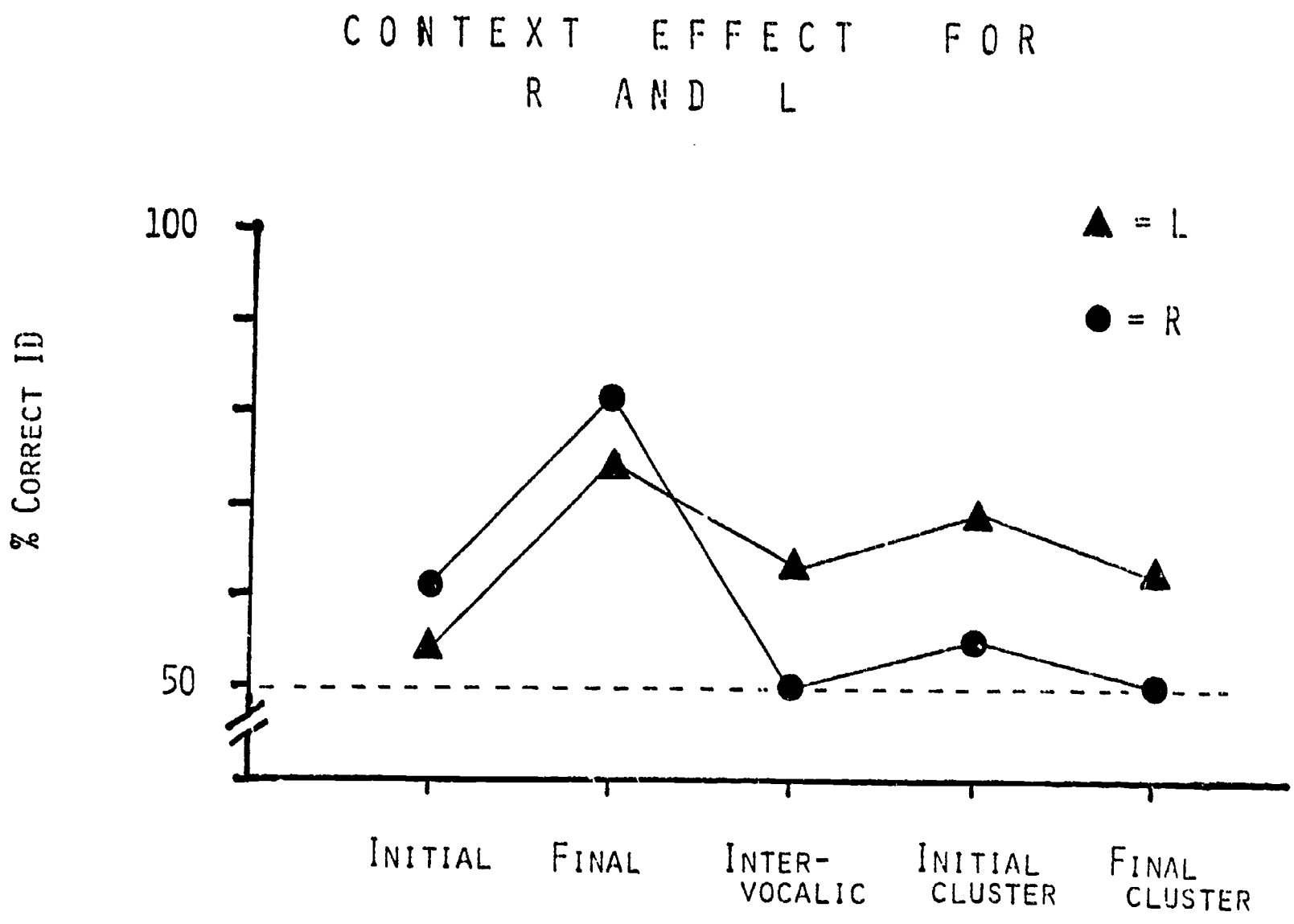


Figure 2. Percent correct identification of /r/ and /l/ displayed separately across the different environments.

of /r/ in tore is considerably longer than that in toward. This might account for the fact that /r/ was so much better perceived in tore than toward.

Insert Figure 3 about here

Although this spectrographic analysis of our tapes is not completed yet, some preliminary results from measurements of about 10 percent of the stimuli produced by both talkers are shown in Table 3. It can be seen that /r/ and /l/ in final position had longer durations than those at any other position--when using our measurement criteria. Because of difficulties in measurement, of course, it will be difficult to make a persuasive case from production data alone for the hypothesis that the context effect in accuracy of English /r/-/l/ perception by Japanese can be predicted from the duration of the acoustic information for those segments. The best evidence about this hypothesis will come from perceptual experiments using synthetic speech or manipulated natural speech. We have several obvious experiments in mind.

Insert Table 3 about here

Thus, these results indicate that there is not only a reliable effect of the context of the /r/and /l/, but there is also an interaction such that particular contexts select for or against either /r/ or /l/. We have no hypothesis at the moment to account for these particular interactions.

Conclusions

In conclusion, then, we have confirmed the difficulty experienced by Japanese speakers in perceiving /r/ and /l/ even when produced carefully by American speakers. Although our subjects did not improve at this task during their classroom training of 9 weeks, Japanese students who live in the United States do greatly improve over time. Over the short term, apparently, progress is slow.

We also found strong effects of the position of the /r/ and /l/ in the word. In fact, /r/ seemed to be more sensitive to the context than /l/. Our acoustic analysis of a sample of the test stimuli suggest a possible account for the context effect in terms of the duration of the acoustic signal corresponding to the segment. Hopefully, from this preliminary work we will be able to achieve some insights into the problems adults have when they attempt to learn the sound system of another language. Obviously, there is an important interaction between their knowledge of the sound system of the language and the particular acoustic correlates that represent phonological contrasts in that language.

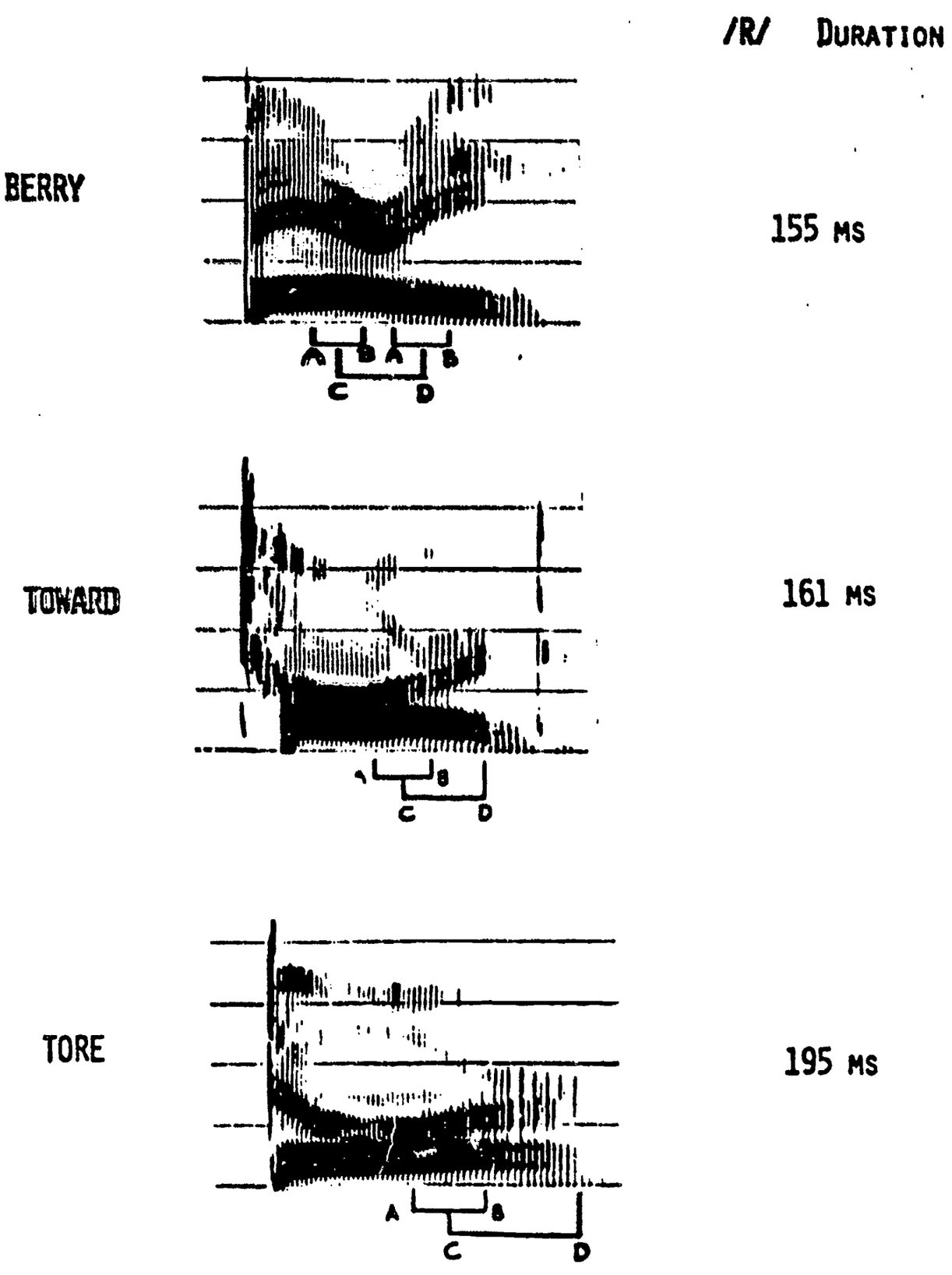


Figure 3. Spectrograms illustrating the acoustic criteria used to measure the durations of /r/ in three different test words.

Table 3

Mean r/l Duration for a Subset of the Test Words

Initial	Final	Intervocalic	Initial Cluster	Final Cluster
108 ms	195ms	121ms	120ms	161ms

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

An Activation Model of Auditory Word Recognition*

Howard C. Nusbaum and Louisa M. Slowiaczek

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

*This research was supported by NIH grant NS-12179 and NIH training grant NS-01734 to Indiana University in Bloomington.

In recent years, speech research has become increasingly concerned with the processes that mediate perception of spoken words. In general, three findings have emerged that characterize auditory word recognition (see Cole & Jakimik, 1980; Foss & Blank, 1980; Grosjean, 1980; Marslen-Wilson & Welsh, 1978). First, spoken words are recognized from left to right; that is, words are recognized in the same temporal sequence by which they are produced. Second, the beginnings of words appear to be far more important for directing recognition than either the middles or the ends of words. Finally, word recognition results from an interaction between bottom-up pattern processing and top-down expectations derived from context and linguistic knowledge.

The evidence that supports these conclusions has come from research using a variety of different experimental procedures. One early set of studies used a mispronounciation detection paradigm in which subjects were instructed to respond to every mispronounced word in spoken sentences (see Cole & Jakimik, 1978, for a review). The results showed that listeners were more accurate in detecting word-initial mispronounciations than in detecting mispronounciations that occurred later in the words (e.g., Cole, Jakimik & Cooper, 1978). In addition, mispronounced words were detected faster when constrained by prior context (Cole & Jakimik, 1978). Moreover, the semantic information in an immediately preceding word was sufficient to enhance detection of a subsequent mispronounced word (Cole & Jakimik, 1978).

Similar results have been obtained with a shadowing procedure. Marslen-Wilson and Welsh (1978) examined the "fluent restorations" produced by subjects shadowing (repeating aloud) speech containing mispronounced words. A fluent restoration occurred whenever a subject restored a mispronounced word to its "normal" correctly pronounced form with no disruption of shadowing. Marslen-Wilson and Welsh found that fewer fluent restorations occurred for mispronounciations in the initial syllable of words compared to the third syllable. Further, more fluent restorations were produced when the prior context was highly constraining.

Using a phoneme monitoring procedure, Foss and Blank (1980) found that the latency to detect word-initial phonemes was enhanced by prior context. In addition, they found that subjects could detect target phonemes in initial position in words and nonwords with equivalent latencies. However, when a target-bearing word was preceded by a nonword, detection latencies were significantly longer than when the target-bearing word was preceded by a word. From these results Foss and Blank argued that the initial portions of words are perceived by strictly bottom-up pattern processes without reference to lexical knowledge. In contrast, the lexical status (i.e., word or nonword) of the utterance preceding a target-bearing word was critical because it affected the listener's ability to locate the beginning of the target-bearing word. Subjects could easily separate the end of one word from the beginning of the next. But when a nonword preceded the target-bearing word, subjects found it hard to determine that the nonword had ended and the subsequent word begun. This slowed detection of a target phoneme that was supposed to be in initial position.

Finally, the gating technique has been used to study the amount of acoustic-phonetic information needed for word recognition. In a study by Grosjean (1980) using this paradigm, a single auditory stimulus was presented on each trial and subjects were asked to identify the stimulus as a particular word. The stimulus consisted of the initial portion of a word's waveform. On the first

trial, for a particular word, subjects would be presented with the first 30 msec of the word. On each successive trial, an additional 30 msec of the word was presented so that subjects heard increasingly larger segments of a word until the entire word was presented. Grosjean found that subjects could correctly identify words at durations that were substantially less than the total duration of the words. This "critical recognition point" is the point at which the recognized word diverges from all other words in the lexicon. Grosjean found that the critical recognition point was affected by linguistic knowledge (e.g., the frequency of occurrence in English of a word) as well as by prior constraining sentential context. Recently, Salasoo and Pisoni (Note 1) have obtained similar results when every word in a sentence was gated. They showed that subjects recognized words using less waveform when words were gated from the beginning of each word compared to a condition where words were gated from the end. Also, the critical recognition points occurred earlier when words were gated in meaningful sentences compared to words gated in syntactically correct but semantically anomalous sentences.

Taken together, the results of these various studies indicate that spoken words are recognized one at a time, using the beginning of words in conjunction with linguistic expectations derived from prior contextual constraints. Attempts have been made to accomodate these findings within a single theory of auditory word recognition -- cohort theory (see Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978; Tyler & Marslen-Wilson, 1982a, 1982b). Cohort theory depicts word recognition as a two-stage process. First, using only the acoustic-phonetic information at the beginning of a word, a set of cohorts or word candidates is activated. These cohorts are all the words in the lexicon that share the same initial phonetic sequence. For a stimulus word to be recognized, it must be contained in this initial set of word candidates. When all the candidates but one are eliminated, the remaining word is recognized. Thus, the second stage of cohort theory describes the process by which competing candidates are eliminated. Cohort members are deactivated by a mismatch with acoustic-phonetic information later in the stimulus word or by a mismatch with contextual expectancies. As more of a stimulus word is heard, more word candidates become inconsistent with the sensory information and prior sentential context, and these candidates are eliminated from further consideration. At some point before the end of the word (i.e., the critical recognition point) all candidates but one are deactivated leaving the recognized word.

To summarize cohort theory, a set of word candidates is instantiated by bottom-up (priority) processing of the sensory information contained in the initial part of a stimulus word. Members of this set are then deactived by an interaction between top-down expectations and bottom-up processing of subsequent acoustic-phonetic information. Beyond this basic description, cohort theory makes several additional assumptions about the time-course of auditory word recognition. First, Marslen-Wilson and Welsh (1978) have stated that once a member of the pool of word candidates is eliminated it "may remain activated for a short period thereafter" (p. 56). More recently, Tyler and Marslen-Wilson (1982a) have asserted that the auditory word recognition process is "optimally efficient" (also see Marslen-Wilson, Note 2). Optimal efficiency in recognition refers to the ability of the system to reject possible word candidates at the very first indication of inconsistency with the input stimulus. Thus, a listener should be able to reject cohort members based on the earliest mismatch of information, without using the redundancy inherent in all levels of spoken

language. This property of optimal efficiency is the opposite of the principle of "least committment" (see Marr, 1982) by which a recognition system refrains from making a decision until all the pertinent information has been evaluated. Finally, Marslen-Wilson (Note 2) has claimed that word recognition decision time is independent of the size of the activated cohort. This means that words with a great deal of initial phonetic overlap with other words should be recognized as quickly as words with a small number of cohort members as long as the critical recognition points (based on acoustic-phonetic information alone) are the same. In essence, this argues that there is no cost or processing load associated with the activation of a set of word candidates.

These assumptions, taken in conjunction with the basic description of word candidate activation and elimination, make cohort theory a fairly complex and powerful account of auditory word recognition. However, because of this complexity and power, cohort theory has not been specified well enough to generate explicit testable predictions (although cf. Tyler & Marslen-Wilson, 1982b). Although this highly interactive theory has been proposed as an alternative to several serial autonomous theories of word perception (see Marslen-Wilson & Tyler, 1980; Tyler & Marslen-Wilson, 1982b), it is vague and imprecise in a number of respects. In order to generate empirically testable hypotheses, certain constraints must be placed on cohort theory, just as Marslen-Wilson and Tyler (1980) made specific assumptions about autonomous theories of word perception to test these theories.

Lexical Activation

One deficiency in the current formulation of cohort theory is that no processing mechanism has been specified to instantiate cohort theory. As a result, it is impossible to make specific predictions about the time course of word recognition. However, a recent model of visual word recognition has provided a mechanism which seems ideal for implementing cohort theory. The interactive activation model proposed by McClelland and Rummelhart (1981; Rummelhart & McClelland, 1982) describes the growth and decay of activation in different cognitive processing units (e.g., words or features). This model developed from the cascaded activation model described by McClelland (1979). The basic assumption of this type of model is that the fundamental processing units in the perceptual system are nodes that may become activated by positive input from other nodes or deactivated by inhibition (see Anderson, Silverstein, Ritz, & Jones, 1977). The first step in implementing cohort theory as an activation model is to specify a differential equation describing the change in activation of a set of word candidates over time. The general form of the equation comes from the interactive activation model proposed by McClelland and Rummelhart (1982). The basic description of the change in activation of a word candidate node is given by:

$$da_i(t)/dt = r_i(t)\ (M_i - a_i(t))$$

In this differential equation, $r_i(t)$ represents a rate modifier, M_i represents the current asymptote the i-th node is driving towards, and $a_i(t)$ is the activation level of the i-th node at time t. A fundamental assumption of

cohort theory is that a large number of word candidates that share word-initial acoustic-phonetic information are activated together. A corollary of this assumption is that at any point in the stimulus word where disconfirming information is encountered, a subset of these candidates will be deactivated together. An additional assumption made in our lexical activation model of cohort theory is that nodes that are activated together from the same initial activation level will have the same growth rate. Similarly, nodes that are deactivated together from the same activation level will decay at the same rate. These assumptions could be changed to attribute different characteristic growth and decay rates to different word nodes, perhaps based on differences in the frequency of occurrence in English. However, for the present purposes, the simplifying assumptions have been adopted. In addition, the model assumes that a node may be in one of three distinctly different states. These states change the value of the rate modifier $r_i(t)$ and the asymptote M_i. As long as a feature of the stimulus matches a feature of the i-th node, $r_i(t)$ will equal a constant C and M_i will be set by:

$$M_i = I \cdot A_i(0)$$

where I represents the activation level of the input feature and $A_i(0)$ is the baseline activation of the node. The constant C that serves as the rate modifier is the same for all nodes.

The second possible state occurs as soon as a mismatch is obtained between the input and a candidate node. The asymptote M_i is set to zero so the unit will turn itself off. Also, the rate modifier $r_i(t)$ becomes a function of the current activation level of the deactivated node and the modifier constant C. This equation is given by:

$$r_i(t) = C(1 + a_i(t))$$

This equation causes the rate of decay to depend on the current level of activation. Thus, the higher the activation, the faster the decay. This allows the system to quickly eliminate any incorrect word candidates, providing a fast error recovery process.

Finally, the third state occurs when the end of the input is reached. At this point the asymptote M_i is set to zero to allow the activation of all nodes (including the recognized word) to decay. In addition, the rate modifier $r_i(t)$ that governs the speed of decay is set to a constant that is different from the growth constant. This insures that the recognized word candidate will remain active in memory for a short period of time following the stimulus. Persistence of the recognized word permits higher level linguistic processes (e.g., semantic integration) to use the product of the recognition stage.

Insert Figure 1 about here

Figure 1 shows the time course of word recognition in the cohort activation model. The horizontal dotted line represents the activation level of the encoded stimulus word for the duration of the input. This input activation level serves as the asymptote for all the word candidates starting at a baseline of zero that match the input features. The activation level of each cohort was computed using a quartic Runge-Kutta numerical approximation (see Gerald, 1978) to the solution of the differential equation for each node. The activation functions for four word cohorts are shown by the solid curves in Figure 1. In this figure, all four cohorts match the first feature of the input word, and thus are activated. The second feature of the stimulus word is consistent with three of the candidates and they receive further activation. However, this feature does not match the fourth cohort and this node is deactivated. Similarly, the second and third features deactivate word candidates until a single candidate remains that is recognized.

Clearly, our activation model captures the essential operating characteristics of cohort theory as well as the specific assumptions of optimal efficiency, residual activation of eliminated candidates, and independence of candidate activation from cohort size. Moreover, this is true for a fairly wide range of parameters (for the specific implementation in this paper, the growth constant was 2.0, the decay constant was 3.0, and the asymptote set by the input was .6). Thus, this activation model of cohort theory can be used as the basis for generating several predictions about auditory word recognition.

Predictions from the Cohort Activation Model

In order to make specific predictions, it is necessary to derive responses from the activation functions. Since response latency is the dependent measure used in many recognition experiments (cf. Marslen-Wilson & Tyler, 1980), it is important to generate hypotheses based on reaction time. With respect to the activation model, two factors are assumed to mediate the speed of word recognition. The first factor is the temporal location (within the stimulus word) of the critical recognition point. This sets a lower bound on the time needed to recognize a word. The critical recognition point is determined by setting a threshold for the separation of the activation of the last word candidate from any deactivated nodes. A threshold of .3 was used for the model represented in Figure 1. With this threshold the input word was recognized after 91% of the stimulus was processed. Thus if we assume that the recognition features are phonemes and the duration of the entire four-phoneme word was 250 msec, the recognition point occurred 228 msec after the start of the word. This recognition point follows the deactivation of word candidates sharing up to three phonemes with the stimulus word. At this critical recognition point the listener can begin a response sequence. In other words, the listener cannot initiate a recognition response until the distractor candidates sharing 75% of the acoustic-phonetic information in the test word are eliminated.

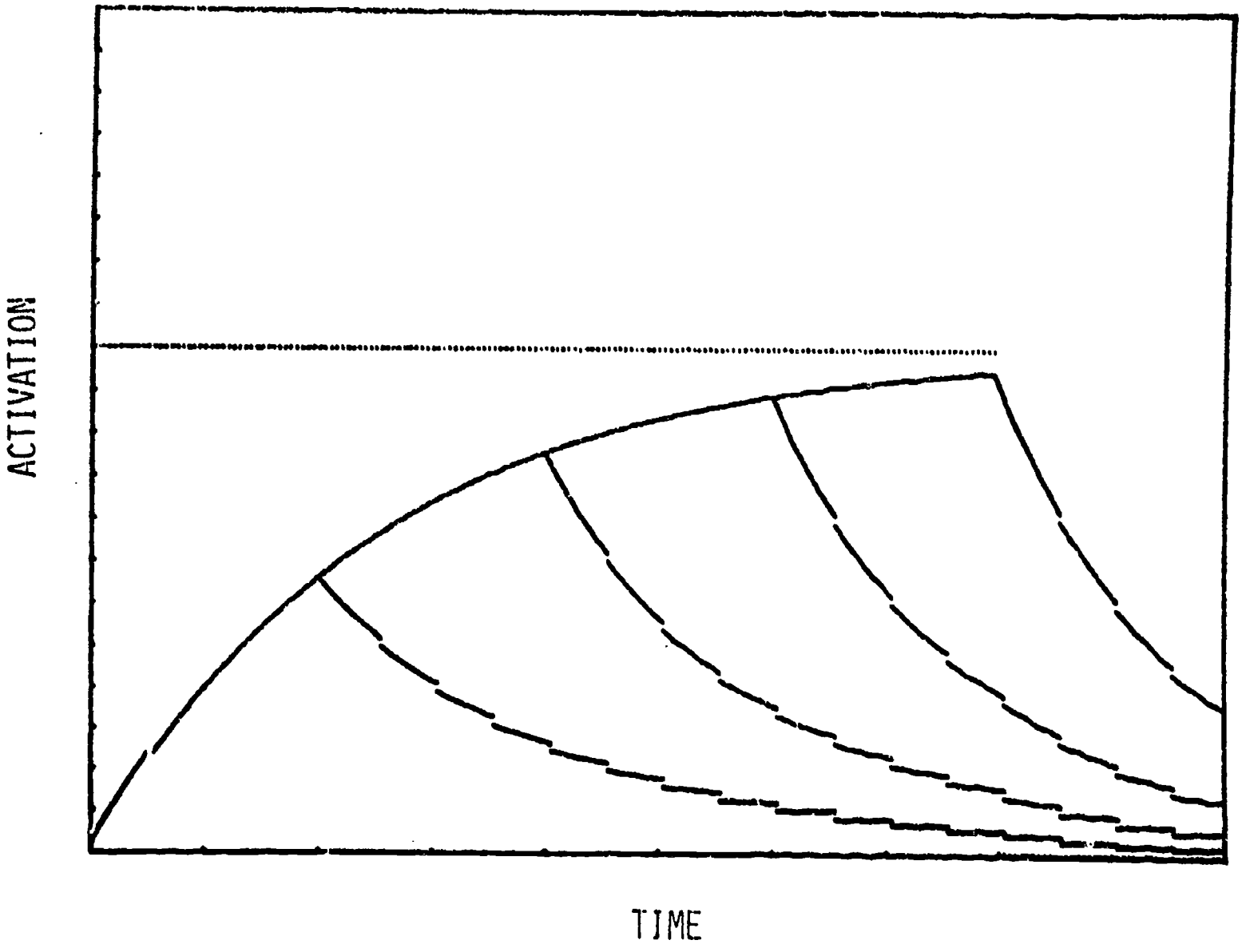


Figure 1. The effects of a single (unprimed) word on cohort activations. The horizontal dotted line represents the activation of the encoded input features. The solid curves show the time-course of activation of four cohorts.

The second factor that affects response time is the amount of activation of the accepted candidate relative to the activation level of the encoded input. To some extent, this factor might be thought of as an index of the confidence a listener has in identifying a word. While the principle of optimal efficiency specifies that a decision is made at the earliest possible moment, this does not mean that listeners are highly confident of such decisions. Indeed, consider the case where a word can be identified upon hearing the second acoustic-phonetic feature in the word. Even though a response can be made at this point, accumulating further evidence should increase confidence in the identification producing a faster response. Thus, in the activation model, decision time is modulated by the relative activation level of the accepted candidate. In the example given in Figure 1, at the critical recognition point the activation level of the recognized node is .03 units below the input activation level. Using a baseline decision time of 450 msec (cf. Tyler & Marslen-Wilson, 1982a), the relative activation level would add 14 msec to this baseline decision time to produce a total decision time of 464 msec. Reaction time from the onset of the stimulus would equal 692 msec which is the sum of the recognition point (228 msec) and the total decision time (464 msec). Of course, this value could be adjusted up or down depending on the selection of various parameters. But by using the same set of parameters for different hypothetical conditions, the relative effects of experimental manipulations on reaction time can be determined.

Let us take a more concrete example to show that the activation model can be used to generate hypotheses about auditory word priming in cohort theory. Recently, priming has been used to investigate the processes that mediate access to word meanings (e.g., Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982; Swinney, 1982). To date, this research has been concerned with the influence of the meaning of a prime word on access to the meaning of a second test word (cf. Meyer, Schvanevelt & Ruddy, 1975). However, it is also possible that the phonological representation of one prime word could facilitate or inhibit recognition of a second test word (cf. Tanenhaus, Flanagan & Seidenberg, 1980).

Indeed, the residual activation levels of the cohort members after a word is recognized suggests that phonological overlap between a prime and test word might facilitate the speed of recognition of a test word compared to an unprimed condition. Furthermore, the relationship between the amount of phonological overlap and the amount of residual activation suggests that reaction time facilitation should increase with increasing amounts of overlap between the prime and test words. Consider the following experiment. On each trial, subjects hear a prime word followed by a test word. On some trials, the prime and test words will be identical, while on other trials, although the prime and test words will be different, they will have the same initial acoustic-phonetic information. On these trials, the prime and test words will share the same initial phoneme or they will share the first two or three phonemes. Thus, if the prime and test words are all four phonemes long there would be four levels of acoustic-phonetic overlap.

Consider, in this context, the effects of a single four-phoneme word on the recognition system (e.g., the prime). This word will be recognized following activation of cohorts that share one, two, three, or all the phonemes in the word. Word candidates that share one phoneme will be deactivated first and candidates that share three phonemes will be deactivated last leaving the

recognized word (see Figure 1). Accordingly, the different cohorts will retain a residual amount of activation corresponding to the point at which elimination occurred. In other words, candidates sharing only one phoneme with the input will have much less residual activation than the cohorts that matched three phonemes of the input. This observation suggests that the recognized word should have the highest level of residual activation after the stimulus word is heard. In turn, the facilitation of recognition of a subsequent word should depend on the amount of phonological overlap with the preceding word. When the prime and test words are the same, recognition should be the fastest, with the three-phoneme overlap condition next fastest and so on.

Insert Figures 2 and 3 about here

The activation model can easily be used to derive these predictions. If the cohort activations in Figure 1 represent the effects of a prime word on the recognition system, the residual activations can be used to produce differential baseline activations when the test word is presented. Figure 2 shows the effects of a test word on cohort activations when the prime and test words are identical. The overall higher activation of the candidate corresponding to the test word results from its initial (residual from priming) activation. Figure 3 shows the effects on activation for a test word that differs from the prime in only the final phoneme. Finally, the results for a two-phoneme overlap between prime and test are shown in Figure 4 and the effects of a one-phoneme overlap condition are displayed in Figure 5. Computing the reaction times with the same threshold used for the unprimed case yields the following results: latency for recognition with a same-word prime is 598 msec; with a three-phoneme overlap the response latency is 666 msec; with a two-phoneme overlap the latency is 681 msec; and finally, with a one-phoneme overlap between prime and test words the response latency obtained is 686 msec. Compared to the unprimed response latency of 692 msec, all the conditions of phonological overlap produce some facilitation. However, comparing the response latencies in the various priming conditions indicates that some of the relative differences may be quite small. For example, the difference between the one-phoneme overlap condition and the two-phoneme overlap condition is predicted to be only 5 msec. This difference might not be observable with human subjects because of other factors producing variability in performance.

Insert Figures 4 and 5 about here

However, the general predictions are quite clear. An activation model of cohort theory that permits residual activation of eliminated candidates predicts priming effects based on phonological overlap. Thus, our instantiation of the

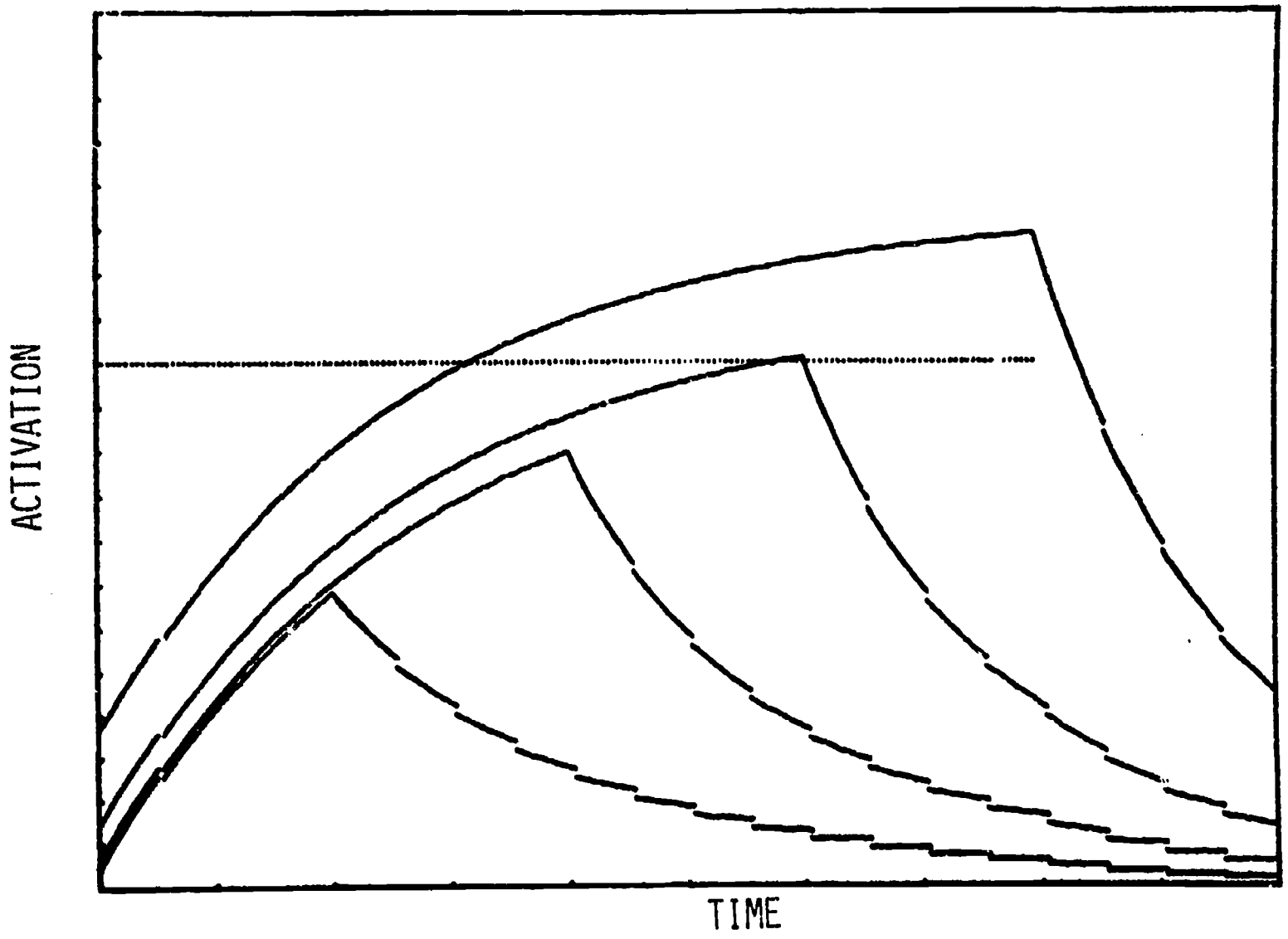


Figure 2. The effects of a test word on cohort activations following a same-word prime.

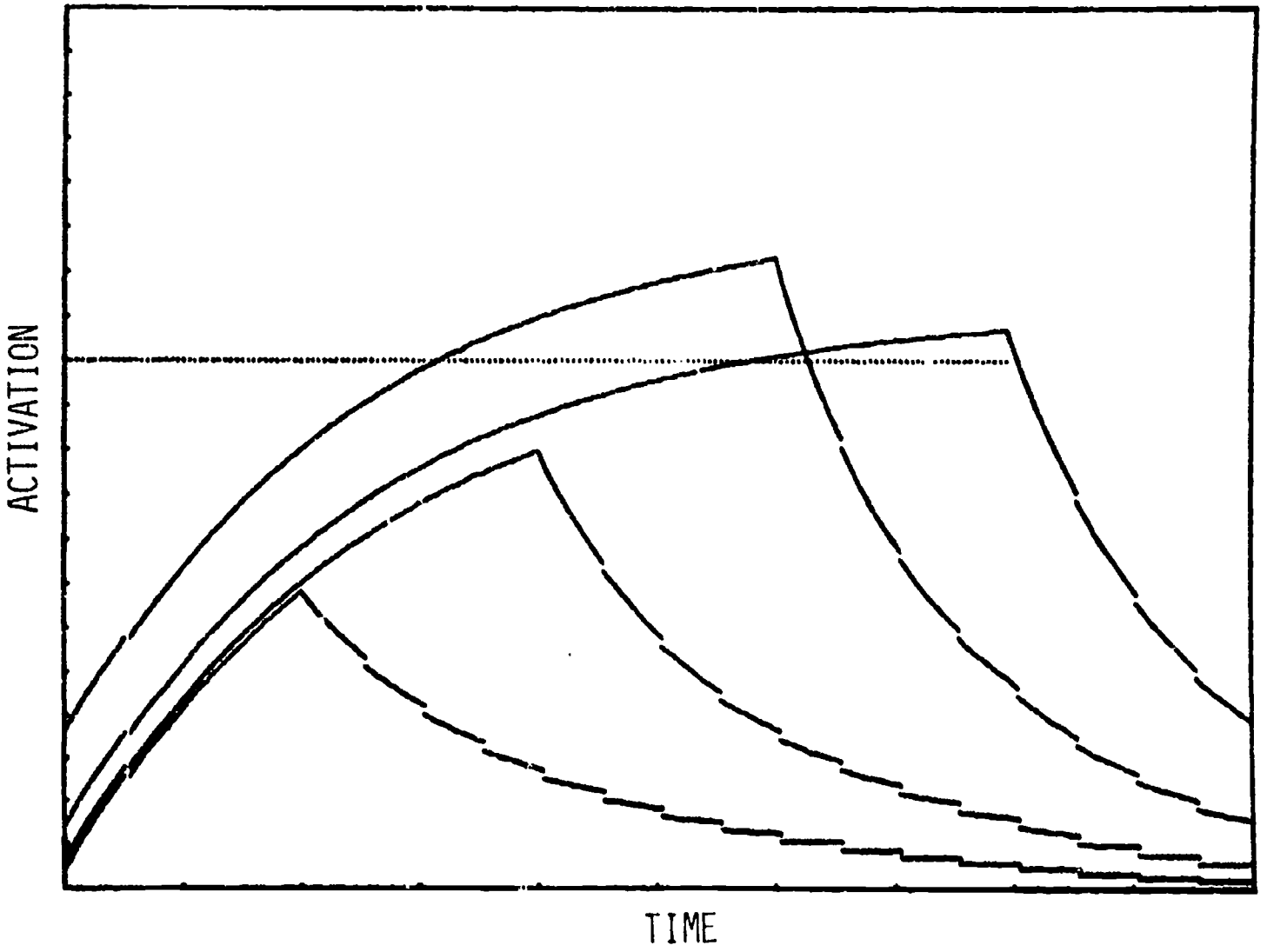


Figure 3. The effects of a test word on cohort activations following a prime that shared its first three phonemes with the test word

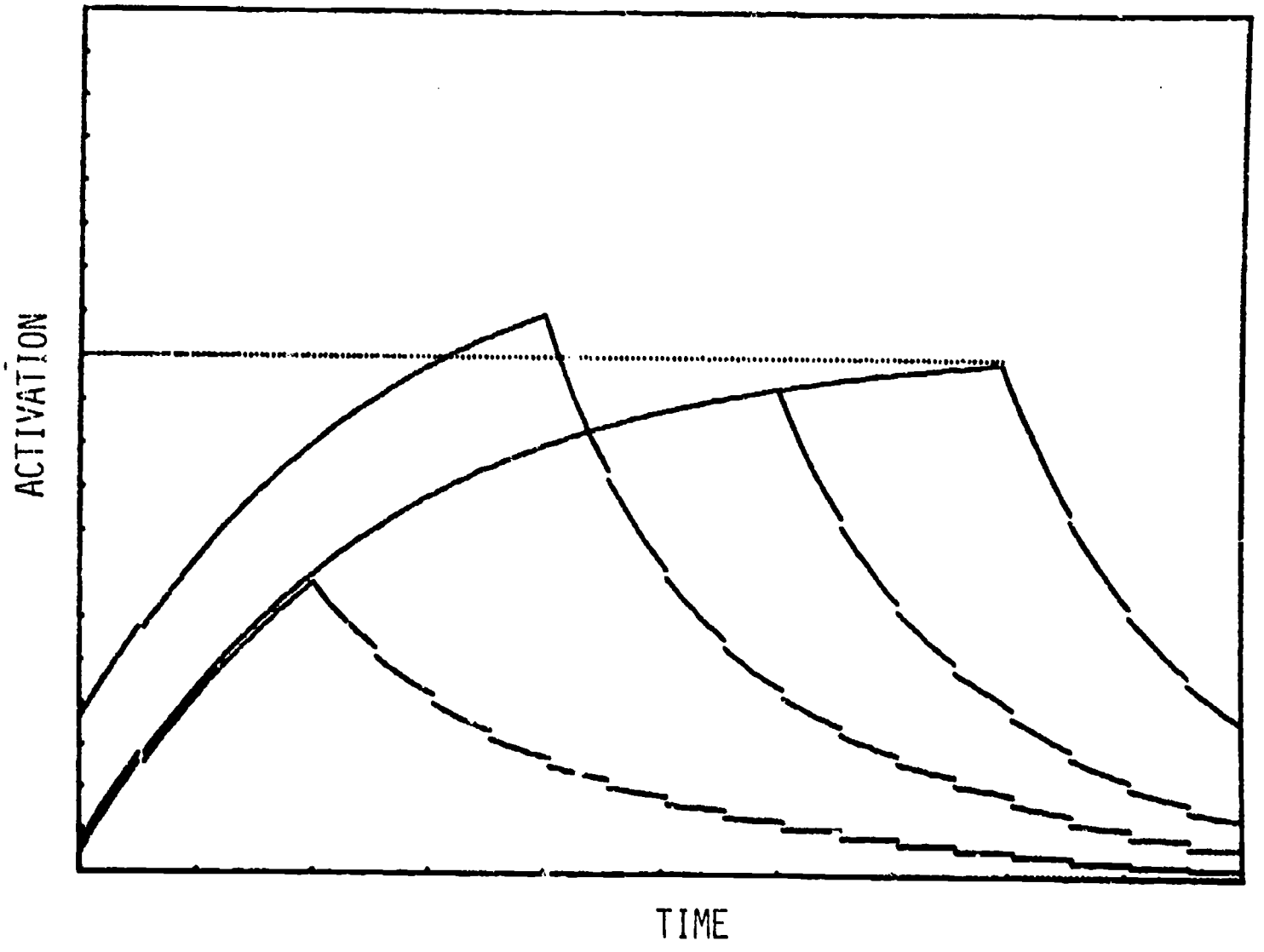


Figure 4. The effects of a test word on cohort activations following a prime that shared its first two phonemes with the test word.

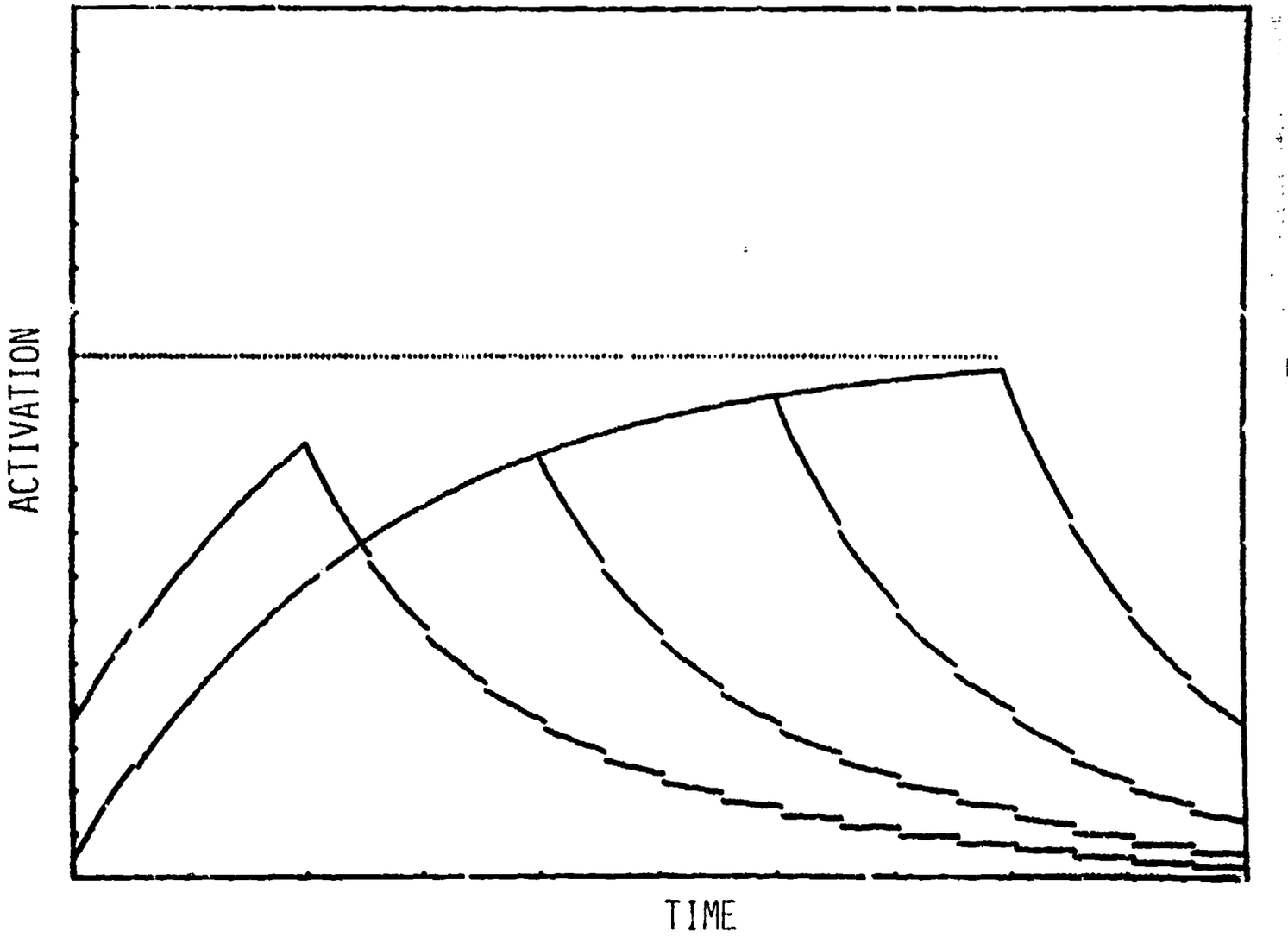


Figure 5. The effects of a test word on cohort activations following a prime that shared only its initial phoneme with the test word.

model generates several precise empirically testable hypotheses that cannot be derived from cohort theory as described qualitatively by Marslen-Wilson and his collaborators (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978; Tyler & Marslen-Wilson, 1982a, 1982b).

Recently, an initial attempt to test these predictions was made by Slowiaczek and Pisoni (Note 3) using phonological priming in a lexical decision task. On each trial, subjects heard a prime and test word or nonword pair and were asked to decide whether the test item was a word or a nonword. Unfortunately, this study did not include a baseline condition of either a test item without a prime or an unrelated prime. However, the results were quite interesting anyway. The same-word prime facilitated lexical decision latency compared to the other conditions, although the other priming conditions did not differentially affect performance. These results might be expected from the cohort activation model if it is assumed the lexical decision task is simply not sensitive enough to reveal some of the smaller RT differences. One result that was unexpected, however, was the finding that the latency to respond to nonwords was also significantly facilitated by a same-nonword prime. This result would not have been predicted by the activation model since all possible word candidates should be eliminated before the nonword is recognized. Thus, nonword recognition should not be affected by priming in this model.

At first glance, the nonword results would seem to disconfirm our initial implementation of the activation model. However, subsequent research is being conducted to further explore this finding. One recent study by Feustel (1982) has suggested that the enhancement effects produced by priming may be the result of activation of episodic codes rather than activation of lexical representations. Before any strong conclusions can be drawn, more sensitive procedures must be employed to assess effects of different amounts of phonological overlap and to separate any potential episodic effects from priming due to lexical activation.

Conclusions

In order to test autonomous theories of word recognition, Marslen-Wilson and Tyler (1980) have made certain constraining assumptions about this class of theories. The results they obtained contradicted predictions made by their version of these theories and, as a consequence, they rejected autonomous theories of word recognition. Clearly, there is a potential problem in constraining a theory in order to test it--the adopted constraints may simply be inappropriate, a point that has been made recently by Norris (1982). However, it is also true that a theory that is untestable is of little scientific value. Thus, while it may be necessary to constrain a theory, it is also important to make the additional assumptions as reasonable and consistent with the spirit of the theory as possible (cf. Tyler & Marslen-Wilson, 1982b). This is what we have tried to do in formulating a mathematical model of cohort theory. The choice of an activation mechanism was motivated by the verbal descriptions of the theory (e.g., Marslen-Wilson & Welsh, 1978) and the computational properties of this type of system. The advantage of this model is that it can generate explicit and novel hypotheses about the behavior of the recognition system described by cohort theory. The disconfirmation of these predictions would therefore provide evidence against the cohort theory just as Marslen-Wilson and Tyler's (1980) test

of their constrained version of autonomous theories led them to reject those theories. We expect to continue our work on models of word recognition, since the recognition of words is assumed to be an important subcomponent of spoken language comprehension. At the present time, theories of word recognition and lexical access are unusually vague and unspecified thereby preventing clear and explicit tests of these theories. Until theories of word perception are sufficiently well-specified to generate unique and testable predictions, the interpretation of experiments will be equivocal (cf. Norris, 1982; Tyler & Marslen-Wilson, 1982b).

Reference Notes

1. Salasoo, A., and Pisoni, D. B. Sources of knowledge in spoken word identification. RESEARCH ON SPEECH PERCEPTION, Progress Report No. 8, Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University, 1982.

2. Marslen-Wilson, W. D. Optimal efficiency in human speech processing. Unpublished manuscript.

3. Slowiaczek, L. M. and Pisoni, D. B. Priming auditory word recognition: Some tests of the cohort theory. RESEARCH ON SPEECH PERCEPTION, Progress Report No. 8, Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University, 1983.

References

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. Distinctive features, categorical perception, and probability learning: Some applications of a neural model. Psychological Review, 1977, 84, 413-451.

Cole, R. A., & Jakimik, J. Understanding speech: How words are heard. In G. Underwood (Ed.), Information processing strategies. London: Academic Press, 1978.

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

Cole, R. A., Jakimik, J., & Cooper, W. E. Perceptibility of phonetic features in fluent speech. Journal of the Acoustical Society of America, 1978, 64, 44-56.

Feustel, T. C. The repetition effect in word identification: Implications for the semantic-episodic distinction. Unpublished doctoral dissertation, Indiana University, 1982.

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

Gerald, C. F. Applied numerical analysis. Reading, Mass.: Addison-Wesley, 1978.

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

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.

McClelland, J. L. On the time relations of mental processes: An examination of systems of processes in cascade. Psychological Review, 1979, 86, 287-330.

McClelland, J. L., & Rumelhart, D. E. An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. Psychological Review, 1981, 88, 375-407.

Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. G. Loci of context effects in visual word recognition. In P. M. A. Rabbitt & S. Dornic (Eds.), Attention and performance V. New York: Academic Press, 1975.

Norris, D. Autonomous processes in comprehension: A reply to Marslen-Wilson and Tyler. Cognition, 1982, 11, 97-101.

Rumelhart, D. E., & McClelland, J. L An inte. tive activation model of context effects in letter perception: Part 2. The contextual enhan ement effect and some tests and exten ions of the model. Psychological Review, 1982, 89, 60-94.

Seidenberg, M.S., Tanenhaus, M. K., Leiman, J. M., Bienkowski, M. Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. Cognitive Psychology, 1982, 14, 489-537.

Swinney, D. A. The structure and time-course of information interaction during speech comprehension: Lexical segmentation, access, and interpretation. In J. Mehler, E. C. T. Walker, & M. Garrett (Eds.), Perspectives on mental representation. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1982.

Tanenhaus, M. K., Flanagan, H. P., & Seidenberg, M. S. Orthographic and phonological activation in auditory and visual word recognition. Memory & Cognition, 1980, 8, 513-520.

Tyler, L. K., & Marslen-Wilson, W. Speech comprehension processes. In J. Mehler, E. C. T. Walker, & M. Garrett (Eds.), Perspectives on mental representation. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1982(a)

Tyler, L. K., & Marslen-Wilson, W. Conjectures and refutations: A reply to Norris. Cognition, 1982, 11, 103-107. (b)

III. INSTRUMENTATION AND SOFTWARE DEVELOPMENT

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

JOT: Improved Graphics Capabilities for KLTEXC*

Bob Bernacki

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

*The development of this software was supported by NIH grant NS-12179 to Indiana University in Bloomington. I thank Tom Carrell for advice and suggestions.

The KLTEXC program is a versatile implementation of the Klatt digital synthesizer developed here at Indiana University (Kewley-Port, 1978). One version of this program (Carrell, 1978) utilized a digitizing tablet for graphic input and a graphic display scope for output. This version, while providing more efficient user interaction with the synthesizer, was rather limited in its capabilities. The graphic interface supported only formants one, two, and three, and required a Polaroid camera mounted on the display scope for hardcopy output.

Subsequent use of this system demonstrated both the usefulness of graphic I/O for synthesis control and the limitations of the system which was then in use. Recent hardware improvements in the laboratory have made possible a much improved graphic subsystem for KLTEXC called "JOT". This paper summarizes the features of this new subsystem.

As in the original JOT package, the major hardware components supporting the graphics features consist of a Summagraphics 2000 digitizing tablet, and a DEC VT-11 graphics system. The original system configuration also included a DEC 11/05 processor and RK05 disc drives. The Summagraphics tablet is an X-Y device of .1 mm resolution on a 60 cm square active surface. A small hand held unit, commonly termed a "mouse", receives the tablet signals and sends its' coordinates to the tablet controller. The tablet controller is serviced by an interrupt routine for low program overhead. The DEC graphics display system consists of a VT-11 DMA display processor, and a VR-17 refresh display scope. The VT-11 display system provides a 1000 point resolution for both the horizontal and the vertical dimensions.

In addition to the tablet and display system, new equipment supports larger parameter buffers, and hardcopy graphics. The current system is centered around a DEC 11/34 processor configured with 80k of extended memory. Hardcopy graphics is now available through a TEKTRONIX 4010 display terminal, and the accompanying 4631 hardcopy unit. Other equipment providing improved operation includes two 80 megabyte CDC 9762 disc drives, and an FP-11A floating point hardware unit. A VRM-11 video monitor is used as an adjunct display to the VT-100 system console (see Forshee, 1979).

Insert Figure 1 about here

The Current Graphics Subsystem

The new version of the JOT command has been considerably expanded in scope and function. In addition to the original JOT functions of parameter entry and display, various new capabilities are provided. The first group, COPY and SCALE, allow for direct transfer and manipulation of parameters over the entire 2.2 second buffer. A second group, GLOBAL formant plot, parameter PLOT, and parameter ERASE, permit rapid configuration of the display grid. The hardcopy command may be used to transfer the VT-11 screen display to hard copy for future reference. Another command, MEASURE, automatizes measurements traditionally performed with a pencil and paper from a ruled spectrogram. These commands and the current global parameters are displayed for the user on the adjunct monitor screen for easy reference. The JOT command display and parameter list is shown in Figure 1.

```
        ---JOT Options---

C = COPY A PARAMETER
E = ERASE PARAMETERS FROM THE GRID
G = DISPLAY/ERASE  F4, F5, F6
I = INITIALIZE TABLET
J = INPUT PARAMETERS FROM THE TABLET
M = MEASURE PARAMETERS ON A SPECTROGRAM
P = DISPLAY PARAMETERS ON THE GRID
S = SCALE A PARAMETER
T = TRANSFER THE DISPLAY TO THE TEK 4010
X = EXIT JOT (RETURN TO KLTEXC COMMAND ENTRY)

    ---JOT GLOBAL PARAMETER LIST---

         F4                  BW4
         F5                  BW5
         F6                  BW6
         FNP                 BWNP
         FNZ (VARIABLE)      BWNZ

    ---JOT VARIABLE PARAMETER LIST---

F0  AV  F1  B1  F2  B2  F3  B3  NZ  AN
AR  AB  AS  AF  A1  A2  A3  A4  A5  A6
```

Figure 1. The JOT Menu Display.

The JOT Display

To implement real-time plotting for the VT-11 display, some custom routines were added to the DEC Fortran graphics package. These routines provided performance levels that otherwise would have been impossible.

Insert Figure 2 about here

The JOT display may be configured for a buffer size of either 1.0 or 2.2 seconds, and the display grid is fixed at 5 msec resolution. Figure 2 shows the display configuration. A floating grid is available in place of the fixed top or bottom grids: it can be positioned anywhere in the vertical dimension at the user's convenience. The JOT parameter entry command uses a full resolution plot of one point for each 5 msec interval. The parameter plot command displays F1, F2,and F3 with connected lines for every 2 points, and the other parameters with every 4 points connected. Accompanying each plot on the screen is a two letter parameter designator to the left of the grid.

Tablet Input

The tablet system is initialized with the JOT command I. The tablet may be configured in one of two tablet modes, freehand and trace. The default mode is freehand, which gives the user a preset tablet work space that does not require the mounting of a spectrogram. In addition, a trace mode is available in which the user must actually set up a spectrogram on the tablet surface. Alignment points at specified positions on the spectrogram are entered as tablet data to permit precise adjustment for skew.

The JOT parameter entry command, J, is used to the trace spectrograms for entry of data into parameter buffers. The tablet data entry routine incorporates several advanced features. These features, in conjunction with the DEC-11/34 CPU and the FP-11A floating point unit, permit rapid tracing to be processed without resorting to interpolation.

The J command permits two modes of operation. In mode one the tablet data is entered into the parameter buffers. Data is entered by a left to right motion of the "mouse" to conveniently allow for immediate backup and correction. Mode two is a free tracking mode that allows the user to reposition the "mouse" anywhere on the tablet without entering data. When mode one is selected after mode two, data may be entered to overwrite previous values. The unidirectional data entry and the free tracking features give the user complete touch-up freedom while avoiding accidental data erasure. When the "mouse" is in proximity to the tablet, the display tracks its position in real-time. Hertz, decibels, and milliseconds are presented as digital meter readouts adjacent to the parameter grids. With these meters, the "mouse" position may be precisely located for data entry.

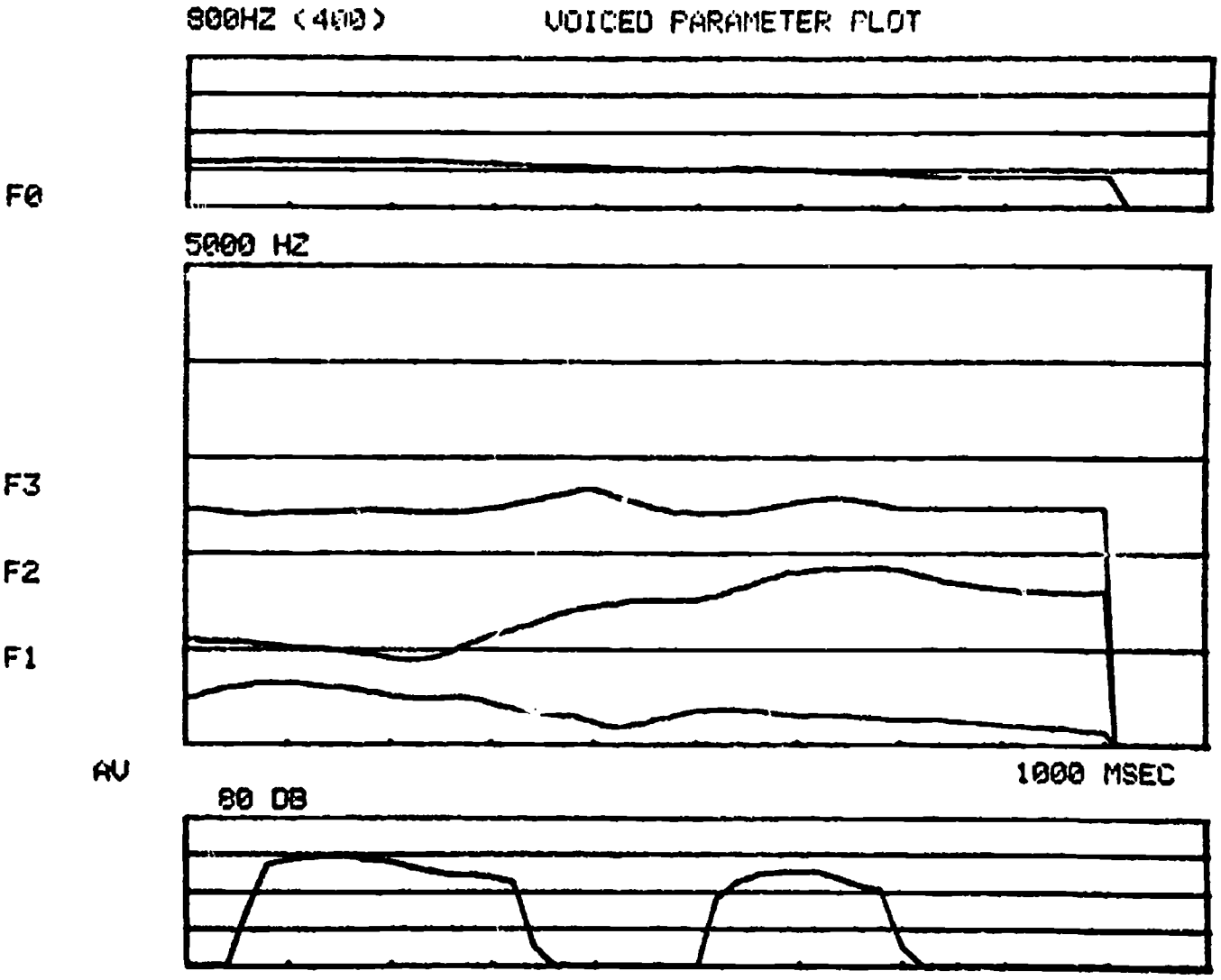


Figure 2. Parameter Display for Vowel Synthesis.

The J command incorporates data processing techniques to provide a high degree of flexibility and performance. Tablet data values are processed with hysteresis and a zone averaging technique. A linear interpolation is performed if the data input rate is faster than the program can track. When data entry is complete, the smoothing provision may be used to remove irregularities that can result from manual tracing. Four levels of smoothing are provided in an interactive system, in which the user may view the results to determine the best level. The smoothing procedure and the display update is accomplished in less than two seconds.

Parameter Manipulations

The C and S commands permit entire parameters to be transfered or modified. The C command transfers the values from a source parameter to a destination parameter (eg: source=AV, destination=AF). This feature is especially useful in conjunction with the scale command when creating the amplitude parameters. The scaling of a parameter has been provided as a universal command to allow for special applications. DB parameters may be scaled by addition or subtraction of dB offsets, and optionally compressed or expanded with a ratio. Frequency parameters are scaled by multiplication with a ratio scale factor ranging from .01 to 100. When scaling is performed, the display updates in real-time to provide the user with immediate visual feedback.

Display Management

Plots of parameters on the display grid are managed with the plot, erase, and global plot commands. Parameter letter pair designations are entered sequentially to plot or erase parameter displays. Each operation is accomplished within seconds. The global command toggles the display of the global parameters F4, F5,and F6 on and off. This display is useful as a reference when creating frication, or when entering F3 data.

The VT-11 display may be transfered to the TEKTRONIX 4010 with the T command. The display can be titled by entering an identifier phrase, and the 4010 screen optionally erased t. permit an overlay of displays. A switch on the TEKTRONIX 4010 transfers the display to a TEKTRONIX 4631 unit to obtain a hardcopy.

Additional JOT Commands

The measure command enables the user to measure frequencies without entering values into the buffer. The values are read directly from the display grid meters, providing a convenient method for determining global parameter values.

The exit command performs two functions. The first is the setting of an end time to truncate the fixed JOT buffers for data transfer to the KLTEXC files. The second function transfers the parameter values, and returns control to the KLTEXC program.

Creating KLTEXC Parameter Files with JOT

In general, narrow bands of formant energy are produced with the cascade resonators driven by the unvoiced AH or voiced AV level control. Figure 2 presents a JOT display of parameters for the voiced portion of a word. The important parameters from any spectrogram may be traced in as follows. Wide bands of unvoiced energy are created with the parallel resonator system which consists of the resonators F1-F6, (Bandwidths B1-B6) and the level controls A1-A6, and AF. Figure 3 shows a display of unvoiced portions of a synthetic word. The AV parameters are traced from an average amplitude spectrogram. This parameter should then copied to the AF, AH, and AN parameters. Scaling should be applied to adjust each of these to appropriate levels, and edited via the freehand tablet command to remove portions of the synthesis where each is inactive. For frication, the AF parameter is copied to the A1 to A6 level controls and scaled where necessary. Bandwidths one through three are entered to broaden the resonators in the frication portions. Formants one through three are traced in next, and the nasal zero, if present, is approximated.

Insert Figure 3 about here

The fundamental frequency may be obtained from a narrow band spectrogram, and the values entered via the freehand tablet mode. The measure command is invoked to measure and record average values for several of the global parameters that will be manually entered on the terminal in KLTEXC. Lastly, the exit command may be used to set an end time for the buffer and return control to the KLTEXC program for waveform synthesis.

Summary

New hardware and software have resulted in graphic I/O rates that keep pace with the user. This increased flexibility permits efficient use of KLTEXC in sentence length synthesis. The modular graphics software, programmed for KLTEXC, will be adapted to other synthesizers in the near future.

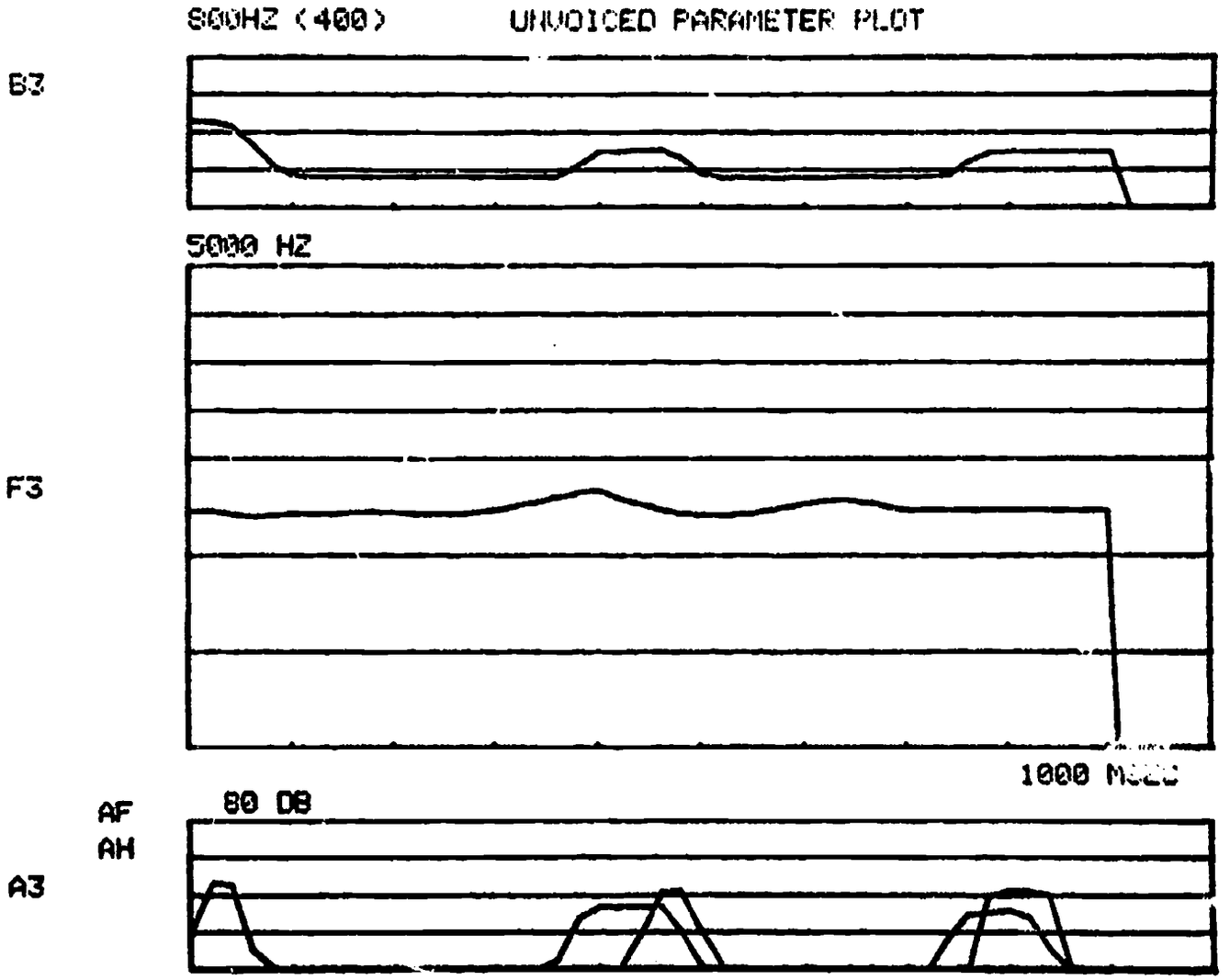


Figure 3. Parameter Display for Fricative Synthesis.

References

and Kewley-Port, D. Graphic Support for KLTEXC. Research on Speech ..ption Progress Report No. 4, 1978, 247-256.

Forshee, J. Computer facilities in the Speech Perception Laboratory. Research on Speech Perception: Progress Report No. 5, 1979, 449-473.

Kewley-Port, D. KLTEXC: Executive Program to Implement the KLATT Software Speech Synthesizer. Research on Speech Perception: Progress Report No. 4 1978, 235-246.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 8 (1982) Indiana University]

EARS: A Simple Auditory Screening Test*

Laurie Ann Walker

Speech Research Laboratory
Indiana University
Bloomington, Indiana 47405

*This work was supported by NIH grant NS-12179 to Indiana University in Bloomington. I thank Diane Kewley-Port for her help and assistance in developing this package.

Introduction

EARS is a program designed to assess the auditory acuity of subjects prior to their participation in speech perception experiments. The program was written by Laurie A. Walker and Diane Kewley-Port for use on the two PDP-11/34 computers in the Speech Research Laboratory. The procedure used in EARS is a modification of a standard audiometric screening test described by Davis (Davis & Silverman,1970). This recently implemented Fortran-IV program promises to be useful in screening subjects for perceptual experiments.

EARS tests each ear at three frequencies (1000, 2000, & 4000 Hz). These frequencies were chosen based on standard reference zero levels (suggested by the International Organization of Standards) and were intended to test the frequency range most important for understanding speech. The frequency range tested by EARS is shown in Figure 1 (adapted from Davis & Silverman, pp.192-3).

Insert Figure 1 about here

Because the 500 Hz tone was found to be nearly inaudible due to the ambient noise level in the subject testing rooms, it was not chosen for testing. The screening is conducted at levels of 20 dB and 40 dB SPL. These levels are presented in a random order intermixed with silent intervals that serve as catch trials.

Before Running EARS

In order to ensure that the output levels at the headphones are correct, the entire system must be calibrated regularly. This calibration involves determining, for example, the necessary attenuation to produce .1 volts at the headphones for a 1000 Hz pure tone. EARS takes into account the attenuation levels measured by this calibration.

In addition to the system calibration, which should take place every two or three months, the manual attenuators must be set each time EARS is run. The manual attenuation levels for each channel (corresponding to each ear) are set by adjusting the appropriate dials on the analog equipment rack. EARS compensates for the settings on the manual attenuators. This compensation enables an experimenter to set the manual attenuation levels that will be appropriate for a perceptual experiment to be run immediately following the hearing test.

Subjects are provided with high quality Telephonics TDH-39 (300z) headphones and response boxes to enter their judgments. The response boxes must have a cue-light and at least two buttons. The left-most button is labelled "NO," and the button next to it is labelled "YES." Prior to testing, subjects are supplied with written instructions concerning the screening test procedure.

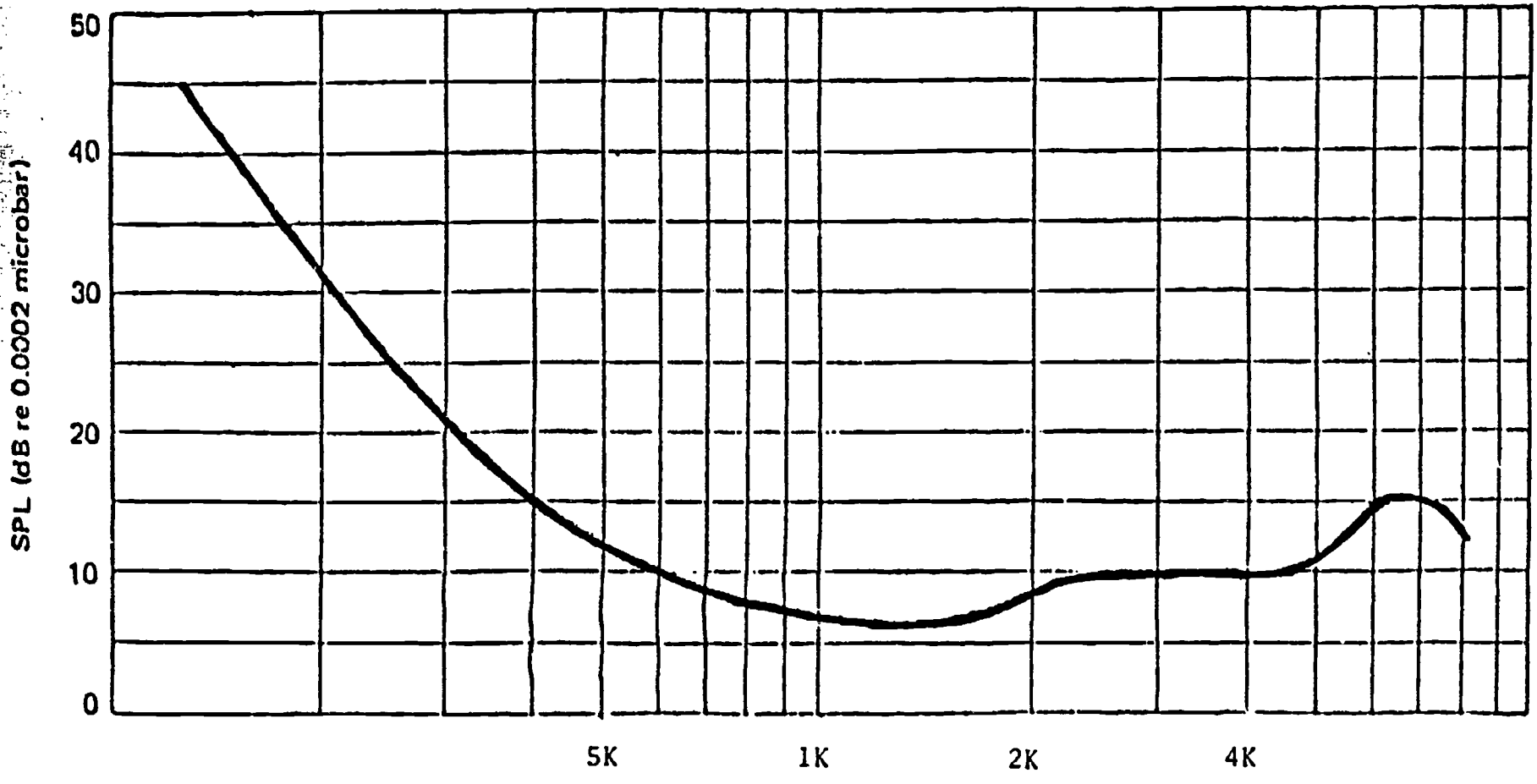


Figure 1. ISO threshold values for pure tone.

Running EARS

Upon entering the EARS program, the user is queried for the manual attenuation levels for each channel that were set in accordance with the subsequent experiment. The user is also asked for session identification information, the session number, the subject numbers, and the number of stimuli. The number of stimuli is usually three. Finally, the user is asked for the response file name.

All other file names associated with the EARS program are contained in DATA statements. These files include the stimulus set file and the random order file. The stimulus set file contains the names of the stimulus files for the three tones, and the random order file contains the order of the presentation of the tone and silent intervals. The random order file is structured as a sequence of blocks, each containing the numbers one through six in random succession. These files are resident on the system disk, along with EARS, and are easily modified. After the user responds to all the questions, EARS waits for the user to initiate the execution of the screening test procedure by typing a carriage return.

Testing Procedure and Instructions to Subjects

Subjects are given a brief description of the event sequence for each trial. They are told that a cue-light will be illuminated at the top of each response box to indicate the beginning of each trial. The subjects are then instructed that the two lights above the response buttons will light up indicating an "observation" interval - the time period when either a tone or a silent interval is presented. Subjects are informed that when the lights go out, the "observation" interval is over, and they have three seconds to respond. Their task is to determine whether they hear a tone during the "observation" interval. If they hear a tone, they press the button labelled "YES" on the response box; otherwise, they press the button labelled "NO." Subjects must respond with a "Yes" or "No" on each trial.

The testing procedure consists of two parts: familiarization and testing. The two parts are executed for each ear, beginning with the left ear. In the familiarization procedure, a 1000 Hz tone is presented at 60 dB SPL. This tone is of sufficient amplitude to orient subjects to the ear being tested. Each subject's response interval is terminated immediately after his/her first response on a given trial, and trial presentation is paced to the slowest subject's responses. If all subjects respond that they have heard the tone, an appropriate message is printed at the experimenter's terminal.

If one or more subjects do not respond correctly, the output level is incremented by 5 dB and the familiarization procedure is repeated. If the output level reaches 80 dB and a subject is still not responding correctly, EARS will print a message at the experimenter's terminal indicating the number of each subject who is not responding correctly. Because it is unlikely that a subject cannot detect a 1000 Hz tone at 80 dB, the familiarization procedure is useful for identifying subjects who have somehow misunderstood the instructions. If a subject fails to respond correctly, the experimenter may provide further instructions concerning the procedure. After the experimenter types a carriage return, the familiarization procedure will start again. When all subjects have responded correctly, the screening test begins immediately.

For the screening test, pure tones at three frequencies are presented in the following order: 1000, 2000, and 4000 Hz. For each tone, there are six trials. On three trials the tone is presented (once at 40 dB and twice at 20 dB); the other three trials are silent intervals. The order of the presentation of the signal or silence is determined by a prearranged random order file. The order varies randomly for each tone. After the screening procedure is completed for the left ear, both parts of the procedure (including familiarization) are repeated for the right ear.

After Running EARS

Analysis of the data is carried out immediately after a given session has been completed. EARS computes the percent correct for each subject for each dB level for each ear, as well as the percent correct for each subject for each tone for each ear. The data for the dB levels is printed out following each session, whereas a more detailed analysis including tone information is available from the disk file. The disk file also contains the session information, the random order of dB levels, and the raw trial-by-trial response data. The disk file may be examined by using the standard TYPE or PRINT commands on the RT-11 operating system.

The experimenter can examine the data from the screening test at any time; but typically this occurs after a perceptual experiment has been completed. The data from a subject who performs poorly on the screening test may simply be excluded from the analysis of the experimental results. In addition, the experimenter may choose to inform the subject, after the experiment, that a hearing impairment was suggested by the initial screening test. The subject is advised to obtain a more complete audiological examination elsewhere. It is emphasized both to subjects who do well and to subjects who do not do well on the hearing test that it is merely a screening test, designed for the purposes of our laboratory, and that the test is not meant to be a complete diagnostic audiological exam.

Summary

EARS provides the user with a fast and simple means of screening subjects for auditory perceptual experiments. In less than ten minutes, EARS can screen six subjects simultaneously, as opposed to the sixty minutes required to test each subject individually with a standard audiometer. Although the frequency range tested by EARS is quite limited, it is the most important range for understanding speech; hence, it is most adequate for our purposes.

References

Davis, H. Audiometry: pure tone and simple speech tests. In Davis, H. & Silverman, S.R. (Ed.), Hearing and Deafness. New York: Holt, Rinehart, & Winston, 1970, 179-220.

IV. Publications

Grunke, M. E. and Pisoni, D. B. Some Experiments on Perceptual Learning of Mirror-Image Acoustic Patterns. Perception & Psychophysics, 1982, 31, 3, 210-218.

Pisoni, D. B., Aslin, R. N., Perey, A. J. and Hennessy, B. L. Some Effects of Laboratory Training on Identification and Discrimination of Voicing Contrasts in Stop Consonants. Journal of Experimental Psychology: Human Perception and Performance, 1982, 8, 2, 297-314.

Brunner, H. and Pisoni, D. B. Some effects of perceptual load on spoken text comprehension. Journal of Verbal Learning and Verbal Behavior, 1982, 21, 2, 186-195.

Walley, A. C. and Pisoni, D. B. Review of J. Morton & J. Marshall (Ed.), "Psycholinguistics." Journal of Communication Disorders, 1982, 15, 63-75.

Pisoni, D. B. Perception of Speech: The Human Listener as a Cognitive Interface. Speech Technology, 1982, 1, 2, 10-23.

Kewley-Port, D. Measurement of formant transitions in naturally produced stop consonant-vowel syllables. Journal of the Acoustical Society of America, 1982, 72, 2, 379-389.

Pisoni, D. B. Speech Technology: The Evolution of Computers that Speak... and Listen. Bloomington: Indiana University Office of Research & Graduate Development, 1982.

Pisoni, D. B. Lexical Access. In A. S. House (Ed.), Project SCAMP 1981: Acoustic Phonetics and Speech Modeling. Princeton, NJ: Institute for Defense Analyses, Communications Research Division, 1982.

Pisoni, D. B. In Defense of Segmental Representations in Speech Processing. In A. S. House (Ed.), Project SCAMP 1981: Acoustic Phonetics and Speech Modeling. Princeton, NJ: Institute for Defense Analyses, Communications Research Division, 1982.

Nusbaum, H. C. and 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: Lockheed, 1982.

Technical Reports:

Aslin, R. N., Pisoni, D. B. and Jusczyk, P. W. Auditory Development and Speech Perception in Infancy. Technical Report No. 4, June 1, 1982.

Manuscripts to be published:

Pisoni, D. B. Some measures of intelligibility and comprehension. In J. Allen (Ed.), Conversion of Unrestricted English Text-to-Speech. 1983 (In Press).

Sinnott, J. M., Pisoni, D. B. and Aslin, R. N. Pure Tone Thresholds in the Human Infant and Adult. Infant Behavior and Development, 1983 (In Press).

Pisoni, D. B. Perceptual evaluation of voice response systems: Intelligibility, Recognition and Understanding. Proceedings of the Workshop on Standarization for Speech I/O Technology. Washington, D. C.: National Bureau of Standards, 1982 (In Press).

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, 1982 (In Press).

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 (In Press).

Pisoni, D. B. and Nusbaum, H. 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 (In Press).

McClasky, C. L., Pisoni, D. B. and Carrell, T. D. Effects of Transfer of Training on Identification of a New Linguistic Contrast. Perception & Psychophysics, 1982 (In Press).

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 (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, 1983 (In Press).

Kewley-Port, D. Time-varying features as correlates of place of articulation in stop consonants. Journal of the Acoustical Society of America, 1983 (In Press).

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 (In Press).

Pisoni, D. B. Categorical perception of speech and nonspeech signals. In S. Harnad (Ed.), Categorical Perception. New York: Cambridge University Press, 1983 (In Press).

V. Speech Research Laboratory Staff, Associated Faculty and Technical Personnel:

(1/1/82 - 12/31/82)

Research Personnel:

David B. Pisoni, Ph.D. -------- Professor of Psychology and Director
Richard N. Aslin, Ph.D. ------- Professor of Psychology
Hans Brunner, Ph.D. ---------- Visiting Assistant Professor
Beth G. Greene, Ph.D. --------- Assistant Research Scientist
Diane Kewley-Port, Ph.D. ------ Research Associate*
Michael R. Petersen, Ph.D. ---- Assistant Professor of Psychology
Eileen E. Schwab, Ph.D. ------- Visiting Assistant Professor
Rebecca Treiman, Ph.D. -------- Assistant Professor of Psychology

Cathy A. Kubaska, Ph.D. ------- NIH Post-doctoral Fellow
Howard Nusbaum, Ph.D. --------- NIH Post-doctoral Fellow

Thomas D. Carrell, M.A. ------- Graduate Research Assistant
Timothy C. Feustel, B.A. ------ Graduate Research Assistant**
Janis C. Luce, B.A. ----------- Graduate Research Assistant
Paul A. Luce, B.A. ------------ Graduate Research Assistant
Peter Mimmack, B.A. ----------- Graduate Research Assistant
Aita Salasoo, B.A. ------------ Graduate Research Assistant
Louisa M. Slowiaczek, B.A. ---- Graduate Research Assistant
Amanda C. Walley, B.A. -------- Graduate Research Assistant

Technical Support Personnel:

Mary Buuck, A.A. -------------- Research Assistant (Infant Laboratory)
Jerry C. Forshee, M.A. -------- Computer Systems Analyst
Nancy J. Layman --------------- Administrative Secretary
David A. Link ----------------- Electronics Engineer

Robert Bernacki --------------- Undergraduate Research Assistant
Mike Dedina ------------------- Undergraduate Research Assistant
Esti Koen --------------------- Undergraduate Research Assistant***
Laurie Ann Walker ------------- Undergraduate Research Assistant

*Currently at Bell Laboratories, Murray Hill, NJ

**Now at Bell Laboratories, Lincroft, NJ

***Now at Control Data Corporation, Minneapolis. MN