RESEARCH ON SPEECH PERCEPTION

Progress Report No. 7

January 1981 - December 1981

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-12179-06

and

National Institutes of Health
Training Grant No. NS-07134-03

Table of Contents

INTRODUCTION

This is the seventh annual progress and status 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 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 other colleagues.

We are distributing progress 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 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

Capacity Demands in Short Term Memory for Synthetic and Natural Speech*

Paul A. Luce
Timothy C. Feustel
David B. Pisoni

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

*This research was supported by NIH Research Grant NS-12179 to Indiana University in Bloomington. An earlier version of this paper was presented at the meetings of the Acoustical Society in Miami, December, 1981.

ABSTRACT

In order to determine the locus of previously demonstrated difficulties in the perception of synthetic speech, three experiments comparing recall for synthetic and natural lists of monosyllabic words were performed. If the perception of synthetic speech requires increased processing capacity in either early encoding stages, rehearsal, or both, recall differences between synthetic and natural speech should arise when short term memory is differentially stressed. In the first experiment, three presentation rates (one, two, and five sec per word) were used to vary the demands placed on short term memory in terms of encoding time. Although free recall was consistently poorer overall for the synthetic lists at all presentation rates, the decrement for synthetic stimuli did not increase differentially with faster rates. A similar pattern of results was obtained in a second experiment in which strings of digits of varying length (zero, three, and six digits) were presented visually for retention prior to free recall of each spoken word list in a pre-load paradigm. However, the number of subjects able to correctly recall all of the digits was considerably lower for the six-digit list relative to the three-digit list when the following word lists were synthetic. In the third experiment capacity demands on short term memory were directly manipulated by requiring subjects to recall lists of natural and synthetic words in the exact order in which they were presented. Differences in ordered recall between the synthetic and natural word lists were substantially larger for the primacy portion of the serial position curve than the recency portion. This result for ordered recall demonstrates that increased processing demands for synthetic speech interfere with rehearsal and subsequent transfer of items to long term memory. Taken together, our results indicate that difficulties observed in the perception and comprehension of synthetic speech are due, in part, to increased processing demands in short term memory. These findings are similar to results previously reported for recall of lists of items presented in noise in demonstrating a trade-off between encoding difficulty and processing capacity in perception and recall of synthetic speech. Implications for the use of synthetic speech in various types of voice response applications are discussed in terms of constraints on the human listener as a cognitive interface.

INTRODUCTION

Over the past few years, the sophistication of voice-response devices using synthetic speech has increased rapidly. Such systems are beginning to be used in reading aids for the blind, in speaking aids for the deaf, in computer-aided instruction, and in a variety of consumer products. Yet despite the wide use of synthetic speech in many voice-response applications and the expectation of even greater use of these devices in the future, there is at present little basic or applied research on the detailed cognitive processes by which we perceive and comprehend synthetically-generated speech.

For a naive listener, synthetic speech often seems difficult to understand. Problems may arise in the recognition of words and the interpretation of the meaning of sentences because of the distracting, mechanical quality of the speech signal. In this paper, we investigated a number of possible explanations for the difficulties typically observed in the perception and comprehension of synthetic speech.

Several investigators (Allen, 1976; Nickerson, 1977) have suggested that prosodic differences between synthetic and natural speech constitute the major difficulty in the comprehension of synthetic speech, particularly fluent synthetic speech. In natural speech, intensity, relative durations of segments and words, and changes in pitch are modulated by a complex set of physiological, phonetic, and linguistic factors that are as yet poorly understood (see Klatt, 1976). To attain high quality speech synthesis, it appears that these same natural variations must also be incorporated into the speech if perception and comprehension are to proceed normally.

Another possible explanation for the difficulties observed in perception of synthetic speech may be found at the relatively early stages of perceptual analysis and encoding at which words are recognized from their phonetic representations (Pisoni, 1981). Synthetic speech is often generated by rules that manipulate only a limited number of the potential acoustic cues to the phonological representation of the message. Thus, perception of synthetic speech may be adversely affected by only a partial specification of the acoustic cues to phonetic segments. This difficulty in early acoustic-to-phonetic recoding may therefore contribute directly to problems in word recognition and the subsequent processes involved in lexical access (see Pisoni, 1981).

Finally, the difficulties observed in the perception and comprehension of synthetic speech may arise from more general constraints on the processing of information in short term memory. In particular, synthetic speech may require more processing capacity than natural speech for maintenance of information in short term memory and subsequent transfer of information to long term memory. Because synthetic speech lacks many of the redundancies inherent in natural speech, difficulties in encoding may give rise to degraded or impoverished items that are difficult to maintain in short term memory. In this way, the perception of synthetic speech may be analogous to the perception of natural speech presented in high levels of noise. Earlier research has, in fact, demonstrated that difficulties in encoding of speech perceived in noise produce subsequent difficulties in rehearsal processes in short term memory and therefore recall of information from long term memory (see Dallett, 1964; Rabbitt, 1968).

A number of recent studies using rule-generated synthetic speech have shown lower performance levels for perception of synthetic speech relative to natural speech. For example, Pisoni and Hunnicutt (1980) performed several experiments on the intelligibility of speech generated by the MITalk unrestricted text-to-speech system [see Allen (1976), Allen (1979), and Allen, Hunnicutt, Carlson, and Granstrom (1979) for a description of the MITalk system]. In their first experiment, Pisoni and Hunnicutt asked subjects to identify a single target word from a set of six phonemically confusable alternatives using the Modified Rhyme Test (House, Williams, Hecker, and Kryter, 1965). Phoneme recognition for the synthetic speech was 93.1% compared to 99.4% for natural speech--a difference of about 6%.

In their second experiment, Pisoni and Hunnicutt presented listeners with either meaningful or syntactically correct but anomalous sentences. The subjects' task was immediate word-for-word recall of the sentences. The results for the meaningful sentences were similar to the results obtained in the first experiment: Recall for the synthetic speech was about 6% lower than for the natural speech. With the anomalous sentences, however, recall performance for the synthetic speech was about 19% lower than for the natural speech.

In addition to the Pisoni and Hunnicutt findings, Jenkins and Franklin (Note 1), using both the VOTRAX and PAVE synthesizers, have recently reported that when subjects were asked to recall the gist of simple stories, recall for the synthetic stories was not demonstrably poorer than recall for the natural stories. This result is consistent with Pisoni and Hunnicutt's finding that the identification of meaningful sentences was not as severely impaired as the identification of syntactically correct but anomalous sentences. Words in meaningful sentences can be recognized correctly by the deployment of several sources of knowledge that the listener has available, such as morphology, syntax, and semantics. In contrast, words in anomalous sentences can only be recognized from detailed analysis of the acoustic-phonetic information in the waveform.

Finally, Pisoni (1981) found that when isolated synthetic and natural words were presented in a lexical decision task, response times for synthetic words and nonwords were, on the average, 140 msec slower than response times for natural words and nonwords. This study demonstrates that for isolated words, significant decrements in performance can be shown for synthetic speech relative to natural speech.

Overall, these recent findings seem to indicate that the processes used to perceive and understand synthetic speech are heavily dependent on the contextual environment in which the synthetic speech is presented. When meaningful sentences or simple passages are used, intelligibility and comprehension appear, at first, to suffer little relative to comparable natural speech controls. This is not the case, however, when isolated words or meaningless sentences are presented. In these cases, listeners do not have top-down contextual support for word recognition and must therefore rely more on the degraded synthetic signal. It is apparent, then, that there are some important difficulties in the perception and comprehension of synthetically generated speech. Our goal in the present paper was to attempt to isolate the locus of these difficulties in the information processing system. More specifically, we were interested in

determining whether the observed performance deficits for synthetic speech could be attributed to (1) encoding difficulties, (2) rehearsal difficulties in short term memory, or (3) a combination of the two.

This last possibility is suggested by several earlier studies on the effects of noise on retention of spoken word lists. These studies have shown that both perception and memory are affected insofar as both require or share limited processing capacity. Dallett (1964) reported two experiments in which subjects were asked to recall a series of digits presented at various signal-to-noise ratios. He found that the intelligibility of the digits reduced short term memory capacity and thereby produced decrements in recall. Rabbitt (1968), employing a similar paradigm, also found that recognition errors and capacity limitations on short term memory were responsible for decreased recall of digits presented in noise. In another experiment in which subjects were required to shadow words presented in noise, Rabbitt (1966) found that subsequent identification of shadowed words suffered only if they were presented in noise. From these results, Rabbitt concluded that degraded input requires "spare capacity" in short term memory, thus supporting the proposal that decrements in recall for degraded stimuli are the result of both encoding difficulties and short term memory limitations.

It is now a well accepted assumption that human short term memory is limited in its capacity to hold and process information (Shiffrin, 1976). If the perception and comprehension deficits observed for synthetic speech are due to encoding difficulties at early processing stages, then there should be measurable increases in the demands that these stimuli place on the limited resources available in short term memory. These additional demands should therefore result in relatively less available processing capacity when the difficulty of the primary task is increased or when a secondary task is added (Posner & Rossman, 1965). If this is the case, then as the difficulty of either the primary or secondary task increases, performance should decrease more for synthetic than for natural speech. To examine this problem, we selected free recall of lists of synthetic and natural words as the experimental task. We chose this paradigm because the difficulty of a free recall task can be easily and reliably manipulated by a number of well understood experimental variables.

EXPERIMENT 1

In Experiment 1 we manipulated the difficulty of the free recall task by varying the presentation rate of the individual words in the lists. It was predicted that, as the presentation rate increased, recall performance for the synthetic lists would decrease more rapidly than recall performance for the natural lists. We expected this outcome because any encoding difficulties entailed by the synthetic stimuli should detract from subjects' ability to rehearse and store the words for later recall.

METHOD

Subjects. The subjects were 72 undergraduates drawn from a paid subject pool at Indiana University. They were paid $3.00 for participating in the experiment. All of the subjects were native speakers of English and reported no hearing or speech disorders at the time of testing. None of the subjects had had any previous exposure to synthetic speech generated by the MITalk system.

Stimuli. The stimuli were six lists of 25 words selected from the Modified Rhyme Test (House, Williams, Hecker, & Kryter, 1965). The lists were constructed so that the words on successive lists differed only by either the initial or final phoneme. There were both natural and synthetic recordings of each of the lists. Altogether, then, 12 lists of words were used. The natural lists consisted of the same test words as the items on the synthetic lists but were recorded by a male speaker.

The test words were first low-pass filtered at 4.8 kHz and digitized via a 12-bit analog-to-digital converter. All stimuli were played back to listeners through a 12-bit digital-to-analog converter which was interfaced to matched and calibrated Telephonics (TDH-39) headphones. The words were presented at a comfortable listening level of 80 dB SPL against a background of wideband Gaussian white noise at 50 dB SPL. Presentation of the stimuli was controlled in real-time by a PDP 11/34 computer.

Procedure. We tested 12 groups of six subjects in a sound treated room used for perceptual experiments. Each subject heard all six lists of words. No subject heard the same list of words spoken in both the natural and synthetic voice. The lists were blocked; half of the subjects heard the natural lists first and half heard the synthetic lists first.

Each of the synthetic and each of the natural lists was presented at each of three presentation rates: one, two, and five sec. Rate was measured between the onsets of successive words in the lists. The order of the lists and the presentation rates were counterbalanced across groups according to a Latin square design.

At the beginning of each experimental session, the subjects heard two short paragraphs spoken in the synthetic and natural voices to familiarize them with the quality of the speech (see Pisoni and Hunnicutt, 1980). In addition, one natural and one synthetic practice list were presented to acquaint the subjects with the details of the experimental procedures. The practice lists consisted of ten words presented at a rate of two items per sec.

Immediately preceding the presentation of each list the subjects heard a 500 msec, 1000 Hz warning tone. Following presentation of the test list, another tone signalled the end of the list and the beginning of a two min recall period. During this period the subjects were required to write down as many of the words from the list as they could recall. A third tone signalled the end of the recall period. There was a short break between the third and fourth lists in the session.

The subjects were told that they need not recall the words in the same order in which they were presented. However, they were strongly encouraged to use the entire recall period to remember as many of the words from the list as possible and to guess if necessary.

RESULTS AND DISCUSSION

Figure 1 shows the mean number of words recalled for the natural and synthetic lists as a function of rate of presentation.

Insert Figure 1 about here

At each presentation rate, natural words were recalled significantly better than synthetic words, $F(1,71)=47.71$, $p<0.01$. There was also a main effect for rate, $F(2,142)=100.28$, $p<0.01$. That is, recall improved for both the natural and synthetic lists as the rate of presentation increased from one to two to five sec. However, no interaction between voice-type (natural vs. synthetic) and rate was observed, $F(1,142)=0.25$, $p<0.78$.

A similar pattern of results was observed when recall intrusions were scored. Any word in the subjects' response protocols which was not on the presented list was scored as an intrusion error unless it was an alternative spelling of a homophone or an obvious misspelling of a presented word as determined by two independent observers. The results for the intrusion analysis are shown in Figure 2.

Insert Figure 2 about here

Again, there was a significant main effect for synthetic vs. natural voice, $F(1,71)=40.11$, $p<0.01$, and rate, $F(2,142)=4.01$, $p<0.02$. The interaction of the two variables was not significant, $F=(2,142)=1.75$, $p>0.18$. Although a slight increase in intrusions is apparent at the five sec interval for the synthetic lists, a Scheffe pairwise comparison showed that this increase was not significant, $S=0.56$, $p>0.25$.

Note that the scales for the mean words recalled in Figure 1 and the mean intrusions in Figure 2 are different. The mean difference in the recall data for the natural and synthetic words across rate was approximately 1.71, whereas the mean difference for the intrusions was approximately 1.19 words.

The overall pattern of these results suggests that subjects were simply misperceiving some of the synthetic words regardless of the presentation rate.

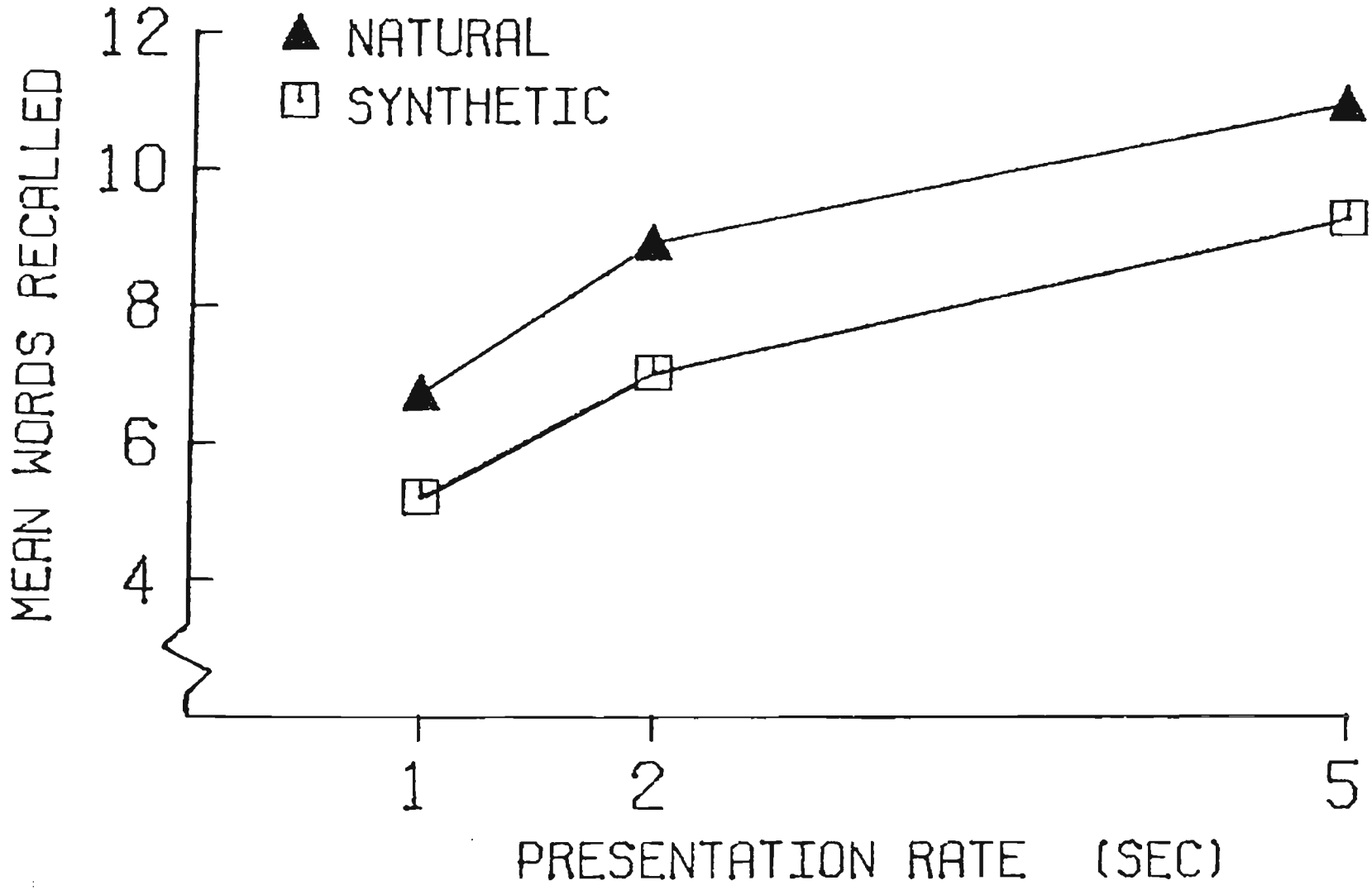


Figure 1. Mean number of natural and synthetic words recalled as a function of presetation rate.

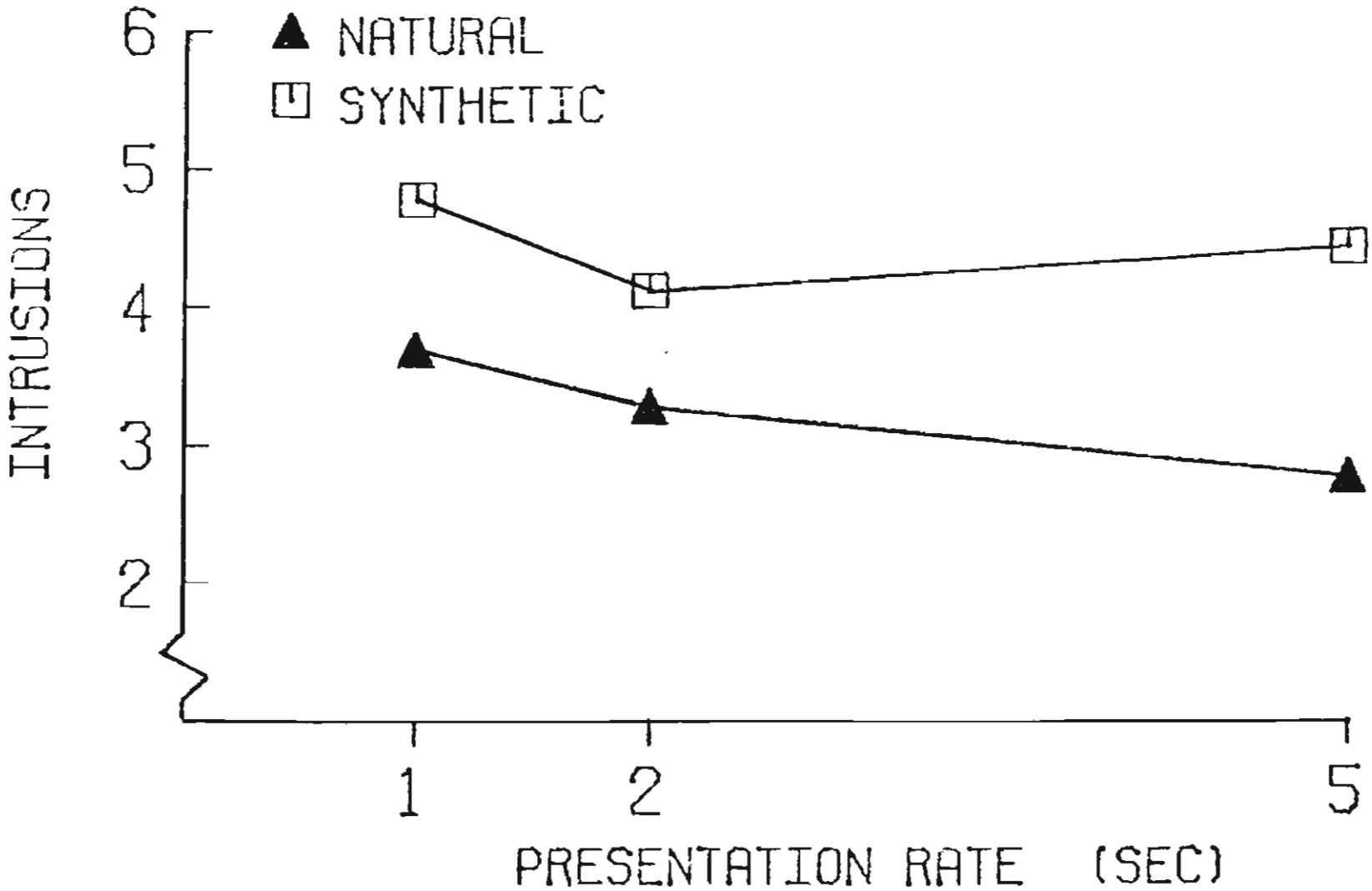


Figure 2. Mean number of intrusions for both natural and synthetic word lists as a function of presentation rate.

This conclusion is supported by the finding that recall of the synthetic items was not differentially affected by increasing presentation rate. There was thus no indication from the presentation rate manipulations used in this experiment that the synthetic words demanded additional processing capacity during encoding or rehearsal. It is possible, however, that the rate of one sec was not fast enough to reveal any encoding and/or rehearsal effects which may have been present. Unfortunately, technical limitations prevented us from increasing the presentation rate beyond one word per second in this experiment. In order to further increase the demands placed on short term memory, we performed a second experiment in which a secondary task was added to the primary recall task.

EXPERIMENT 2

Experiment 2 employed a memory pre-load technique originally developed by Baddeley and Hitch (1974). This technique consists of loading short term memory with a short list of items which the subjects are asked to actively maintain (i.e. to rehearse) throughout the primary task. Baddeley and Hitch found this technique to be useful in assessing short term memory demands for such primary tasks as reasoning, sentence comprehension, and free recall. We used this pre-load technique to determine if the synthetic word lists would place increased demands on encoding and/or rehearsal processes in short term memory when the subjects were simultaneously engaged in another task requiring processing capacity in short term memory.

METHOD

Subjects. The subjects were 120 undergraduates from Indiana University. Some of the subjects received credit for an introductory psychology course, others were paid $3.00 for their participation. All of the subjects met the same criteria for participation as those in Experiment 1.

Stimuli. The natural and synthetic stimuli were the same words as used in Experiment 1. List length, however, was reduced to fifteen words per list for the six lists. The order of the words within each list was random.

Procedure. As in Experiment 1, each subject listened to three synthetic and three natural word lists. However, prior to the presentation of each word list the subjects saw either zero, three, or six digits, one at a time, on a CRT video display monitor (GBC model MV-10A). The monitor was located approximately 42 cm from the subject. Each digit, sampled without replacement from the digits one through nine, remained on the screen for two sec. The interval between presentation of the digits was one sec. The presentation rate for the words was fixed at two sec.

The placement of warning tones was the same as in Experiment 1. However, an additional tone was added to the experimental procedure to indicate the beginning of the digit presentation. The recall interval was also reduced to 90 sec. Counterbalancing was the same as in Experiment 1 with the digit pre-load manipulation substituted for rate.

The subjects were instructed to remember the digits in the exact serial order in which they were presented on the CRT screen. After the word list was presented, the subjects were first required to write down the digit list and then to recall as many of the words from the test list as they could remember. In order to ensure that the subjects would maintain the digits throughout presentation of a given word list, they were told that their recall of the test words could only be scored if all of the digits were recalled in the exact order in which they were presented.

Before the test lists were presented, the subjects heard the same two paragraphs as in Experiment 1 and two practice lists, one synthetic and one natural. The practice lists were ten words long and were preceded by a pre-load of three digits. As in the experimental lists, the rate was set at two sec.

RESULTS AND DISCUSSION

Because there were two dependent variables of interest in this experiment--word recall and digit recall, the analysis of the data has been broken down into two parts.

Word recall. Figure 3 presents the mean words recalled as a function of pre-load condition.

Insert Figure 3 About Here

As in Experiment 1, the natural word lists were recalled better overall than the synthetic word lists across all three pre-load conditions, $F(1,119)=106.95$, $p<0.01$. A main effect of pre-load condition was also observed, $F(2,238)=37.36$, $p<0.01$, with mean word recall for both lists decreasing with increasing pre-load. However, no interaction between voice-type and pre-load was observed for word recall, $F(2,238)=1.11$, $p>0.33$.

Figure 4 presents the mean number of intrusions as a function of pre-load condition.

Insert Figure 4 About Here

Again, intrusions were significantly greater for the synthetic than the natural word lists, $F(1,119)=75.57$, $p<0.01$. However, no main effect of pre-load was observed, $F(2,238)=2.59$, $p>0.07$, nor was there an interaction between these two variables, $F(2,238)=0.02$, $p>0.80$.

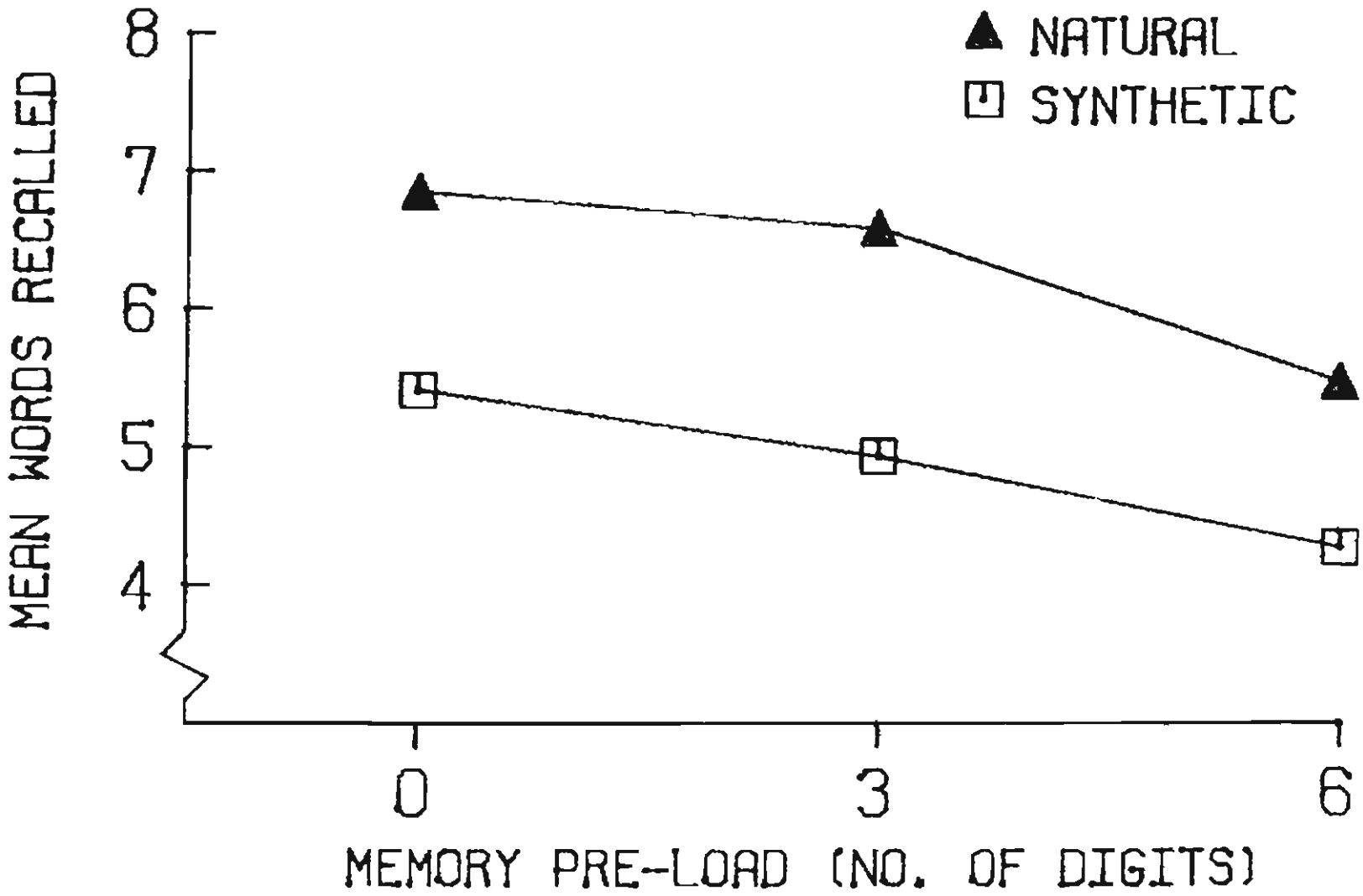


Figure 3. Mean number of natural and synthetic words recalled as a function of memory pre-load.

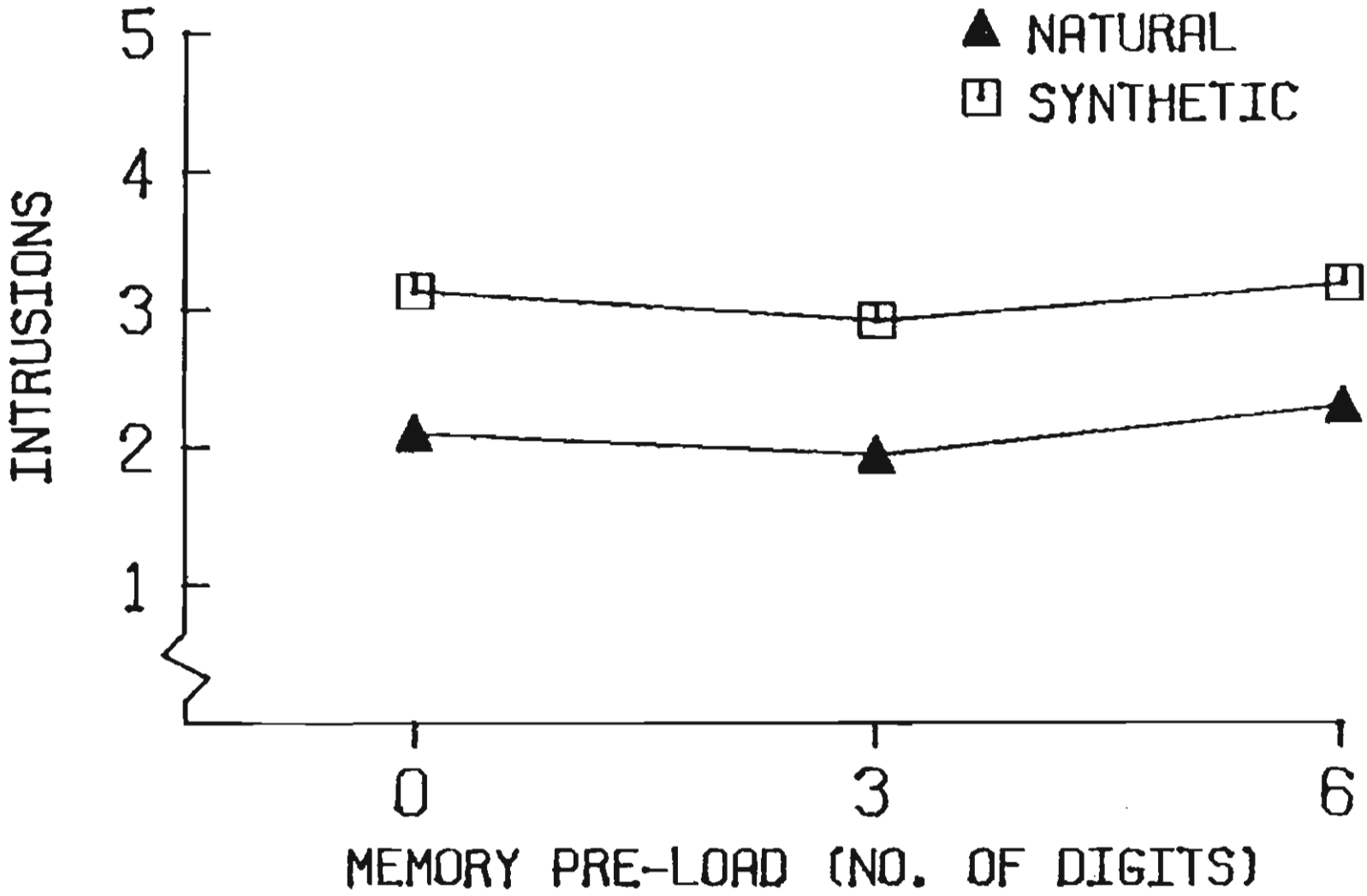


Figure 4. Mean number of intrusions for both natural and synthetic word lists as a function of memory pre-load.

Aside from the apparent failure to find a main effect for pre-load in the intrusion data, the results for the primary word recall task are consistent with those obtained in Experiment 1. The synthetic word lists were recalled more poorly overall than the natural word lists; however, no differential effects of the digit pre-load manipulation were observed across the natural and synthetic lists.

Digit recall. Three different analyses of the pre-load digit recall data were performed. The first analysis, shown in Figure 5, was carried out on the number of subjects who correctly recalled all of the digits in the exact order in which they were presented for the two conditions in which the load items were present.

Insert Figure 5 About Here

The interaction that we expected to find for word recall is clearly present for recall of the pre-load digits: The number of subjects correctly recalling the digits decreased more rapidly for recall of items from the synthetic lists than from the natural lists as the pre-load on short term memory increased from three to six items, $z=1.63$, $p<0.05$, one-tailed.

The second analysis performed on the digit recall is shown in Figure 6.

Insert Figure 6 About Here

In the upper panel, the average percentage of the digits recalled is plotted as a function of load without regard for the order in which the pre-load digits were recalled. In the third analysis, shown in the lower panel of Figure 6, the digits were scored as correct only if they were recalled in the same serial position in which they were originally presented.

Proportions for both sets of data were transformed via an Arcsin transformation with a correction for small N. An analysis of variance on the transformed data showed significant main effects in the item-only condition for voice-type and pre-load ($F(1,119)=6.31$, $p<0.02$ and $F(1,119)=163.03$, $p<0.01$, respectively). The predicted interaction, although in the right direction, did not reach significance, $F(1,119)=2.47$, $p>0.11$. Thus, the data reveal a trend for the digit recall to be poorer under the high pre-load condition for the synthetic relative to the natural word lists, although the effect is not reliable.

In the item-and-position analysis, only the effects of pre-load were significant, $F(1,119)=200.81$, $p<0.01$. The effects of voice-type and the voice-type by pre-load interaction were not significant ($F(1,119)=3.25$, $p>0.07$ and $F(1,119)=1.78$, $p>0.18$, respectively).

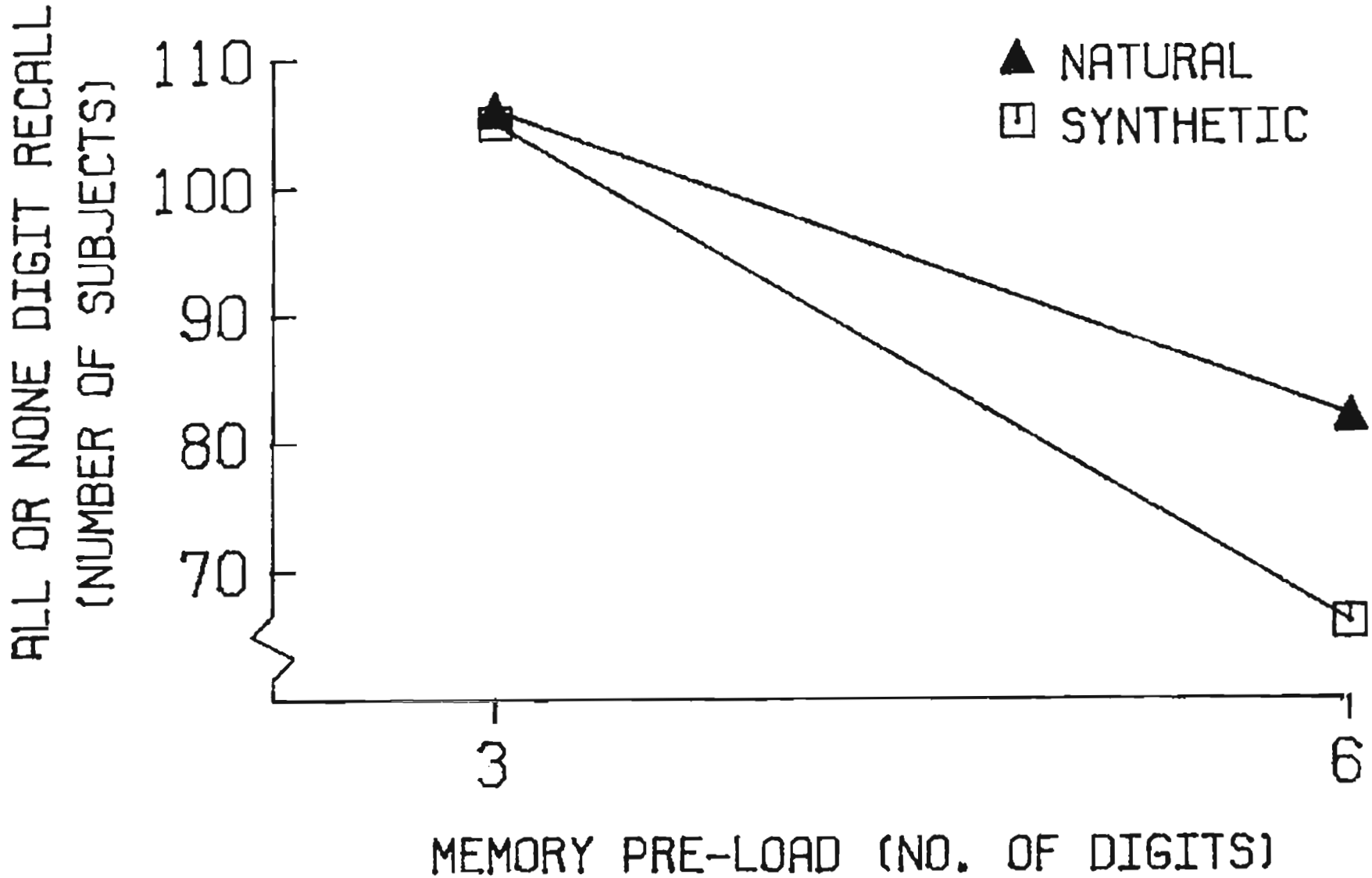


Figure 5. Number of subjects correctly recalling all of the digits as a function of memory pre-load.

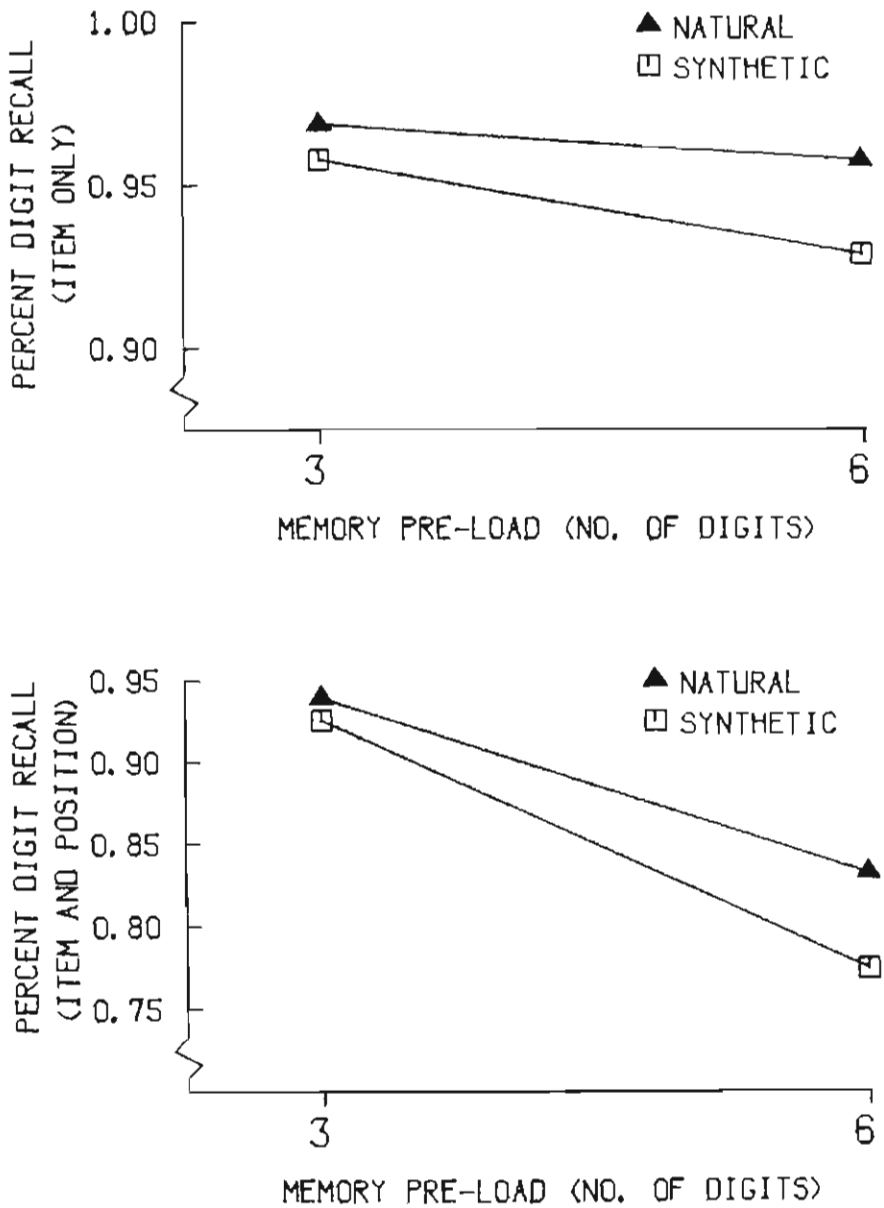


Figure 6. Percentage of digits recalled as a function of memory pre-load. The upper panel shows percent correct when the digits were scored without respect to the positions in which they were recalled. The lower panel shows the percentage correct for the digits scored according to position.

In summary, the analyses for the digit recall show a tendency for performance in the six-digit pre-load condition relative to the three-digit pre-load condition to be poorer for the synthetic than the natural word lists. These findings suggest that perception of synthetically produced word lists may interfere with the subjects' ability to maintain information in short term memory (see Posner & Rossman, 1965). Moreover, the greater effects of the synthetic words lists relative to the natural word lists appears to occur only under conditions where memory stress is present. The obvious interpretation of these effects is that the subjects "borrow" from the capacity needed for maintenance rehearsal of the digits in order to encode or rehearse the synthetic word lists for later recall (Rabbitt, 1968). That is, encoding and subsequent processing of lists of synthetic words in short term memory appear to require more capacity or allocation of resources than encoding and subsequent processing of lists of natural words.

Although the results from the digit pre-load experiment indicate that encoding and/or rehearsal processes are differentially stressed by synthetic speech, the main finding that supports this claim--number of subjects recalling all of the digits--is at best a crude measure of the capacity demands of synthetic speech. To obtain stronger evidence for the increased capacity demands of synthetic speech, we performed a third experiment in which subjects were required to recall the synthetic and natural word lists in the exact order in which the lists were presented. In an ordered serial recall task the subject must encode not only specific items but also additional contextual information about the location of the items in the list.

EXPERIMENT 3

Experiment 3 employed a serial ordered recall task. We reasoned that if synthetic speech places greater demands on encoding processes, rehearsal processes, or both, then requiring serial ordered recall would differentially affect the primacy portion of the serial positions curve for the synthetic word lists and the natural word lists. We based this prediction on the hypothesis that increased demands on encoding and/or rehearsal processes arising from the synthetic speech would cause fewer items presented early in the synthetic lists to be transferred to long term memory (Baddeley and Hitch, 1974). That is, increased demands on encoding and/or rehearsal processes should adversely affect transfer of information to long term memory of the synthetic words relative to the natural words. This reduced capacity should be manifested by poorer recall performance in the primacy portion of the serial position curve for the synthetic speech.

METHOD

Subjects. The subjects were 72 undergraduates from Indiana University. They received credit for an introductory psychology course for their participation. All of the subjects met the same criteria for participation as those in Experiments 1 and 2.

Stimuli. The test stimuli were the same as those used in Experiments 1 and 2. List length, however, was reduced to ten words per list for the six lists. The order of the words within each list was random.

Procedure. As in Experiments 1 and 2, subjects listened to three natural and three synthetic word lists. Again, subjects never heard both a natural and synthetic token of the same word. The subjects were instructed to recall the words in the exact order in which they were presented and to leave blank any spaces on their answer sheets that corresponded to words they were unable to recall.

The placement of the warning tones was the same as in Experiment 1. Two practice lists of five words each were presented prior to the presentation of the ten experimental lists. The words in both the practice and experimental lists were presented at a rate of two sec and the recall periods were 90 sec in length. Counterbalancing was the same as in Experiment 1.

RESULTS AND DISCUSSION

Figure 7 presents the overall serial position curves for both the natural and synthetic word lists. Serial position is given on the abscissa and the probability of correct recall is given on the ordinate.

Insert Figure 7 about here

The serial position curves were obtained by scoring an item as correct only if it was recalled in the position in which it was presented. The number of correct responses for each serial position for each voice was then summed across all subjects and an overall percent correct score was obtained.

As in Experiments 1 and 2, the natural word lists were recalled better overall than the synthetic word lists, $F(1,71)=43.23$, $p<0.01$. When the first and second halves of the curves were collapsed across the synthetic and natural word lists and compared, a significant interaction of list half by serial position within each list half was obtained, $F(4,284)=196.92$, $p<0.01$. This interaction confirms that there were significant recency and primacy effects across both the synthetic and natural word lists. However, a significant interaction between voice-type, list half, and serial position, $F(4,284)=2.65$, $p<0.05$, indicated that the primacy portion of the curve for the synthetic word lists showed lower recall scores relative to the natural word lists than did the recency portion of the curve. That is, there was a greater difference in recall between synthetic and natural words for the primacy portions of both curves relative to the recency portions of the curves.

The difference observed in the primacy portions of the serial position curves for the natural and synthetic word lists clearly demonstrates that

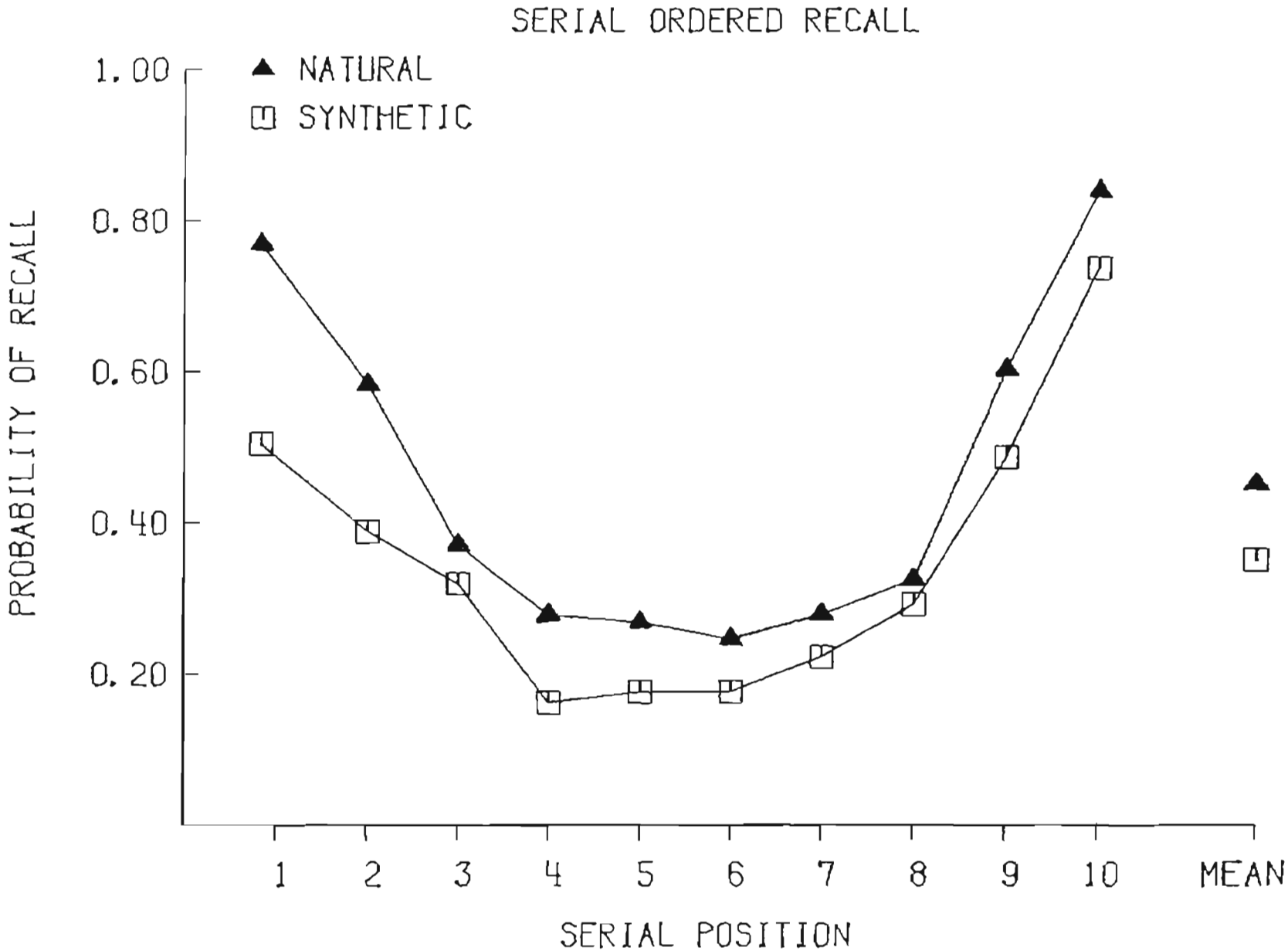


Figure 7. Overall serial position curves for the natural and synthetic word lists.

synthetic speech places increased demands on encoding and/or rehearsal processes in short term memory. Because the perception of synthetic speech decreases processing capacity in short term memory, successful transfer of items from short to long term memory appears to be more adversely affected for the synthetic words than for the natural words. Considering the extensie literature on the limited processing capacity of short term memory in human information processing, the results from the serial-ordered recall experiment as well as from our earlier digit pre-load experiment are not at all surprising. However, the results of these experiments provide strong evidence that some of the difficulties observed in the perception and comprehension of synthetic speech do, in fact, lie in increased demands in encoding and/or rehearsal processes in short term memory. That is, the present results demonstrate that synthetic speech is difficult to perceive and understand, relative to natural speech, in part because it affects the allocation of limited processing resources in short term working memory.

GENERAL DISCUSSION

Two conclusions can be drawn from the results of these experiments. First, the large constant overall decrement observed in recall performance across all three experiments is probably due, in part, to a failure to encode of some of the acoustic-phonetic information in the synthetic words themselves. That is, some of the observed performance differences lie in relatively early stages of pattern recognition required for word identification. This conclusion is supported by the fact that recall intrusions were, on the average, between one and two words more frequent for the synthetic than natural word lists. Although an extensive analysis of the intrusions has not been performed, it was apparent during scoring of the data that most intrusions for the synthetic word lists were words that differed from list items by only a single phoneme (e.g., "boil" was frequently recalled as "oil"). In other words, the intrusions were acoustically (phonologically) based and were a result of early perceptual confusions or misperceptions.

A second, and perhaps more important, conclusion is that at least some of the observed difficulties in the perception and retention of synthetic speech are clearly due to increased processing demands for these items in short term memory. This conclusion is supported by the results from the digit recall in Experiment 2 and the serial ordered recall in Experiment 3. Although the serial-ordered recall data clearly demonstrate the role of increased processing demands in the recall of synthetic word lists, it is not clear why the increased processing demands in Experiment 2 were manifested only in the digit recall.

One account of the failure to find differential effects of synthetic speech on word recall in Experiment 2 comes from Rabbitt (1968): Rehearsal of the pre-load items which are supposed to be actively maintained in short term memory may be inhibited by the increased processing demands required to encode and rehearse the list of synthetic items. In several experiments on the effects of noise on short term memory, Rabbitt (1968) found that digits from the early part of a list were recalled more poorly when the digits from later serial positions had to be identified in noise. This result was obtained regardless of whether or not the items in the early part of the list were presented in noise. More

important, however, the opposite effect was not observed; that is, recall of items in the second half of the list was not affected when items in the first half were presented in noise. Rabbitt's results are consistent with the view that the process of recognizing digits through noise reallocates processing capacity for efficient retention and rehearsal of items in immediate memory.

Our results for the serial-ordered recall data are consistent with Rabbitt's earlier findings. The effects of the digits embedded in noise from later serial positions on recall of the digits from earlier serial positions is directly analogous to the effects we observed in the primacy portion of the serial position curve for the synthetic words. The synthetic words were, in a sense, acting as if they were "noisy" items by placing increased demands on encoding and/or rehearsal processes because they were initially more difficult to encode and identify.

Synthetic words and natural words presented in noise may thus be poorly recalled for two quite different reasons. First, the items may be poorly recognized at the time of encoding because of impoverished acoustic-phonetic information due to masking or perceptual confusions. Second, and perhaps most relevant for our purposes, items that are poorly encoded may interfere with the rehearsal and subsequent retention of other items in active short term memory, whether these items are visually-presented digits or auditorily-presented synthetic words.

The results of these recall experiments are closely related to several other recent findings concerning the intelligibility, perception, and comprehension of synthetic speech generated by the MITalk text-to-speech system. In particular, Luce (Note 3) has found several interesting differences in comprehension between synthetic and natural speech when subjects are required to answer various types of questions after listening to passages of fluent connected speech. He found that subjects perform more poorly for synthetic passages on comprehension questions designed to probe the content of a given passage. However, subjects hearing synthetic passages perform better than those hearing natural passages on questions that probe retention of the surface structure of the passages. These comprehension results suggest that the subjects' attention is somehow directed more toward the superficial (surface) properties of the actual speech signal in the synthetic speech condition than to the properties of the message in the natural speech condition (see Aaronson, 1976).

In another study, Pisoni and Koen (Note 4) have recently found differences in intelligibility of natural and synthetic words presented in noise at several different signal-to-noise ratios. Intelligibility of synthetic words is affected by noise more than the same naturally produced words. Thus, the effects of noise produce a greater decrement on recognition of the synthetic items presumably because they contain fewer redundant acoustic cues to support recognition of the phonetic structure.

In summary, the present experiments indicate that synthetic speech places increased processing demands on encoding and/or rehearsal processes in short term memory than does natural speech. Moreover, our results show that traditional experimental paradigms in memory research can be advantageously applied to the

assessment of the intelligibility and perceptual processing of synthetic speech. We believe that increased processing demands for the encoding and rehearsal of synthetic speech signals may place important constraints on the use of various voice-response devices in high information load conditions, particularly in conditions requiring differential allocation of attention among several sensory inputs. In applications such as aircraft cockpits or complex command-control displays, voice-response systems using synthetic speech should be carefully considered in terms of the potential interactions of specific tasks, perceivers, and signal quality.

REFERENCE NOTES

1. 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.

2. Jenkins, J. J., & Franklin, L. D. Recall of passages of synthetic speech. Paper presented at the meeting of the Psychonomic Society, Philadelphia, November, 1981.

3. Luce, P. A. Comprehension of fluent synthetic speech produced by rule. Journal of the Acoustical Society of America, 1982, 71, UU11.

4. Pisoni, D. B., & Koen, E. Intelligibility of natural and synthetic speech at several different signal-to-noise ratios. Journal of the Acoustical Society of America, 1982, 71, UU1.

REFERENCES

Aaronson, D. Performance theories for sentence encoding: Some qualitative observations. Journal of Experimental Psychology: Human Perception and Performance, 1976 2, 42-55.

Allen, J. Synthesis of speech from unrestricted text. Proceedings of the IEEE, 1976, 4, 433-442.

Allen, J. Conversion of unrestricted English text to speech. Forthcoming research monograph based on chapters prepared for a special summer session course held at MIT, June 25-29, 1979.

Allen, J., Hunnicutt, S., Carlson, S., & Granstrom, B. MITalk-79: The 1979 MIT text-to-speech system. In J. J. Wolf & D. H. Klatt (Eds.), Speech communication papers presented at the 97th meeting of the Acoustical Society of America. New York: Acoustical Society of America, 1979, 507-510.

Baddeley, A. D., & Hitch, G. Working memory. In G. H. Bower (Ed.), The psychology of learning and memory, Vol. 8, 1974, 47-90.

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

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

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

Klatt, D. H. Software for a cascade/parallel format synthesizer. Journal of the Acoustical Society of America, 1980, 67, 971-995.

Nickerson, R. S. Characteristics of the speech of deaf persons. The Volta Review, 1975, 77, 342-362.

Pisoni, D. B., & Hunnicutt, S. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. In 1980 IEEE International Conference Record on Acoustics, Speech and Signal Processing, April, 1980.

Posner, M. I., & Rossman, E. Effect of size and location of informational transforms upon short-term retention. Journal of Experimental Psychology, 1965, 67, 496-505.

Rabbitt, P. Recognition memory for words correctly heard in noise. Psychonomic Science, 1966, 6, 383-384.

Rabbitt, P. Channel-capacity, intelligibility and immediate memory. Quarterly Journal of Experimental Psychology, 1968, 20, 241-248.

Shiffrin, R. M. Capacity limitations in information processing, attention, and memory. In W. K. Estes (Ed.), Handbook of learning and cognitive processes, Vol. 4, 1976, 177-236.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

Identification and discrimination of rise time: Is it categorical or noncategorical?

Diane Kewley-Port and David B. Pisoni

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. An earlier version of this paper was presented at the meetings of the Acoustical Society in Chicago, April, 1982.

ABSTRACT

Previous studies have reported that rise time of sawtooth waveforms may be discriminated in either a categorical-like manner under some experimental conditions or according to Weber's law under other conditions. In the present experiments, rise time discrimination was examined with two experimental procedures: the traditional labeling and ABX tasks used in speech perception studies and an adaptive tracking procedure used in psychophysical studies. Rise time varied from 0 ms to 80 ms in 10 ms intervals for sawtooth signals of one second duration. Discrimination functions for subjects who simply discriminated the signals on any basis whatsoever as well as functions for subjects who practiced labeling the endpoint stimuli as "pluck" and "bow" before ABX discrimination were not categorical in the ABX task. In the adaptive tracking procedure, the Weber fraction obtained from the JND's of rise time was found to be a constant above 20 ms rise time. The results from the two discrimination paradigms were then compared by predicting a JND for rise time from the ABX discrimination data by reference to the underlying psychometric function. Using this method of analysis, discrimination results from previous studies were shown to be quite similar to the discrimination results observed in this study. Taken together, the results demonstrate clearly that rise time discrimination of sawtooth signals follows predictions derived from Weber's law.

INTRODUCTION

Within the last few years there has been a renewed interest in the identification and discrimination of nonspeech sounds. The theoretical motivation for studying perception of nonspeech signals has changed somewhat from earlier studies which simply attempted to demonstrate that "speech is special" (Liberman, Harris, Hoffman, and Griffith, 1957; Mattingly, Liberman, Syrdal and Halwes, 1971; Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967). Instead, a number of current investigations have been more concerned with finding natural sensitivities in the auditory system for nonspeech signals which might also serve as the underlying basis for acoustic features in the perception of speech sounds (Miller, Pastore, Wier, Kelly and Dooling, 1976; Pisoni, 1977; Schouten, 1980; Stevens, 1982). One type of nonspeech signal studied thus far has manipulated the timing of events at stimulus onset associated with the feature of voice onset time. Several other studies have examined amplitude at onset. In particular, these studies have investigated variations in rise time at stimulus onset for sawtooth waveforms.

In a widely cited study, Cutting and Rosner (1974) reported the first compelling evidence for categorical-like perception of nonspeech stimuli. They employed the typical procedures used in speech perception experiments -- labeling in combination with ABX discrimination. Subjects used the labels of "pluck" and "bow" to identify stimuli generated by a Moog synthesizer. The discrimination functions were non-monotonic and showed a typical categorical peak in the middle of the rise time continuum. The generality of the perception results they obtained was extended to infants (Jusczyk, Rosner, Cutting, Foard and Smith, 1977) and adults in cross-continua adaptation studies (Remez, Cutting and Studdert-Kennedy, 1980).

Recently, however, Cutting and Rosner's results (1974) have been brought into question by Rosen and Howell (1981). These authors failed to replicate the earlier categorical perception results of Cutting and Rosner. Rosen and Howell's rise time stimuli were generated digitally by a computer rather than with analog techniques. Having failed to replicate the original finding, Rosen and Howell then measured the actual Cutting and Rosner stimuli and discovered that the differences in rise time were not equally spaced as had been assumed. In a series of experiments, Rosen and Howell then determined that categorical discrimination functions could be obtained with stimuli containing rise time intervals similar to Cutting and Rosner's. However, they could not obtain categorical-like discrimination functions with stimuli spaced with equal intervals. In fact, discrimination of equal-interval stimuli appeared to follow Weber's law in their ABX experiments. That is, the discrimination was good for short rise times and decreased as rise times became longer. The interpretation of the discrimination data in terms of Weber's law was in good agreement with an earlier psychophysical study conducted by van Heuven and van den Broeke (1979) who reported constant Weber fractions calculated from an indirect measure of the JND's of rise time.

Cutting (1982) has recently conducted several experiments in response to the criticisms raised by Rosen and Howell (1981). In two experiments he replicated Rosen and Howell's result that sawtooth stimuli differing in equal linear increments of rise time are not categorically perceived. Cutting proposed an analysis of the ABX discrimination functions which attempted to show that a

constant Weber fraction model was not fully supported by the data either. The validity of this analysis, however, can be questioned. Cutting compared obtained discrimination functions to another function predicted by a Weber fraction model. The predicted function assumed that the Weber fraction was constant over the entire rise time continuum ranging from 10 to 50 ms. However, the Weber fraction is not usually constant over the entire stimulus range (c.f. Geschelder, 1976), and, in fact, van Heuvan and van den Broeke (1979) showed that the Weber fraction increased rapidly for rise times less than 20 ms. Thus, the less than perfect fit of Cutting's predicted functions based on Weber's fraction may be attributed to his assumption of a constant Weber fraction for discrimination of rise time when in fact such an assumption is questionable.

In Cutting's report (1982), the investigation of rise time was examined further in a rather unusual experiment. He claimed that "if a Weber-fraction view is correct, stimuli generated with equal logarithmic increments of rise time ought to yield a flat discrimination function." He then picked an arbitrary logarithmic increment for generating a rise time continuum and examined the stimuli with standard identification and ABX discrimination procedures. Analysis of the discrimination results showed that these logarithmically spaced stimuli were generally perceived categorically. From these results, Cutting concluded that there is a "general fickleness of the phenomenon of categorical perception in any stimulus domain" and that rise time can be perceived categorically.

Despite Cutting's reply to Rosen and Howell, several basic issues remain in understanding discrimination of rise time in both speech and nonspeech signals. The perception of rise time at the onset of a waveform is a complex auditory event. Subjects discriminate pairs of rise time stimuli somewhat differently depending on whether they are asked to judge their relative loudness or their rate of onset (Gershuni and Zaboeva, 1962; Nabelek, 1965). While the categorical perception of other complex acoustic stimuli such as stop consonants differing in voice onset time has been demonstrated in high uncertainty paradigms (Liberman et al., 1957), noncategorical perception has also been obtained in low uncertainty paradigms (Sachs and Grant, 1976; Carney, Widin and Viemeister, 1977) or when the procedures emphasize attention to stimulus differences (Pisoni and Lazarus, 1974). Previous studies of rise time discrimination of sawtooth stimuli have not specified the stimulus conditions under which categorical perception of rise time can be observed. If rise time discrimination follows Weber's law and is similar to discrimination of other auditory signals, we should be able to demonstrate this experimentally using standard psychophysical techniques.

In the present report, we sought to improve our understanding of the discrimination of rise time in nonspeech signals. In the first experiment, we conducted an independent replication of one of Rosen and Howell's (1981) experiments for a linearly incremented rise time continuum. Given the apparent controversy surrounding Rosen and Howell's results, we wanted to replicate the ABX discrimination functions for a newly generated set of sawtooth stimuli. Furthermore, we wanted to separate the possible effects of prior labeling experience on discrimination from a subject's sensory capacity to discriminate rise time differences on any basis whatsoever.

The second study employed a more sensitive psychophysical task to directly estimate the JND's for rise time for the same sawtooth stimuli. By calculating the Weber fraction from these data, rise time discrimination can be compared to

the discrimination of simpler auditory properties such as frequency and intensity. Finally, the third experiment directly compared the discrimination results obtained in the ABX and the psychophysical task. The subjects who participated in the previous psychophysical task were recalled for this experiment to participate in the ABX discrimination task. The results of the two studies were compared by referencing both sets of data to the underlying psychometric function. Using this analysis, we found that discrimination of rise time follows Weber's law, a finding that was in close agreement with Rosen and Howell's earlier data.

I. EXPERIMENT 1: IDENTIFICATION AND ABX DISCRIMINATION

A. Method

1. Stimuli

The stimuli generated for this experiment were intended to replicate as closely as possible the stimuli generated by Rosen and Howell (1981, Exp. 3). A set of nine sawtooth stimuli was generated digitally on a PDP-11/34 computer. Each signal was synthesized by adding together in phase the first eleven sinusoidal components of a sawtooth waveform using a modified version of the program SOUND (Lovell and Carterette, 1972). The fundamental frequency was 300 Hz and the amplitude of the harmonics decreased in 6 dB steps from the 66.23 dB (11 bit) maximum at the fundamental. The amplitude envelope impressed at the onset of each stimulus rose linearly from 0 to its maximum in 1, 10, 20, 30, 40, 50, 60, 70 and 80 ms of time. The 1 ms ramp interacted with the fundamental of the stimulus so that this stimulus reached its maximum amplitude in about 3 ms. This stimulus will be nominally referred to as the 0 ms rise time stimulus in keeping with previous studies. The remaining stimuli reached maximum amplitude at precisely the durations specified by the 10 ms interval onset ramps. From the maximum, a second linear decay ramp was impressed over the remaining waveform which reached 0 in 1.02 seconds. Thus, the overall length of the stimuli covaried with the duration of the onset ramps from 1020 ms to 1100 ms as in the previous studies reported by Rosen and Howell (1981) and Cutting and Rosner (1974). Figure 1 shows the first 100 ms of the 0, 10 and 80 ms stimuli.

Insert Figure 1 about here

The stimuli were output under computer control at a 10 KHz sampling rate through a 12 bit D/A converter. Stimuli were low-passed once at 4.8 KHz, and then filtered through two cascaded low-pass filters (Krohn-Hite model 3202R) at 3500 Hz with a total roll-off of 48 dB per octave. The stimuli were then amplified to a comfortable listening level and routed to TDH-39 headphones located in a quiet testing room. A constant output level was calibrated across experimental tasks.

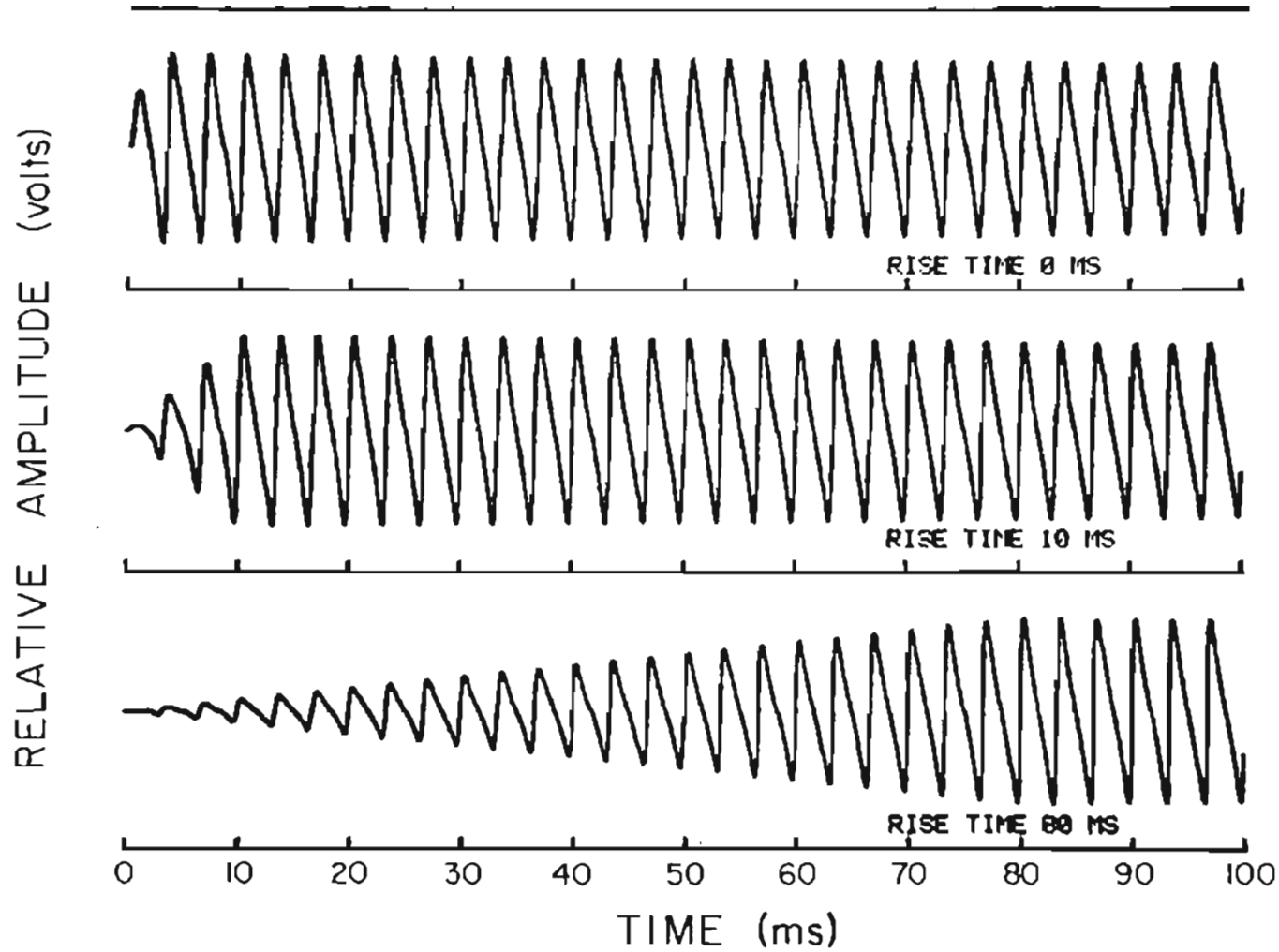


Figure 1. Waveforms of the first 100 ms of the 0, 10 and 80 ms rise time stimuli.

2. Procedure: Identification preceding ABX

All experimental tasks were controlled by a PDP-11/34 laboratory computer. Subjects were run in small groups in a quiet room equipped with six individual cubicles interfaced to the computer. Responses were collected on two-button response boxes equipped with feedback and cue lights.

The experiment consisted of two one-hour sessions conducted on successive days. The procedure was divided into three general parts: familiarization, identification testing and discrimination testing. Separate written instructions were read by subjects before each task. Four tasks were run on Day 1. The first task introduced subjects to the 0 and 80 ms rise time stimuli, and their respective labels of 'pluck' and 'bow.' In this task the 0 and 80 ms stimuli were presented in sequential order 10 times each. The second task involved training subjects to identify these two "endpoint" stimuli as 'pluck' and 'bow' using feedback. Each stimulus was presented 20 times in a random order. After completing this task, the procedure was repeated with a different random order but without feedback after each trial. The results of this task were used to assess whether subjects met a pre-determined criterion of 90% correct over the 40 trials to be included in subsequent parts of the experiment.

Following this initial training phase on Day 1, the identification task was run. Subjects were presented with the full set of 0 to 80 ms rise time stimuli in 2 blocks of 90 trials each. The stimuli were presented in a random order with no feedback. Subjects were asked to judge each stimulus as either 'pluck' or 'bow.'

On Day 2 the subjects who passed the criterion initially received a warm-up sequence containing 10 replications of the 0 and 80 ms stimuli for labeling with feedback. Subjects were then given the identification task a second time with all 9 stimuli presented 10 times each in a random order.

The final task was a standard ABX discrimination test using feedback for the correct response. The response buttons were relabeled as 'first sound' and 'second sound'. All seven two-step pairs from the rise time continuum were presented 20 times each in all possible ABX triads for a total of 140 trials. Stimuli within the triad were separated by 1 second each. Presentation of each ABX triad was paced to the slowest subject in a group.

3. Procedure: ABX only

The same ABX discrimination test previously described was employed here. However, no training or identification tasks preceded the ABX task. Written instructions referred to the stimuli only as complex sounds generated by a computer. The response buttons were labeled 'first sound' and 'second sound.' Subjects in this task were presented with 2 blocks of 140 trials each, a total of two times the number of ABX trials that the previous subjects received. As before, the ABX sequences were presented with feedback indicating the correct response on each trial. Finally, subjects answered written questions concerning the nature of the stimuli at the end of the experimental session to assess their subjective impressions of the sounds and to determine what strategies they may have used in the discrimination task.

4. Subjects: Identification preceding ABX

Thirty-eight subjects were recruited from a paid laboratory subject pool. Subjects were native speakers of English with no reported history of hearing or speech disorders. Subjects were naive with respect to this experiment and had no previous experience in psychoacoustic experiments. They were paid $5.00 per hour for each testing day.

B. Results: Identification preceding ABX

In order to have an equal number of subjects in each condition to facilitate statistical analysis, we determined in advance of testing that results would be analyzed from 15 subjects in each group. Thus, subjects with the poorest performance in the ABX task from each group were eliminated until 15 subjects remained. In the identification preceding ABX procedure, 21 subjects participated in the first day of testing. Three of these subjects could not identify the pluck and bow stimuli at the 90% criterion level and were excused from testing after Day 1. Of the 18 subjects who completed both days, the three worst subjects on the ABX task with an average of 55% correct performance (chance = 50%) were dropped in the final analyses.

Results from the second day of testing averaged over all 15 subjects are displayed in Fig. 2. Identification results obtained on Day 1 were considered as practice with the labeling task and were not analyzed further. Percent identification of each stimulus using the "pluck" response is plotted in Fig. 2 as a function of rise time duration. The overall shape of the identification function indicates that subjects classified the stimuli into two reasonably discrete categories of plucks and bows. Since this experiment was a replication of Rosen and Howell's Experiment 3, we may compare the present results to their findings (see their Fig. 3). From inspection, the two identification functions appear to be quite similar and have approximately the same crossover point between the pluck-bow categories at about 30 ms rise time.

Insert Figures 2, 3, and 4 about here

Figures 3 and 4 show the individual identification and ABX discrimination functions separately for each of the 15 subjects. Almost all subjects showed sharp labeling functions using 100% pluck and bow responses for the end point stimuli. These consistent identification functions are no doubt the result of the criterion testing and practice given to subjects on Day 1. The identification functions compare quite favorably with those obtained from continua such as VOT and place of articulation (see for example, Abramson and Lisker, 1970 and Pisoni, 1971). Given these labeling functions, if listeners perceive rise time categorically, we would certainly expect to find categorical ABX functions from these subjects.

Results from the ABX task are plotted as percent correct discrimination of each two-step pair as a function of the average rise time value of the pair.

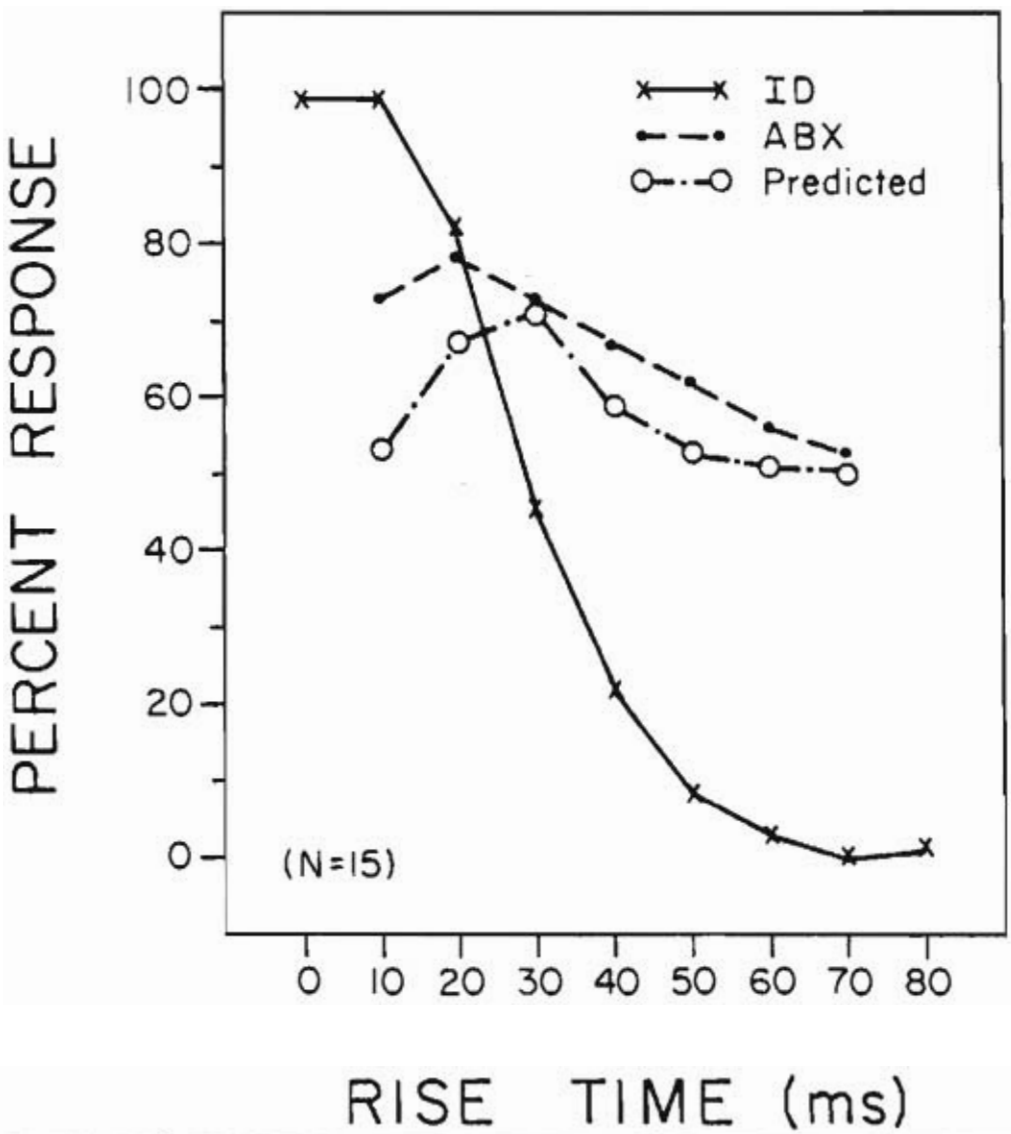


Figure 2. Group results from the identification and ABX tests for subjects trained in using the 'pluck' and 'bow' labels before ABX discrimination. The identification function is plotted as percent of pluck responses. The ABX obtained and predicted functions are plotted as percent correct discrimination (see text).

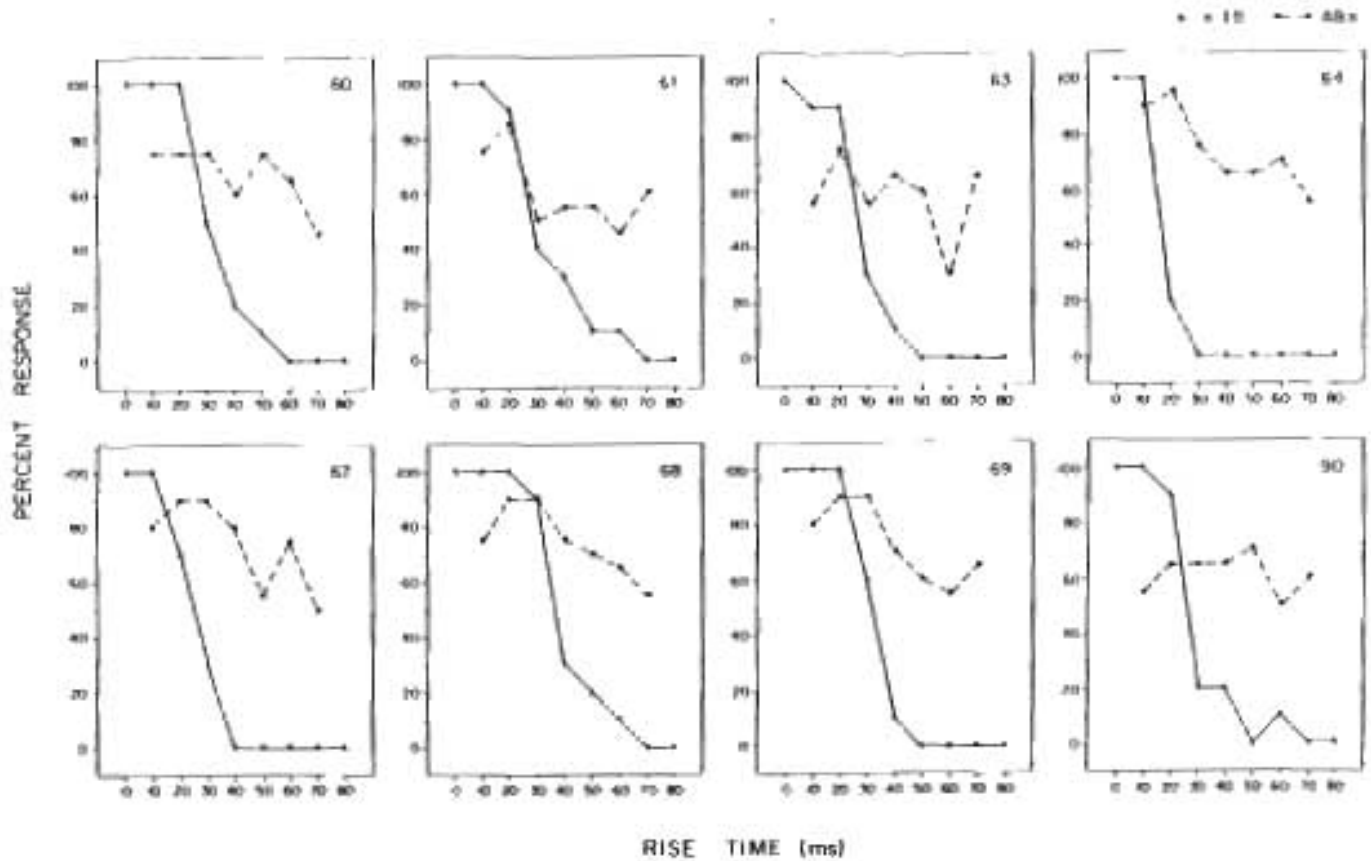


Figure 3. Individual data for subjects in the identification and ABX tasks from Experiment 1.

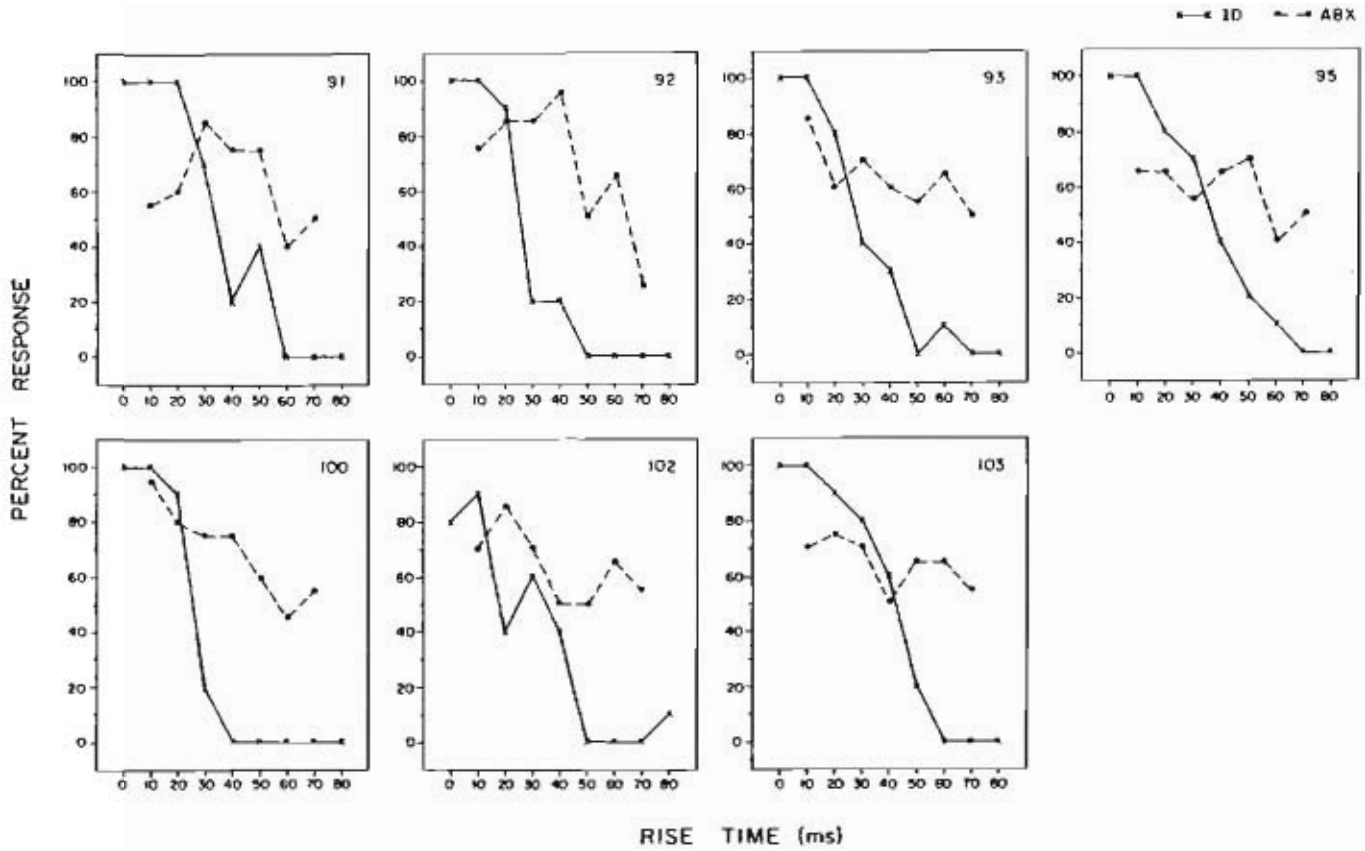


Figure 4. Individual data for subjects in the identification and ABX tasks from Experiment 1.

Almost none of the individual subjects shown in Figures 3 and 4 produced categorical-like discrimination functions. Even subjects like S64 and S100 who show sharp and consistent identification functions failed to display the characteristic peaks and troughs in the ABX discrimination function that are correlated with changes in labeling performance. In fact, the most striking aspect of the individual ABX functions is the large subject-to-subject variability. Thus, subjects who reliably classified rise time stimuli into pluck and bow categories did not discriminate rise time in a categorical manner.

The ABX function averaged over all 15 subjects is shown in Fig. 2. This function shows an overall decrease in discrimination from 73 to 53 percent correct as rise time increases in duration. The predicted discrimination performance from individual identification functions using the Haskins formula (Pisoni, 1971) is also plotted in Fig. 2 for comparison. As expected, the predicted discrimination function is quite categorical. However, as shown in the figure, the predicted and obtained functions do not match very closely and by a within subject analysis of variance they were significantly different ($F(1,14) = 48.98$, $p < .001$). Thus, from these results we conclude that differences in rise time are <u>not</u> categorically perceived. Figures 3 and 4 display the individual ABX results for the 15 subjects. The same general trend can be observed here as well.

The group ABX discrimination function compares favorably with the results reported by Rosen and Howell (1981) in Experiment 3 of their study. Both discrimination functions fell from around 75% correct discrimination for the 0-20 ms pair to near chance as rise time increased. The most obvious difference between the functions is the slight 5% increase in our discrimination results at 20 ms whereas Rosen and Howell's function decreased monotonically. (We will say more about this difference below in describing the ABX only results.)

Rosen and Howell suggested that their ABX discrimination results can be predicted from Weber's law. Since the precise relation between discrimination performance in an ABX task and JND's for rise time predicted by Weber's law was unknown, Rosen and Howell (1981) employed the signal detection analysis developed by Macmillan, Kaplan and Creelman (1977) to predict Weber functions. Since we directly compare the ABX results with Weber fractions in Experiments 2 and 3 below, we suggest the following as a simple test of Weber's law for ABX functions. If the linear correlation of ABX performance with rise time is high and negative, Weber's law holds. The linear correlation of the group data in Fig. 2 was -.95. The linear correlation of Rosen and Howell's ABX function calculated from data in their Table 1 was -.96. In contrast, the linear correlation of an idealized categorical ABX function is zero. Therefore, the discrimination results obtained in this experiment as well as in Rosen and Howell's suggest that discrimination of rise time follows predictions derived from Weber's law.

We conclude that the identification and discrimination results obtained in this study closely replicate the results reported earlier by Rosen and Howell (1981) using an equivalent rise time continuum. The subjects in the present study were trained to label the stimuli consistently as pluck or bow. Although these subjects identified the rise time stimuli into discrete perceptual categories, their ABX discrimination functions were clearly not categorical by the traditional definition of the phenomenon (see Studdert-Kennedy, Liberman, Harris and Cooper, 1970).

C. Results: ABX Only

In order to find out whether ABX discrimination of rise time stimuli would be different for subjects not previously introduced to the pluck and bow labels, data were also collected from 17 subjects in the ABX only procedure. The two poorest subjects with an average performance of 50% correct were not included in the final analyses. Group results for the remaining 15 subjects in the ABX only procedure are displayed in Fig. 5. The left hand panel of Fig. 5 shows the results obtained separately from block 1 and block 2 of this experiment. Discrimination improved overall from block 1 (64.2%) to block 2 (67.9%). The two ABX functions were reliably different from one another by a within subjects analysis of variance ($F(1,14) = 6.81$, $p < .03$).

Insert Figures 5, 6, and 7 about here

ABX discrimination results are plotted separately for individual subjects in Figures 6 and 7. While some improvement in discrimination can be seen for all subjects, learning effects from block 1 to block 2 were small overall. Large subject-to-subject variability can be observed in the individual ABX functions which are similar to that shown in Figures 3 and 4. No clear pattern of results appears in the individual subject functions. Since these subjects received feedback for the correct response throughout the experiment and were all performing above chance, these results suggest that discrimination of rise time in this test format is a difficult task.

The purpose of collecting the ABX only data was to compare it to the ABX data following identification in the previous experiment (hereafter called ABX after ID). Each block in the ABX only procedure had the same number of triads as in the ABX after ID task. Block 2 of the ABX only data is most comparable to the ABX after ID data because both groups had similar amounts of prior listening experience with the rise time stimuli. The discrimination results from these two blocks of ABX trials are plotted together in the right hand panel of Fig. 5. These discrimination functions are very similar overall and did not differ from one another in an analysis of variance. There was, of course, a significant difference in discrimination across rise time pairs ($F(6,168) = 25.11$, $p < .01$). Thus, although individual ABX functions showed considerable variability, the average functions obtained for the two groups of 15 subjects were comparable. We conclude, therefore, that subjects' ability to label these rise time stimuli as plucks or bows apparently has little, if any, influence on their ability to discriminate differences along the continuum.

A detailed examination of the shape of the ABX only function (block 2) revealed a slight peak in discrimination for the 10-30 ms stimulus pair. It is not clear what caused this peak. Since only slightly more than half of the 30 subjects actually showed an increase from the first to the second stimulus pair, perhaps the peak is nothing more than a sampling problem. Whatever the reason, the overall shape of the discrimination functions are clearly <u>not</u> categorical. The ABX only function had a linear correlation of -.94 with stimulus duration, a value quite close to the -.95 value for the ABX after ID function. Thus, the

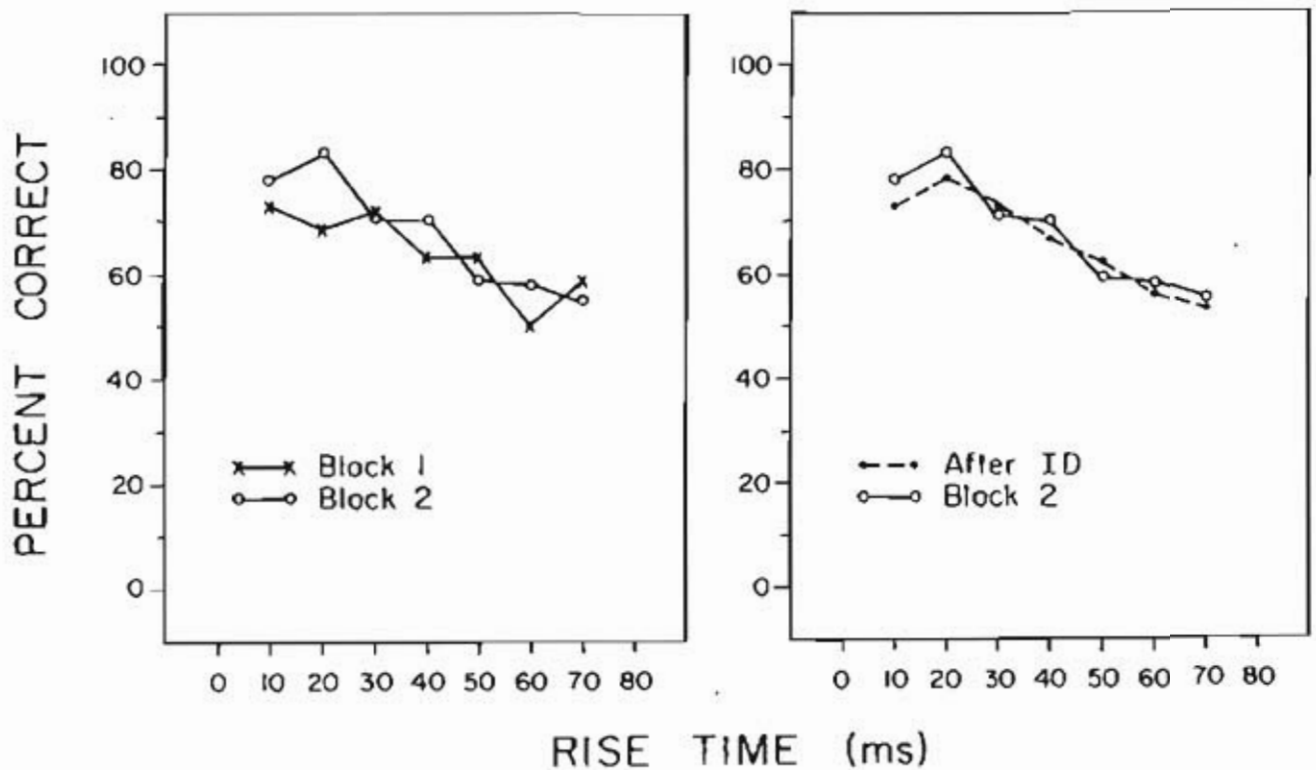


Figure 5. Group results from the ABX task plotted as percent correct discrimination. The left panel shows the results from Block 1 and Block 2 for subjects in the ABX only condition. The right panel compares Block 2 of the ABX only results with the ABX after ID results shown in Fig. 3.

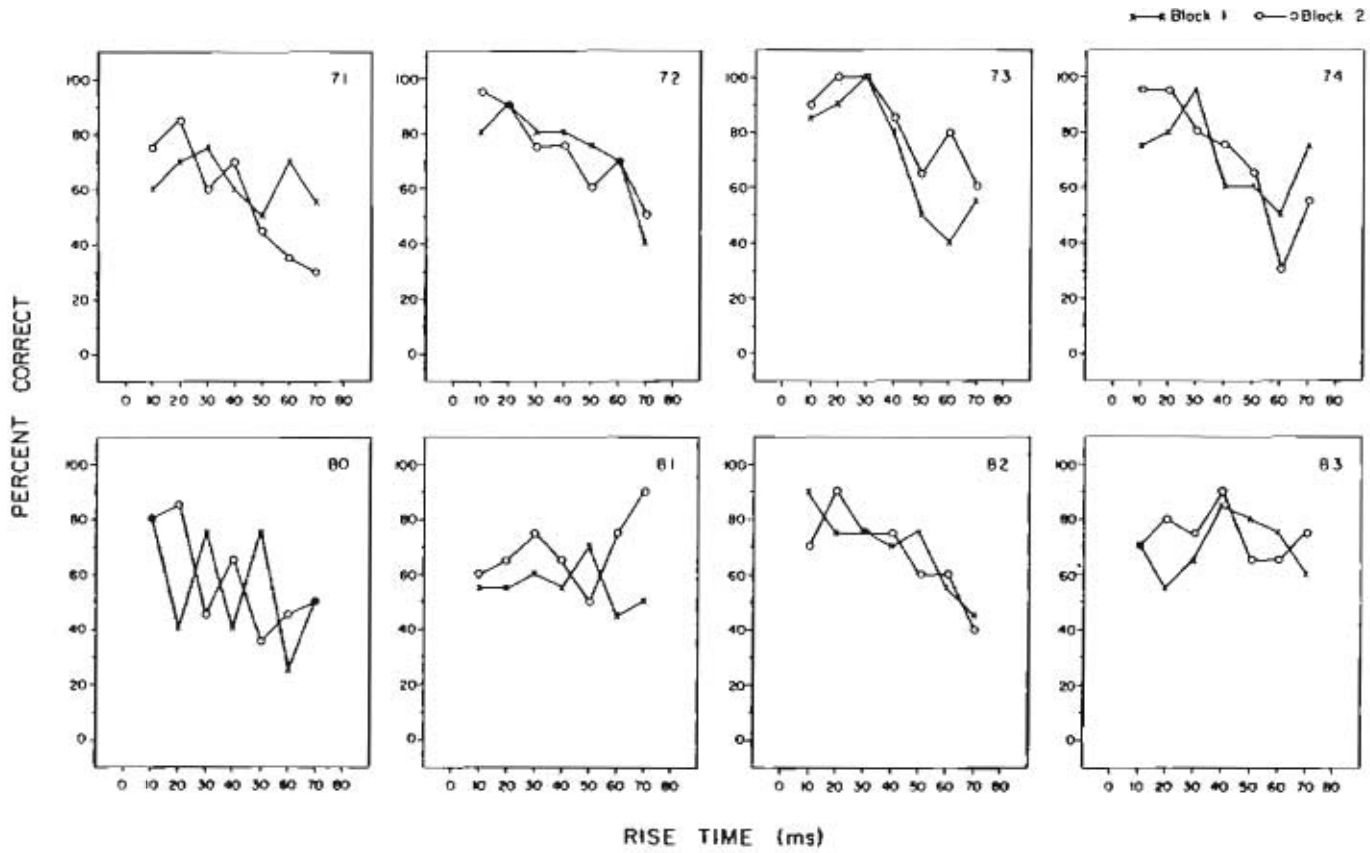


Figure 6. Individual data for Block 1 and Block 2 for subjects in the ABX only condition shown in the left panel of Fig. 5.

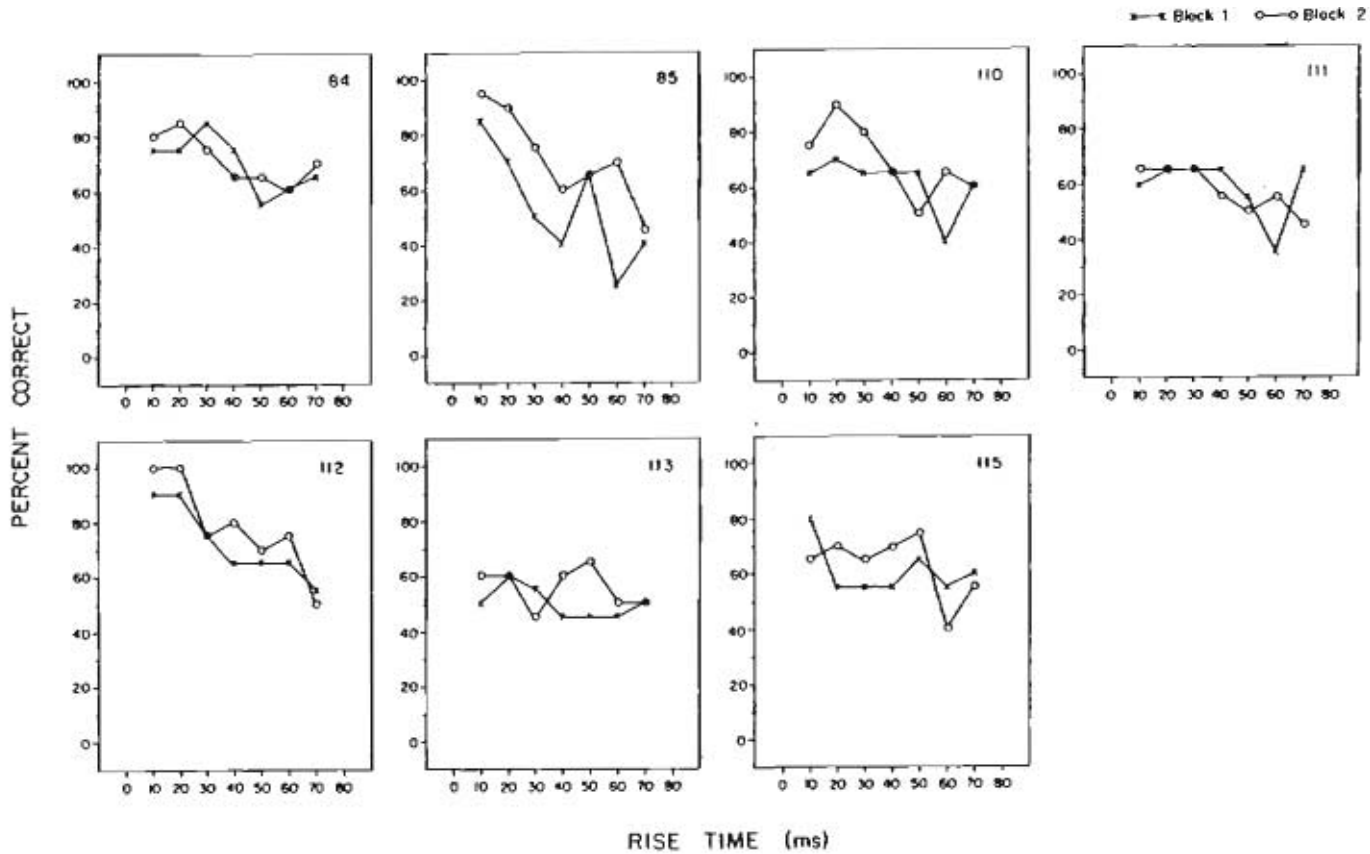


Figure 7. Individual data for Block 1 and Block 2 for subjects in the ABX only condition shown in the left panel of Fig. 5.

results from both ABX conditions clearly demonstrate that Weber's law holds for discrimination of rise time in sawtooth signals.

D. Discussion: Experiment 1

The present study provides an independent replication of Rosen and Howell's recent study (1981, Exp. 3). Cutting (1982) has also replicated the overall results of this study for a rise time continuum with linear increments of rise time. His stimuli differed considerably, however, in parameters such as period, size of the rise time increments and range of rise times presented. The overall results of these three independent studies indicate that subjects can reliably classify rise time stimuli into discrete perceptual categories such as pluck and bow. However, subjects do not discriminate these stimuli in the ABX task in a categorical manner. Instead, their discrimination performance is consistent with predictions derived from Weber's law. The motivation behind the next experiment was to test this hypothesis more directly using an appropriate psychophysical procedure for estimating JND's of rise time.

II. EXPERIMENT 2: DISCRIMINATION USING AN ADAPTIVE PROCEDURE

A widely used adaptive tracking procedure was chosen for this study. The transformed up-down procedure developed by Levitt (1970) for estimating the JND of rise time at a 70.7% correct level of discrimination was modified to run on-line from a minicomputer. Pilot studies with this procedure indicated that subjects initially found the discrimination task very difficult. Since our goal was to compare the JND's of rise time to the results obtained in the ABX task, the experiment was designed with the following task variables in mind. The ABX discrimination paradigm may be considered a very high uncertainty discrimination task (Harris, 1952). Sachs and Grant (1976) have shown that task uncertainty can influence the shape of the observed discrimination function. In particular, they obtained categorical-like discrimination functions for stimuli in a high uncertainty task, but not in a low uncertainty task. Thus, we decided to arrange our task as much as possible towards higher uncertainty conditions so that any tendency towards categorical perception of rise time might be revealed. Subjects were not highly practiced as is typically the case in psychophysical experiments and we began collecting data as soon as their discrimination performance stabilized. The rise time stimuli were randomly selected from the continuum during each testing session. The overall goal of this experiment was to examine the shape of the underlying discrimination functions obtained from this psychophysical procedure for the rise time durations previously examined in the identification and ABX tasks of Experiment 1.

A. Method

1. Stimuli

Forty-five additional sawtooth stimuli were digitally generated. Each sawtooth waveform was generated from 11 harmonics as in the previous experiment. The duration of all waveforms, however, was fixed at one second. The duration of the linear envelope impressed at the onset of the stimuli varied from 10.0 ms to

156.7 ms in 3.3 ms increments (i.e., the period of the 300 Hz fundamental). The amplitude of the stimulus then decayed linearly to 0 over the remainder of the waveform. The eight stimuli from 10 ms to 80 ms in 10 ms increments were the base stimuli used to estimate the JND's in discrimination. Rise time stimuli shorter than 10 ms were not used in this experiment because of the "thumpy" quality of these stimuli compared to longer rise time stimuli (see Rosen and Howell, 1981, p. 163).

All stimuli were output and low-pass filtered in the same way as in the previous experiment. Subjects listened to stimuli over one pair of matched, calibrated TDH-39 earphones in a sound-treated room (IAC model 401-A). The output voltage was calibrated to a constant level before each experimental session.

Rise time duration at the earphones was carefully measured for each of the base stimuli. The earphone was placed in an artificial ear connected to a 1 inch Bruel and Kjaer condenser microphone and sound level meter (model 2203) using the C weighting. The first 100 ms of each waveform was displayed and photographed on a Tektronix storage oscilloscope (model S103N). The only distortion observed in the photographs was a baseline offset probably due to bias in the earphone or microphone. The shape of the amplitude envelope at stimulus onset was clearly linear. The rise time values of the base stimuli were measured independently by two people and found to be the same as the nominal values, i.e., 10 to 80 ms.

2. Subjects

Four students working in the laboratory were recruited as paid subjects for the experiment. Subjects had little or no experience in listening to rise time stimuli and none of them had previously participated in a psychoacoustic experiment. All subjects were audiometrically tested for the octave frequencies from 500 to 8000 Hz using a Grason-Stadler Model 1701 audiometer. Audiograms for both ears were within normal ranges of individuals in their age group of 19 to 23 years.

3. Procedure

Using a two-interval forced choice task with feedback, the subjects were asked to indicate which item of each pair of stimuli had the "more gradual onset." For each run of the tracking procedure, a base stimulus and an associated initial stimulus were selected. The longer rise time of the initial stimulus was chosen to be about four times the JND's estimated by van Heuven and van den Broeke (1979). For the first five reversals in a run, a 10 ms step size was used. The step size was then decreased to 3.3 ms for two test reversals. Finally, the JND was estimated from the mid-run average of the next 10 reversals at the $p(c) = .707$ level of correct responses. On the average one JND estimate involved about 60 trials.

Stimulus presentation and data collection were under the control of a PDP-11/34 computer. Responses were collected from button presses on a response box equipped with feedback and cue lights. Subjects ran individually in approximately one hour sessions. A new random order of presentation of each of the eight base stimuli was used each day. At the beginning of each day, each subject listened as long as he wanted to the pair of initial and base stimuli

used in the first run. Subjects were tested daily until the JND estimates stabilized. The experimental data were then collected on the next four consecutive days. Three of the four subjects ran for a total of eight days, the fourth subject ran for an additional day to collect further data.

B. Results

JND estimates for each of the four subjects are plotted in Fig. 8 as the difference in rise time (ΔRT) for each base stimulus. ΔRT was calculated from each trial as the difference between the base stimulus rise time and the mid-run average rise time. Going from left to right in Fig. 8, the first panel displays the mean of the ΔRT values, the second panel the median, and the third panel the range of the maximum and minimum ΔRT values obtained. Each panel shows a clear overall trend of increasing values of ΔRT as the rise time of the base stimuli increased from 10 to 80 ms. Individual subject differences can be observed but they are not excessive given the modest amount of training prior to data collection. The rather extreme values in the range of ΔRT indicates that rise time discrimination in this task was very difficult for subjects, particularly for long duration rise time stimuli. Since subjects occasionally produced "wild points" in the tracking procedure, the median discrimination functions are probably a better estimate of rise time JND's in this experiment than the mean functions.

Insert Figure 8 about here

Combined functions for the four subjects are shown in Fig. 9 for the means and Fig. 10 for the medians. The discrimination functions shown in the left hand panels of these two figures indicate that ΔRT increases with rise time. The slight decrease in ΔRT from 10 to 30 ms of about 1.3 ms is less than the step size of 3.3 ms between stimuli and was contributed primarily by one subject, JL. The drop in ΔRT at 80 ms for the median function was also contributed by only one subject, SG. Overall, these functions clearly show that subjects discriminate rise time differences well at short rise time durations, whereas they do much more poorly and display larger amounts of variability at longer rise time durations.

Insert Figures 9 and 10 about here

If Weber's law is applicable to rise time discrimination, then the ΔRT functions for the mean and median should closely approximate a straight line. The high linear correlations of $r = .91$ (mean) and $r = .88$ (medians) suggest that these functions are reasonable approximations to straight lines. In addition, we have calculated the Weber fraction as ΔRT/rise time and displayed these values in the right hand panels of Figures 9 and 10. These functions are typical of

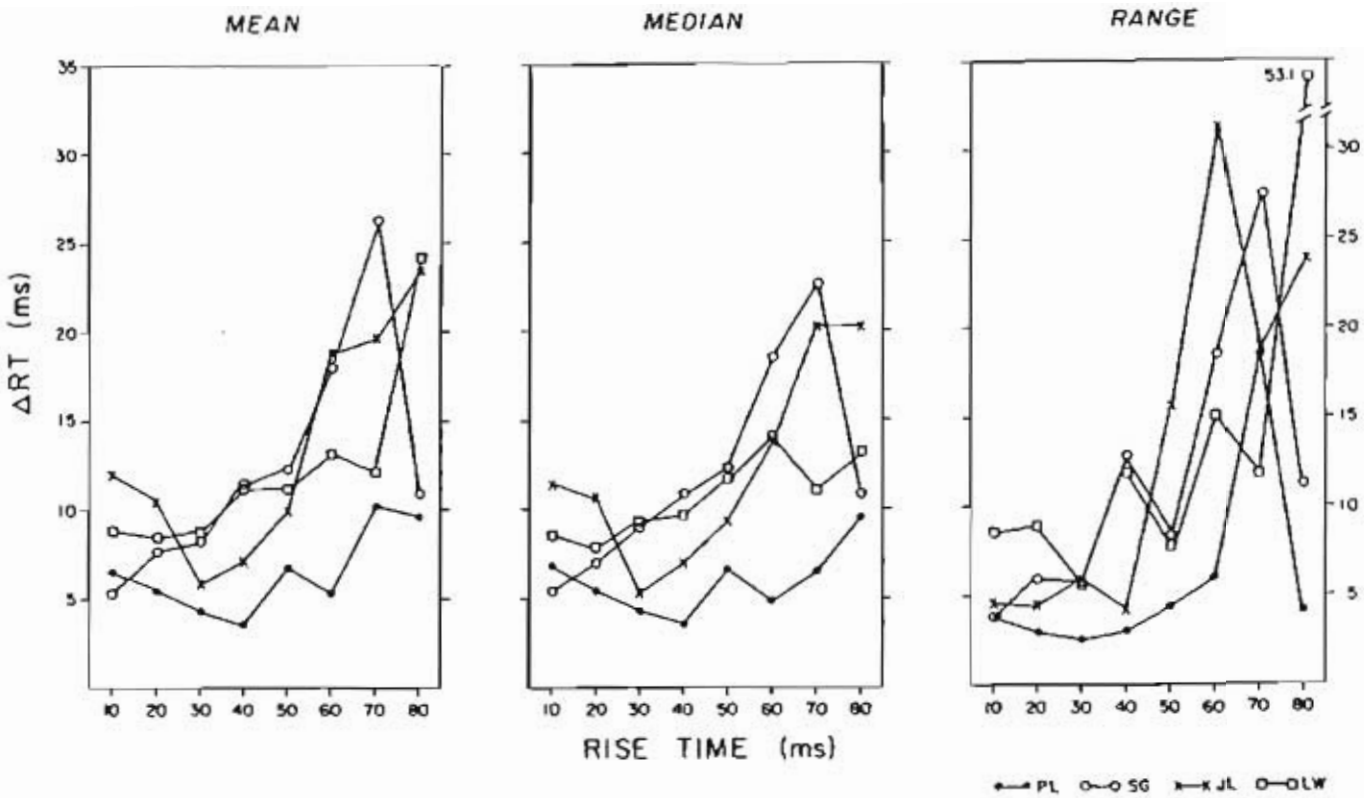


Figure 8. Estimates of the JND's for rise time (ΔRT) as a function of rise time for each of four subjects in Experiment 2. The panels display the mean, median and range of ΔRT values for the four subjects estimated in the adaptive tracking procedure (see text for details).

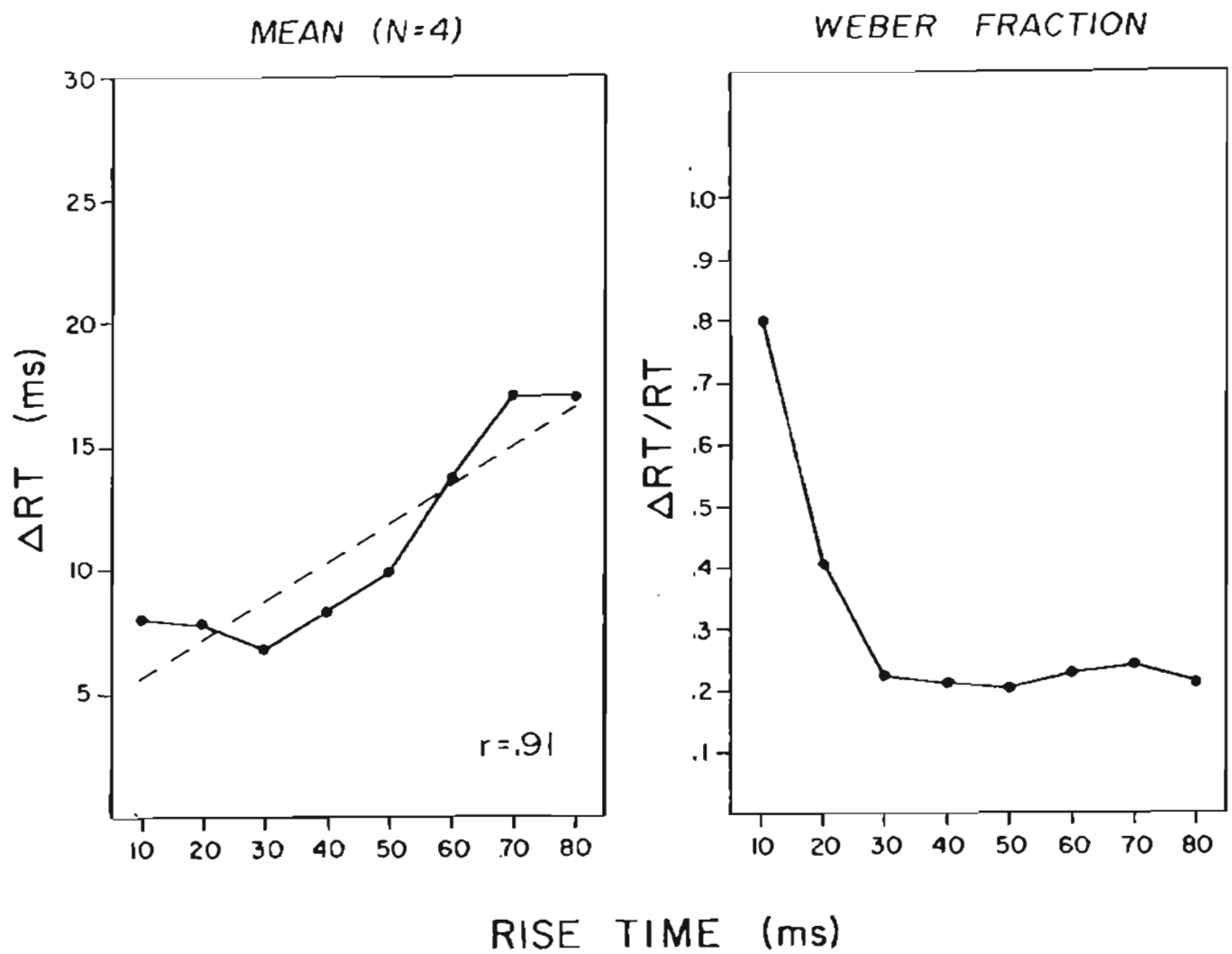


Figure 9. Averaged results from the means shown in Figure 8. The left panel shows the ΔRT mean function and its associated regression line, the right panel displays the Weber fraction calculated from the ΔRT mean function as ΔRT/RT.

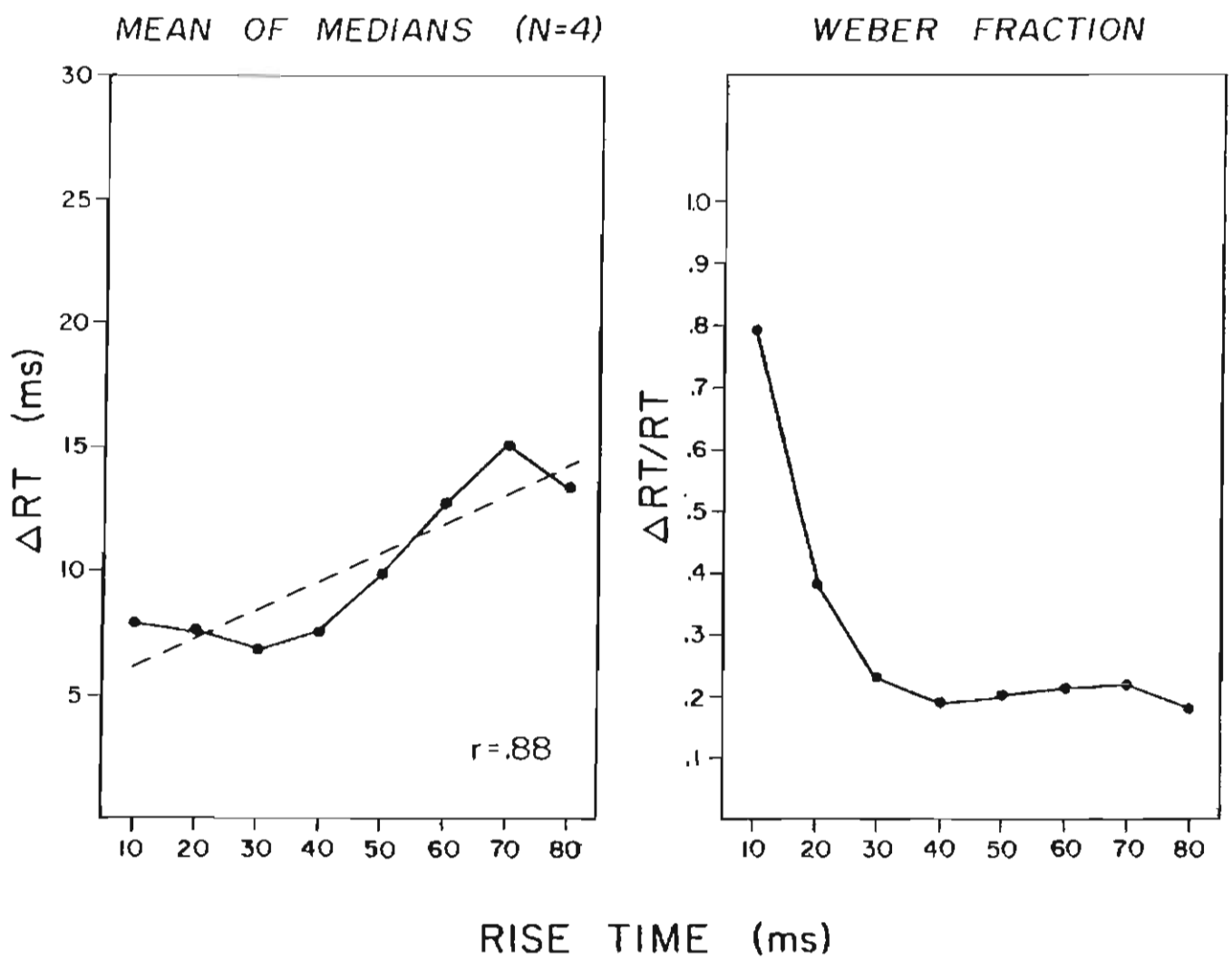


Figure 10. Averaged results from the means shown in Figure 8. The left panel shows the ΔRT median function and its associated regression line, the right panel displays the Weber fraction calculated from the ΔRT median function as ΔRT/RT.

discrimination data obtained for simple auditory dimensions and other sensory continua where the Weber fraction is large at short stimulus values and then becomes constant over a wide range of longer stimulus values (Gescheider, 1976). Nearly constant values of the Weber fraction were obtained between 30 and 80 ms rise time for both the mean (= .22) and median (= .21) functions.

C. Discussion

The JND's estimated directly in the adaptive tracking procedure demonstrate that the discrimination of rise time follows predictions derived from Weber's law. Our results are in good overall agreement with those reported recently by van Heuvan and van den Broeke (1979) who estimated JND's using an indirect method of adjustment. The main difference in the results between the two studies occurred for base rise time values between 0 and 30 ms. In our results, ΔRT was fairly constant in this range while van Heuvan and van den Broeke reported increasing ΔRT values. As discussed below, this difference may be due to the differences in step size between stimulus pairs. Our step size was 3.3 ms compared to approximately 1 ms for their 1000 Hz sine wave. It is also possible that a floor effect is present in estimating ΔRT which depends on stimulus step size.

Although we found results that supported predictions derived from Weber's law, our experiment was arranged to include several high uncertainty task variables. For similar task variables, Sachs and Grant (1976) have reported non-monotonic, V-shaped functions for measuring the JND of voice onset time. In the present experiment, only one subject, JL, produced a V-shaped function. As a partial check on the influence of task uncertainty and subject strategies, we asked subject JL to return so we could measure the JND for the 10 ms rise time stimulus in a blocked, low uncertainty test at the end of the experiment. Our measurement of her ΔRT fell from approximately 12 ms to 5 ms. This subject commented that her overall listening strategy had not worked well for these somewhat "thumpy" stimuli in the usual random order presentation. Interestingly, Rosen and Howell (1981) have reported previously that their subjects appeared to have different listening strategies for the short rise time stimuli which influenced the ABX discrimination results. Perhaps subjects in these discrimination experiments should have been specifically instructed to discriminate rise time differences according to a criterion based on either loudness or rate of onset as was done in several earlier experiments (Gershuni and Zaboeva, 1962; Vigran, Graevenes and Arnesen,1964; Nabelek, 1965). Thus, at least some of the variability obtained in our ΔRT estimates may be attributable to the influence of task uncertainty as well as subjects' listening strategies.

Given the outcome of Experiments 1 and 2 we were not able to account for the categorical-like discrimination results Cutting (1982) obtained with an equal interval logarithmic continuum. Therefore, we sought to develop a method for directly comparing the results from the equal logarithmic and equal linear increment continua by referring to the underlying psychometric function. In order to develop this method, the next experiment was designed to specify the relation between functions obtained in the ABX and adaptive tracking procedures.

III. EXPERIMENT 3: IDENTIFICATION AND DISCRIMINATION USING EXPERIENCED LISTENERS

All four subjects who ran in the adaptive tracking task of Experiment 2 were recalled after 4 weeks and asked to participate in both the identification and ABX tasks used in Experiment 1 so that the results from both procedures could be compared. While these subjects were experienced listeners of rise time stimuli, they were not familiar with the pluck and bow labels used in the identification testing in Experiment 1. A warm-up sequence of 10 presentations of the 0 and 80 ms stimuli with feedback was used to acquaint these subjects with the pluck and bow labels. The same set of nine stimuli generated for Experiment 1 were used here. As before, all four subjects from Experiment 2 were paid to participate in this experiment.

A. Results and discussion

Identification and ABX discrimination functions averaged over the four subjects are presented in Fig. 11 using the same format as for Experiment 1 (see Fig. 2). Individual functions are plotted in Fig. 12. The identification functions shown here are slightly noisier than those observed in Experiment 1. This is probably true because these subjects had no prior experience with the identification task. All four subjects, however, were able to categorize the endpoint stimuli as plucks or bows 100% of the time with reasonably sharp crossovers between the response categories. Overall, however, these identification functions are quite similar to those obtained in Experiment 1.

Insert Figures 11 and 12 about here

The individual ABX discrimination functions shown in Fig. 12 also reveal a great deal of subject-to-subject variability. While these subjects performed similarly to those in Experiment 1, it was surprising to us that these experienced listeners did not respond more uniformly. The average ABX function is shown in Fig. 11. While discrimination of rise time decreased as rise time increased, this function is somewhat different from the ABX functions in Experiment 1 shown in Fig. 5. The present function has more inflections and reveals higher levels of discrimination for the longer, 30 to 70 ms stimulus pairs. Nonetheless, there is a strong linear component to the discrimination function with $r = -.84$. As shown clearly in Fig. 11, the predicted ABX discrimination function does not resemble the observed function. The observed ABX discrimination functions for these four subjects are consistent with the interpretation we gave in Experiment 1 based on predictions from Weber's law.

The purpose of this experiment was to directly compare the ABX discrimination results with the JND estimates obtained for the same subjects in the previous experiment. This comparison can be made by interpreting the results from each experiment in terms of the underlying psychometric function. First, we need to accept the following assumptions about the psychometric function. A psychometric function for rise time was not estimated in the adaptive tracking

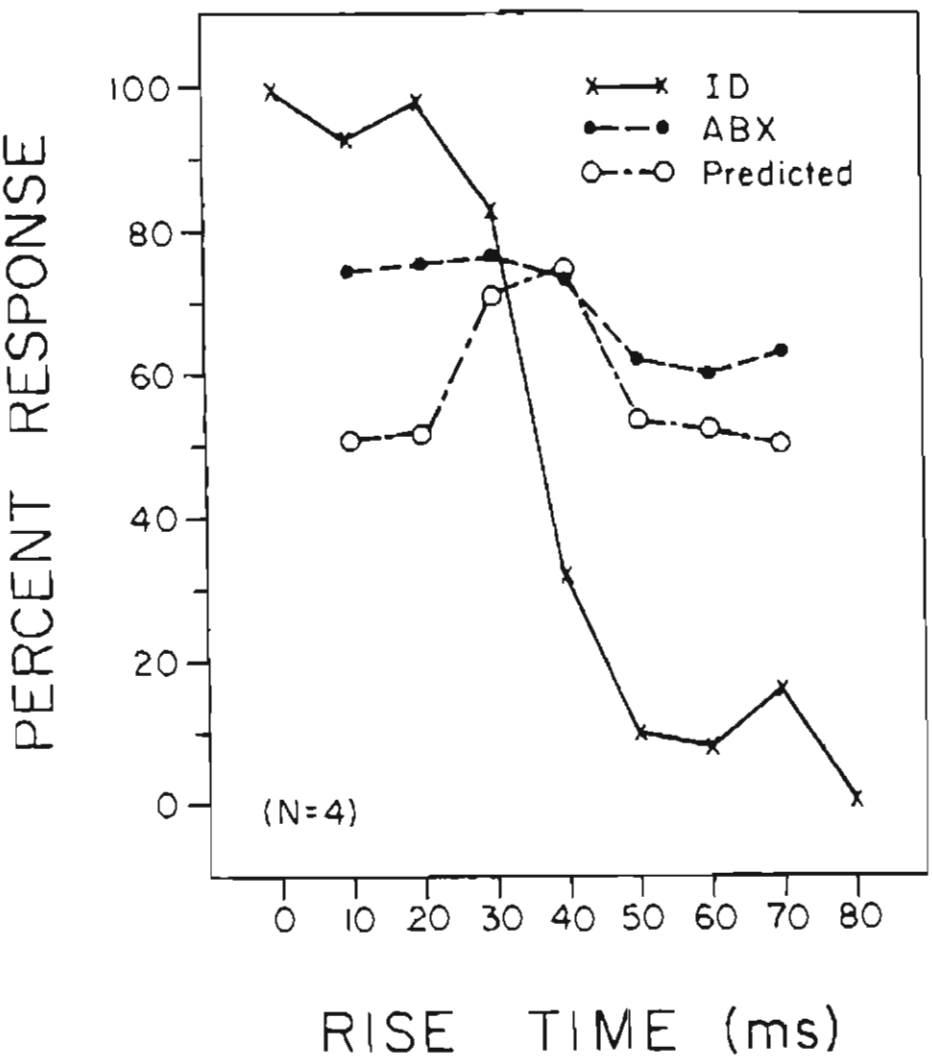


Figure 11. Group results for the four experienced listeners for the identification and ABX tasks. Data are plotted in the same way as in Fig. 2.

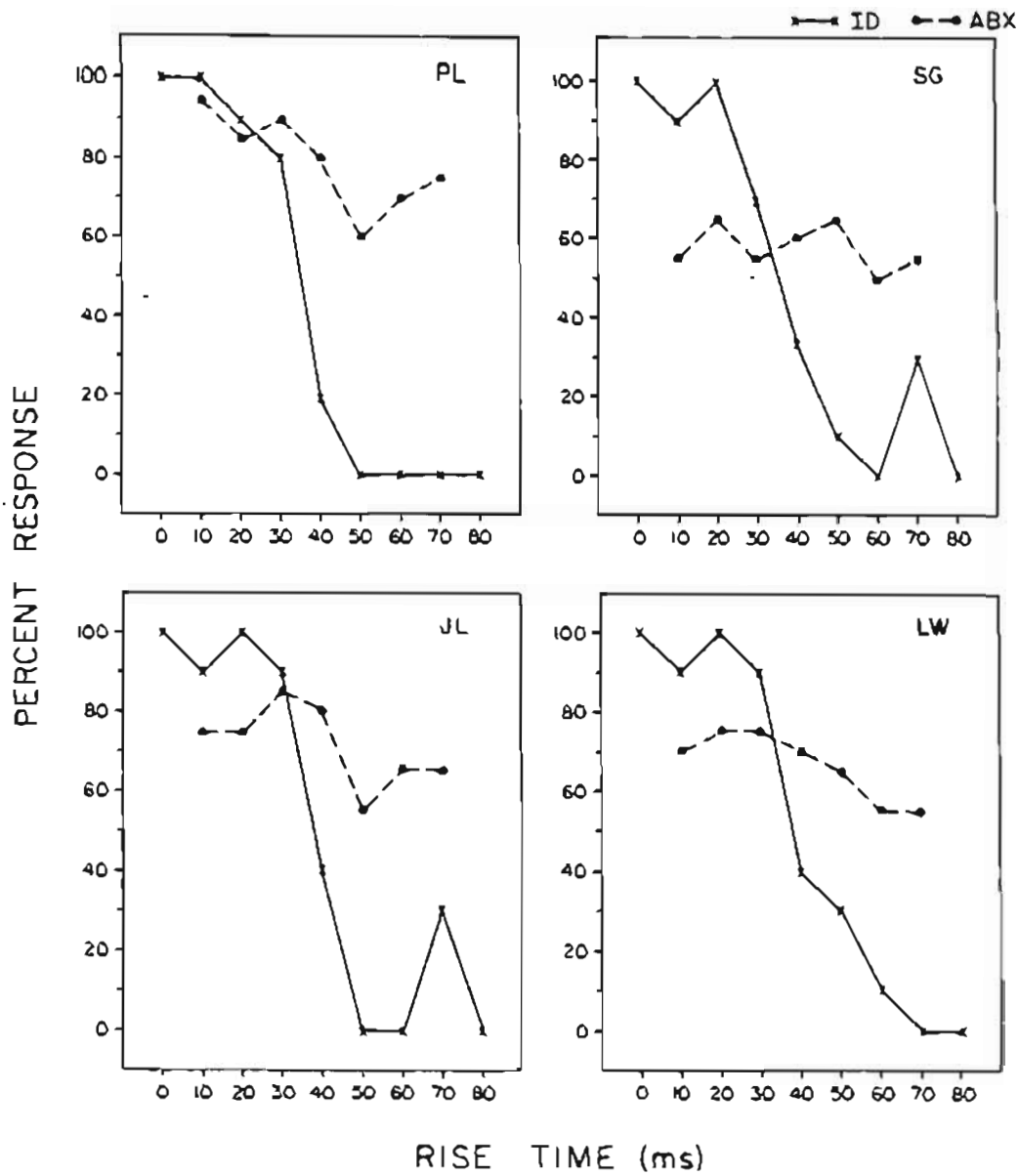


Figure 12. Individual results for the four experienced listeners in the identification and ABX tasks.

procedure used in this research. We may reasonably assume, however, that psychometric functions for rise time are monotonically increasing functions. Since discrimination of rise time as measured by the adaptive tracking procedure, or by the method of adjustment (van Heuvan and van den Broeke, 1979), is similar to the discrimination of other sensory signals, we will assume that the psychometric function resembles the usual ogive shape obtained when percent correct response is plotted as a function of ΔRT (c.f. Levitt, 1970 and Gescheider, 1976). The important assumption for our analysis is that the psychometric function may be approximated by a straight line over some range of percent correct. In this analysis, the range will be between 55% and 95% correct responses which is appropriate for an ogive where chance is 50% correct. A further assumption we make is that similar psychometric functions also underlie rise time discrimination in the ABX task.

Given the above assumptions, a predicted ΔRT for rise time can be calculated from the percent correct response in the ABX discrimination task when the percent correct response lies on the linear portion of the psychometric function. The equation for computing this straight line is:

$$p = a \cdot \Delta RT + b \qquad (1)$$

where p is percent correct. In the ABX task, the ΔRT for each discrimination pair was held constant at a 20 ms step size and a percent correct response was obtained. In the tracking procedure, the percent correct response was held constant at 70.7% correct and a ΔRT was obtained. Thus, we may obtain an estimated ΔRT in the ABX task at the 70.7% correct level of performance by solving the two simultaneous Eqs. (1) in terms of the intercept, b. The result is Eq. (2) where the slope of the psychometric function, a, is still a parameter and p is the obtained percent correct in the ABX task.

$$\Delta RTpred = 20 + \frac{(70.7 - p)}{a} \qquad (2)$$

While the slope a was not determined in these analyses, some reasonable bounds on the values it can take may be estimated from data obtained in Experiment 2. A slope of 1 means a percent correct increase by 1% for each increase in ΔRT of 1 ms. This shallow slope implies that an additional ΔRT of 25 ms would be necessary to go from 70% correct on the psychometric to 95% correct. Referring to the functions displayed in Fig. 8, it is obvious that 25 ms is too large. A value of 5, on the other hand, is better but still too steep. If we pick an intermediate value for a, say a=3, this will allow us to solve for ΔRTpred in Eq. (2). Although the slope of the psychometric becomes shallower at longer stimulus durations, we will set a to a constant here. Later we show that if a becomes smaller, the results in this comparison improve even more.

Figure 13 shows three estimates of ΔRT as a function of rise time. From the adaptive tracking procedure, the median function from Fig. 10 is plotted as the best estimate of rise time discrimination. Plotted for the same four subjects are the ΔRTpred values calculated from the ABX results using Eq. (2), labeled

ABX, N = 4. Note that the ABX procedure did not include the 80 ms rise time value. For further comparison, the ΔRTpred values for the fifteen subjects in the ABX after ID procedure from Experiment 1 are also plotted, labeled ABX, N = 15. The 70 ms point was dropped on the N=15 function because ABX performance was only 53% correct.

The shape of the predicted and obtained ΔRT functions from the four experienced listeners are remarkably similar, although they are displaced from each other by about 12 ms. The Pearson product moment correlation coefficient r was used to assess whether the shape of the two functions is similar. In this case, we interpret r^2 in percent to mean the extent to which the shape of one function, as measured by its variance about its mean, can be predicted from the other function. r^2 is a measure of the similarity of shape which is not sensitive to displacements between the functions' means nor any differences in the absolute values that these functions take about their means.

Insert Figure 13 about here

The ΔRTpred function correlated with the median function from the same four subjects had an r = .85 and r^2 =.72. That is, 72% of the variance of the ΔRTpred curve in the ABX experiment can be predicted from the shape of the median function. In Fig. 13, the ABX, N = 15 function from Experiment 1 and the present, ABX, N = 4 function lie very close to one another. Correlation of the ΔRTpred values from 10 to 60 ms rise time was r = .92 indicating that 84% of the shape of one function is predictable from the other. Furthermore, the correlation r of the ABX, N = 15 function with the median function was also high at .89. Taken together, this analysis indicates that there is a strong relation between the overall shape of discrimination functions obtained from the ABX paradigm and the adaptive tracking method of estimating ΔRT. We assume that this is true, in part, because the adaptive tracking task was biased towards higher stimulus uncertainty conditions making it more similar to the high uncertainty ABX task.

We can now consider how estimating the slope a as a constant equal to 3 in Eq. (2) affects the above results. Varying a between 1 and 5 as a constant in Eq. (2) will only change the spread of the ΔRTpred values about the mean of the function, and therefore will not affect the Pearson r. On the other hand, if a were decreased as rise time increases to model the usual trend of the psychometric function becoming more shallow, the ΔRTpred values for longer rise times would become larger. In this case, the shape of the ΔRTpred function would more closely resemble the median function and the correlations between the two sets of data presented above would improve.

Given the above analysis for predicting JND's from the ABX task, we turn to the recent findings reported by Cutting (Fig. 4, 1982) in which he claimed that categorical functions can be obtained for some rise time continua. In his Experiment 3, 12 subjects listened to two different rise time continua. The first continuum had equal linear step sizes (16 ms) and the obtained ABX discrimination function was similar to those of Rosen and Howell (1981) and our

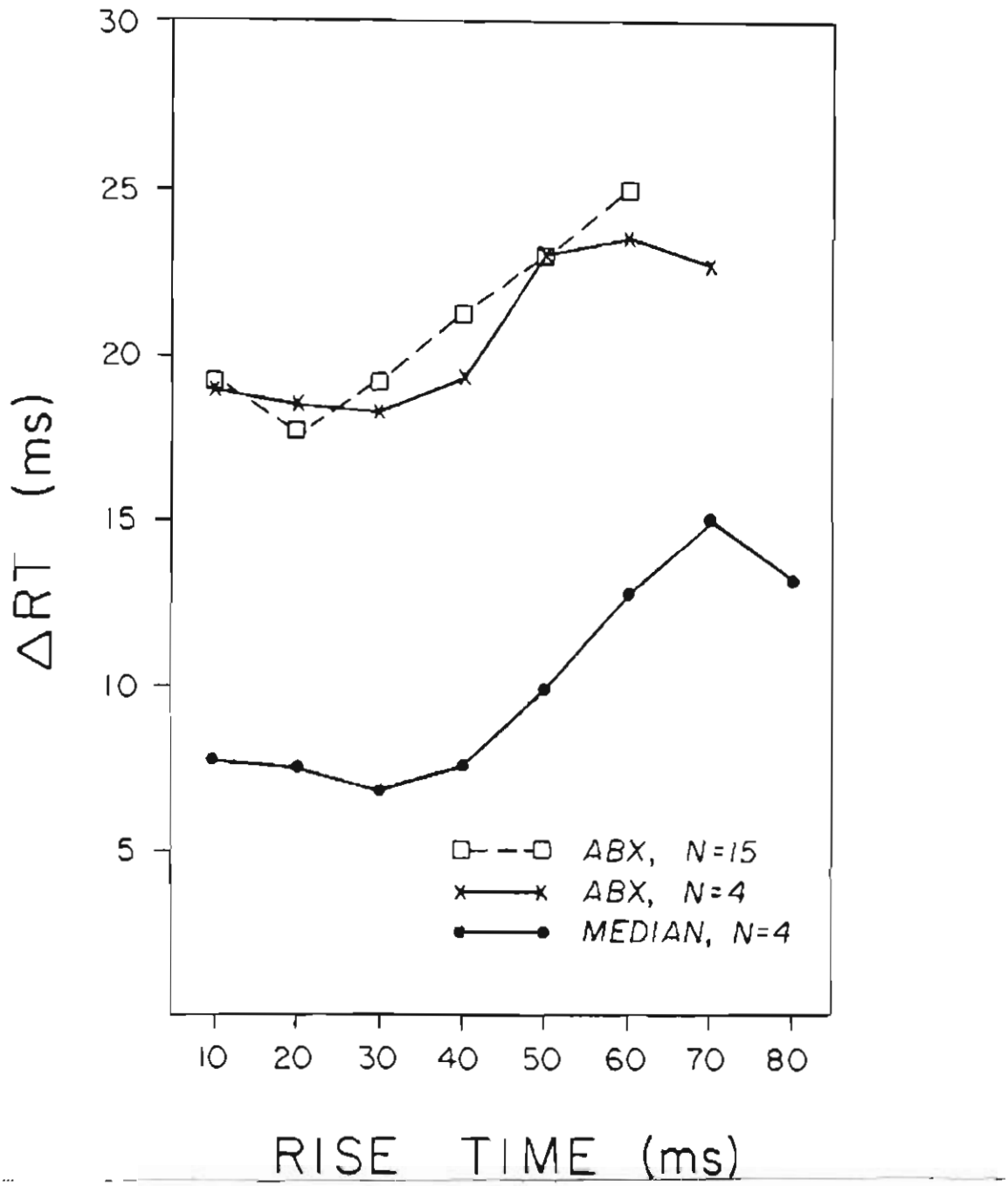


Figure 13. Predicted values of ΔRT calculated from Eq. 2 are shown for the ABX after ID results from 15 naive listeners (see Fig. 2) and the four experienced listeners (see Fig. 11). The ΔRT median function for the same 4 experienced listeners (see Fig. 10) is also shown here for comparison.

Experiment 1 (Fig. 5). The other continuum had arbitrarily chosen equal logarithmic step sizes ranging from 3.6 ms to 10.8 ms as rise time increased. The obtained two-step ABX discrimination function for this continuum had a slightly more categorical shape. While Cutting offered no explanation why the same subjects discriminated similar stimuli according to Weber's law on the one hand or a more nearly categorical basis on the other hand, we believe we can resolve this paradox. To do this, we calculated a ΔRTpred for Cutting's stimuli from Eq. 2 by substituting the ΔRT step size he used in his experiments for the 20 ms step size we used. In his ABX experiments, the step size for the linear stimuli was held constant at 16 ms; the step size for the logarithmic stimuli varied from 7.9 ms to 19.7 ms. The ΔRTpred for both continua are plotted in Fig. 14 as a function of the average value for the linear stimulus pairs (labeled JEC linear) and the intermediate rise time value for the two-step log pairs (labeled JEC, log). The first pair of log stimuli and the last pair of linear stimuli were dropped from this analysis because percent correct response fell below 55%. The ΔRTpred for the four experienced listeners and the median JND estimates from Fig. 13 are also plotted in Fig. 14 for comparison.

Insert Figure 14 about here

Several interesting results are displayed in this figure. First, both the linear and the log functions are nearly straight lines with linear regression correlations of .99 and .97 respectively. The JEC linear function lies on top of the ΔRTpred function from our ABX, N = 4 experiment. The JEC log function begins with ΔRTpred values midway between our two functions and then increases linearly to reach the same ΔRTpred values achieved by ABX, N = 4 subjects. The JEC log function shows almost no evidence of the peaked, categorical function shown in Cutting's Fig. 4. Although the range of rise times used by Cutting was slightly smaller than those used here, our ΔRT functions were quite linear over the range encompassed by Cutting's log continuum. As one would expect, the Pearson correlation of the JEC log function with our median function was very high, r = .99. On the other hand, the range covered by the JEC linear continuum included an inflection point on our median ΔRT function, so the Pearson r was low, .69. As we mentioned earlier, however, the JEC linear function itself is monotonically increasing and highly linear (regression r = .99). Overall, we conclude that the results for both Cutting's linear and log continua are fully replicated by the findings of our experiments. Furthermore, we claim that all of Cutting's results are in complete agreement with an account of rise time discrimination that is based entirely on predictions derived from Weber's law.

Since the four ΔRT functions in Fig. 14 are quite similar in shape, how can we account for the displacement between these functions? One reasonable explanation lies in the step size between the rise time stimulus pairs. In the tracking procedure, the step size was 3.3 ms. In our ABX task, the step size was 20 ms while in Cutting's linear task it was 16 ms. The variable step sizes for Cutting's logarithmic stimuli ranged between 7.9 ms and 19.7 ms. Results shown in Fig. 14 show that the ΔRTpred functions for the 20 and 16 ms step sizes overlap. The JEC log function achieved similar ΔRTpred values for the last two log pairs having rise time step sizes of 16.4 and 19.7 ms. The rapidly

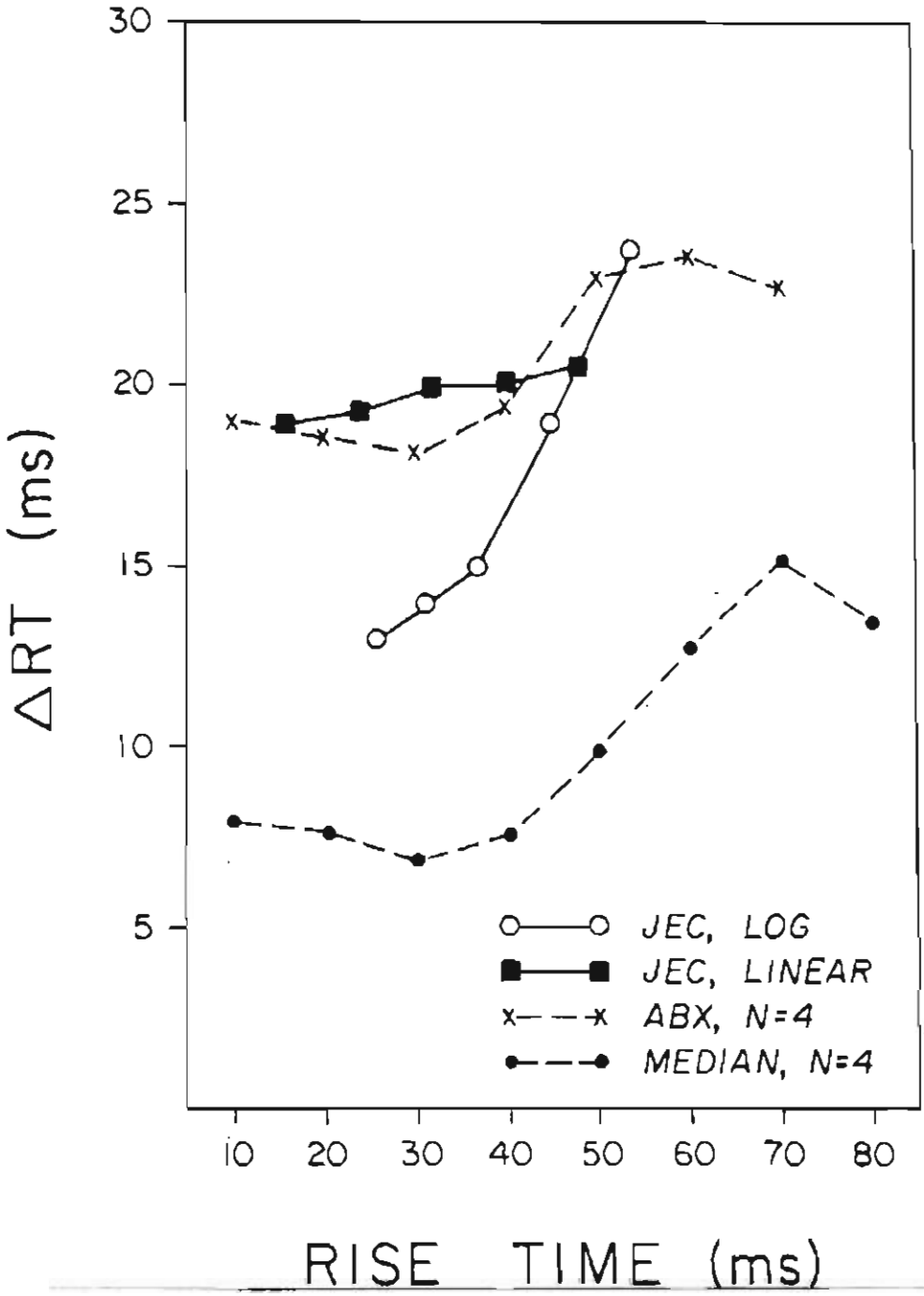


Figure 14. Predicted values of ΔRT calculated from Eq. 2 are drawn in solid lines for results reported by Cutting (1982) for two rise time continua. The filled squares show results for a continuum with equal linear increments whereas the open circles display the results for a continuum with equal logarithmic increments. The two ΔRT functions from Fig. 13 are redrawn with dashed lines for comparison.

increasing ΔRTpred values for the short rise time pairs suggests the existence of a family of functions for ΔRT which falls between the median function (at 3.3 ms) and the ABX, N = 4 function (at 20 ms) according to the step size used in the experiment. This account also provides a firm basis for explaining the differences in JND's between our results and those reported by van Heuvan and van den Broeke (1979). Specifically, they obtained an increasing ΔRT function for rise times less than 30 ms, whereas our function was flat. This difference is probably attributable to differences in step size. These investigators used 1 ms differences for their stimuli; we used a longer 3.3 ms step size for our stimuli which may have introduced a floor effect resulting in a constant ΔRT for small rise time values.

IV. CONCLUSIONS

Several conclusions can be drawn from the present set of experiments on rise time discrimination. First, the discrimination of rise time differences for sawtooth waveforms appears to be typical of discrimination of differences for other sensory continua -- the discrimination follows predictions derived from Weber's law. In this study we showed that an interpretation based on Weber's law can be derived from ABX discrimination results as well as by directly estimating the JND's of rise time with an adaptive psychophysical procedure. Correlational analyses confirmed that the predicted JND functions from the ABX data were similar to the JND functions obtained from the tracking procedure, except for an offset attributed to the step size of the rise time continuum. Since Weber's law clearly applied to the JND's for rise time estimated directly, we also claim that Weber's law applies to the discrimination functions obtained from the ABX paradigm. Negatively sloping ABX functions for longer rise time values were obtained for all the equal-interval rise time continua investigated by Rosen and Howell (1981) and even Cutting (1982) himself. While these authors assumed that Weber's law should imply negatively sloping discrimination functions, we believe the results of the present study have now clearly demonstrated this to be the case.

An interpretation of the ABX discrimination functions in terms of predictions based on Weber's law is important because the ABX paradigm is often used to compare speech and non-speech continua for the purpose of determining natural sensitivities in the auditory system (cf. Stevens, 1982). In the past, if the ABX discrimination functions were monotonic and did not show the predicted discrimination peaks and troughs suggestive of categorical perception, no specific alternative explanations of the discrimination of the stimuli were proposed. Now it is clear that negatively sloping ABX functions are evidence that Weber's law holds for the underlying sensory discrimination being examined.

The non-categorical perception of rise time differences for sawtooth waveforms is surprising, however, in light of Delgutte's (1981) recent study of the pattern of auditory-nerve firings to speech sounds. Rise time at the onset of fricative sounds is an acoustic correlate for the distinction between the speech categories of /t// and /// (Gerstman, 1957). Furthermore, there is evidence that this distinction may be categorically perceived for a /t// to /// rise time continuum (Cutting and Rosner, 1974). Delgutte (1981) observed differences in the patterns of discharge rates between /t// and /// which were interpreted to be possible physiological responses associated with the

categorical perception of these sounds. Delgutte's findings suggest the intriguing possibility that there may be differences in perception of rise time between noise in fricative waveforms and periodic waveforms such as sawtooths and sinusoidal signals. Indeed, van Huevan and van den Broeke (1979) using the method of adjustment showed that the JND estimates for rise time in sinusoidal signals were somewhat better than for noise signals. Thus, further investigation of the discrimination of rise time in aperiodic waveforms such as fricatives is evidently warranted based on the present results with periodic signals.

Finally, we wish to reexamine Cutting's (1982) remarks concerning categorical perception. We strongly disagree with Cutting's claim that categorical perception as assessed by identification and ABX paradigms is a "fickle" phenomenon which may "appear and disappear" in curious ways as task variables are manipulated experimentally. The effects of stimulus uncertainty (Sachs and Grant, 1976; Carney, Widin and Viemister, 1977), and memory load (Fujasaki and Kawashimi, 1970; Pisoni, 1971; Pisoni, 1973) on speech sound discrimination are reasonably well understood at this time. An analysis of discrimination in the ABX paradigm and its relation to other discrimination paradigms has been undertaken as well by Macmillan, Kaplan and Creelman (1977). Cutting's remarks about the relations between identification and discrimination were directed primarily to difficulties in interpreting his recent results with stimuli which were based on arbitrarily selected equal logarithmic intervals. While his initial analysis of the ABX discrimination results for the log stimuli appeared categorical, our re-analysis of his data demonstrated that the peaks in the discrimination functions were actually artifacts of the unequal rise time intervals he selected. Thus, contrary to Cutting's remarks, the present experiments demonstrate that when careful attention is given to task variables, the ABX paradigm is valid and useful for measuring discrimination of sensory continua, whether speech and nonspeech, and for making comparisons between various continua. More importantly, however, is the resolution of what appeared to be a conflict in the literature between Rosen and Howell and Cutting. Our results demonstrate clearly that rise time is discriminated in a manner that is consistent with predictions derived from Weber's law. The earlier reports in the literature demonstrating categorical-like discrimination of rise time appear to be due primarily to artifacts in the spacing of signals used to measure discrimination.

ACKNOWLEDGMENTS

We wish to thank our colleagues Donald E. Robinson, Thomas E. Hanna and Robert Gilkey for many fruitful discussions concerning the design and analysis of the adaptive tracking experiments. We also thank Paul A. Luce for helping to run subjects in the experiments. A preliminary report of these findings was presented before the Society at the 103d meeting in Chicago, IL, April, 1982. The research reported here was supported by the National Institutes of Health, Research Grant NS-12179and the National Institutes of Mental Health, Research Grant MH-24027 to Indiana University, Bloomington, IN.

REFERENCES

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

Carney, A. E., Widin, G. P., and Viemeister, N. F. (1977). "Noncategorical perception of stop consonants differing in VOT," J. Acoust. Soc. Am., 62, 961-970.

Cutting, J. E. (1982). "Plucks and bows are categorically perceived, sometimes," Percept, Psychophy., 31, 462-476.

Cutting, J. E., and Rosner, B. S. (1974). "Categories and boundaries in speech and music," Percept. Psychophy. 16, 564-570.

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

Fujisaki, H., and Kawashima, T. (1970). "Some experiments on speech perception and a model for the perceptual mechanism," Annual Report of the Engineering Research Institute, 29, Tokyo: University of Tokyo, Faculty of Engineering.

Gershuni, G. V., and Zaboeva, N. V. (1962). "Evaluation of functional significance of auditory system responses to exponentially increasing wide-band noises and tones," Fiziol. Zh. SSSR 48, 1178 (Translation T275-T289).

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

Geocheider, G. A. (1976). Psychophysics: Method and Theory (John Wiley & Sons, New York).

Harris, J. D. (1952). "Remarks on the determination of a differential threshold by the so-called ABX technique," J. Acoust. Soc. Am. 24, 417.

van Heuvan, V. J. J. P., and van den Broeke, M. P. H. (1979). "Auditory discrimination of rise and decay times in tone and noise bursts," J. Acoust. Soc. Am. 66, 1308-1315.

Jusczyk, P. W., Rosner, B. S., Cutting, J. E., Foard, C. F., and Smith, L. B. (1977). "Categorical perception of non-speech sounds by two-month old infants," Percept. Psychophy. 21, 50-54.

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

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). "The discrimination of speech sounds within and across phoneme boundaries," J. Exp. Psych. 54, 358-368.

Lovell, J. D., and Carterette, E. C. (1972). "Digital generation of acoustic stimuli," Behav. Res. Meth. & Instru. 4, 151-155.

Macmillan, N. A., Kaplan, H. L., and Creelman, C. D. (1977). "The psychophysics of categorical perception," Psych. Rev. 84, 452-471.

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

Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. M., and Dooling, R. M. (1976). "Discrimination and labeling of noise-buzz sequences with varying noise lead times: An example of categorical perception," J. Acoust. Soc. Am., 60, 410-417.

Nabelek, I. (1965). "Discriminability of the rise time for noise and tone pulses," 5th Congress International D'Acoustique, Liege, B28.

Pisoni, D. B. (1971). "On the nature of categorical perception of speech sounds," Supplement to Status Report on Speech Research (SR-27), New Haven: Haskins Laboratories.

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

Pisoni, D. B. (1977). "Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stops," J. Acoust. Soc. Am. 61, 1352-1361.

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

Remez, R. E., Cutting, J. E., and Studdert-Kennedy, M. (1980). "Cross-series adaptation using song and string," Percept. Psychophy. 27, 524-530.

Rosen, S. M., and Howell, P. (1981). "Plucks and bows are not categorically perceived," Percept. Psychophy. 30, 156-168.

Sachs, R. M., and Grant, K. W. (1976). "Stimulus correlates in the perception of voice onset time (VOT): II. Discrimination of speech with high and low stimulus uncertainty," J. Acoust. Soc. Am. 60, S91(A).

Schouten, M. E. H. (1980). "The case against a speech mode of perception," Acta Psychologica 44, 71-98.

Stevens, K. N. (1982). "Constraints imposed by the auditory system on properties used to classify speech sounds: Data from phonology, acoustics, and psychoacoustics," in The Cognitive Representation of Speech, edited by Meyers, T. F., Laver, J. and Anderson, J. (North Holland, Amsterdam, in press).

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

Vigran, E., Graevenes, K., and Arnesen, G. (1964). "Two experiments concerning rise time and loudness," J. Acoust. Soc. Am. 36, 1468-1470.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants

Amanda C. Walley and Thomas D. Carrell

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

This research was supported, in part, by NIH Research Grant NS-12179, NIMH Research Grant MH-24027 and NINCDS Research Grant HD-11915 to Indiana University in Bloomington. An earlier version of these findings was presented at the meeting of the Acoustical Society in Los Angeles, November, 1980.

Stevens and Blumstein have proposed that the global shape of the CV syllable onset spectrum provides the listener with a primary and contextually invariant cue for place of stop consonant articulation. Contextually variable formant transitions are, in contrast, claimed to constitute secondary cues to place of articulation that, during development, are learned through their co-occurrence with the primary spectral ones. In the two experiments reported here, these claims about the relative importance of the onset spectrum and formant transition information were assessed by obtaining adults' and young children's identifications of synthetic stimuli in which these two potential cues specified different places of articulation. In general, the responses of both adults and children appeared to be determined by the formant transitions of the stimuli. These results provide little support for the claim that sensitivity to the global properties of the onset spectrum (as described by Stevens and Blumstein) underly place of articulation perception or for Stevens and Blumstein's primary vs. secondary cue distinction. Rather, these findings are consistent with the view that dynamic, time-varying information is important in the perception of place of articulation.

INTRODUCTION

A number of studies that have investigated the role of the release burst and formant transitions in signalling place of stop consonant articulation in CV syllables have indicated that these acoustic features vary as a function of the identity of the following vowel (e.g., Cooper, Delattre, Liberman, Borst and Gerstman, 1952; Dorman, Studdert-Kennedy and Raphael, 1977; Liberman, Delattre, Cooper and Gerstman, 1954; but see Cole and Scott, 1974). For example, the second formant transition, which has been thought by some investigators to be important in carrying information about place of articulation, falls following consonantal release in /du/, but rises in /di/, and the starting frequencies of the second-formant transition differ between the two syllables (Cooper et al., 1952; Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967). Indeed, the relative importance of burst and transition information in specifying place of stop consonant articulation may also vary in a context-dependent manner (e.g., Cooper et al., 1952; Dorman et al., 1977). The contextual variability of the burst and formant transitions associated with different places of articulation would seem to require the postulation of active perceptual mechanisms capable of using higher-level linguistic knowledge to interpret the speech waveform in a contextually appropriate manner (e.g., Liberman et al., 1967; Stevens and House, 1972).

However, the absence of acoustic-phonetic invariances observed in previous speech investigations may not be an inherent characteristic of the speech signal. Instead, this "lack of invariance" may be indicative of a failure to find an appropriate psychological description of the speech waveform. With respect to place of articulation, Fant (1960; 1973) and Stevens and Blumstein (1978; 1981; Blumstein and Stevens, 1979; 1980) have, for example, argued that the burst and transition information in the first 10-30 ms of the CV syllable, which are the result of one articulatory gesture, provide a single, integrated cue to consonantal identity that is independent of the following vowel context. As Stevens and Blumstein (1978) have pointed out, although bursts and formant transitions are visually distinctive in spectrographic displays, the auditory system does not necessarily process these features independently of one another. Rather, these acoustic segments might, in the early stages of auditory analysis, combine in such a way that they provide the basis for the constancy of the place of articulation percept.

Recent attempts to model the initial stages of speech processing, which employ more sophisticated methods of analysis than traditional spectrographic ones, have, in fact, met with some success in identifying invariant acoustic correlates of place of articulation (e.g., Blumstein and Stevens, 1979; Kewley-Port, 1982b; Searle, Jacobson and Kimberly, 1980; Searle, Jacobson and Rayment, 1979; Stevens and Blumstein, 1978). The work of Stevens and Blumstein has been particularly influential in promoting the notion that such correlates exist in the speech waveform and that these properties mediate the perception of place of articulation (see also Blumstein and Stevens, 1980; Stevens and Blumstein, 1981). They have proposed that the relative slope and diffuseness of energy in the short-term spectrum sampled at consonantal release in a CV syllable specify place of articulation in a context-independent manner: as shown in Figure 1, labials may be

---

Insert Figure 1 About Here

---

characterized by a diffuse-flat or -falling spectrum, alveolars by a diffuse-rising spectrum, and velars by a prominent mid-frequency spectral peak (cf. Jakobson, Fant and Halle, 1952). Within Stevens and Blumstein's model, the onset spectrum of a stop CV syllable is obtained by integrating energy over the first 25.6 ms of the syllable and using linear prediction analysis. The criterial properties of the onset spectrum then are determined by the stop release burst and the initial portions of the first three formant transitions; the same spectral shapes are observed when only the formant transitions are present in a stimulus, but these shapes are enhanced by the presence of the burst (e.g., Stevens and Blumstein, 1978). Thus, the global, context-independent properties associated with different places of articulation do not depend on fine details of the initial part of the waveform.

According to Stevens and Blumstein's theory, sensitivity to the global properties of the onset spectrum provides the primary basis for the adult listener's ability to differentially identify place of articulation for stop CV syllables. In contrast, formant transition information (e.g., starting frequency and formant trajectories) constitutes a secondary, context-dependent cue to place of articulation, which may be invoked in the absence or distortion of the primary, invariant spectral cues. Thus, adults are able to identify place of articulation for two-formant stimuli (Cooper et al., 1952; Delattre, Liberman and Cooper, 1955), even though these stimuli lack the primary spectral properties (e.g., Stevens and Blumstein, 1978). However, formant transitions are not essential for accurate place categorization; their main "function" is instead to provide a smooth and continuous change between the onset spectrum and the following vowel.

The ability to use secondary, context-dependent formant transitions in place of articulation perception is one that is, Stevens and Blumstein suggest, acquired in development by virtue of their co-occurrence with the primary spectral cues. The prelinguistic infant's discrimination of place differences (e.g., Bush and Williams, 1977; Eimas, 1974; Moffitt, 1971; Morse, 1972) is explained by assuming that the auditory system is innately endowed with feature detectors which are sensitive to the invariant spectral properties associated with different places of articulation.[1] Thus, the properties of the onset spectrum are asserted to be primary for the perception of place of articulation in that: 1) for the adult, they constitute the major and most reliable basis for the perception of place and 2) developmentally, they are used prior to formant transition information and sensitivity to these properties is innate.

In order to provide empirical support for their claim that it is sensitivity to the global properties of a stimulus' onset spectrum which mediates the perception of place of articulation, Stevens and Blumstein (1978) conducted a study in which adult listeners were required to identify synthetic CV syllables from several continua. The stimuli within a continuum varied in place of

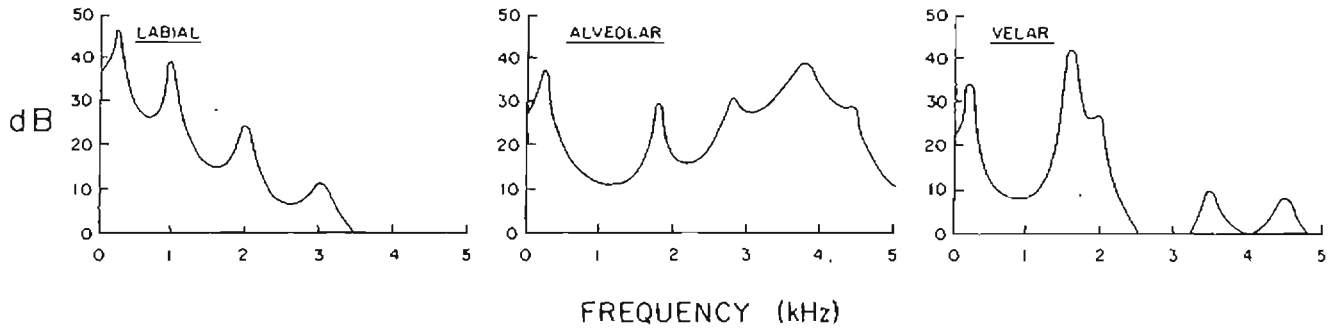


Figure 1. Theoretical onset spectra of labial, alveolar and velar stop consonants (after Stevens and Blumstein, 1978).

articulation (/b-d-g/ or /b-g/) in one of three vowel contexts (/i/, /a/ or /u/) and contained either bursts + transitions, transitions only, or bursts only. It was found that the best exemplars (according to subjects' identifications) of each place category for the burst + transition and the transition-only continua had onset spectra which possessed the proposed primary, context-independent cues. These spectral shapes were not observed for stimuli that were inconsistently identified -- i.e., for category boundary stimuli. (Although the burst-only stimuli possessed distinctive onset spectra, they contained an additional and conflicting discontinuity and, according to Stevens and Blumstein, were therefore inconsistently identified.)

Stevens and Blumstein interpreted their results as supporting the notion that identification of place of articulation is achieved through the operation of auditory feature detectors which are sensitive to the gross, invariant properties of the onset spectrum. Yet, whether or not sensitivity to these spectral properties is primarily responsible for mediating perception of place remains unresolved since, in their burst + transition and transition-only stimuli, these properties were confounded (as they undoubtedly are in natural speech) with other acoustic properties, such as burst frequency and formant transitions, that may also provide the listener with information about place of articulation. So, for example, the spectral properties that Stevens and Blumstein consider to be distinctive for each place category may have co-occurred with what is perceptually optimal formant transition information for the same place categories, and nondistinctive onset spectra may have been confounded with ambiguous formant transitions. On the basis of Stevens and Blumstein's results, the relative importance of a stimulus' onset spectrum, burst frequency, formant transitions and VOT in specifying its place of articulation is not clear. Their results merely serve to illustrate that those properties of the onset spectrum which they claim underly the perception of place of articulation are associated with different places categories and do not require the conclusion that these properties are necessary or even sufficient cues for the specification of place of articulation.

In a more recent series of experiments, Blumstein and Stevens (1980) presented listeners with truncated versions of synthetic burst + transition and transition-only CV syllables containing either moving or "straight" transitions. Subjects were able to reliably identify the place of articulation of these brief stimuli, when their onset frequencies corresponded to those of the best exemplars of each place category (according to the identification functions obtained in their previous study; Stevens and Blumstein, 1978). From these results Stevens and Blumstein concluded that sufficient information is contained in the first 10-20 ms of the CV waveform to cue place of stop consonant articulation (although the effective durations of their truncated stimuli were longer than they suggest; see Kewley-Port, 1980). Moreover, because subjects were able to identify stimuli with straight transitions, Stevens and Blumstein also argued that formant motions are not essential for place of articulation perception. According to Stevens and Blumstein, these results, are consistent with the notion that it is the global properties of the initial portion of the waveform which are most important in the perception of place of articulation.

It is not clear, however, that Stevens and Blumstein's conclusions are warranted, since, for example, formant starting frequencies and onset spectra were still confounded in their straight transition stimuli. A more general problem with this study, as well as their previous one (Stevens and Blumstein, 1978), is that the strategy for investigating place of articulation perception has been to vary burst, VOT and formant transition information and to assess the effects of these manipulations on perception, rather than to directly manipulate properties of the onset spectrum. Such an approach can provide only indirect support for the claim that place perception is primarily mediated by the detection of the relative slope and diffuseness of spectral energy at stimulus onset. A strong test of their theory "...would be to determine whether perception of place of articulation depends on attributes of the gross shape of the spectrum at onset, independent of fine details such as burst characteristics and formant onset frequencies" (Stevens and Blumstein, 1978, p. 1367).

In the present study, we adopted such an approach in evaluating two of the major claims of Stevens and Blumstein's theory -- i.e., that global properties of the CV syllable onset spectrum are primary in the adult's and the developing child's perception of place of articulation. These claims were assessed by obtaining adults' and young children's identifications (in Experiment 1 and Experiment 2, respectively) of synthetic CV syllables in which formant transition information specified one place of articulation, but in which the onset spectrum specified a different (or conflicting) place of articulation. This manipulation was, following the suggestion of Stevens and Blumstein (1978), achieved by varying the relative amplitudes of the formants at stimulus onset. By examining how listeners identify such stimuli, we hoped to determine the relative contribution of spectral information at stimulus onset and transition information to the perception of place of articulation.

1. EXPERIMENT 1: ADULT PERCEPTION

If Stevens and Blumstein's claims concerning the perceptual primacy of the onset spectrum are correct, one might expect that stimuli in which onset spectra and formant transitions conflict would be identified by adults in a manner consistent with the properties of the onset spectra. For example, a stimulus which has an onset spectrum appropriate for a /d/ (i.e., a diffuse-rising shape), but which has the rising formant transitions characteristic of /b/ in the context of /a/, should be identified as /da/. However, if information residing in a stimulus' formant transitions is relatively more important or salient to listeners, this stimulus should be identified as a /ba/. Of course, it is also possible that both types of cues are important for the perception of place of articulation or that, through the amplitude manipulation, other important characteristics of the stimuli will be distorted and the stimuli will simply be rendered ambiguous. In this case, adults might not be able to identify the stimuli consistently.

A. Subjects

Thirty-six Indiana University students, who had little or no experience in listening to synthetic speech, served as subjects in the experiment. Three additional subjects were excluded from the data analysis.[2] Subjects were either paid for their participation or received credit towards an introductory psychology course in which they were enrolled. Each subject was a native speaker of English and reported no history of speech or hearing disorder at the time of testing.

B. Stimuli

Nine CV syllables (three control and six experimental or "conflicting cue" stimuli) were constructed for each of the two vowel conditions (/a/ and /u/) using a modified version of the Klatt (1980) software digital speech synthesizer (see Kewley-Port, 1978). Each stimulus was synthesized with the digital resonators connected in parallel. In contrast to synthesis in the cascade branch of the synthesizer, where formant amplitudes are calculated automatically, synthesis in the parallel mode allows the experimenter to explicitly set and therefore have some control over the relative amplitudes of the formants. This synthesis strategy was, of course, necessary for construction of the conflicting cue stimuli. The same strategy was employed in the construction of the control stimuli. (An earlier pilot study indicated that identification accuracy for the control stimuli did not depend on whether they had been synthesized in the cascade or in the parallel mode.) All the stimuli were synthesized at a 10 KHz sampling rate, output through a 12-bit D-A converter and low-pass filtered at 4.8 KHz.

1. Control Stimuli

The control stimuli were modelled after the best exemplars of each place of articulation category from Stevens and Blumstein's (1978) transition-only /ba-da-ga/ and /bu-du-gu/ continua - i.e., after stimulus 1, 8 and 12, and stimulus 1, 7 and 13, respectively.[3] Thus, for one vowel condition of the present experiment, the center frequencies (and bandwidths) of the formants of the control stimuli were appropriate for the steady-state vowel /a/ and were set at 720 (50), 1240 (70), 2500 (110), 3600 (170) and 4500 (250) Hz for F1, F2, F3, F4 and F5, respectively. Each control stimulus in this set had the same starting frequency of 220 Hz for F1, but F1 had varying transition durations of 20, 35 and 45 ms for /ba/, /da/ and /ga/, respectively. The starting frequencies of F2 and F3 were 900 Hz and 2000 Hz for /b/, 1700 Hz and 2800 Hz for /da/, 1640 Hz and 2100 Hz for /ga/ and the transition durations for F2 and F3 were 40 ms. The trajectory of each formant from its starting frequency to the steady-state value for the vowel was linear. F4 and F5 were steady-state formants and thus had no formant transitions. The three control stimuli, /ba/, /da/ and /ga/, will be referred to as stimulus number 1a, 2a and 3a, respectively.

For the other vowel condition /u/, the formant frequencies were set at 370-300 Hz, 1100-1000 Hz, 2350 Hz, 3200 Hz and 4500 Hz, for F1, F2, F3, F4 and F5, respectively. (The first two formants changed over the frequency range

indicated to simulate diphthongization.) The bandwidths of the formants corresponded to those values used in the /a/ condition. The starting frequency (and duration) of the F1 transition was 180 Hz (15 ms) for all three control stimuli. The starting frequency of F2 and F3 was 800 and 2000 Hz for /bu/, 1600 and 2700 Hz for /du/, and 1400 and 2000 Hz for /gu/. The duration of these transitions was 40 ms for all stimuli. F4 and F5 were steady-state formants and did not have any transitions. The three control stimuli, /bu/, /du/ and /gu/, will be referred to as stimulus 4a, 5a and 6a, respectively.

The duration of voicing for each stimulus in both vowel conditions (and thus total stimulus duration) was 255 ms. The excitation source for the vowel began abruptly, such that the first glottal pulse coincided with the beginning of the formant transitions. The amplitude of voicing was constant for 220 ms and fell linearly to 0 dB over the last 35 ms of a stimulus. The fundamental-frequency contour began at 103 Hz, rose linearly to 125 Hz in 35 ms, fell first to 94 Hz in 180 ms and then to 50 Hz in 40 ms.

The onset spectra of the control stimuli were obtained using linear prediction analysis similar to that of Stevens and Blumstein (1979). A general purpose digital signal processing program, SPECTRUM (Kewley-Port, 1979), preemphasised, windowed and analyzed each stimulus waveform with 14 linear prediction coefficients using the autocorrelation method. The 25.6 ms window, which was positioned at stimulus onset, was an extended half-Hamming window; the first 12.8 ms of the window was rectangular, the following 12.8 ms was a half-Hamming window. This window differed slightly from the extended half-Kaiser window employed by Stevens and Blumstein.

The shape of the onset spectrum of each control stimulus (see the top of Figure 2 for the control stimuli in the /a/ condition, the top of Figure 3 for those in the /u/ condition) was consistent

---

Insert Figure 2 and Figure 3 About Here

---

with the stimulus' place of articulation as specified by its formant transitions. Each onset spectrum was accepted by the appropriate Stevens and Blumstein template and rejected by the other two templates (for a detailed description of these templates and matching criteria, see Blumstein and Stevens, 1979). Template fitting was achieved for purposes of the present experiment by obtaining hard copies from a Tektronix 4010-1 of the onset spectra and then comparing them visually with transparencies of Stevens and Blumstein's templates. In order to achieve the desired fits, the relative amplitudes of the formants were set to those values specified in Table I. A value of 66 dB was chosen as a baseline

---

Insert Table I About Here

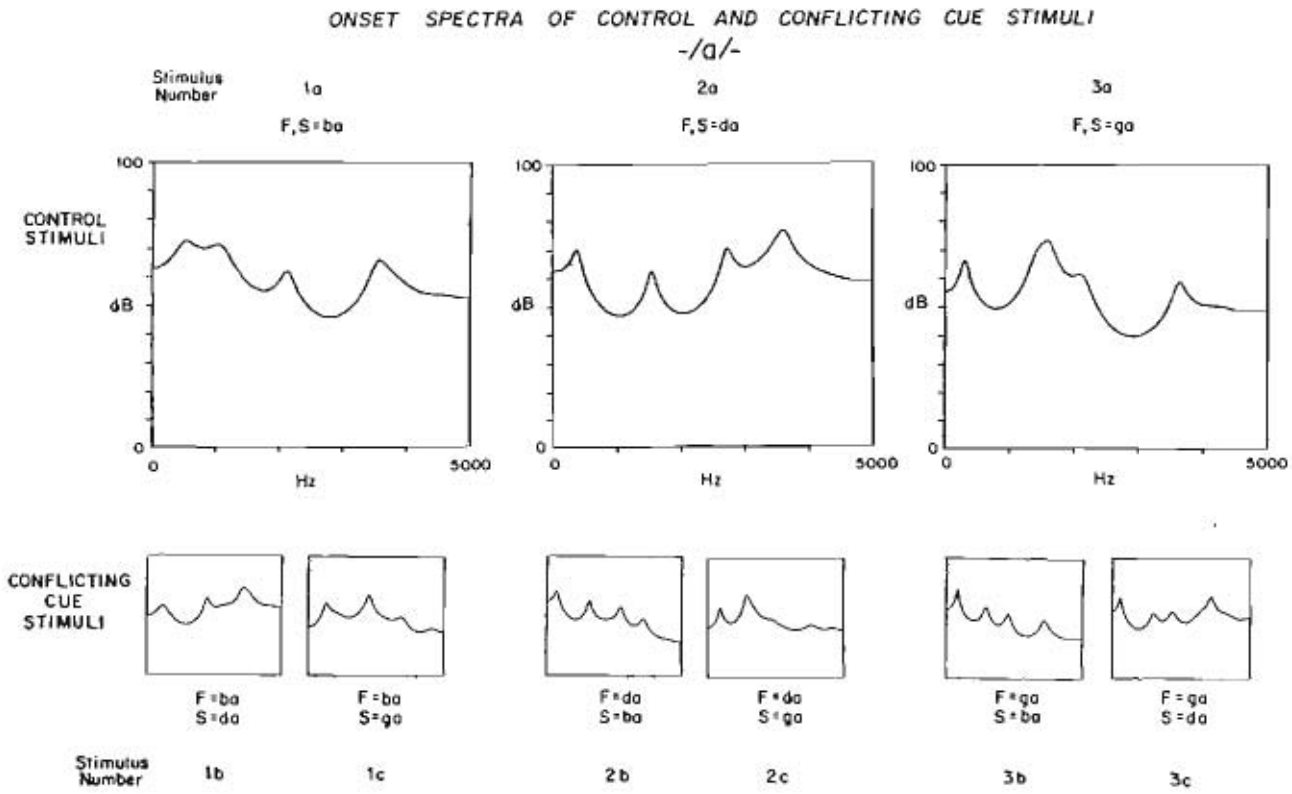


Figure 2. Onset spectra of the control and conflicting cue stimuli in the /a/ condition. Note that the amplitude scale in this figure differs from that in Figure 1.

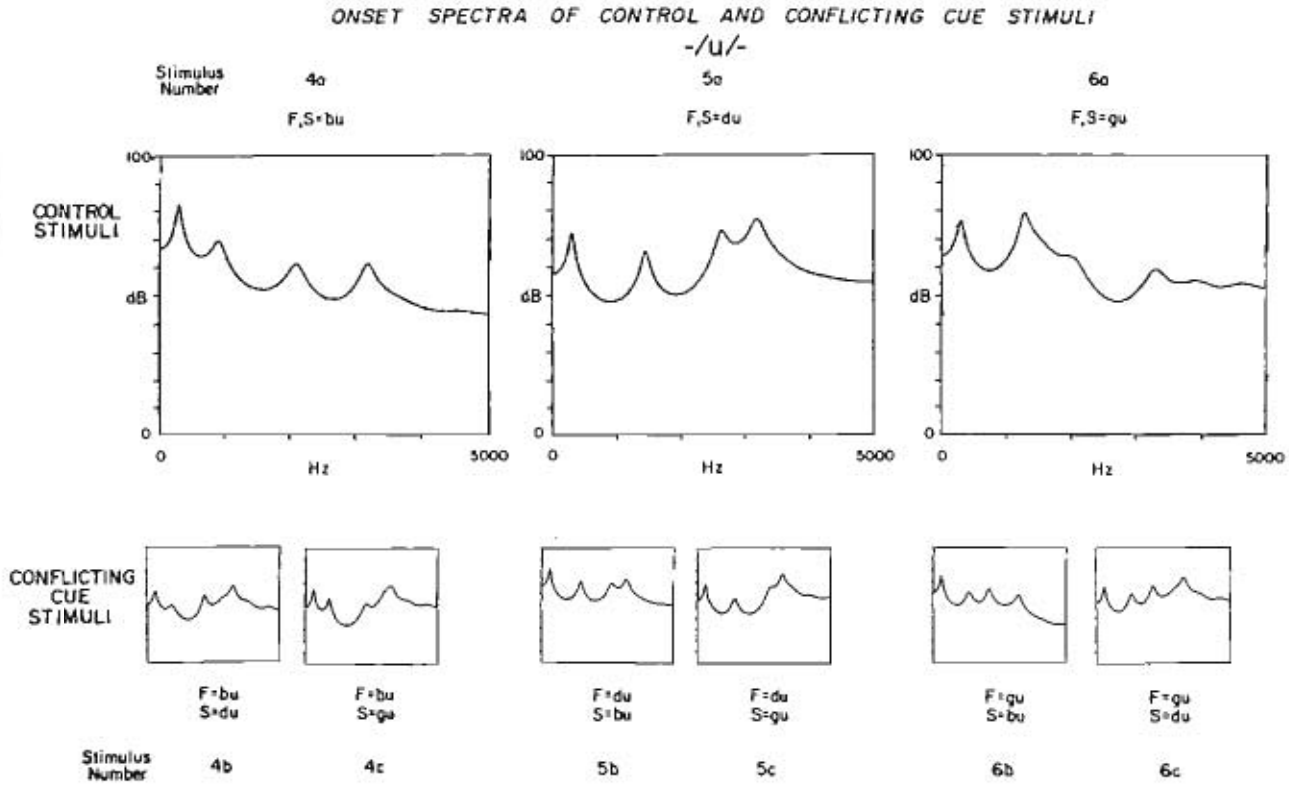


Figure 3. Onset spectra of the control and conflicting cue stimuli in the /u/ condition. Note that the amplitude scale in this figure differs from that in Figure 1.

Table 1. Relative formant amplitudes (dB) of control and conflicting cue stimuli in the /a/ and /u/ conditions.

	Formant				
Stimulus No.	F1	F2	F3	F4	F5
/a/					
1a	66	66	66	66	66
1b	50	41	66	72	66
1c	51	45	69	45	33
2a	66	60	66	74	66
2b	66	60	54	45	41
2c	54	66	36	36	36
3a	61	74	61	57	66
3b	66	57	54	47	47
3c	61	54	57	66	63
/u/					
4a	66	66	69	66	66
4b	46	46	62	69	66
4c	47	50	54	72	69
5a	57	57	61	66	66
5b	66	63	59	61	66
5c	54	50	52	66	66
6a	60	72	60	51	54
6b	60	56	63	56	66
6c	50	54	64	69	66

from which to vary the relative amplitudes of the formants. In cases where the amplitude of a given formant relative to other formant amplitudes was altered, the amplitude of that formant was constant for the first 25 ms of the stimulus (i.e., for about the duration of the analysis window) and then rose (or fell) linearly back to the baseline amplitude over the next 25 ms of the stimulus.

2. Conflicting Cue Stimuli

Two experimental or conflicting cue stimuli were derived from each control stimulus. Thus, a total of 12 conflicting cue stimuli (6 for each vowel condition) were constructed. A conflicting cue stimulus differed from the original control stimulus only in that its formant amplitudes were manipulated (see preceding section) such that its onset spectrum specified a place of articulation different from (and thus in conflict with) that specified by its formant transitions. (The formant amplitudes employed to achieve this manipulation are shown in Table 1; the onset spectra of the conflicting cue stimuli are shown in Figures 2 and 3 beneath the control stimuli from which they were derived.) As an example, the conflicting cue stimulus 1b has the same spectro-temporal specifications as control stimulus 1a and is, therefore, by this particular description a /ba/. However, the onset spectrum of stimulus 1b is appropriate for a /da/. Each conflicting cue stimulus had an onset spectrum which was accepted by one of the Stevens and Blumstein templates, but which was rejected by the other two templates -- including the one which was actually appropriate according to the stimulus' formant transitions. Thus, stimulus 1b was, for example, rejected by the labial template.

C. Procedure

Half of the subjects were assigned to the /a/ condition, the other half to the /u/ condition. The subjects were told that they would hear computer-generated versions of the syllables "ba", "da" and "ga" (or "bu", "du" and "gu") and that they were to identify each stimulus as quickly as possible by pressing one of three labelled response buttons. A cue-light signalled the presentation of a stimulus.

The subjects first listened to a block of stimuli, in which each of the three control stimuli for that condition were presented ten times and each of the six conflicting cue stimuli were presented five times in random order (such that subjects heard an equal proportion of control and conflicting cue stimuli). This block of trials served to familiarize the subjects with the stimuli and to allow them to practice responding. These data were not analyzed. In the testing phase of the experiment, subjects were presented with 24 repetitions of each of the six conflicting cue stimuli and 48 repetitions of each of the three control stimuli in random order. The stimuli were presented over matched and calibrated TDH-39 headphones at approximately 80 dB SPL with a maximum inter-trial interval of three seconds. Stimulus presentation and response collection were conducted on-line and controlled by a PDP-11 computer. The entire experimental session lasted approximately 45 minutes.

D. Results

1. Group Analysis

The subjects tested in the /a/ condition failed to respond on only 0.3% of the trials (for both control and conflicting cue trials). In the /u/ condition, subjects failed to respond for 0.4% of the trials (for both control and conflicting cue trials).

Shown on the leftmost side of each panel in Figures 4 and 5

Insert Figure 4 and Figure 5 About Here

---

are the group mean proportions of correct and incorrect identifications for each control stimulus in the /a/ and /u/ vowel condition, respectively. As one can see, the mean proportion of hits for each control stimulus is very high. These identification responses could be based on information provided by a given stimulus' formant transitions or on the gross properties of that stimulus' onset spectrum. Correct responses for control stimuli are therefore labelled F,S in the figures, incorrect responses are labelled O for "other" (i.e., the control stimulus was labelled according to one of the other two possible place categories).

The pattern of identification responses observed for the conflicting cue stimuli in the /a/ condition is represented in Figure 4. The mean proportion of identifications by formant transitions (which is denoted by F in the figure), by onset spectrum (S), and the mean proportion of "other" (O) responses (i.e, according to the third possible place category) for each conflicting cue stimulus are shown to the right of the control stimulus from which it was derived. Subjects categorized these conflicting cue stimuli by formant transitions reliably more often than expected by chance for five of the six conflicting cue stimuli (two tailed, $t_{(17)}$ = 2.10, 2.57, 109.00, 8.23, 7.84, for stimulus 1b, 1c, 2b, 3b, and 3c, respectively; $p < .05$ in all cases). Categorization by onset spectrum was not reliably greater than chance for any of these five stimuli. Indeed, categorization on this basis was reliably less than chance in most cases ($t_{(17)}$ = -4.36, -165.5, -3.71, -5.09, for stimulus 1b, 2b, 3b, 3c; $p < .001$). The remaining conflicting cue stimulus (2c) was reliably identified by its onset spectrum ($t_{(17)}$ = 7.30; $p < .001$) and identification of this stimulus on the basis of its formant transitions was significantly less than would be expected by chance ($t_{(17)}$ = -4.86; $p < .001$). The significance of these tests is indicated by an asterisk (*) above the appropriate column (F or S) in Figure 4.

In the /u/ condition (see Figure 5), all six of the conflicting cue stimuli were reliably identified according to their formant transitions. (For stimulus 4b, 4c, 5b, 5c, 6b and 6c, $t_{(17)}$ = 2.14, 2.38, 13.35, 18.27, 7.97, 7.51, respectively; $p < .05$ in all cases.) Identifications by onset spectrum were reliably below chance for stimulus 4b, 5b, 5c, 6b and 6c ($t_{(17)}$ = -5.59, -6.13,

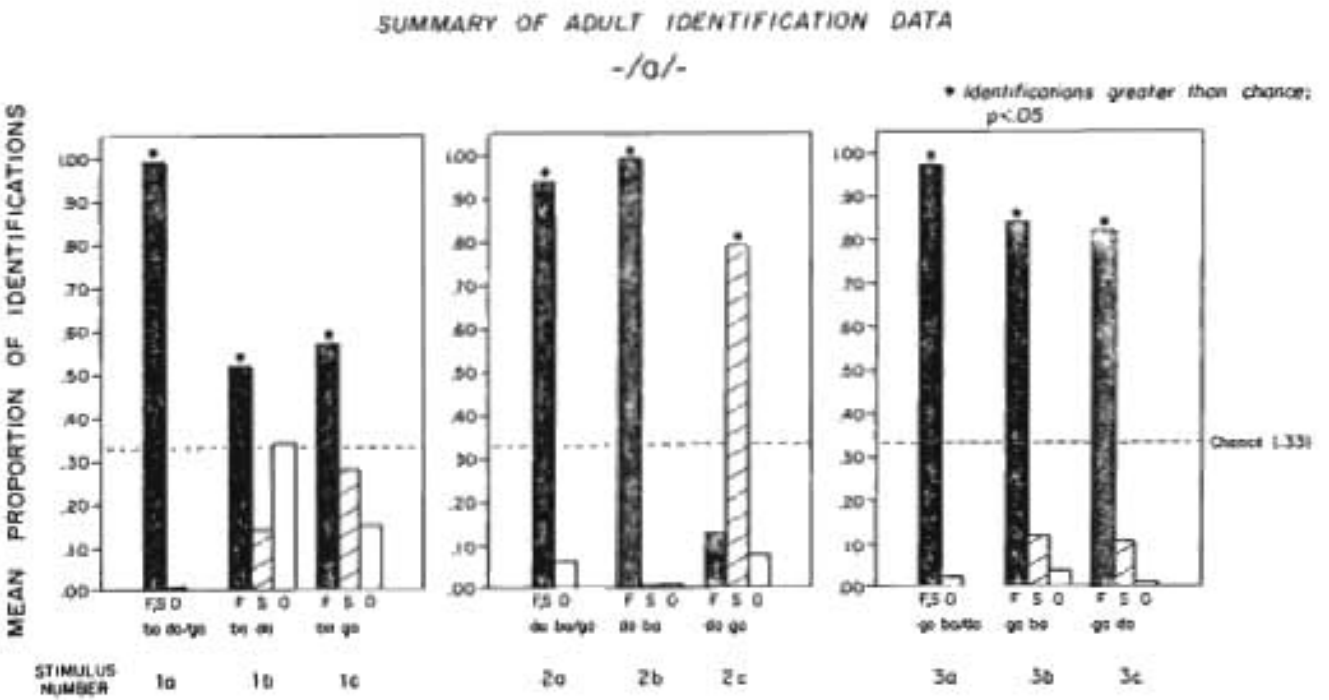


Figure 4. The grouped adult identification data obtained for the /a/ condition. F refers to identifications according to formant transitions, S to those according to onset spectrum and O to those according to the other possible place response(s) for a given stimulus. F,S refers to the same response for a given control stimulus (1a, 2a or 3a).

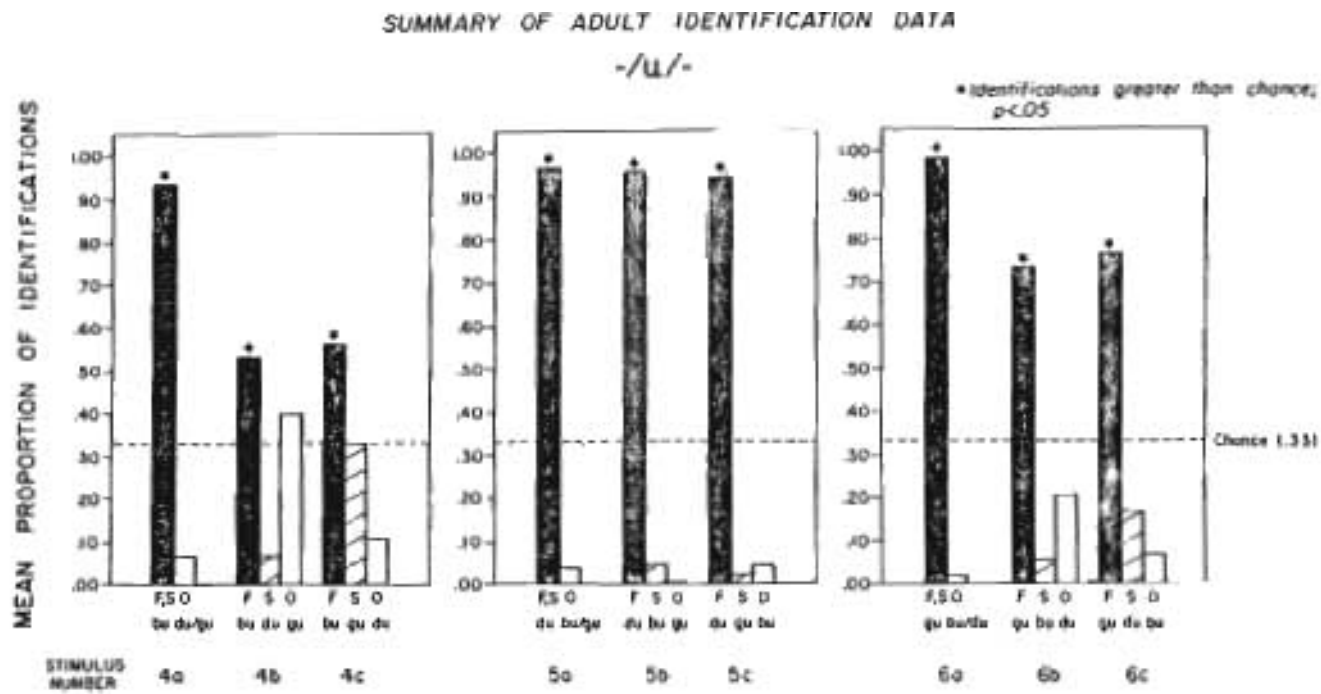


Figure 5. The grouped adult identification data obtained for the /u/ condition. F refers to identifications according to formant transitions, S to those according to onset spectrum and O to those according to the other possible place response(s) for a given stimulus. F,S refers to the same response for a given control stimulus (4a, 5a or 6a).

-39.63, -72.71, -3.74, respectively; $p < .002$ in all cases) and at chance for stimulus 4c.

By this analysis of the group data, subjects appeared to rely primarily on formant transition information for identifying place of articulation in the conflicting cue stimuli. Nevertheless, in some instances, the overall mean proportion of identifications was considerably less than for the control stimuli -- a result which suggests that subjects experienced difficulty in making their identifications. This difficulty could be explained in at least two ways. Assuming, for example, that both formant transition information and the shape of the onset spectrum normally (e.g., for the control stimuli) provide independent cues for place of articulation and that both cues are adequately specified within the experimental stimuli, it could be that the two cues conflict with one another in the experimental stimuli and that it is therefore difficult for subjects to selectively attend to either cue. Alternatively, it could be the case that the stimuli are simply rendered ambiguous with the amplitude manipulation. Even though the onset spectrum and formant transitions are still specified in the stimuli, perhaps neither constitutes a good description of the acoustic information that subjects actually use to identify place of articulation and we inadvertently altered some other critical information. These issues were pursued by examining the performance of individual subjects.

2. Individual Subject Analysis

In the analysis of the data of individual subjects, an identification "rule" was inferred from the pattern of a subject's identifications for each of the nine stimuli in the two vowel conditions. A rule was defined as the identification of a stimulus according to a particular place category on 60% or more of the trials for that stimulus. (The probability of doing so by chance is less than .01 by a Chi-square test.) At the left of each of the three panels in Figure 6 is shown the number of

---

Insert Figure 6 About Here

---

subjects that correctly identified each control stimulus in the /a/ condition by this criterion. Here the rule is referred to as F,S, since subjects could have used either formant transitions, the onset spectrum, or both to identify a stimulus. To the right of the control stimulus in each panel is shown the number of subjects that identified the two conflicting cue stimuli (derived from that control stimulus) according to a formant transition rule (F), an onset spectrum rule (S), or with the "other" third response (O). Also shown for each control and conflicting cue stimulus in the figure is the number of subjects who did not respond according to one of these rules, but instead divided their responses among the three place categories and thus responded inconsistently (I). Individual subjects' rule use in the /u/ condition is represented in Figure 7.

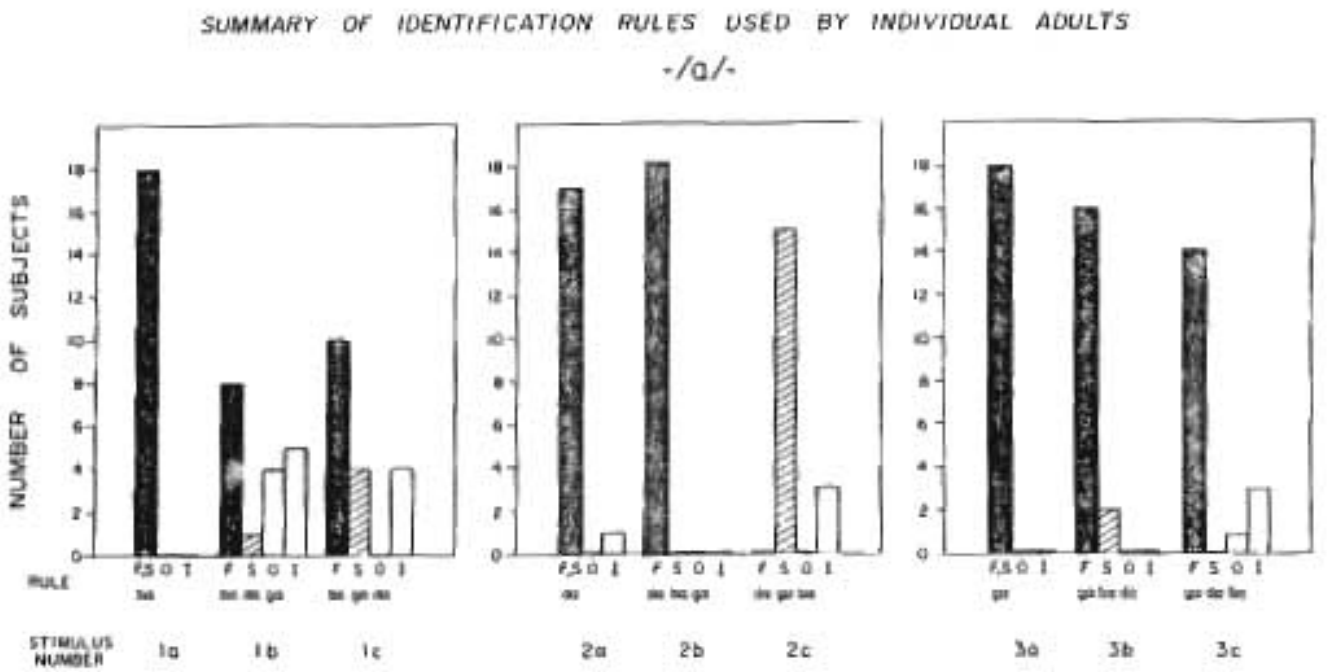


Figure 6. Identification rule choice of individual adult subjects tested in the /a/ condition. I refers to inconsistent responding for a given stimulus.

Insert Figure 7 About Here

From Figures 6 and 7 it can be seen that although placing formant transitions and onset spectra in conflict with one another did increase inconsistencies in the responses of some subjects, the majority of subjects were very consistent in using either a formant- or an onset spectrum-based rule -- even for stimulus 1b, 1c, 4b and 4c. This suggests that the poorer identification seen in the group data for these stimuli was not due to the ambiguous nature of the stimuli, but rather may be attributed to individual differences in the type of cue used to identify place of articulation. For example, 10 subjects consistently identified stimulus 4c (in the /u/ condition) by a formant transition rule, whereas 5 subjects consistently identified this same stimulus according to its onset spectrum (i.e., as /gu/). This sort of finding would appear to indicate that both cues may be used by adults and that neither is primary in the sense that, in the face of a conflict between the two cues, one alone is always selected as the basis for stimulus identification. Nevertheless, it is quite obvious from this analysis, as well as the group one, that formant-based responses to the conflicting cue stimuli predominated.

E. Discussion

From the results of Experiment 1, it is apparent that subjects were able to identify the conflicting cue stimuli either on the basis of formant transition information or the shape of their onset spectra (a finding which was both subject- and stimulus-dependent). To this extent, the results are consistent with Stevens and Blumstein's contention that the two cues co-exist in the stop CV waveform and support the perception of place of articulation. However, neither cue is sufficient in itself to account for the identification data; i.e., when the "primary" and "secondary" cues conflicted, subjects did not rely exclusively on one of these cues to identify all the stimuli. This fact lends little support to the utility of a primary vs. secondary cue distinction. Moreover, the results indicate that formant transition information may contribute in an important way to the specification of place of articulation. Despite the fact that the "primary" properties of the onset spectrum were present in a given conflicting cue stimulus, listeners' identifications of the stimulus generally agreed with its place of articulation as specified by the "secondary" formant transitions. This finding argues strongly against Stevens and Blumstein's claim that the global properties of the onset spectrum provide the major basis for the perception of place of articulation in syllable-initial stop consonants.

EXPERIMENT 2: CHILDREN'S PERCEPTION

Stevens and Blumstein (1978; 1981; Blumstein and Stevens, 1979; 1980) have proposed that innate sensitivity to the relative slope and diffuseness of spectral energy at stimulus onset allows the infant to derive the appropriate

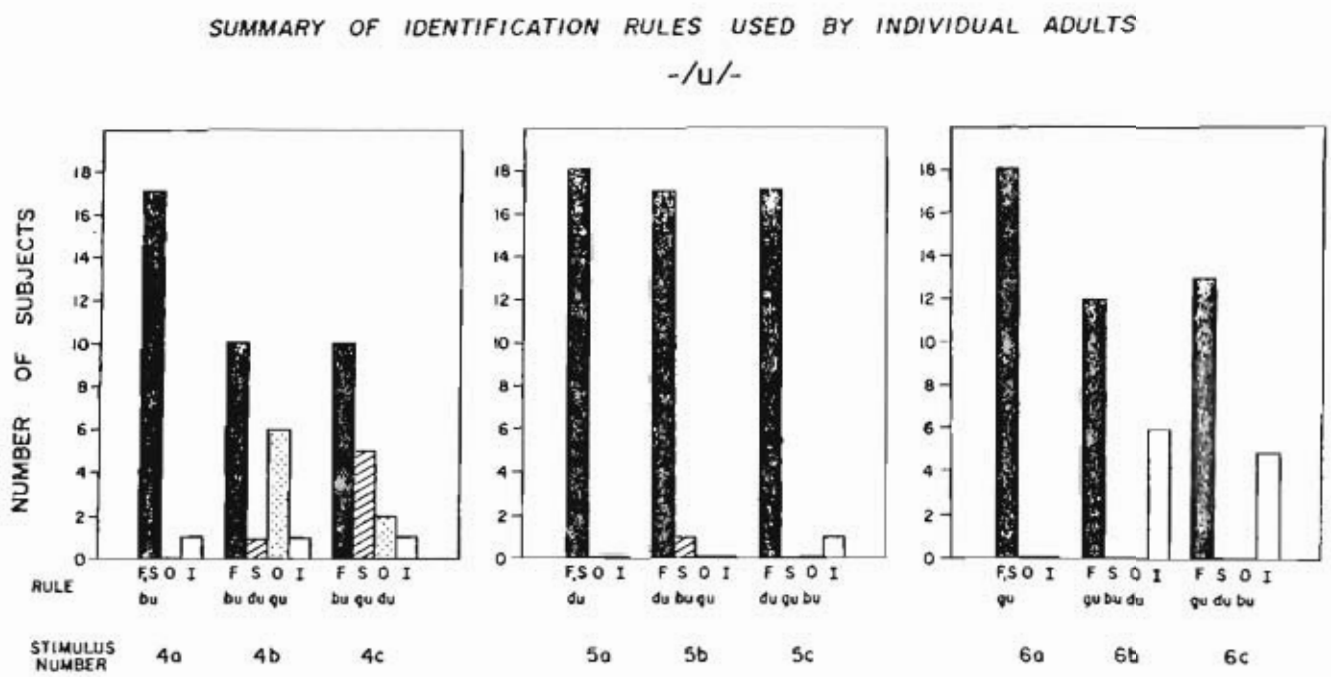


Figure 7. Identification rule choice of individual adult subjects tested in the /u/ condition.

place categories from the stop CV waveform. In contrast, the ability to use formant transition information for stop consonant identification is, according to their account, one which is acquired in development -- through a process similar to incidental learning; because secondary, context-dependent formant transitions normally co-occur with the primary, invariant properties of the onset spectrum, the developing child eventually comes to be able to use the secondary cue (e.g., in the absence or distortion of the primary one).

There has to date been little empirical research directed at evaluating Stevens and Blumstein's developmental claims (but see Aslin and Walley, 1980).[4] In Experiment 2, we attempted to do so by studying how young children (of about 5 years of age) identify stop consonant syllables in which the properties of the onset spectrum and formant transition information conflict with each other. The results of Experiment 1 indicated that formant transitions do not assume merely secondary importance relative to the onset spectrum in adult perception. Yet, if formant transition cues are learned only incidently, it might be expected that the ability to use them would not be fully developed in young children. Therefore, young children might experience greater difficulty in using this information when it conflicts with the properties of the onset spectrum (as in the conflicting cue stimuli of Experiment 1) than when it does not conflict (as in the control stimuli of Experiment 1).

A. Subjects

Ten children (mean age = 5 years, 1 month; range = 4,1 - 6,0) served as subjects in the experiment. Six additional children (mean age = 4,5; range = 4,0 - 5,5) participated in the experiment, but their data were excluded from the analysis.[5] Subjects were obtained by placing an advertisement in a local newspaper and were paid for their participation. None of the subjects, according to their parents' report, had any history of hearing or speech disorder.

B. Stimuli

The stimuli used in this experiment were the nine (three control and six conflicting cue) stimuli of the /a/ condition in Experiment 1.

C. Procedure

Each subject was tested individually in two sessions on separate days (with not more than one day intervening between sessions). A subject was seated in front of a response box which had three "labelled" response buttons. The labels for the response buttons were three visually distinctive, cartoon faces. The subject was told that he/she would later listen to a computer making the sounds "ba", "da" and "ga" and that he/she was to press the nose of the person that made the sounds as they were presented.

1. Training

In the training phase of Session 1, the subject was first taught to associate the syllables "ba", "da" and "ga" with a certain face and response button. The experimenter (ACW), who was seated beside the subject, explained which sound each of the three people made and then produced orally five tokens of each of the three syllables in a predetermined random order. The subject was asked to press the nose of the person who made each sound. Responses were recorded manually by the experimenter, who provided the subject with feedback on each trial. If a child responded correctly on 14 of the last 15 of these oral training trials (a maximum of 45 trials were given), he/she advanced to the second phase of training. Otherwise, the subject was considered to have failed training and the session was terminated. On the second day of testing, this part of training was eliminated; the child was simply reminded which sound each of the three people made and was "retrained" with 15 presentations of the synthetic control stimuli (see below).

In the second phase of training, the subject was presented over headphones with five tokens of the three control stimuli (i.e., stimulus 1a, 2a and 3a) in a predetermined random order. Again the child was asked to indicate which person had made a sound by pressing one of the three response buttons. The experimenter, who also listened to the stimuli over headphones, observed and recorded the subject's responses manually and provided feedback. The same criterion as that used in the first phase of training was adopted to determine whether or not the subject advanced to the testing phase of the experiment.

2. Testing

In the testing phase of each session, the subject heard six presentations of each of the nine (three control, six conflicting cue) stimuli in random order. Thus, in the two testing sessions of the experiment, the subject heard each stimulus twelve times. The child was told that he/she should continue to indicate what sound was heard by pressing one of the three response buttons and that a light would come on to indicate when a sound was about to occur.

The inter-trial interval was controlled by the experimenter and was therefore variable (although the maximum interval was 30 seconds). This arrangement allowed the child to make comments and to ask questions during testing and permitted the experimenter to encourage the child to complete the task in an attentive manner. All other aspects of stimulus presentation and response collection were controlled on-line by a PDP-11 computer. The synthetic stimuli were presented over matched and calibrated TDH-39 headphones at approximately 80 dB (SPL). Each testing session lasted approximately 20-30 minutes.

D. Results

1. Group Analysis

As was the case for the adults tested in Experiment 1, the children tested in the present experiment failed to respond on only a very small number of trials (0.5% of the conflicting cue trials).

The mean proportion of correct identifications by children for each of the control stimuli in Experiment 2 was very high. These proportions are shown, together with the mean proportion of incorrect responses for a given control stimulus, to the left of each panel in Figure 8. The mean proportion of identifications

---

Insert Figure 8 About Here

---

according to formant transitions (F) and to onset spectrum (S) and the mean proportion of "other" responses (O) for each conflicting cue stimulus are displayed to the right of the control stimulus from which it was derived. Children identified five of the six conflicting cue stimuli according to their formant transitions reliably more often than would be expected by chance (two-tailed, $t_{(9)}$ = 10.88, 4.05, 172.94, 353.89, 147.37, for stimulus 1b, 1c, 2b, 3b and 3c, respectively; $p < .01$ in all cases). Categorization by onset spectrum was not, therefore, greater than chance for any of these stimuli; for four of these five stimuli, identifications by onset spectrum were significantly below chance ($t_{(9)}$ = -3.62, -765.00, -391.25, -57.89, for stimulus 1b, 2b, 3b, 3c; $p < .01$). Conflicting cue stimulus 2c was reliably categorized according to its onset spectrum ($t_{(9)}$ = 30.00; $p < .001$) and was categorized by its formant transitions at a level below chance ($t_{(9)}$ = -24.18; $p < .001$). From this analysis, it appears that children, like the adults in Experiment 1, relied primarily on formant transition information in order to identify the conflicting cue stimuli.

2. Individual Subject Data

As in Experiment 1, an identification rule was inferred from the pattern of each subject's identification responses for each of the nine stimuli. A rule was defined as the identification of a stimulus according to a particular place category on 67% or more of the trials for that stimulus. (The probability of doing so by chance is less than .02 by a Chi-square test.[6]) The number of subjects that correctly identified a given control stimulus is shown to the left of each of the three panels in Figure 9. To the

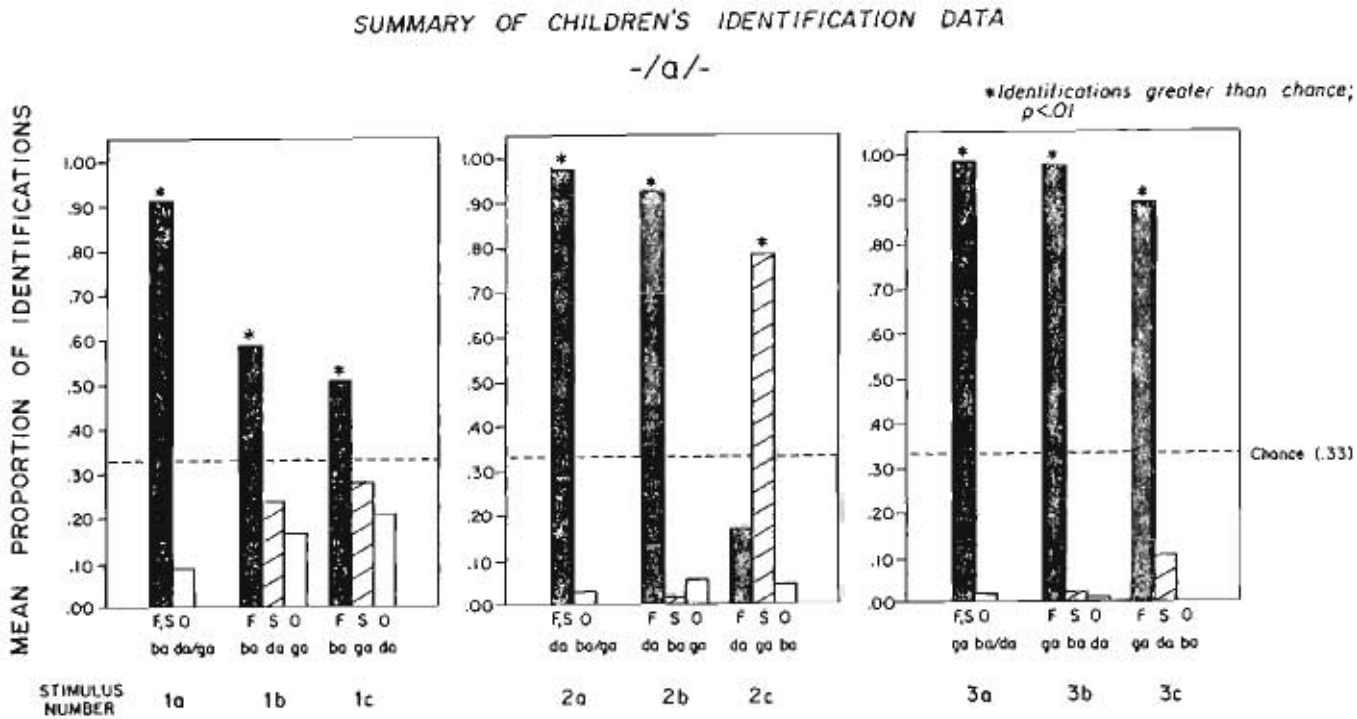


Figure 8. The grouped identification data obtained for children. F refers to identifications according to formant transitions, S to those according to onset spectrum and O to those according to the other possible place response(s) for a given stimulus. F,S refers to the same response for a given control stimulus (1a, 2a or 3a).

Insert Figure 9 About Here

---

right of a control stimulus is shown the number of subjects that identified the two conflicting cue stimuli derived from that control stimulus by a formant transition rule (F), an onset spectum rule (S), or by the third possible place category (O). The figure also shows the number of subjects who identified each control and conflicting cue stimulus in an inconsistent manner (I).

The conflict between the formant transitions and onset spectrum of the experimental stimuli produced a slight increase in inconsistent responses for some of the subjects for three of the stimuli (1b, 1c and 2c), but, in general, subjects responded in a very consistent manner. Thus, the young subjects in this experiment did not appear to experience any extensive difficulty in making their identifications. Their responses were, for the most part, in agreement with the place of articulation of a stimulus as specified by its formant transitions. However, identification responses for stimulus 2c did appear to be based on information in the onset spectrum of that stimulus.

E. Discussion

The group and individual identification results obtained in Experiment 2 with children were virtually identical to those obtained in Experiment 1 with adults in the /a/ condition; i.e., although stimulus 2c was identified according to its onset spectrum, overall the children's responses for conflicting cue stimuli appeared to be determined by formant transitions. These responses were quite consistent as indicated by the number of subjects that identified each stimulus according to a rule. Our findings concerning the consistency of children's identification responses may be contrasted with the inconsistent responses by adults for similar stimuli recently observed by Blumstein, Isaacs and Mertus (1981). The performance of the children tested in Experiment 2 is also at variance with the results of a study conducted by Elliott, Longinotti, Meyer, Raz and Zucker (1981), in which age-related differences in the ability to label place of articulation for synthetic voiced stop consonants without bursts were observed. These developmental differences appear to depend in part on the fact that young children require relatively greater stimulus intensities to perform accurately in various listening tasks employing speech and nonspeech stimuli (see Elliot, Longinotti, Clifton and Meyer, 1981). However, it is quite evident from our results that there were no substantial differences in the ability of adults and children to label the control and/or the conflicting cue stimuli when these stimuli were presented at the same intensity to both groups of subjects.[7]

The identification results observed here for young children argue against Stevens and Blumstein's claim that formant transitions constitute a secondary cue to place of articulation which is acquired in development by virtue of the co-occurrence of formant transitions and primary information contained in the CV

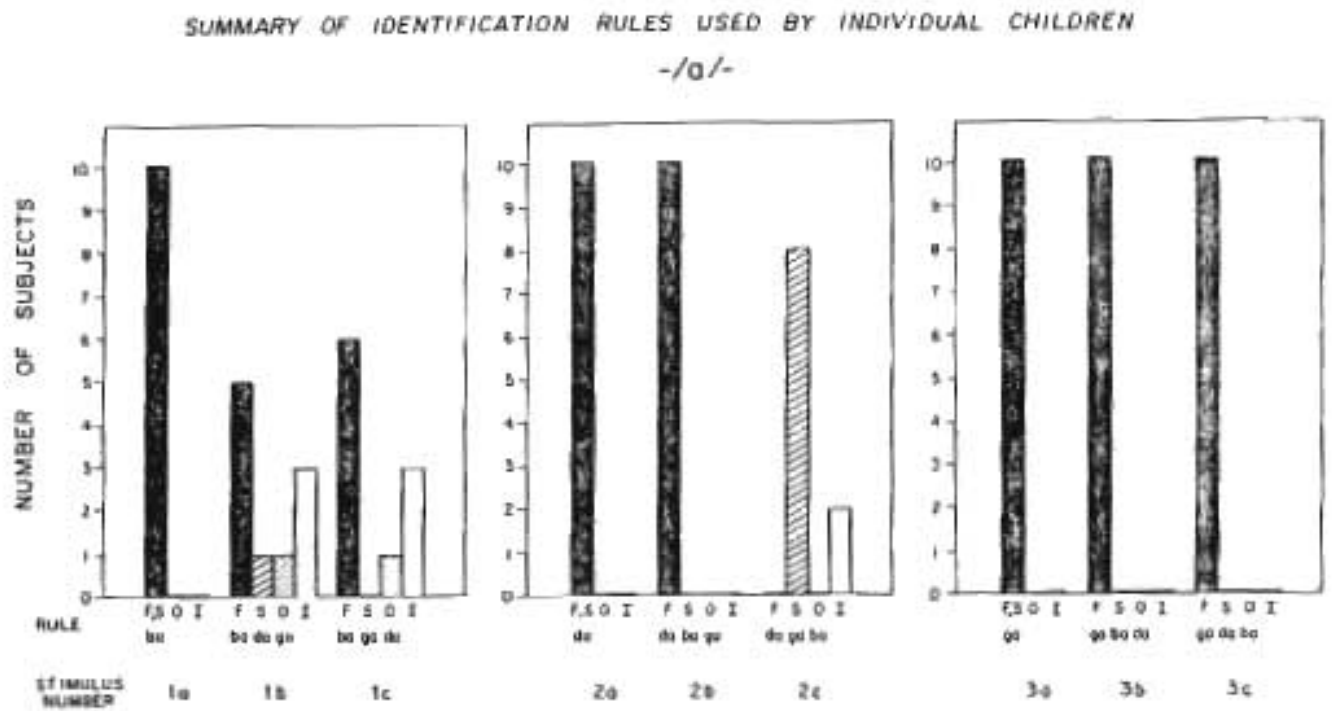


Figure 9. Identification rule choice of individual children. I refers to inconsistent responding for a given stimulus.

syllable onset spectrum. If this were true, one might expect, as suggested earlier, that children would be unable to respond on the basis of formant transition information in a consistent manner. Apparently, children had no difficulty using such information to identify place of articulation in these synthetic stop consonants. It could be argued that the ability to use this information is completely acquired in children of the age that we tested. Yet, since Stevens and Blumstein's theory asserts that properties of the onset spectrum are primary for the adult's perception of place of articulation, it would still predict that children's responses should have been consistent with these properties. This prediction was not upheld. Rather, formant-based responses predominated for the children tested in this experiment.

III. GENERAL DISCUSSION

The results obtained in Experiments 1 and 2 of the present study provide little empirical support for Stevens and Blumstein's claim (1978; 1981; Blumstein and Stevens, 1979; 1980) that sensitivity to certain distinctive and invariant properties of the CV syllable onset spectrum provides the primary basis for place of articulation perception in the adult and the young child. Nor do these results support the claim that formant transition information is, relative to the onset spectrum, secondary in signalling place of articulation. When adults and children were required to identify synthetic stimuli in which formant transitions conflicted with place of articulation as specified by the onset spectrum, their responses were, in fact, almost entirely determined by the formant transitions. Moreover, both adults' and children's responses were made in a very consistent manner indicating that the conflict which was physically present in the stimuli did not produce any extensive perceptual conflict (see, however, Blumstein et al., 1981). The consistency with which children identified these conflicting cue stimuli would also seem to argue against the notion that formant transition cues have only a secondary, developmental status.

It is possible that the critical properties of the onset spectrum were not specified optimally in our conflicting cue stimuli. With our manipulation of relative formant amplitudes, we might have been more successful in matching some of the stimuli to Stevens and Blumstein's templates than we were for others. However, if we had succeeded to varying degrees in matching the stimuli to the appropriate templates (and if, in fact, these templates incorporate properties that are important for the identification of place of stop consonant articulation), many more inconsistencies in responding should have been observed in our data. Moreover, since Stevens and Blumstein assert that it is the global properties of the onset spectrum which uniquely and invariantly characterize a given place of articulation, identification performance should not depend on such precise matches. (If it does, then the properties of the onset spectrum do not constitute the sort of robust cues that Stevens and Blumstein have intended them be.) The same argument would seem to rule out the possibility that Stevens and Blumstein's templates are themselves merely incorrect in their local details.

The finding of the present study that adults' and children's responses to our conflicting cue stimuli agreed overall with place of articulation as specified by formant transition information is consistent with the position of

Liberman et al. (1967). These investigators have maintained that the direction and extent of the second and third formant transitions are important cues for place of articulation. Because this position derives its main support from the results of studies with synthetic speech, it is important to note that studies with natural speech have shown that formant transition information is sufficient for distinguishing place of articulation only within certain vowel contexts (e.g., Dorman et al., 1977; Kewley-Port, 1982a; Ohde and Scharf, 1977).[8] In natural speech, formant transitions are not only context-dependent correlates of place of articulation, but, within particular vowel contexts, they may be insufficient cues for distinguishing place of articulation. Dorman et al. (1977) did find that tokens of natural labial and alveolar voiced stops, in the context of /a/ and /u/, were identified by listeners with moderate to high accuracy when only voiced transitions were retained in the stimuli. The identification responses obtained in the present study for the synthetic conflicting cue stimuli are, therefore, consistent with these observations using natural speech.[9]

In light of the general context dependency and nondistinctiveness of formant transitions observed in these studies of natural speech, we would not wish to argue then that formant transition information constitutes the best description of the acoustic information which listeners use to identify place of articulation in normal speech processing. Moreover, although the majority of our adult and child subjects were able to identify our conflicting cue stimuli quite consistently, this was not true of all subjects. With respect then to invariant characterizations of the acoustic/auditory correlates of place of articulation, the demonstration that adults' and children's identification responses were made on the basis of transitional information in the onsets of the conflicting cue stimuli is perhaps most compatible with Kewley-Port's (1980) account of the acoustic cues for place of articulation perception in stops. Kewley-Port's analysis employs dynamically changing spectral properties to describe place of articulation and retains information about the fine temporal structure of the initial portion of the CV waveform. However, her characterization of the difference between /b/ and /d/ in terms of spectral tilt, like Stevens and Blumstein's, were not generally supported by our research. Therefore, although recent investigations, such as those of Stevens and Blumstein and of Kewley-Port, have met with some success in describing invariant acoustic correlates of place of articulation, the results reported here indicate that these characterizations still do not adequately capture the stimulus information listeners may use in the perception of different place of articulation categories.

ACKNOWLEDGMENTS

This research was supported in part by NIH research grant NS-12179-05 and NIMH research grant MH-24027-06 to Indiana University, Bloomington and by a doctoral fellowship awarded to the first author by the Research Council of Canada. The results of Experiment 1 were presented in preliminary form at the 100th Meeting of the Acoustical Society of America in Los Angeles, California in November, 1980. We gratefully acknowledge the valuable contributions of L. B. Smith and D. B. Pisoni to this research. We also thank D. Kewley-Port and our reviewers for their helpful comments on an earlier draft of this paper. Reprints may be obtained from the first author at the address above.

FOOTNOTES

[1]Although previous demonstrations of the infant's ability to discriminate place of articulation differences in stop consonants would appear to lend support to the argument that some invariant property must exist in the acoustic events associated with each place of articulation category, it is important to realize that these studies have not shown that infants possess the perceptual abilities which Stevens and Blumstein imply they do; i.e., these studies have not shown, for example, that infants perceive syllables such as /du/ and /di/ as being similar with respect to their initial consonants, but only that they are capable of discriminating place of articulation differences within a particular vowel context. The existence of invariant acoustic correlates certainly need not be assumed in order to account for this ability. Fodor, Garrett and Brill (1975) have, it is true, reported evidence of perceptual constancy in infants for voiceless stop consonants differing in place of articulation, but their evidence is rather weak (see Aslin, Pisoni and Jusczyk, 1982) and their findings have not been replicated (cf. Katz and Jusczyk, 1980).

[2]One of these subjects failed to respond on 22 of the test trials. The other two subjects responded in what was considered to be an inappropriate manner for control stimulus 4a. This stimulus which, according to either its formant transitions or onset spectrum, should be identified as /bu/, was identified as /gu/ more than 60% of the time by one of these subjects. The other subject identified this stimulus in what was determined to be an unacceptably inconsistent manner.

[3]We chose to model and manipulate Stevens and Blumstein's transition-only stimuli rather than the more natural burst + transition stimuli. This was done in order to avoid having to decide which, if any, properties of the burst should, in addition to those of the formant transitions, be modified to achieve the desired combination of the onset spectrum and formant transition cues. It was felt that which of these two types of stimuli (i.e., burst + transition or transition-only) was selected for testing Stevens and Blumstein's theory should not be crucial to the outcome of the present experiment, since Stevens and Blumstein claim that their transition-only stimuli have distinctive and context-independent onset spectra and their shapes are merely enhanced by the presence of the release burst.

[4]Aslin and Walley (1980) found that six-month-old infants were able to discriminate two-formant tokens of /da/ and /ga/, even though the onset spectra of these stimuli did not have those properties which Stevens and Blumstein propose characterize alveolar and velar place of articulation and which they maintain mediate place discrimination in infants.

[5]Three subjects failed to pass training with the orally produced stimuli and one subject failed to pass with the synthetic stimuli in less than 45 trials. Therefore, these subjects were not tested. The remaining two children completed testing, but their data were excluded from the analysis because they responded in an unacceptably inconsistent manner for one of the control stimuli.

[6]This level of alpha was chosen for the analysis of the children's data because they were given 12 rather than the 24 or 48 trials per stimulus that the adults received. Therefore, it was not possible to obtain the resolution necessary to classify the children as having used a particular rule for alpha = .01 without being much more stringent than in the analysis of the adult data.

[7]In a 3-way ANOVA (PLACE x CONSISTENCY x AGE), we examined the consistency with which individual children and adults identified each of the nine stimuli in the /a/ series, using 1 - Relative H as a measure of the amount of uncertainty present in a given subject's categorization of a stimulus (see Attneave, 1959; Garner, 1962). No main effect of Age was obtained, nor did Age interact significantly with any other factor, indicating that children and adults did not differ in their responding for any of the stimuli.

[8]Kewley-Port (1982a) did observe that if the onset frequencies of the second and third formant transitions were combined in a two-dimensional, F2 x F3 space, then, given vowel context, place of articulation categories could be distinguished from one another. However, it is not clear how the auditory system would represent this information.

[9]Identification of velar tokens in the Dorman et al. study was very poor when only the voiced transitions were retained -- a result which is not consistent with the performance for velars in our study. Superior performance in our study might be attributed to the fact that the second and third formant transitions of synthetic velar stimuli come artificially close together in comparison to those of natural tokens (Kewley-Port, 1980). This feature may somehow provide listeners with information by which to identify a synthetic stimulus as velar.

REFERENCES

Aslin, R. N. and Walley, A. C. (1980). "Infants' discrimination of cues to place of articulation in stop consonants," J. Acoust. Soc. Am. Suppl. 1 68, S10-S11.

Aslin, R. N., Pisoni, D. B. and Jusczyk, P. W. (1982). "Auditory development and speech perception in infancy," in M. M. Haith and J. J. Campos (Eds.), Infancy and the Biology of Development. Volume 2 of Carmichael's Manual of Child Psychology, 4th Edition (P. H. Mussen, Series Ed.). (New York: Wiley), in press.

Attneave, F. (1959). Applications of Information Theory to Psychology. (New York: Holt, Rinehart and Winston).

Blumstein, S. E., Isaacs, E. and Mertus, J. (1981). "The role of the gross spectral shape as perceptual cues to place of articulation in stop consonants," J. Acoust. Soc. Am. Suppl. 1 70, S32.

Blumstein, S. E. and Stevens, K. N. (1979). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. 66, 1001-1017.

Blumstein, S. E. and Stevens, K. N. (1980). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. 67, 648-662.

Bush, L. and Williams, M. (1978). "Discrimination by young infants of voiced stop consonants with and without release bursts," J. Acoust. Soc. Am. 63 (4), 1223-1226.

Cole, R. A. and Scott, B. (1974). "The phantom in the phoneme: Invariant cues for stop consonants," Percept. Psychophys. 15, 101-107.

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

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

Dorman, M. F., Studdert-Kennedy, M. and Raphael, L. J. (1977). "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues," Percept. Psychophys. 22, 109-122.

Eimas, P. D. (1974). "Auditory and linguistic processing of cues for place of articulation by infants," Percept. Psychophys. 16 (3), 513-521.

Elliott, L. L., Longinotti, C., Clifton, L. and Meyer, D. (1981.) "Detection and identification thresholds for consonant-vowel syllables," Percept. Psychophys. 30 (5), 411-416.

Elliott, L. L., Longinotti, C., Meyer, D., Raz, I. and Zucker, K. (1981). "Developmental differences in identifying and discriminating CV syllables," J. Acoust. Soc. Am. 70 (3), 669-677.

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

Fant, G. (1973). "Stops in CV-syllables," in G. Fant (Ed.), Speech Sounds and Features. (Cambridge, Mass.: MIT Press), pp. 110-139.

Fodor, J. A., Garrett, M. F. and Brill, S. L. (1975). "Pi ka pu: The perception of speech sounds by prelinguistic infants," Percept. Psychophys. 18, 74-78.

Garner, W. R. (1962). Uncertainty and Structure as Psychological Concepts. (New York: Wiley).

Jakobson, R., Fant, C. G. M. and Halle, M. (1952). Preliminaries to Speech Analysis. (Cambridge, Mass.: MIT Press).

Katz, J. and Jusczyk, P. W. "Do six-month-olds have perceptual constancy for phonetic segments?" Paper presented at the International Conference on Infant Studies, New Haven, April, 1980.

Kewley-Port, D. (1978). "KLTEXC: Executive program to implement the KLATT software speech synthesizer," Res. Speech Percept.: Prog. Rep. No. 4, Depart. of Psychol., Indiana University, 235-245.

Kewley-Port, D. (1979). "Spectrum: A program for analyzing the spectral properties of speech," Res. Speech Percept.: Prog. Rep. No. 5, Dept. of Psychol., Indiana University, 475-492.

Kewley-Port, D. (1980). "Representations of spectral change as cues to place of articulation in stop consonants," Res. Speech Percept.: Techn. Rep. No. 3, Depart. of Psychol., Indiana University.

Kewley-Port, D. (1982a). "Measurement of formant transitions in naturally produced stop consonant-vowel syllables," J. Acoust. Soc. Am. 72 (2), 379-389.

Kewley-Port, D. (1982b). "Time-varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am., in press.

Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. 67, 971-995.

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

Liberman, A. M., Delattre, P. C., Cooper, F. S. and Gerstman, L. J. (1954). "The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monongr. 68 (8, Whole No. 379), 1-13.

Moffit, A. R. (1971). "Consonant cue perception by twenty-four-week-old infants," Child Dev. 42, 717-731.

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

Ohde, R. N. and Sharf, D. J. (1977). "Order effect of acoustic segments of VC and CV syllables on stop and vowel identification," J. Speech Hear. Res. 20, 543-554.

Searle, C. L., Jacobson, J. Z. and Kimberly, B. P. (1980). "Speech patterns in the 3-space of time and frequency," in R. A. Cole (Ed.), Perception and Production of Fluent Speech. (Hillsdale, N. J.: Erlbaum), pp. 73-102.

Searle, C. L., Jacobson, J. Z. and Rayment, S. G. (1979). "Stop consonant discrimination based on human audition," J. Acoust. Soc. Am. 65, 799-809.

Stevens, K. N. and Blumstein, S. E. (1978). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. 64, 1358-1368.

Stevens, K. N. and Blumstein, S. E. (1981). "The search for invariant acoustic correlates of phonetic features," in P. D. Eimas and J. Miller (Eds.), Perspectives on the Study of Speech. (Hillsdale, N.J.: Lawrence Erlbaum Associates), pp. 1-38.

Stevens, K. N. and House, A. S. (1972). "Speech perception," in J. Tobias (Ed.), Foundations of Modern Auditory Theory: Volume II. (New York: Academic Press), pp. 3-62.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants*

Diane Kewley-Port
David B. Pisoni

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

and

Michael Studdert-Kennedy

Queens College and Graduate Center of
The City University of New York, New York, NY 10036
and Haskins Laboratories, New Haven, CT 06510

*Earlier reports of these experiments were presented before the Acoustical Society of America at the 100th meeting in Los Angeles, and the 102nd meeting in Miami Beach. This research was supported, in part, by the National Institutes of Health, Research Grant NS-12179 and the National Institute of Mental Health, Research Grant MH-24027 to Indiana University in Bloomington, and, in part, by the National Institute of Child Health and Development, Research Grant HD-01994 to Haskins Laboratories, New Haven, CT.

ABSTRACT

Two recent accounts of the acoustic cues which specify place of articulation in syllable-initial stop consonants claim that they are located in the initial portions of the CV waveform and are context-free. Stevens and Blumstein [J. Acoust. Soc. Am., 64,1358-1368 (1978)] have described the spectral properties of these cues as static, while Kewley-Port [J. Acoust. Soc. Am., in press (1982)] describes these cues as dynamic. Three perceptual experiments were conducted to test predictions derived from these accounts. Experiment 1 demonstrated that acoustic cues for place of articulation are located in the initial 20 to 40 ms of natural stop-vowel syllables. Next short synthetic CV's modeled after natural syllables were generated using either a digital, parallel-resonance synthesizer in Experiment 2 or linear prediction synthesis in Experiment 3. One set of synthetic stimuli preserved the static spectral properties proposed by Stevens and Blumstein. Another set of synthetic stimuli preserved the dynamic properties suggested by Kewley-Port. Listeners in both experiments could identify place of articulation significantly better from stimuli which preserved the dynamic acoustic properties of stop waveforms than from the static onset spectra. Evidently the dynamic structure of the initial stop-vowel articulatory gesture can be preserved in context-free acoustic cues which listeners use to identify place of articulation.

INTRODUCTION

The search for acoustic cues to place of articulation in initial stop consonants has focused on two different sources of information in the stop-vowel syllable: context-free and context-dependent. The context-free source has usually been taken to be the stop release burst, represented as a single, static spectral section and described in the terminology of distinctive feature theory (Halle, Hughes and Radley, 1957; Fant, 1960; Stevens, 1975; Stevens and Blumstein, 1978). The context-dependent source has usually been taken to be the stop-vowel formant transitions, represented as dynamic variations in the distribution of energy within restricted regions of the spectrum and described with an emphasis on their origins in the stop articulatory gesture (Liberman, Delattre, Cooper, and Gerstman, 1954; Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967; Liberman and Studdert-Kennedy, 1978). Finally, some investigators have studied the combined role of release bursts and formant transitions, stressing variation in the effective weights of the two cues as a function of context (Cooper, Delattre, Liberman, Borst and Gerstman, 1952; Fischer-Jorgensen, 1972; Dorman, Studdert-Kennedy and Raphael, 1977).

The research reported here attempts to combine a view of stop-consonant place information as invariant, context-free and largely located in the release burst with a view of the information as dynamic rather than static. The assumption of invariance derives, in part, from the acoustic theory of speech production (Fant, 1960; Stevens and Blumstein, 1978), according to which invariant acoustic properties, associated with particular places of articulation, should lie in the brief initial portions of a CV waveform, encompassing both burst and transition. The assumption of dynamic rather than static properties derives from previous work of Kewley-Port (1980, 1982b), and is intended as an explicit alternative to the assumptions of Stevens and Blumstein (1978, 1981) who have recently argued for invariant cues to consonant perception.

Stevens and Blumstein have suggested that the auditory system integrates spectral energy over approximately 20 ms at the onset of a release burst. This static spectrum is said to contain the gross spectral properties that distinguish among places of articulation. Blumstein and Stevens (1979) tested spectral templates, visually matched to a single, static spectrum positioned at the burst onset, and obtained 84% correct place identification. By contrast, Kewley-Port (1980; 1982b) noted that, since short-term spectral integrations in the auditory system are rapidly updated to preserve temporally changing spectral information (Schroeder, Atal and Hall, 1979), the running spectra of linear prediction analysis might provide a more appropriate visual representation of the rapidly changing auditory information. Kewley-Port (1982b) defined three time-varying features to distinguish place of articulation visually in the running spectra, and showed that judges could use these features to identify place of articulation 88% correctly from running spectral displays of the first 40 ms of CV syllables.

A more detailed look at the specific spectral and temporal properties of these two proposals for place cues reveals other interesting similarities and differences. The invariant acoustic cues for place proposed by Stevens and

Blumstein (1981) may be distinguished on the basis of spectral properties alone. There is no temporal dimension because spectral energy is integrated over a fixed 25.6 ms window. The gross shapes of the onset spectra are described as diffuse-rising for bilabials, diffuse-falling for alveolars and compact for velars. In contrast, the three time-varying features proposed by Kewley-Port (1980; 1982b) have both spectral and temporal dimensions. The first feature is the spectral tilt of the burst observed in approximately the first 5 ms of the stop-vowel waveform. Tilt of burst is described as rising for alveolars, flat or falling for bilabials, and having unspecified tilt for velars. Thus, the spectral properties of the burst for bilabials and alveolars have the same description for Kewley-Port as for Stevens and Blumstein. The temporal dimension, however, is different because Stevens and Blumstein integrate energy over a longer window. Thus, onset spectra usually contain some vowel transition information, whereas the initial 5 ms window of Kewley-Port represents only the burst spectrum.

The second time-varying feature is the presence or absence of mid-frequency peaks extending in time for at least 20 ms. Extended mid-frequency peaks are reliable acoustic correlates of velar place, while quickly dissipating peaks are not (Fant, 1968, p. 223). Clearly, the spectral quality of the extended mid-frequency peaks is the same as that of Stevens and Blumstein's compact onset spectra. However, while temporal extension is essential for Kewley-Port, it is irrelevant for Stevens and Blumstein, since their basic premise is that spectral energy is integrated over a fixed time window. Nonetheless, they have stated (Blumstein and Stevens, 1980, p. 661) that perhaps "a longer time is necessary to build up a representation of the 'compact' onset spectrum in the auditory system." The studies reported here investigate the necessity of longer duration waveforms for velars in some detail.

The third time-varying feature for Kewley-Port (1982b) is late onset of an F1 peak relative to the burst. This is essentially a measure of VOT (voice onset time), previously shown to correlate with place of articulation (Lisker and Abramson, 1964). In the running spectral analysis, the late onset feature is viewed as a secondary feature specifying velar place of articulation. This feature assumes that details of the change from frication to voicing play a role in place identification. The feature has no counterpart in Stevens and Blumstein's account, because the fixed time window integrates energy over both the fricative and voiced portions of the CV waveform.

To examine the similarities and differences between these sets of static and dynamic acoustic properties, three perceptual experiments were conducted. In Experiment 1, we investigated how much stimulus information is needed from the initial portions of natural stop waveforms to identify place of articulation accurately. More specifically, we sought to determine whether place of articulation could be identified accurately from 20 ms waveform segments as Stevens and Blumstein have claimed. Experiments 2 and 3 were designed to examine whether static or dynamic acoustic properties are used by listeners to identify place in short stop consonant-vowel waveforms. These experiments compared the perception of natural speech segments with two sets of synthetic stimuli. One set of synthetic stimuli modeled the Stevens and Blumstein onset spectra; the other was patterned after the time-varying features proposed by Kewley-Port.

Experiment 2 used a digital, parallel resonance synthesizer; Experiment 3 replicated the results using linear prediction synthesis. The overall goal of this research was to determine whether the acoustic structure that supports the perception of place of articulation in English stop consonants is more properly described as static or dynamic in nature.

I. EXPERIMENT 1: IDENTIFICATION OF TRUNCATED NATURAL CV's

The purpose of Experiment 1 was to determine the duration of the initial portion of a stop-vowel syllable necessary to identify place of articulation accurately. A related, though much less detailed, study was carried out recently by Tekieli and Cullinan (1979). They measured the minimum duration necessary for listeners to identify consonants and vowels correctly from the electronically gated, initial portions of CV syllables, spoken by a single talker. The authors did not score identification of place of articulation separately from voicing, but their results indicate that place of articulation, averaged over the six English stops, was identified correctly more than 95% of the time from 30 ms waveform durations.

Experiment 1 of the present study was nearly completed when Blumstein and Stevens (1980) reported a series of experiments examining durations in initial portions of synthetic CV stimuli. They concluded (p. 660) that "information with regard to place of articulation for a voiced stop consonant resides in the initial 10-20 ms of a consonant-vowel syllable." Unfortunately, the stimuli and procedures employed in their study make it difficult to generalize the results to the identification of naturally produced CV's. The bursts in their synthetic stimuli were acoustically impoverished since they were generated with only one formant of excited energy. Furthermore, Blumstein and Stevens (1980) always presented their results in terms of the duration of the voiced portion of the stimuli. Since the stimuli resembling natural CV's contained aperiodic bursts preceding the voiced portions, the shortest, so-called 10 ms stimuli actually varied in duration from 20 ms for /ba/ to 35 ms for /gi/. Therefore, a reliable estimate of the minimum duration necessary to identify place of articulation in natural CV's cannot be determined from the results of the Blumstein and Stevens study.

In Experiment 1, then, we sought to determine directly how much place information actually resides in the early portions of natural stop consonant syllables. The present experiment examines this question with naturally spoken consonant-vowel syllables obtained from two male talkers in five different vowel contexts. The aperiodic and following waveform segments were edited and measured digitally by computer. Naive listeners were required to identify place of articulation in short truncated CV's at various waveform durations.

A. Method

1. Stimuli: CV syllables

A set of 30 CV syllables spoken by two male talkers (RP and TF) was chosen from a larger set of utterances used in an earlier experiment (Kewley-Port, 1981b). The syllables consisted of all combinations of initial /b,d,g/ and the vowels /i,e,a,o,u/, one syllable from each talker. These syllables were read in the carrier sentence, "Teddy said CV" from random lists in a sound attenuated room and recorded on an Ampex AG-500 tape recorder. The sentences were low-pass filtered at 4.9 kHz and digitized with a 12 bit A/D converter at a 10.0 kHz sample rate using a PDP 11/05 computer; they were then edited so that only the target CV was permanently stored on disk.

Before the set of stop-vowel syllables was edited further for this experiment, we checked that listeners could correctly identify the consonants in the full syllables. A computer program was used to randomize and output the 30 full syllables through a 12 bit D/A converter for recording on audio tape. The tape consisted of 10 blocks of the 30 CV's for a total of 300 trials. Six naive subjects listened to the tape over headphones in a quiet room. All subjects were paid for their services. Subjects were given instructions to write down the letter which corresponded to the consonant they heard at the beginning of each syllable. The response set, therefore, was the open set of all English consonants. Results showed that subjects correctly identified the stop consonants in the full syllables at a level of 99.8% correct, with no consonant responses other than b, d or g. Evidently, all 30 original CV syllables may be considered good exemplars of the stop consonant the talkers had intended to produce.

2. Stimuli: Waveform editing

Each of the 30 original CV's was then edited digitally to retain only the initial portions of the waveforms. Five different cuts were made at zero crossings to produce five truncated stops from each original CV syllable. For /d/ and /g/, the first cut was made just before the first voicing pulse. This aperiodic portion of the waveform, containing the stop release burst and aspiration, will be referred to as the burst. Its mean duration was 14 ms for /d/ and 21 ms for /g/. The second cut included the burst and the first pitch pulse. For /b/, it was not always possible to obtain a burst-only waveform portion because voicing was occasionally continuous from the carrier phrase, "Teddy said" into the voiced stop syllable. Thus, the first waveform cut for /b/ included the burst and the first pitch pulse with a mean duration of 13 ms. The next cut included the burst and two pitch pulses. The remaining cuts for all stops were made so that the waveform segments included the burst plus 3, 5, or 7 pitch pulses. The total number of test stimuli produced by this editing procedure was 150, with durations ranging from 6 to 111 ms in length.

After the data were collected and analyzed for this experiment, the initial pattern of results suggested that one stimulus should be reexamined to determine if a waveform editing error had occurred. This stimulus was the burst plus one pitch pulse /da/ from speaker RP. Examination of the waveform on the CRT

revealed that the last digitized point missed the zero crossing by about 30%. Although the resulting click could not be easily heard in the 19 ms stimulus, it was perceptible and did appear to interfere with the subjects' correct identification of the stimulus as alveolar.

3. Procedure

The experimental session began with a brief familiarization task, followed by the identification test. Audio tapes were produced by a computer program that selected the digital waveforms on disk and then output the stimuli through the D/A converter. The identification tapes consisted of six blocks of all 150 truncated stops. Stops were randomized within blocks, with three seconds between stimuli and seven seconds after each block of 50 stimuli. Two tapes were made for the familiarization task preceding consonant identification. The first tape contained a subset of 60 of the 150 truncated stops in a sequence from /b/ to /d/ to /g/, half from each talker. The second tape contained 25 additional truncated stops in a random sequence, several selected from each talker.

The identification test was given on two days. Day 1 included the short familiarization tasks plus the forced choice identification test for the first three blocks of test trials. Responses were always recorded by hand on prepared answer forms. For consonant identification, the responses were: b, d, g, p, t, k. Although all stimuli were edited from voiced stop consonants, pilot work had indicated that naive subjects were more comfortable identifying the shortest waveforms with the voiceless stop responses, p, t, and k.

Subjects listened to the stimuli through TDH-39 earphones in a quiet testing room. Audio tapes were played back on an Ampex AG-500 tape recorder. A comfortable listening level for the brief stimuli was selected and a single repeated stimulus recorded on each tape was used to calibrate the listening level for all tapes. Separate written instructions were given for each task. Subjects were contacted through a laboratory subject pool and were paid $6 for two days of testing. Subjects were phonetically naive and had no known history of a hearing or speech disorder at the time of testing as assessed by a pretest questionnaire.

B. Results

Ten subjects participated in the identification task. One subject skipped so many responses on the first day that she was asked not to return for the second day of testing. Thus, results were analyzed from nine subjects, providing a total of 54 data points for each truncated stimulus. Responses were scored as correct when place of articulation was correctly identified regardless of the voicing feature. Collapsing over all responses and stimuli, subjects identified place of articulation correctly on 93.2% of all trials. This level of performance is surprising when we consider that over half of the stimuli were shorter than 45 ms. Relatively little effect of learning could be observed from Day 1 to Day 2, a change in the mean percent correct from 92.8% to 93.7%. Identification of the stimuli from the two talkers was the same at 93% correct overall.

Insert Figure 1 about here

Results for Experiment 1 are summarized in Fig. 1. Identification functions averaged across vowels are plotted separately for each stop and each talker according to the number of pitch pulses in the stimuli. The functions are very similar for both talkers and therefore provide an internal replication of the basic results.

Identification performance for /b/ starts with an average of 90% correct identification for the burst plus one pitch pulse stimulus and rises to nearly 100% correct with the next pitch pulse. The identification functions for /d/ are similar to those for /b/, but they rise somewhat more gradually to 100% correct. The identification performance for /g/ differs from both /b/ and /d/. For the burst-only segment, identification is not very accurate with performance at about 70% correct. Furthermore, /g/ identification functions never reach the 100% correct level even for the longest stimuli, as do the /b/ and /d/ functions.

Insert Figures 2, 3 and 4 about here

To examine in more detail the relations between the durations of the truncated stops and the correct identification of place, the results are plotted separately for all 30 CV's in Figures 2, 3 and 4. Identification performance for all vowel contexts of /b/, shown in Fig. 2, is similar to the average functions shown in Fig. 1. Individual functions for /d/ in Fig. 3 are also quite similar to the average /d/ functions with two exceptions. First, the /do/ burst was identified correctly only 44% of the time for speaker TF. The short 6 ms duration alone cannot explain this poor level of identification, since there were two /b/'s whose duration was also 6 ms but these were identified 86% correctly. The second exception was the /da/ stimulus from talker RP which elicited the only non-monotonic identification function in the experiment. Apparently, the waveform editing error described above reduced the correct identification of the alveolar stop.

The results for /g/ demonstrated substantially more vowel context dependency effects than did those for /b/ or /d/ as can be seen in Fig. 4. The most unusual identification functions were obtained for the /gi/ stimuli. Identification was poor at all waveform durations and was less than 50% for the longest (93 ms) stimulus. The identification of /ge/ was also quite poor for the burst-only segments, but increased with three pitch pulses to better than 95% correct. The identification functions for the back vowels with /g/ are quite similar to those for /b/ and /d/. Apparently, /g/ before front vowels, with a more palatal place of articulation, was more difficult for subjects to identify than /g/ before back vowels, with a velar place of articulation.

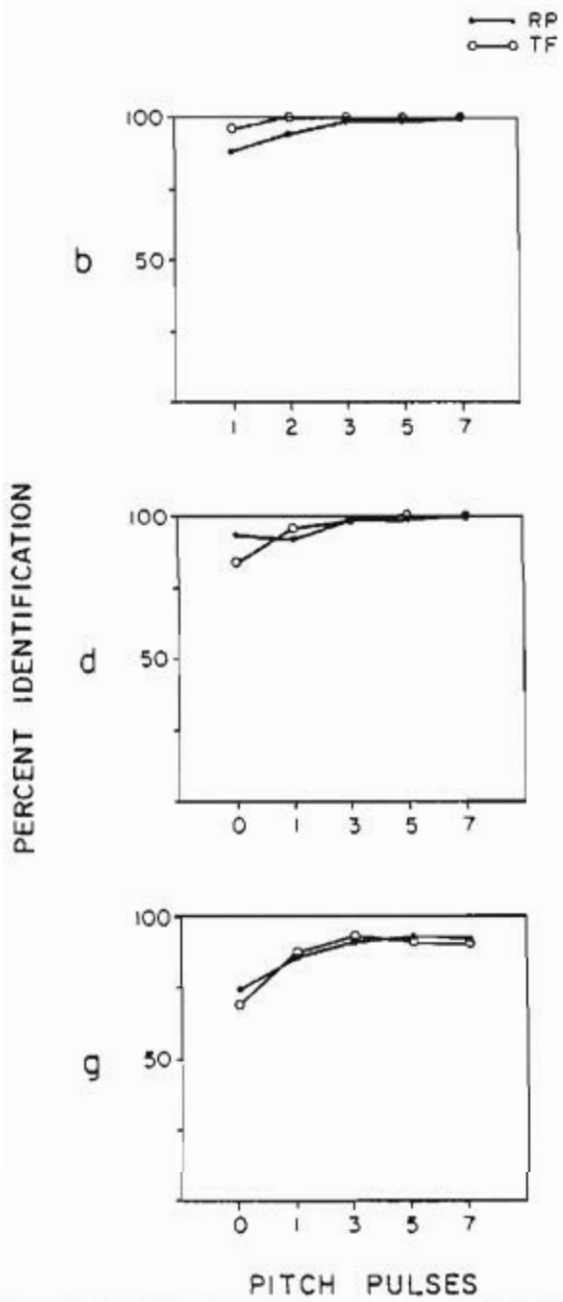


Figure 1. Percent correct consonant identification for truncated waveforms as a function of the number of pitch pulses. Each panel presents the data by consonant averaged over five vowels shown separately for talkers RP and TF.

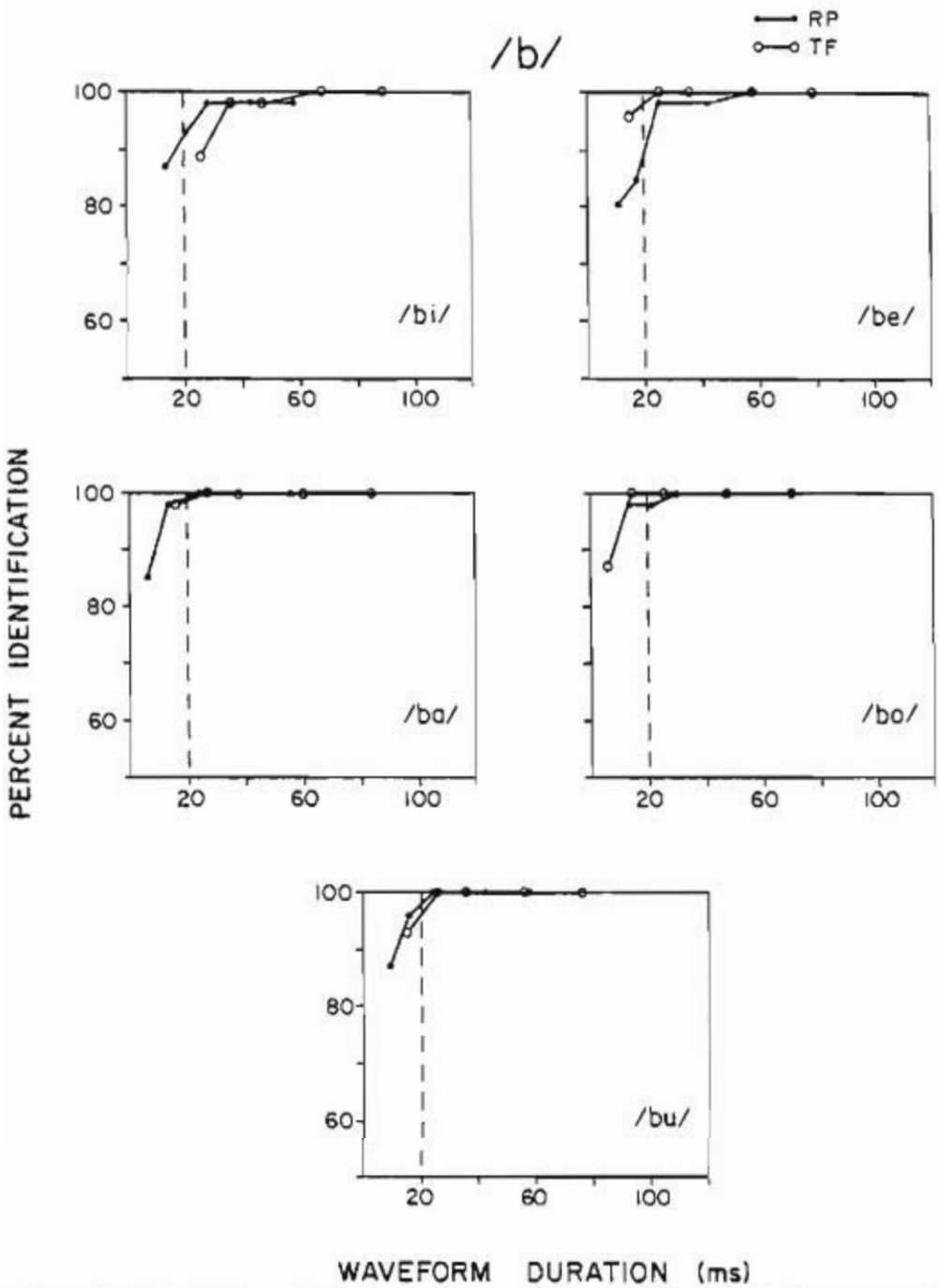


Figure 2. Percent correct consonant identification for all bilabial stops as a function of signal duration plotted separately for talkers RP and TF.

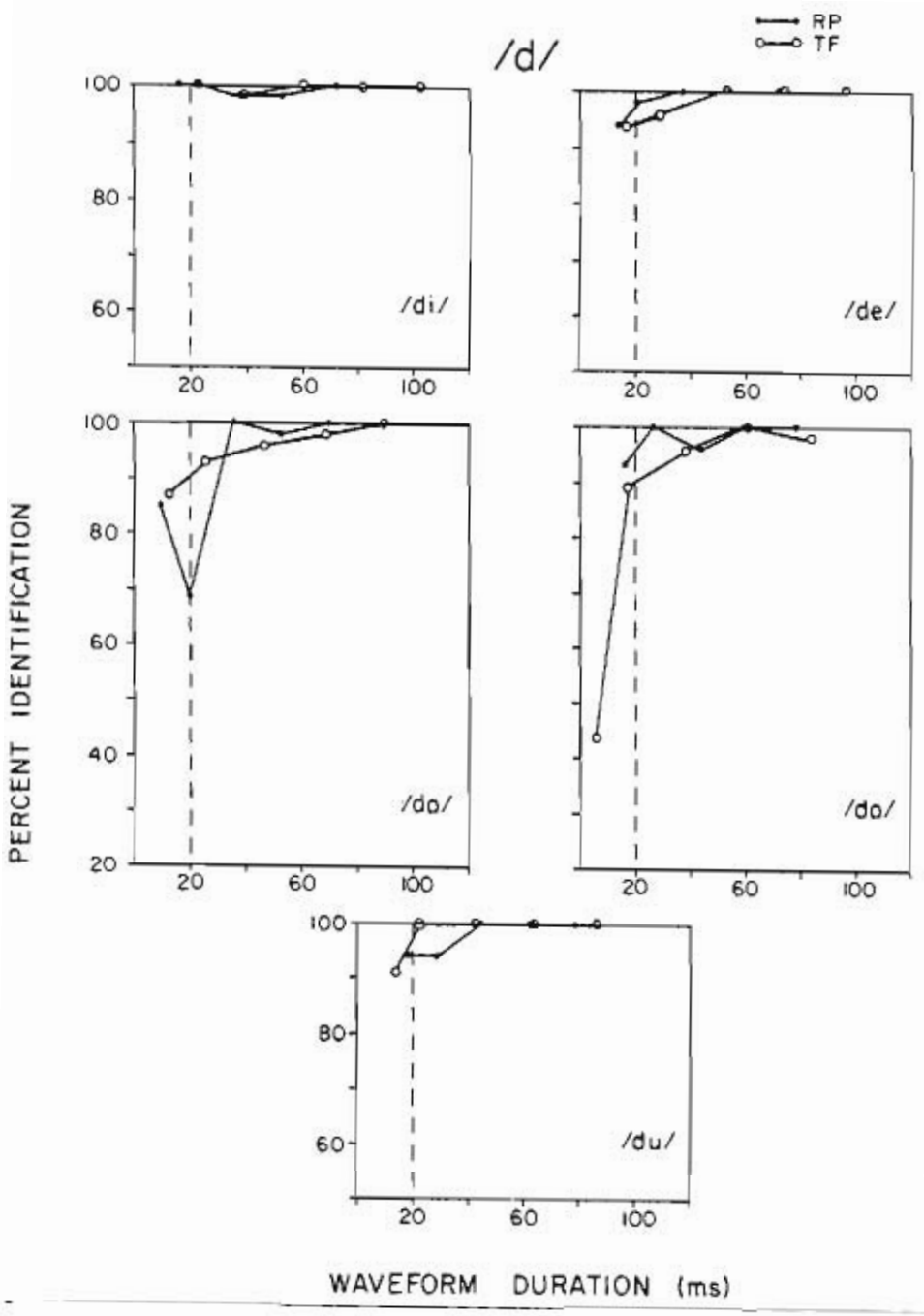


Figure 3. Percent correct consonant identification for all alveolar stops as a function of signal duration plotted separately for talkers RP and TF.

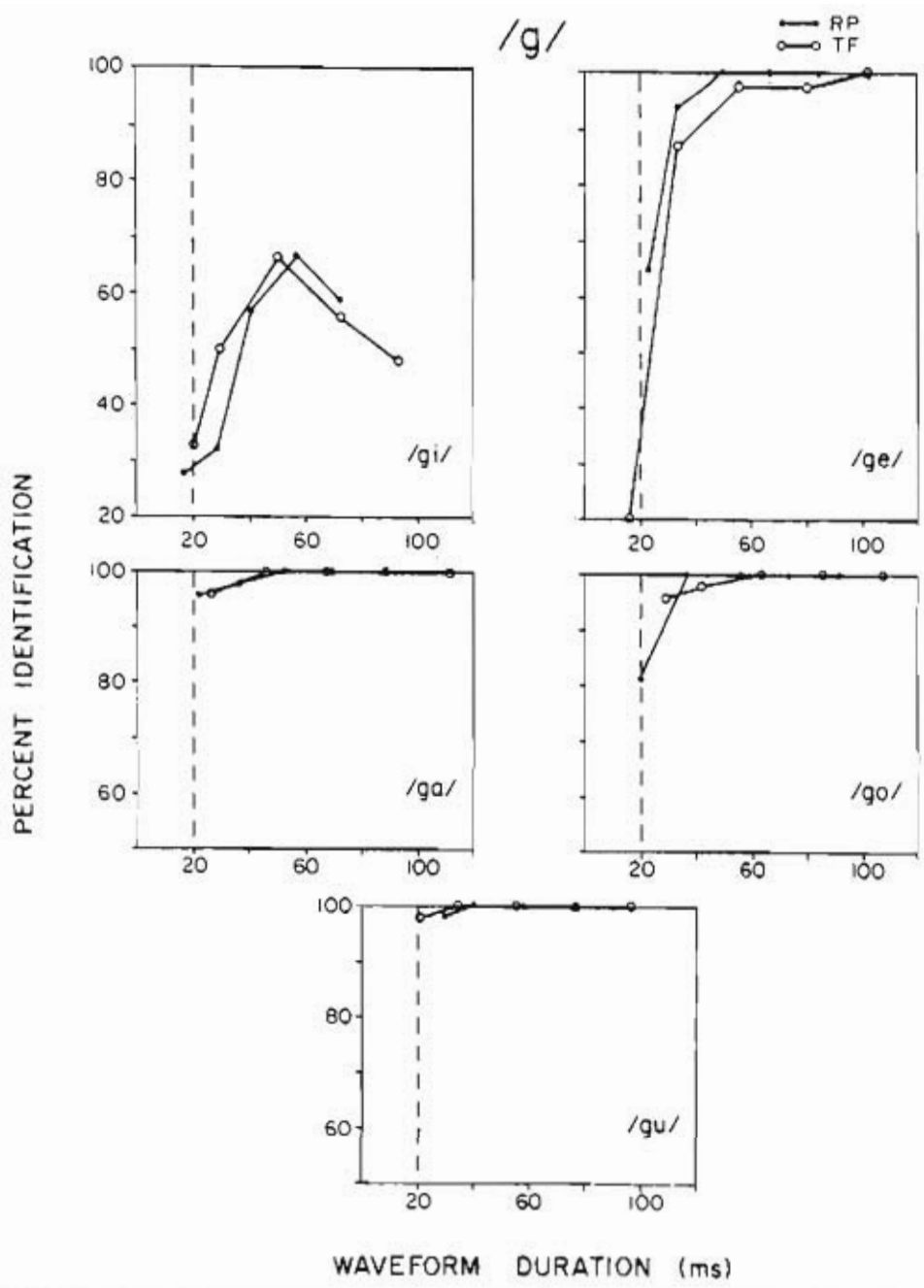


Figure 4. Percent correct consonant identification for all velar stops as a function of signal duration plotted separately for talkers RP and TF.

To summarize, naive subjects were able to identify place of articulation correctly from the initial portions of natural stop CV syllables. The shortest waveform (burst-only for /d/ and /g/ or burst plus one pitch pulse for /b/) was identified with greater than 80% accuracy for 24 out of the 30 CV's examined. The identification functions for /gi/ differed substantially from all others; identification performance was less than 60% correct even for the longest duration stimuli.

C. Discussion

The results from Experiment 1 can be used to evaluate both the Blumstein and Stevens' (1979) static onset spectra hypothesis and Kewley-Port's (1982b) proposed time-varying features. Blumstein and Stevens tested their onset spectra hypothesis experimentally using visual templates designed to fit spectral sections of the first 25.6 ms of a stop waveform. After windowing, the resulting spectra had an effective duration of 20 ms. To facilitate comparisons, a dashed vertical line has been drawn in Figures 2, 3 and 4 at 20 ms. The intersection of an identification function with the dashed line indicates the predicted response accuracy for identification of place from the first 20 ms of a test stimulus. The 20 ms identification values were estimated separately for all 30 CV's (smoothing over the bad /da/ stimulus for EP). These values were then averaged across vowels and talkers for each consonant and are shown in the first column of Table I.

Insert Table 1 about here

As shown in the table, the first 20 ms of a stop waveform contains sufficient place information for /b/ (96% correct) and /d/ (94% correct), but not for /g/ (73% correct). Errors for /g/ were not uniformly distributed but occurred mostly for the syllables /gi/ and /ge/; that is, for /g/ before front vowels. Thus, the results for /b/ and /d/ appear to be consistent with Stevens and Blumstein's hypothesis of a fixed time window. The results for /g/ are clearly not compatible with this hypothesis.

Consider now an alternative hypothesis that it is the dynamic changes in the distribution of spectral energy over time that specify differences in place of articulation. The most important feature for distinguishing /b/ from /d/ in Kewley-Port's three feature system is the tilt of the spectrum at burst onset. The feature system also requires that the extended mid-frequency peaks occurring in the running spectra of /b/ or /d/ be absent. Since information about spectral tilt of the burst and absence of the mid-frequency peaks resides in the earliest portions of the stop waveform, this feature system implies that identification of /b/ and /d/ should be quite good for very short truncated stimuli. The high level of identification performance observed for the first 20 ms of /b/ and /d/ as shown in Table I is consistent with this hypothesis.

Table I. Percent correct identification estimated at 20 and 40 ms stimulus durations averaged across vowels and talkers.

	Duration	
Consonant	20 ms	40 ms
b	96	99
d	94	98
g	73	90

On the other hand, consider the time-varying features for /g/. Both the definitions of late onset of F1 and extended mid-frequency peaks imply that more than 20 ms of a stop waveform is needed to identify /g/ correctly. The identification functions shown in Fig. 4 indicate that relatively high levels of identification of /g/ were achieved at durations of 40 to 50 ms for all vowels except /i/. In order to quantify the identification performance of /g/ at longer waveform durations, 40 ms identification values were located on Figures 2, 3 and 4. Forty millisecond values were chosen here because the duration of spectral information displayed in the eight frames in Kewley-Port's (1982b) running spectra experiment was 40 ms. The 40 ms identification values for each consonant were averaged over vowels and talkers and are presented in Table I. These results indicate substantial improvement in identification of /g/ from the 20 ms point at 73% correct to the 40 ms point at 90% correct. In the case of /b/ and /d/, performance was close to the 100% asymptote level.

To summarize, the results from naive listeners' identification of truncated CV syllables showed that /b/ and /d/ could be identified at 94% or better from the information contained in the first 20 ms of the waveform. We conclude that either the Stevens and Blumstein's (1978) onset spectrum or the running spectrum of Kewley-Port (1982b) is adequate to specify invariant place cues for /b/ and /d/. Velar stops, on the other hand, are poorly identified (73%) from the first 20 ms of waveform, and evidently require longer waveform durations for accurate identification. From this result we conclude that the time-varying spectral features proposed by Kewley-Port are likely to prove more successful overall at specifying reliable information for place of articulation in CV syllables. By corollary, Stevens and Blumstein's fixed 20 ms integration time is less likely to capture the acoustic cues for specifying velar place, and their later hypothesis (1980) of "a longer time" for velars identification is more plausible.

The contradiction between the supposed fixed time window for onset spectra and this modified hypothesis for velars is, however, difficult to reconcile with the proposal that onset spectra are detected by innate property-detecting mechanisms in the human auditory system (Stevens and Blumstein, 1978, p.1367; Blumstein and Stevens, 1979). These property detectors are intended to specify place of articulation as a phonetic feature within distinctive feature theory (Chomsky and Halle, 1968). They are therefore insensitive to details of acoustic structure or phonetic context, being designed to specify place of articulation not only for initial stops, but for final stops and for nasal consonants as well. The role of a fixed integration window in Stevens and Blumstein's innate property detector theory is to fit them for this function by eliminating the temporal dimension and making the onset spectra more abstract. The results of the present experiment, and those of Blumstein and Stevens themselves (1980), suggest that the temporal dimension should not be disregarded and that static onset spectra alone are not sufficient to specify cues for place of articulation in stop consonants (see also the recent findings of Walley and Carrell, 1982).

II. EXPERIMENT 2: IDENTIFICATION OF SHORT SYNTHETIC CV's

The purpose of Experiment 2 was to make a more direct test of the inferences drawn from Experiment 1 by perceptual study of synthetic CV syllables constructed according to the rival static and dynamic patterns. The idea behind the experiment was, in fact, alluded to earlier by Stevens and Blumstein themselves:

> A stronger test of (the) theory would be to determine whether perception of place of articulation depends on attributes of the gross shape of the spectrum at onset, independent of fine details such as burst characteristics and formant onset frequencies (1978, p. 1367).

Since naive subjects were reasonably successful in identifying place in the truncated natural speech stimuli of Experiment 1, we adopted the design of that experiment for use here. Two sets of truncated synthetic stimuli were constructed to test the static and dynamic hypotheses. In the present experiment, subjects first participated in an identification task using truncated natural speech CV's. Subjects who could identify place correctly for the natural speech stimuli were then required to identify stimuli from both synthetic stimulus sets.

In addition to an identification response for each stimulus, we also gathered a confidence rating to indicate whether a subject thought his response was a guess, was surely correct, or was somewhere between the two. We were particularly interested in whether subjects would be as confident about the correct identification of the synthesized stimuli as they were about the natural stimuli.

A. Stimuli: Synthesis parameters

The success of this experiment depends on clearly stated and executed principles for synthesizing the two stimulus sets, one derived from Stevens and Blumstein static onset spectra (hereafter called S+B) and the other from time-varying features in running spectra (hereafter called RS). Each synthesized syllable was modeled after the appropriate natural syllable by visual spectral matching techniques. Spectral analysis of the natural stimuli was carried out by linear prediction analysis as implemented in the SPECTRUM program (Kewley-Port, 1979). The S+B and RS stimuli were synthesized on the KLATT digital synthesizer (Klatt, 1980) as implemented in the KLTEXC program (Kewley-Port, 1978).

Three independent variables, consonant type, vowel type and duration were manipulated in Experiment 2, in addition to stimulus set type. Three consonants /b,d,g/ were paired with the three vowels /i,a,u/ to produce a base set of nine CV syllables, each at three waveform durations: 20, 30 and 40 ms. These durations spanned the 20 ms duration at which /b/ and /d/ were accurately identified in Experiment 1, and the 40 ms duration at which /g/ (except for /gi/) was accurately identified.

The natural stimuli consisted of the nine base syllables spoken by talker RP. Each syllable was analyzed by the SPECTRUM program to determine that the proposed features for place of articulation could be correctly identified by both the Blumstein and Stevens' templates (1979) and the running spectral feature analysis. The SPECTRUM program was used to calculate the onset spectrum by a procedure identical to that used by Blumstein and Stevens (1979), with the exception of substituting a one-half Hamming window for a one-half Kaiser window. All onset spectra were visually examined on a CRT graphics display. Only natural CV's whose spectra clearly fit the overall template descriptions of diffuse-rising, diffuse-falling or compact, and which appeared to meet all the template rules described by Blumstein and Stevens (1979), were accepted. The running spectral display for each syllable was also examined to see that it contained good exemplars of the time-varying features. For velar syllables an additional criterion was checked: the onset spectrum for velars was required to include an F1 peak, so that the resulting synthesized waveforms would all have a voiced component. If all criteria were not met, another utterance spoken by talker RP in the same recording session was selected. In the end, five of the nine natural CV's, /bi,di,ba,bu,gu/ were taken from the stimuli used in Experiment 1. The natural syllables were then edited digitally at zero crossings to approximate the 20, 30 and 40 ms waveform durations as closely as possible. The average durations proved to be 21 ms, 30 ms and 39 ms respectively.

For the synthetic stimuli, the overall strategy was to keep as many of the synthesis parameters as possible the same between the S+B and BS sets, while incorporating differences in the static versus dynamic acoustic properties. To accomplish this, the KLATT synthesizer was configured as a parallel formant synthesizer with six formants (Klatt, 1980). Glottal resonance characteristics were shaped for talker RP's /i/ and then kept constant. Nasal resonances were not used. The fundamental frequency was set to a constant 100 Hz. Synthesis parameters always terminated exactly at the 20, 30 or 40 ms durations. Fortunately, no synthesized waveforms had appreciable baseline offsets since the fundamental was 100 Hz.

The synthesis of the S+B stimuli was accomplished in several steps. With the KLATT synthesizer, it was possible to generate a steady-state stimulus such that its spectrum at any point matched the overall shape of the calculated 25.6 ms onset spectrum of the original natural CV. However, it was not apparent whether these steady-state stimuli would be perceived as stop-consonants or as vowels. In the speech perception literature, it has generally been assumed that a rising F1 transition is an important manner cue for the class of stop consonants (Delattre, Liberman and Cooper, 1955; Stevens and House, 1956; Fant, 1960, Liberman et al., 1967). For this reason, Blumstein and Stevens (1980, p. 651) used rising F1's in their synthesis of otherwise steady-state consonant stimuli. However, Kewley-Port (1982a) observed that a rising F1 transition cannot always be measured in stop-vowel syllables, in particular, for the vowels /i/ and /u/. Since it was not clear what kind of F1 transitions should be used in the synthetic S+B stimuli, we carried out a pilot experiment using the S+B stimuli synthesized with and without F1 transitions, to determine if listeners would judge stimuli with F1 transitions as more stop-like than stimuli with F1 steady-states (see Kewley-Port, 1980 for a more detailed description of this study). Since the outcome showed a slight advantage for stimuli containing F1

transitions, we used the F1 transition values measured from the natural CV's as synthesis parameters for the S+B stimuli.

Insert Figure 5 about here

Figure 5 shows the match between the onset spectra of the final synthesized S+B stimulus for /ga/ and the natural /ga/ stimulus. A good spectral match was obtained by adjusting only the bandwidth and amplitude parameters calculated from the original linear prediction spectrum. The voicing source (AV) was always set to its maximum value.

Synthesis parameters for the RS stimuli were derived from previous studies of RP's stop-vowel syllables. The running spectra, as shown in Fig. 6, were produced by calculating the linear prediction smoothed spectra at 5 ms intervals following the procedures described in Kewley-Port (1982b). The first frame shows the stop release burst which was always positioned at the center of the 20 ms Hamming window. The other sources of parameter information were the average formant transitions and VOT values calculated for 5 repetitions of each of RP's syllables (Kewley-Port, 1982a, Appendix). The synthesis procedures always started by carefully matching the burst frame because it contains the spectral tilt information in the running spectra analysis. After the burst frame, the average VOT values were approximated as /b/ = 0 ms, /d/ = 10 ms and /g/ = 20 ms using the aspiration source in the synthesizer. The spectral shape of these and succeeding frames was determined by inserting the average values of the transition parameters (frequency and duration) after the release burst. The amplitude of the voicing source always started at 5 dB below the maximum, and increased to the maximum in 5 ms. In addition, careful spectral matching of the voiceless frames for /g/ was often needed, in order to preserve the mid-frequency peaks following the burst frame. Figure 6 shows the running spectrum for the natural /ga/, and the final synthesized RS /ga/. The synthesis matching procedure focused on the spectral tilt of the burst frame, and on the mid-frequency peaks feature of the voiceless frames, but not on other spectral properties of these frames.

Thus, the overall procedure for generating the RS stimuli was not based on frame-by-frame spectral matching of the natural and synthetic stimuli. While the burst frames were spectrally matched, which meant one frame for /b/ and /d/, and the four voiceless frames for /g/, the remaining synthesis parameters were determined essentially by rule using average values of the formant transitions. The final RS stimuli were then synthesized at the 20, 30 and 40 ms durations by deleting parameter lines in the formant transitions which extended beyond the specified values.

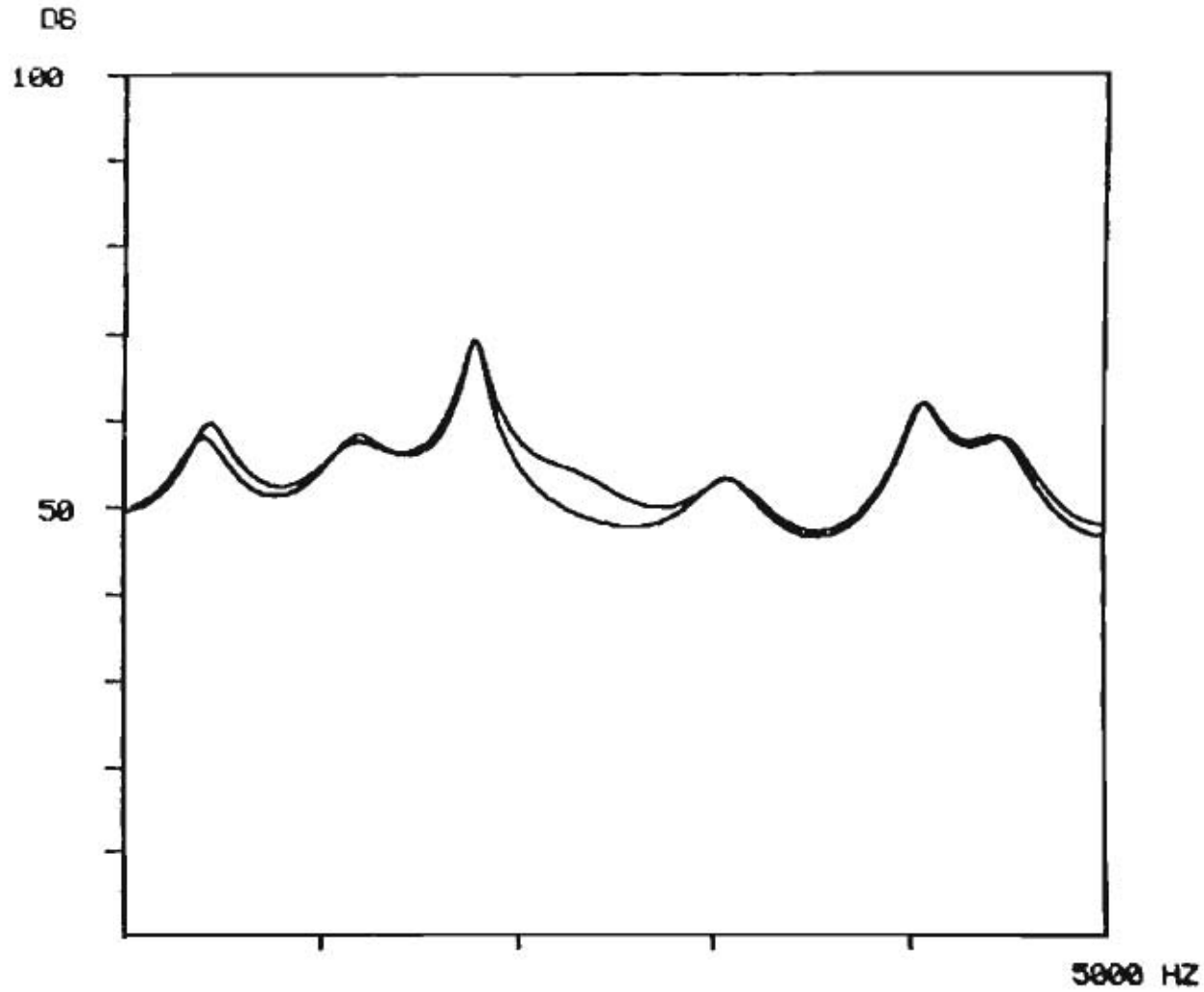


Figure 5. Onset spectra in relative dB for the first 25.6 ms of the natural speech /ga/ and the S+B synthetic /ga/.

Insert Figure 6 about here

Whatever merits or weaknesses there were in the synthesis of the S+B and RS stimuli, the procedures for both sets were developed on the basis of visual spectral matching alone. We should note, incidentally, that in the 20 ms stimuli, the spectral information contributed by the following vowel context was equally represented in both the RS and S+B sets.

B. Procedure

The testing procedures were controlled on-line by a PDP-11/05 computer. Stimuli were output at 10 kHz sampling rate through 12-bit D/A converters, low-pass filtered at 4.8 kHz and then presented over TDH-39 earphones. The output amplitude was calibrated across sessions at a comfortable listening level. Button press responses were collected from up to six subjects at a time and stored on disk. The basic design of the present experiment was intended to follow closely that of Experiment 1, allowing for the appropriate changes from audio tape recordings and written responses to an on-line computer controlled perceptual experiment.

The testing procedures were spread out over two days. Day 1 served to screen subjects audiometrically, and to train and test them with the natural speech stimuli. On Day 2, subjects identified only the synthetic stimuli. Subjects were excluded from the experiment after Day 1 if they identified the natural stimuli essentially at chance, that is to say, if each of the nine natural speech syllables, averaged over all durations, was identified correctly less than 40% of the time.

Responses were collected on a seven button response box with feedback lights. Because voicing was not an experimental variable, the consonant response buttons were labeled "B/P" for bilabial, "D/T" for alveolar, and "G/K" for velar. Three additional buttons were labeled with the confidence rating responses as "Very sure", "Sure" or "Guess". The confidence rating categories were defined as follows on the written instruction sheets:

Confidence Ratings:

++ Very sure the consonant was correctly identified.

+ Reasonably sure the consonant was correctly identified.

- Consonant response represents only a chance guess.

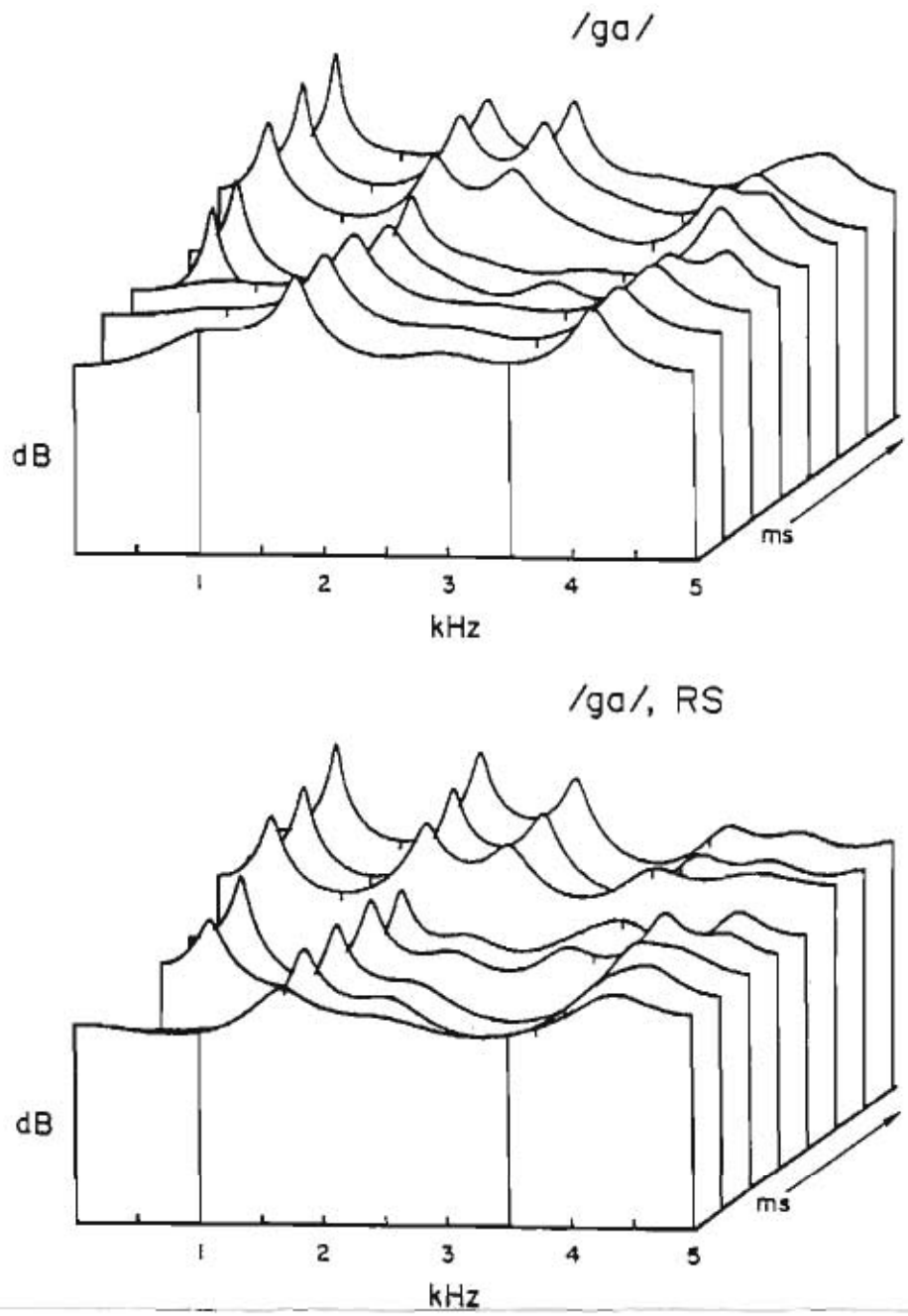


Figure 6. Running spectral displays of the first 4.0 ms of the natural speech /ga/ (top panel) and the RS synthetic speech /ga/ (bottom panel).

Each trial was signaled by a cue light and both a consonant response and a confidence rating were collected for each stimulus on each trial in the experiment.

On Day 1, each subject was screened by an audiometric test for the octave frequencies from 500 to 8000 Hz at a sound pressure level of 20 dB (ANSI-1969) using a Grason-Stadler Model 1701 audiometer. Two familiarization tasks were then conducted using the 27 truncated natural speech stimuli. The first, called "cued familiarization," required listening to all 27 natural stimuli once. In the second familiarization task, two repetitions of each natural stimulus were presented to the subjects randomly with feedback. Subjects were then instructed on how to carry out the identification task and enter their confidence ratings. For the natural set of stimuli, subjects heard five blocks of two repetitions of each stimulus for a total of 10 trials per stimulus. Stimulus presentation was paced to the slowest subject's responses. No feedback was given in this test.

Subjects returning on Day 2 were not told about differences in the nature of the stimuli to be presented. Instructions merely stated, "You will be listening to additional short consonant stimuli one-at-a-time." No familiarization trials were presented on Day 2 and subjects began the identification test directly, having previously listened to only the natural speech stimuli on Day 1. The 27 S+B stimuli and the 27 RS stimuli were fully randomized within one stimulus block of 54 trials. Subjects listened to ten blocks of stimuli. Each subject therefore provided a total of 10 responses to each stimulus. Subjects were obtained through a laboratory subject pool and were paid $3 a day for testing. None of the subjects who participated in Experiment 1 was contacted for this experiment.

C. Results

Of the twenty-one subjects tested, one subject did not pass the audiometric screening test and was dropped from further testing. Ten subjects did not achieve the 40% correct level of performance with the natural stimuli on Day 1 and were asked not to return for testing on Day 2. Thus, data were collected from all three sets of stimuli for 10 subjects, resulting in 100 data points per stimulus.

To assess the contributions of the variables to identification of place of articulation, a four-way analysis of variance over stimulus type, consonant, vowel and stimulus duration was conducted. When appropriate, one-way analyses of variance were calculated using a Scheffe post-hoc analysis at the $p < .05$ level. Probability levels greater than .05 were considered non-significant.

Insert Table II about here

Table II. Percentages of response measures collected in Experiment 2 presented for each stimulus type averaged over vowel and consonant type, and stimulus duration.

Response Measures	Natural Speech	Synthetic, RS	Synthetic, S+B
Correct consonant identification	94	78	68
Guess rating (-)	4	3	11
Sure rating (+)	21	21	26
Very sure rating (++)	75	76	63

A summary of the results obtained for the three stimulus sets is shown in Table II. The first row presents the percent correct consonant identification for each stimulus set averaged over the consonant and vowel types, and stimulus duration. As in Experiment 1, subjects showed high levels of accuracy in identifying place from very brief initial portions of natural stop consonants. The overall percent correct was 94%. Performance levels for both sets of synthetic stimuli fell below that of the natural speech stimuli. Consonants were identified better with the RS stimuli (78% correct) than with the S+B stimuli (68% correct). The stimulus types, natural, RS and S+B, formed three distinct ordered categories in the post-hoc analysis ($F(1,2)=49.74$, $p < .001$). Consonants were identified more accurately for the natural speech stimuli than for the RS synthetic stimuli, and more accurately for the RS stimuli than for the S+B synthetic stimuli, respectively.

The next three rows in Table II summarize the confidence rating results, computed as the percentage of each confidence rating response obtained for each stimulus set. These results show that subjects were equally confident of their responses to the natural and RS stimuli, but were less confident of their responses to the S+B stimuli. Evidently, there were aspects of the natural and RS stimuli that subjects judged as similar to each other, but dissimilar to the S+B stimuli.

Insert Figure 7 about here

The main effects of each of the experimental variables -- consonant type, vowel type and stimulus duration -- on percent correct consonant identification for each stimulus set are shown in Fig. 7. Performance decreased from the natural to the RS and then again from the RS to the S+B stimuli across all variable types but one ($F(2, 8) = 78.28$, $p < .001$). For /b/ syllables, the S+B set was better than the RS set.

The effects of each experimental variable were then examined in more detail. No differences were observed in identification performance among /b/, /d/ and /g/ ($F(2, 8) = 1.05$, NS) nor among the three waveform durations ($F(2, 8) = .88$, NS). For the vowels, differences in performance level were significant ($F(2, 8) = 67.63$, $p < .001$): consonants were identified better before the vowel /a/ (93% correct) than before /u/ (80% correct) or before /i/ (68% correct). Vowel type formed three distinct groups in the post-hoc analysis ($F(1, 2) = 42.26$, $p < .001$). Vowel context, therefore, had an important effect on the identification of consonants. It is interesting that the vowel /a/, often used in synthetic speech perception studies of stop-vowel syllables, provided the most reliable context for stop identification.

The identification results are broken down separately by each experimental variable in the nine panels of Fig. 8. This figure illustrates the major sources of the variation observed in the average identification functions shown in Fig. 7.

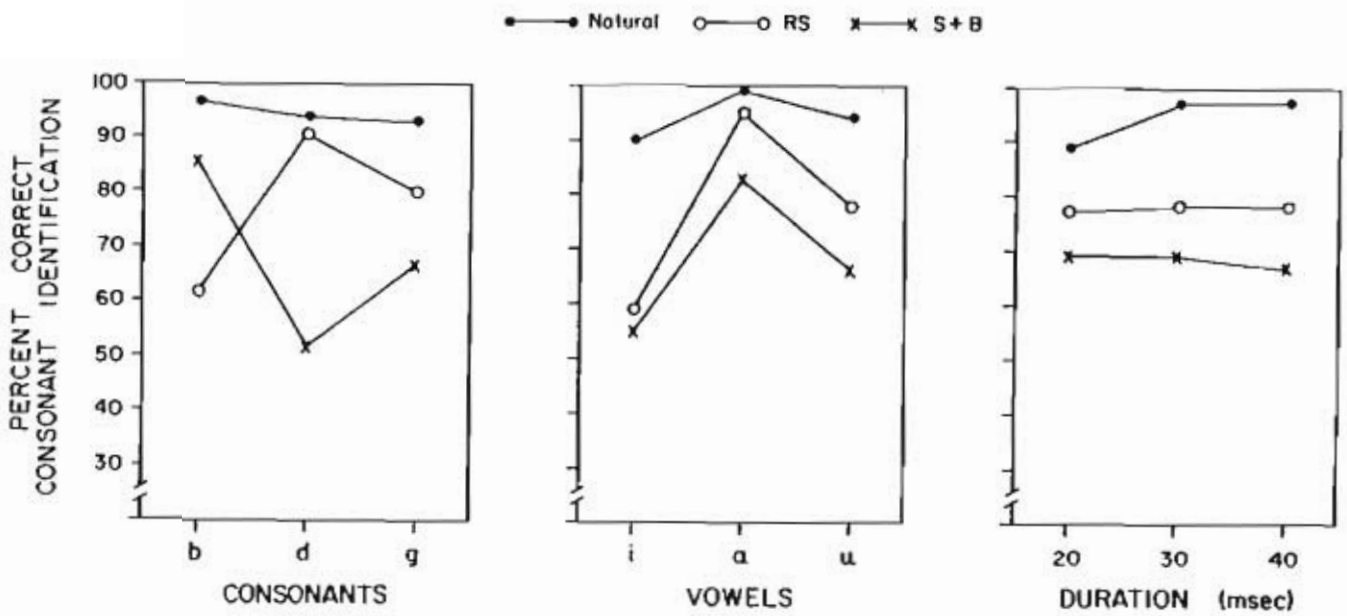


Figure 7. Percent correct consonant identification shown separately for each stimulus type, natural, RS and S+B. Each panel displays results for one independent variable averaged over the other two experimental variables.

Insert Figure 8 about here

The ordered performance levels of the natural, RS and S+B stimuli were preserved for all conditions except for the consonant /b/ and for the syllable /gi/. Figure 8 shows that consonant identification was better for the S+B compared to the RS sets for /bi/ and /bu/, although not for /ba/. A post-hoc spectral analysis of the RS /bi/ and /bu/ stimuli revealed an additional variable previously overlooked in our earlier analyses. Although the spectral tilt and shape of the burst frames were carefully matched, the relative levels of signal energy in the burst frame differed between the natural and RS stimuli: the bursts for RS /bi/ and /bu/ were 6 dB higher than the burst for their natural counterparts. The higher amplitude bursts resulted from a synthesis rule implementing a constant voicing source for all synthesized syllables. This rule, intended to keep the source constant across place, was arbitrary and evidently resulted in an artifact in these stimuli.

Further study of burst energy for five repetitions of talker RP's stops with 8 vowel contexts revealed systematic differences in amplitude for place of articulation. These differences were not represented in either the RS or SB sets of stimuli. We do not, of course, know for certain that place of articulation in RS /bi/ and /bu/ was misidentified because of the unnaturally loud bursts. Nevertheless, this possibility served as part of the motivation for the next experiment.

The effect of vowel context was not uniform across the consonants (see Figure 8). In fact, consonant and vowel types interact ($F(4, 24) = 26.22$, $p < .001$), as do consonant type, vowel type and stimulus set (stimulus X consonant, $F(4, 24) = 36.63$, $p < .001$; stimulus X vowel, $F(4, 24) = 7.07$, $p < .001$; stimulus X consonant X vowel, $F(8, 32) = 10.77$, $p < .001$). Evidently, individual consonant-vowel syllables contributed differentially to the overall effects of the ordered difference in performance levels between the natural, RS and S+B stimuli.

As shown in Fig. 8, variation in waveform durations had a very small effect on identification performance. From Experiment 1, we had expected that several of the natural stimuli (especially /gi/) would show an improvement in consonant identification with stimulus duration. This result was, in fact, observed since /bi/, /du/ and /gi/ all showed increased identification performance with longer duration waveforms. For the synthetic RS and S+B stimuli, however, consonant identification did not improve with longer stimulus durations.

Finally, if the confidence rating scale was reliable, higher confidence ratings should correlate with an increase in correct responses. We examined this relation by calculating the conditional probability of obtaining a correct response, C, for each rating category, Rj, as P(C/Rj). These conditional probabilities are plotted in Fig. 9.

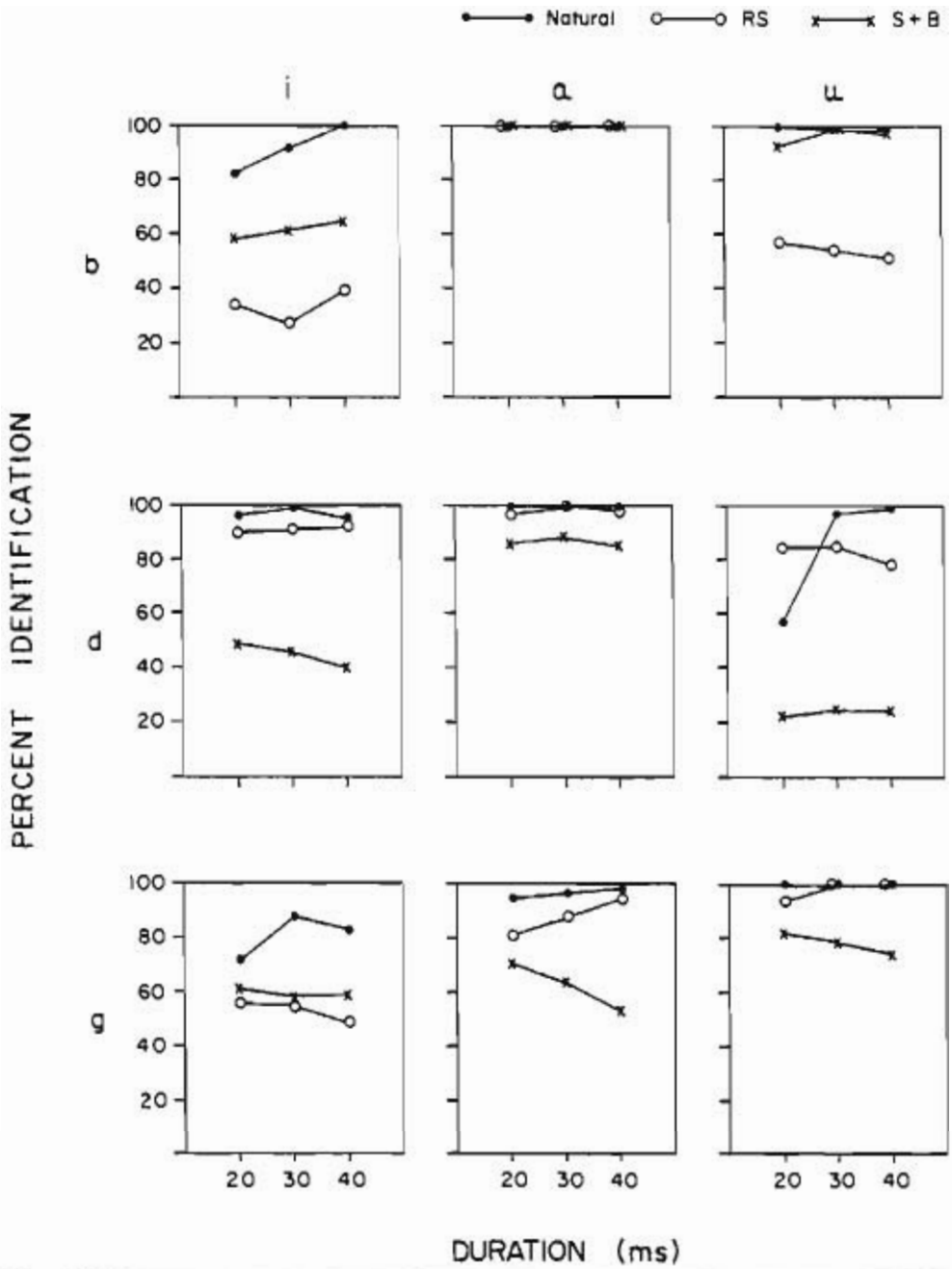


Figure 8. Percent correct consonant identification plotted separately for each experimental variable, stimulus type, stop vowel context, stimulus duration and consonant.

Insert Figure 9 about here

Subjects' confidence ratings were highly correlated with their ability to identify place of articulation correctly in the three sets of stimuli. The conditional probabilities displayed here are, of course, a combined measure of the correct identification of place and the selected confidence rating category. (These measures are also reported separately in Table II.) This combined measure confirms the rank order effects observed across the three stimulus sets. The natural stimuli had higher conditional probabilities than either of the two synthetic stimulus sets. For all confidence rating categories, however, subjects were more likely to identify place of articulation correctly from the RS stimuli than from the S+B stimuli.

D. DISCUSSION

The results of Experiment 2 showed that listeners correctly identified place on 68% of the S+B stimulus trials. Of the nine S+B syllable types examined, only three, /ba/, /bu/ and /da/, were identified better that 85% correct. The remaining six CV's averaged only 55% correct identification. Evidently, subjects cannot reliably identify place of articulation from information contained in only the overall gross shape of the onset spectra of short stop CV waveforms.

This conclusion is strengthened by the comparison between the S+B stimuli and the RS stimuli. Since the RS stimuli were modeled after the same natural speech stimuli as the S+B stimuli, their onset spectra should be good exemplars of the onset spectra for the three places of articulation. A post-hoc comparison of the onset spectra for the RS stimuli and the S+B stimuli showed that the RS onset spectra were, indeed, good place exemplars--in fact, for half of the comparisons the onset spectra for the RS and S+B stimuli were virtually indistinguishable. Thus, both the static onset spectra properties and the dynamic running spectral features are present in the RS stimuli. Since subjects correctly identified place on 78% of the RS trials, they gained 10% in identification on performance from the added dynamically changing place information. Ten percent is not a large overall increase in performance, but the sources of the increase, as shown in Fig. 8, bring the differences between the two stimulus sets into sharper focus. For the three S+B syllables with high levels of identification, (/ba, bu, da/) two of the RS stimuli, /ba/ and /da/, also had high levels of identification. But for the six S+B stimuli with poor levels of identification, four of the RS stimuli showed substantial improvement (/di,du,ga,gu/). In fact, the average identification scores for these six stimuli increased from 55% for the S+B set to 75% for the RS set, a gain of 20%.

Moreover, if we omit RS /bi/ and /bu/ (for which a synthesis error was made) the RS stimuli were identified correctly on 87% of all trials. For half of the CV's, the identification functions for the natural speech and RS stimuli were virtually indistinguishable (see Fig. 8). Since the synthesis procedures for the

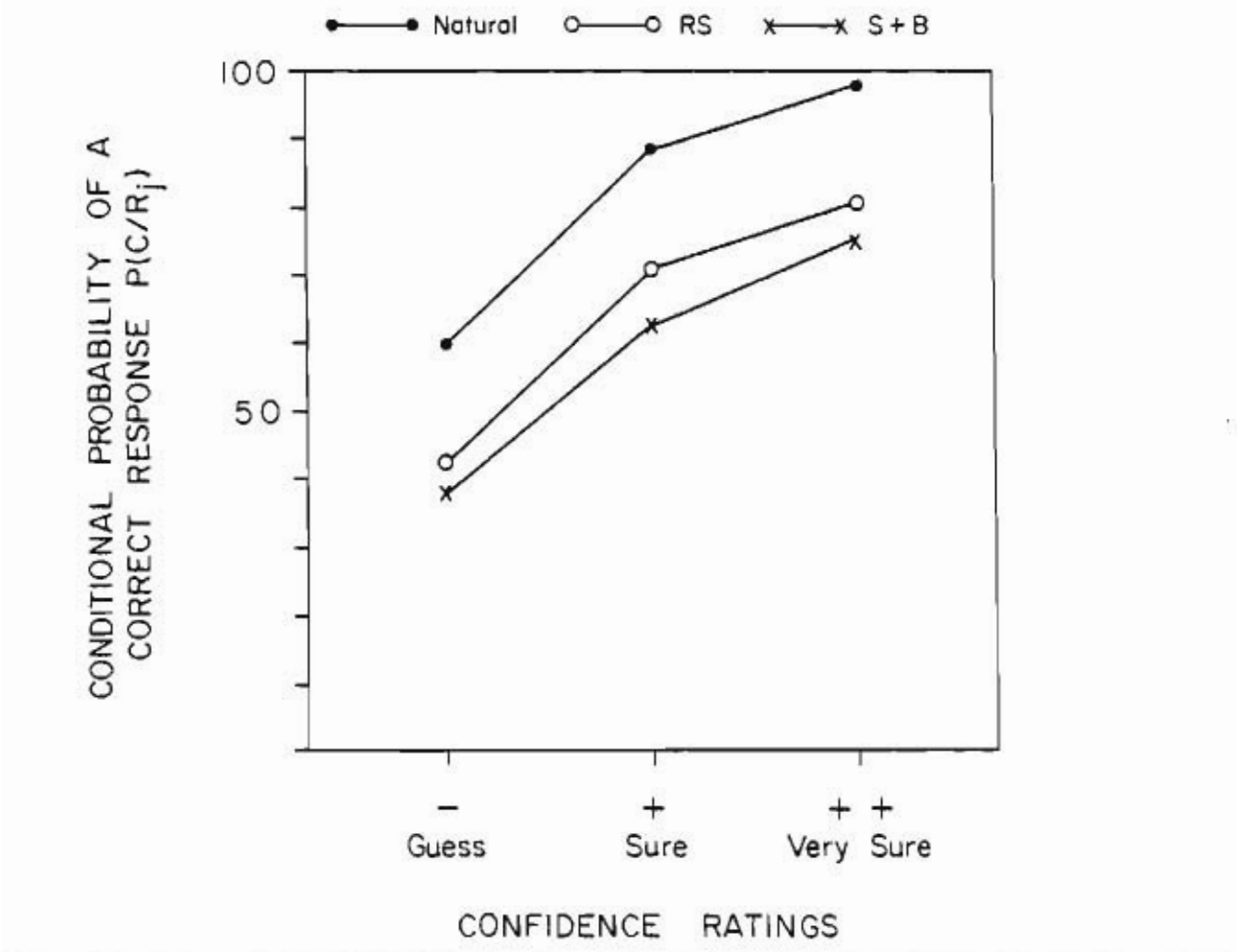


Figure 9. Conditional probability of a correct response obtained for each confidence rating category plotted separately by stimulus type, natural, RS and S+B averaged over stop consonants, vowels, and duration.

RS stimuli were primarily rule governed rather than based on frame-by-frame spectral matching, the difference of 7% between 87% for the RS stimuli and 94% for natural stimuli suggests that the synthesis rules did a reasonably good job of capturing the dynamic acoustic information that specifies place of articulation in these syllables.

Finally, the confidence ratings as shown in Table II establish that subjects were equally confident about identification of the natural and RS stimuli, but were less confident of their responses to the S+B stimuli. Since the natural stimuli and the RS stimuli preserve both the presumed properties of the static onset spectra and the fine temporal details of the early portions of the waveform, while the S+B stimuli preserve only the static properties of the shape of the onset spectra, the grouping of the confidence ratings would seem to reflect the gain of a dynamic over a static description of stop consonant onset.

III. EXPERIMENT 3: IDENTIFICATION OF SHORT LPC SYNTHESIZED CV'S

The previous experiment tested the adequacy of onset spectra as acoustic cues to place of articulation in stop-vowel syllables. While the outcome was clear, the experiment could be criticised on two counts. First, the visual matching procedure employed in synthesis might be subject to experimenter bias. Second, unnatural relative levels of energy present in the bursts of all the synthetic stimuli might have influenced the results. The present experiment was designed to replicate the previous experiment using procedures designed to counter these criticisms.

Two sets of stimuli were synthesized from linear prediction coefficients derived from the same natural stimuli as used in Experiment 2. One set was synthesized from the coefficients specifying only the onset spectra as defined by Stevens and Blumstein: spectral information was constant throughout the stimulus. These stimuli will be referred to as the onset spectra (OS) stimuli. The other set was synthesized with the linear prediction coefficients used to produce running spectra (see Fig. 6). These coefficients were updated every 5 ms. These synthesized stimuli will be referred to as the time-varying (T-V) stimuli.

In linear prediction synthesis, fundamental frequency and overall amplitude are controlled independently of the analysis coefficients. For both the OS and T-V stimuli, the fundamental frequency was set to 100 Hz. The relative amplitude of both sets was adjusted so that the energy present in the first 20 ms of each syllable in the three stimulus sets was equal. Thus, these stimuli were synthesized by computer algorithms quite independently of the experimenter except for matching RMS energy in the first 20 ms. Furthermore, using this procedure any effect that burst energy might have as an acoustic cue to place of articulation would be held constant across all stimulus types.

In order to test the major difference between Stevens and Blumstein's (1978) onset spectra account and Kewley-Port's (1982b) time-varying feature analysis, we will focus our analyses on the 20 ms stimuli. The onset spectra, burst energy and vowel context information was the same for each 20 ms CV in the natural, OS and T-V sets. (In contrast, the 40 ms natural and T-V sets preserved more vowel

context information than the 40 ms OS set.) Thus, the primary difference between the two, 20 ms synthetic stimulus sets, was the static or dynamic quality of the spectral information used to specify place of articulation.

A. Stimuli

Since there was no overall improvement in consonant identification over the three stimulus durations examined in Experiment 2, only the 20 and 40 ms stimuli were used in this experiment. Otherwise, the same natural CV syllables used in Experiment 2 were used here.

The linear prediction coefficients previously calculated for synthesizing the S+B and RS stimuli in Experiment 2 were used again in this experiment for synthesizing the OS and T-V stimuli. A program (MODSYN) was written by the first author for synthesizing waveforms from a set of reflection coefficients. Arbitrary pitch and amplitude parameters could also be input. The TWOMVL algorithm of Markel and Gray (1976, sec. 5.5.2) was implemented with a pulse generator as the pitch source or a pseudo-Gaussian random number generator as the frication source to synthesize the waveforms. An OS stimulus was first synthesized from the coefficients for the onset spectra along with a gain estimate and a pitch of 100 Hz. After synthesis, the energy in the first 20 ms of the stimulus was calculated. If this energy did not match the calculated energy in the first 20 ms of the natural syllable within 1 dB, the input gain was adjusted appropriately and the OS stimulus was resynthesized. The energy in the first 20 ms of the natural syllables relative to the 12 bit waveform was measured as: /bi/ = 77 dB, /ba/ = 78 dB, /bu/ = 72 dB, /di/ = 68 dB, /da/ = 72 dB, /du/ = 67 dB, /gi/ = 61 dB, /ga/ = 61 dB, and /gu/ = 63 dB. The T-V stimuli were synthesized from the coefficients of the running spectra at a 5 ms frame rate with a relative gain value estimated algorithmically for each frame. The same VOT values used in Experiment 2 were employed here. Voiceless frames used the fricative source while voiced frames were synthesized with a pitch of 100 Hz. When necessary, the first 20 ms of energy in each T-V stimulus was set to match the natural stimulus by adjusting the relative gain for all frames by a fixed amount.

B. Procedure

The procedure of Experiment 3 was identical to that of Experiment 2 except that one-third fewer stimuli were presented because the 30 ms duration was dropped. Both identification and confidence rating responses were collected over two days of testing. Subjects were obtained through a laboratory subject pool and paid $3.50 a day. None of the subjects who participated in the first two experiments was contacted for this study.

C. Results and Discussion

All seventeen subjects participating in the experiment passed a screening audiometric test. One failed to return on the second day, and six others did not achieve the 40% correct level of performance on the natural speech stimuli. Thus, 10 subjects participated in both days of testing, resulting in 100 data points per stimulus. The same statistical analyses performed in the previous experiment were used here.

Table III summarizes the results. The findings are similar to those of Experiment 2, shown in Table II. Consonants in natural CV stimuli were again identified 94% correctly for place, while the synthesized stimuli were identified correctly less often. As before, the stimulus sets formed three distinct groups with the natural stimuli identified most accurately, T-V stimuli identified less accurately (87%) and the OS stimuli even less accurately (59%) ($F(1,2) = 80.0$, $p < .0001$). The confidence ratings were also similar to those previously collected. Subjects rated the natural speech stimuli and T-V stimuli at roughly the same high levels of confidence, while the OS stimuli were rated at a lower overall level of confidence.

The main effects of each of the stimulus variables in terms of percent correct consonant identification are shown in Fig. 10. In these data, the main result of decreasing performance levels of identification from natural to T-V and from T-V to OS stimuli was preserved across all the variable types ($F(2,7) = 142.4$, $p < .001$). Thus, the results of Experiment 3 clearly replicate the main stimulus manipulations found in Experiment 2.

Insert Table III and Figure 10 about here

We can compare the overall main effect of stimulus type across Experiments 2 and 3 directly in the following way. In both experiments, stimulus type was described as natural, dynamic or static. The natural stimuli presented in both experiments were in fact the same signals. The dynamic stimuli refer to the RS (running spectra) stimuli in Experiment 2 and the T-V (time-varying) stimuli in this experiment. Both sets preserved the dynamic variation in the stop-vowel acoustic structure. The S+B (Stevens & Blumstein) and OS (onset spectra) stimuli are called static because they preserved the same onset spectral information, but differed primarily in energy level and the presence or absence of an F1 transition. A two-way analysis of variance of the three stimulus types across both experiments showed that the main effect of stimulus type in the two experiments was indeed very similar. That is, the identification of the three stimulus categories was significantly different ($F(2,3) = 113.6$, $p < .0001$), while results between experiments were essentially the same ($F(1,3) = .002$, NS).

While the identification of the natural stimuli in both experiments was clearly the same (94%), there were several differences between the synthetic stimulus sets. The dynamic T-V set showed a significant 9% improvement over the

Table III. Percentage of response measures collected in Experiment 3 for each stimulus type averaged over vowel and consonant type, and stimulus duration.

Response Measures	Natural Speech	Synthetic Time-Varying (T-V)	Synthetic Onset Spectra (OS)
Correct Consonant			
identification	94	87	59
Guess rating (-)	9	5	30
Sure rating (+)	23	28	37
Very sure rating (++)	68	67	33

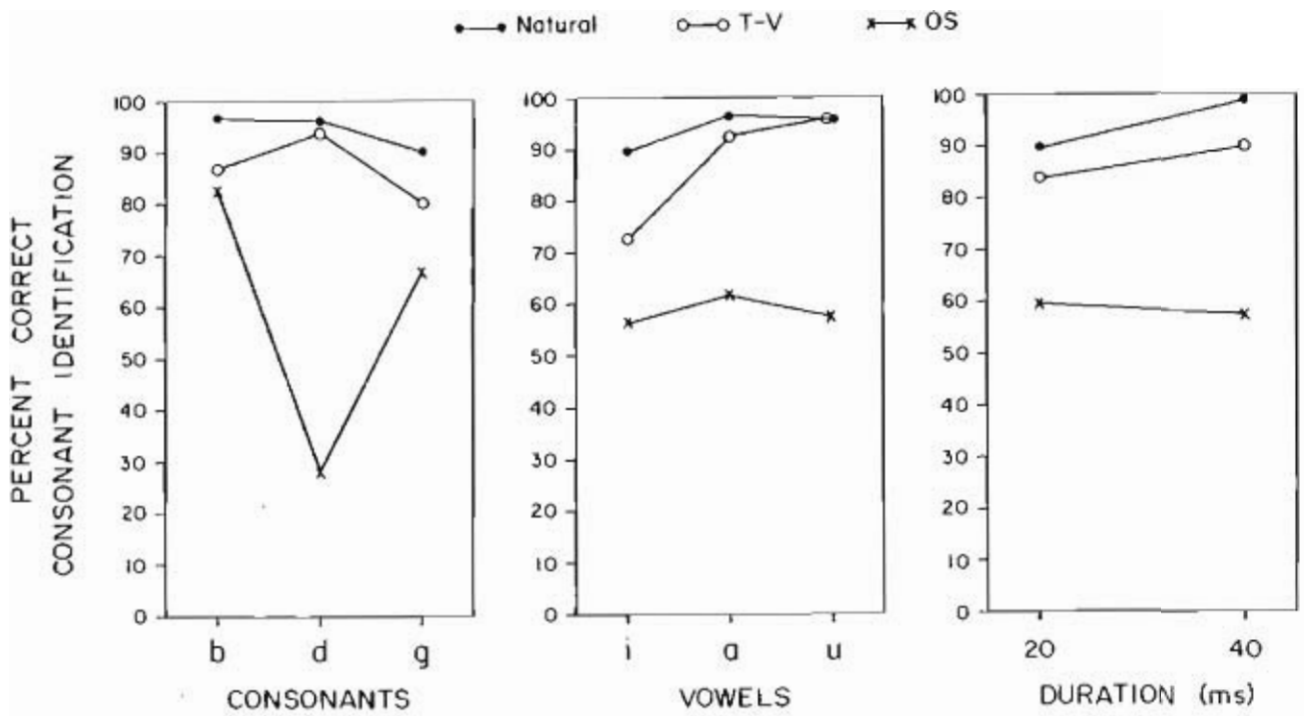


Figure 10. Percent correct consonant identification shown separately for each stimulus type, natural, T-V and OS. Each panel displays results for one independent variable averaged over the other two experimental variables.

RS set (t = 3.06, p <.002). Comparing Fig. 7 with Fig. 10, we can see that the main reason for this improvement is the better identification of bilabials for the T-V stimuli. This result confirms our previous hypothesis that unnatural burst amplitudes in the synthetic RS stimuli in Experiment 2 depressed performance for bilabial identification. Another improvement in the dynamic stimuli in Experiment 3 was an increase in consonant identification from the 20 to the 40 ms stimuli. This increase, shown in Fig. 10, parallels the improvement in the natural stimuli, giving rise to an overall significant effect of duration in the experiment ($F(1,7)$ = 4.9, p < .05). Taken together, the effects of the main variables on the natural and dynamic stimuli in Experiment 3 are more similar than they were in Experiment 2. In other words, the linear prediction synthesis with natural burst amplitudes produced more natural sounding stimuli than the parallel resonance synthesis procedures.

Insert Figure 11 about here

By contrast, the static OS stimuli were identified significantly less well, by roughly 9%, than the S+B stimuli of Experiment 2 (t = 2.49, p < .02). The source of this decrease in performance becomes evident when we break down percent correct consonant identification separately for each experimental variable in the nine panels of Fig. 11. The results from the two experiments, as shown in Figures 8 and 11, are similar, apart from the already mentioned differences in the main variables. One other striking difference appears for the /ba/ and /da/ static onset spectra stimuli: correct identification drops by an average of 36% across experiments. The primary difference between these stimulus sets was the presence of F1 transitions in the S+B stimuli but not the OS stimuli. The pilot study carried out in connection with Experiment 2 showed that of the nine syllables investigated, most improvement in identification of the F1 transition stimuli was obtained for the syllables /ba/ and /da/. The present findings confirm the perceptual importance of the F1 transition in identification of the consonant in these syllables.

The results displayed in Fig. 11 also reveal that the 7% decrease in identification between the natural and T-V stimuli was largely due to poor identification of two T-V syllables, /bi/ and /gi/. In strong contrast, a large difference (35%) between the OS stimuli and the natural stimuli was observed for 6 of the 9 syllables (all but /ba/, /bu/ and /ga/). Thus, for most of the short syllables, linear prediction synthesized stimuli preserving detailed waveform structure were identified at roughly the same levels as the natural syllables. Synthesized syllables preserving only the static onset spectra information, however, were identified much more poorly overall than the natural stimuli.

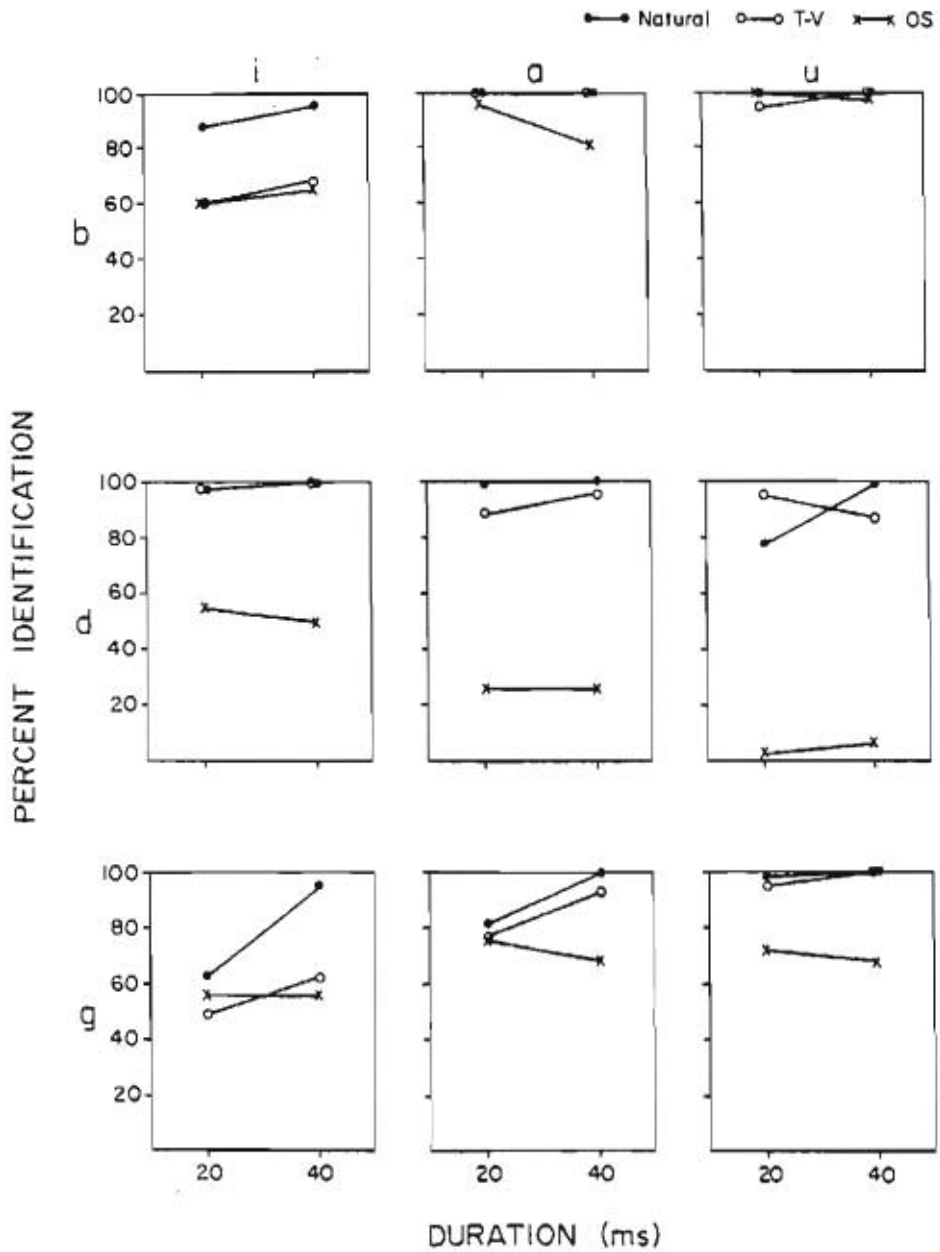


Figure 11. Percent correct consonant identification plotted separately for each experimental variable, stimulus type, stop vowel context, stimulus duration and consonant.

Insert Table IV about here

The conditional probabilities, P(C/Rj), of obtaining a correct consonant identification for a given confidence rating were also calculated for each stimulus type in this experiment. The relation of the P(C/Rj) values across stimulus types for this experiment was the same as that observed in the previous experiment (Fig. 9). Therefore, the values for this experiment are presented only in tabular form in Table IV. Subjects again reported high levels of confidence in their responses to the natural and dynamic T-V stimuli, but relatively lower levels of confidence to the static (OS) stimuli.

Earlier we noted that the 20 ms stimuli in this experiment can be used to evaluate the predictions derived from the hypotheses of Stevens and Blumstein (1978) and Kewley-Port (1982b) concerning acoustic cues for place of articulation. Each 20 ms syllable in the three stimulus sets had the same onset spectra. For the natural and T-V stimuli, the additional temporal variations in spectral energy from the burst into the voicing onset were also present. Although a 20 ms stimulus is unusually brief, our subjects indicated they were much less confident overall of their identification of the static OS stimuli than of the spectrally changing stimuli (for example, 28% guesses for 20 ms OS stimuli compared to 10% guesses averaged over natural and T-V 20 ms stimuli). Thus, subjects clearly perceived a difference in quality between the spectrally static and dynamically changing stimuli, even though they were only 20 ms in duration.

We also examined whether there were any differences in the identification of 20 ms stimuli from each stimulus type due to differences in place of articulation. A two-way analysis of variance for the 20 ms stimuli alone showed significant differences for both stimulus type and place of articulation ($F(2,4) = 31.5$, $p < .001$ for stimulus; $F(2,4) = 11.0$, $p < .001$ for place). Identification of the natural and T-V stimuli were grouped together in the post-hoc analysis, while the OS stimuli were separately categorized ($F(1, 2) = 24.1$, $p < .001$). (Percent correct identification for natural = 90%, for T-V = 84% and for OS = 60%.)

The Stevens and Blumstein onset spectra hypothesis would predict equal levels of identification for all three places of articulation, or perhaps /g/ might be worse than /b/ and /d/, if it takes longer time to build up an unambiguous /g/ onset spectrum (cf. Blumstein and Stevens, 1980, p. 652). The time-varying feature analysis of Kewley-Port (1981b) predicts that identification of 20 ms stimuli should be similar and high for /b/ and /d/, but somewhat lower for /g/. This prediction was confirmed by the data shown in Fig. 12. Figure 12 shows percent change in consonant identification over stimulus duration in separate panels for each stimulus type. Percent identification of /b/ and /d/ were not different for either the natural ($t = .7$, NS) or T-V ($t = 1.34$, NS) stimuli using a one-way analysis of variance with a priori contrasts; therefore, these values were averaged together and plotted using the b-d symbol on Fig. 12. Further analysis confirmed that /b/ and /d/ were identified correctly more often

Table IV. Probability of obtaining a correct response when a confidence rating response was chosen, P(C/Rj), for each stimulus type.

	Stimulus Type		
Confidence Ratings	Natural	T-V	OS
Guess rating (-)	71	63	46
Sure Rating (+)	90	83	62
Very Sure Rating (++)	97	91	69

than /g/ for both the natural ($t = 2.63$, $p < .01$) and T-V ($t = 2.57$, $p < .01$) sets. Thus, for the spectrally changing T-V stimuli, our predictions were fully supported.

Insert Figure 12 about here

While our account makes no prediction about the steady-state OS stimuli, Stevens and Blumstein would presumably expect the same pattern of results. However, these predictions were not confirmed for the OS stimuli: /b/ was identified significantly better than /d/ ($t = 7.59$, $p < .0001$), but identification scores averaged over /b/ and /d/ were lower than for /g/. Furthermore, /b/ and /g/ were grouped together in a post-hoc analysis (average identification 76%). /d/ was identified correctly much less often (28%) ($F(1, 2) = 30.4$, $p < .001$). Thus, predictions derived from Stevens and Blumstein's hypotheses concerning the relative levels of identification of the three places of articulation were not supported.

As previously mentioned in the discussion of Experiment 1, Kewley-Port's (1982b) hypothesis of time-varying features predicts that identification of velar place of articulation should improve to relatively high levels with an increase in duration from 20 to 40 ms, while performance for /b/ and /d/ should show little change. On the other hand, since 20 ms velars were poorly identified in this experiment, Stevens and Blumstein would presumably predict that velar identification should improve at a 40 ms duration, as the compact shape of the velar onset spectra becomes more salient. These predictions were tested by analysis of variance. In a three-way analysis of place of articulation (/b/ vs /d/), stimulus duration and stimulus type (natural vs. T-V), there was a small, though significant, difference of 4% in identification between the natural and T-V stimuli ($F(1,3) = 6.84$, $p = .01$), but no significant difference for place of articulation ($F(1,3) = 1.61$, NS) or for duration ($F(1,3) = 2.70$, NS) (see in Fig. 12). In a separate analysis of stimulus duration by stimulus type for /g/, there was a significant improvement in identification for the longer stimuli ($F(1,2) = 7.75$, $p < .006$), but no difference between the natural and T-V stimuli ($F(1,2) = 3.99$, NS). These results are entirely consistent with the time-varying features hypothesis. By contrast, the prediction derived from Stevens and Blumstein's approach of improved velar identification with increased duration was not supported for the onset spectra stimuli, since the correct identification for OS /g/ stimuli fell slightly from the 20 to 40 ms durations (Fig. 12). We conclude that a static representation of onset spectra in stop consonants is clearly less successful in characterizing the acoustic cues to place of articulation than a time-varying, running spectrum. This outcome is in good overall agreement with the results of the previous experiments.

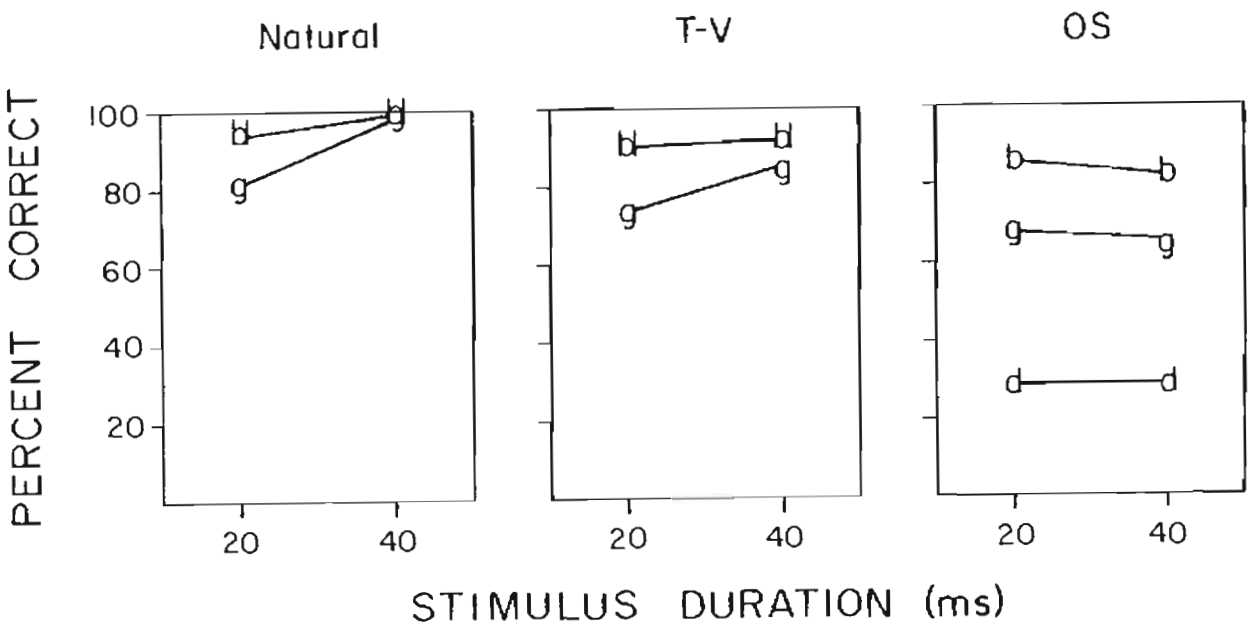


Figure 12. Percent correct consonant identification by stimulus duration displayed separately for each stimulus type. Consonant responses for /b/ and /d/ were averaged together for the natural and T-V stimulus types.

IV. GENERAL DISCUSSION

For Stevens and Blumstein (1978, 1981) the "primary" and invariant correlates of place of articulation are located in a single, static spectral section at syllable onset. "Secondary" correlates are to be found in the time-varying, context-dependent formant transitions (Blumstein and Stevens, 1979, p. 1015). In fact, formant transitions are said merely to "provide the acoustic material that links the transient events at the onset to the slowly varying spectral characteristics of the vowel" (Blumstein and Stevens, 1980, p. 660). This point of view is essentially identical to that of Cole and Scott (1974). However, the theoretical arguments of Fant (1968, 1973) and the results of the running spectral analysis presented earlier by Kewley-Port (1982b) demonstrate that the distinction between primary, static properties and secondary, dynamic properties is arbitrary and empirically unjustified (cf. also the perceptual study by Walley and Carrell, 1982).

Moreover, the conclusion to be drawn from the present experiments is that, while invariant acoustic correlates of place of articulation may indeed be found at syllable onset, they are not adequately described by static spectral sections. Rather, the primary perceptual correlates are time-varying spectral properties that reflect the movements of articulatory release--whether the rapid release typical of labial and alveolar stops or the slower release typical of velars (Fant, 1968, p. 223).

More generally, the present experiments demonstrate that to identify the invariant with the static, and the dynamic with the contextually-dependent, is false. If the spectral correlates of a phonetic segment derive from an articulatory gesture, and if the essence of gesture is structural change, then the spectral correlates must reflect that change. But, of course, patterns of change may be invariant across a variety of contexts, so that the recognition of the dynamics of articulation does not imply lack of invariance in the acoustic signal.

ACKNOWLEDGEMENTS

We are grateful to Katherine S. Harris, Dennis H. Klatt and Lawrence J. Raphael for their comments on earlier versions of this work. Some of these experiments were presented before the Society at the 100th meeting in Los Angeles, CA and the 102nd meeting in Miami Beach, FL. This research was supported by the National Institutes of Health, Research Grant NS-12179 and the National Institute of Mental Health, Research Grant MH-24027 to Indiana University in Bloomington, IN, and, in part, by the National Institute of Child Health and Development, Research Grant HD-01994 to Haskins Laboratories, New Haven, CT.

REFERENCES

Blumstein, S. E., and Stevens, K. N. (1979). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. 66, 1001-1017.

Blumstein, S. E., and Stevens, K. N. (1980). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am., 67, 648-662.

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

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

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

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

Dorman, M. F., Studdert-Kennedy, M., and Raphael, L. J. (1977). "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues," Percept. Psychophy. 22, 109-122.

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

Fant, G. (1968). "Analysis and synthesis of speech processes," in Manual of Phonetics edited by B. Malmberg (North-Holland, Amsterdam), 173-277.

Fant, G. (1973). "Stops in CV-syllables," in Speech Sounds and Features edited by G. Fant. (MIT, Cambridge, MA), pp. 110-139.

Fischer-Jorgensen, E. (1972). "Tape cutting experiments with Danish stop consonants in initial position," Annual Report VII, Institute of Phonetics, University of Copenhagen, 104-175.

Halle, M., Hughes, G. W. and Radley, J. P. A. (1957). "Acoustic properties of stop consonants," J. Acoust. Soc. Am., 29, 107-116.

Kewley-Port, D. (1978). "KLTEXC: Executive Program to implement the KLATT software speech synthesizer," Res. Speech Percept. Prog. Rep. No. 4, Dept. of Psychol., Indiana University, 235-246.

Kewley-Port, D. (1979). "SPECTHUM: A program for analyzing the spectral properties of speech," Res. Speech Percept.: Prog. Rep. No. 5, Dept. of Psychol., Indiana University, 475-492.

Kewley-Port, D. (1980). "Representations of spectral change as cues to place of articulation in stop consonants," Res. Speech Percept. Tech. Rep. No. 3, Dept. of Psychol., Indiana University.

Kewley-Port, D. (1982a). "Measurements of formant transitions in naturally produced stop consonant-vowel syllables," J. Acoust. Soc. Am., in press.

Kewley-Port, D. (1982b). "Time-varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am., in press.

Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am., 67, 971-995.

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

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954). "The role of consonant-vowel transitions in the perception of the stop and nasal consonants," Psychol. Monogr. 68 (8, Whole No. 379), 1-13.

Liberman, A. M., and Studdert-Kennedy, M. (1978). "Phonetic Perception," in Handbook of Sensory Physiology: Perception, edited by R. Held, H. Liebowitz and H. L. Teuber (Springer-Verlag, New York) pp. 143-179.

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

Markel, J. D., and Gray, A. H. (1976). Linear Prediction of Speech (Springer-Verlag, New York).

Schroeder, M. R, Atal, B. S. and Hall, J. L. (1979). "Optimizing digital speech coders by exploiting masking properties of the human ear," J. Acoust. Soc. Am., 66, 1647-1652.

Stevens, K. N. (1975). "The potential role of property detectors in the perception of consonants," in Auditory Analysis and Perception of Speech, edited by G. Fant and M. A. A. Tatham (Academic Press, New York), 303-330.

Stevens, K. N., and Blumstein, S. E. (1978). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. 64, 1358-1368.

Stevens, K. N. and Blumstein, S. E. (1981). "The search for invariant acoustic correlates of phonetic features," in Perspectives on the Study of Speech edited by P. D. Eimas and J. Miller, (Lawrence Erlbaum Associates, Hillsdale, NJ) pp. 1-38.

Stevens, K. N., and House, A. S. (1956). "Studies of formant transitions using a vocal tract analog," J. Acoust. Soc. Am. 28, 578-585.

Tekieli, M. E. and Cullinan, W. L. (1979). "The perception of temporally segmented vowels and consonant-vowel syllables," J. Speech Hear. Res., 22, 103-121.

Walley, A. C., and Carrell, T. D. (1982). "Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants," J. Acoust. Soc. Am., in press.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

Perceptual Evaluation of Voice Response Systems: Intelligibility, Recognition & Understanding*

David B. Pisoni

Speech Research Laboratory
Psychology Department
Indiana University
Bloomington, Indiana 47405

*This paper was presented at the Workshop on Standardization for Speech I/O Technology, National Bureau of Standards (ICST) Gaithersburg, Maryland, March 18-19, 1982. A version is to appear in the proceedings of the conference.

INTRODUCTION

Within the next few years, we can expect to see an extensive proliferation of very sophisticated speech processing devices in various types of man-machine communication systems. These speech processing devices will affect almost every aspect of our daily lives from voice warning and feedback systems in automobiles and household appliances to voice data entry systems using conventional telephone lines. At the present time, there is already a fairly wide diversity of low-cost consumer-oriented products on the market. And, it will not be long before much more specialized devices such as low-cost reading machines for the blind, speaking aids for the deaf and voice-controlled instructional systems will be widely available to a large number of people who have special needs.

Although the field of voice technology has progressed at a substantial rate over the last few years, the same observation does not apply as easily to questions concerning how human observers process (i.e., perceive, encode and interpret) synthetic speech. At the present time, there is relatively little known about the perception and comprehension of synthetic speech, particularly synthetic speech generated automatically by rule. Moreover, and perhaps even more importantly, there has been little if any basic research on the processing of synthetic speech by human observers under conditions of information overload; that is, under conditions where the listener's attentional demands are severely constrained by the presentation of multiple inputs from various sensory modalities such as those encountered, for example, in an aircraft cockpit, a complex command-control environment or computer-assisted instructional application.

In the sections below, I will summarize some of the recent efforts we have made in evaluating the quality of the speech generated by one particular text-to-speech system, the MITalk system (Allen, 1981). Our evaluation surveyed several domains including detailed examination of phoneme intelligibility, word recognition in sentences and comprehension of relatively long passages of fluent connected speech. More recent work has looked at more detailed questions surrounding the mental processing time required to initiate manual responses to synthetic speech. Finally, in another series of experiments, we have examined the cognitive processes used to perceive and remember synthetic and natural speech under conditions of information overload where the observer is required to carry out several complex cognitive tasks at the same time.

BASIC CONSTRAINTS ON PERCEPTION

It seems appropriate at this point to summarize a few preliminary assumptions about the way cognitive psychologists have looked at the performance of human observers in various types of information processing tasks. In any context, whether it be a laboratory experiment or a practical application requiring voice I/O, we assume that an observer's performance reflects the operation of three basic constraints: (1) limitations on the human observer, (2) constraints on the speech signal and (3) the specific task demands.

The first constraint deals with the limitations of the human information processing system to perceive, encode, store and retrieve information presented to the sensory modalities. Because the nervous system cannot maintain all aspects of sensory stimulation and must integrate energy over time, very severe processing limitations have been found in the capacity to encode and store raw sensory data. To circumvent these capacity limitations, sensory information must be recoded or transformed into more abstract forms for storage in memory and

subsequent processing operations (Lindsay & Norman, 1977). The bulk of research on this problem has identified short-term memory (STM) as the major source of the limitation on processing sensory input (Shiffrin & Schneider, 1977). The amount of information that can be processed in and out of working memory is severely limited by the subjects' attentional state, his past experience and the quality of the stimulus display presented to him.

The second constraint concerns the structure of the speech signal. Speech signals represent the physical manifestation of a complex and hierarchically organized system of linguistic rules that, on the one hand, map sounds on to meanings, and on the otherhand, map meanings back on to sounds. At the lowest level in this system, the distinctive properties of the signal are constrained by vocal tract acoustics and articulation. The choice and arrangement of speech sounds in words is constrained by the phonological rules of language. In turn, the arrangement of specific words in sentences is constrained by the syntax. Finally, the meaning of individuals words and the overall meanings of sentences is constrained by semantics. The contribution of these various sources of knowledge to perception will vary substantially from isolated words to sentences to passages of fluent continuous speech.

Finally, the third constraint deals with the specific task demands confronting the observer. Humans are unusually flexible in their ability to develop specialized perceptual and cognitive strategies that can maximize performance under differing task demands. There is now substantial behavioral data in the literature demonstrating the powerful effects of perceptual set, the role of differential instructions, subject expectancies and the influence of long-term familarity, and practice on a variety of perceptual and cognitive tasks. These studies demonstrate that observers are capable of varying the "depth of processing" that a stimulus receives depending on the requirements of the task. In short, human observers are capable of adopting quite different strategies depending on the needs and requirements of the information processing tasks presented to them.

It should be clear from the brief description that an observer's performance in any psychological experiment or for that matter any practical application, will necessarily entail the interaction of all three types of constraints, although to different degrees depending on the specific listening conditions. In the sections below, I will summarize several recent studies that have attempted to make use of these differential constraints on performance to develop appropriate experimental methodologies for evaluation of synthetic speech over a wide range of task demands.

INITIAL PERCEPTUAL EVALUATION OF MITALK

Several years ago we carried out an evaluation of the intelligibility and comprehension of the synthetic speech produced by MITalk: The MIT Unrestricted Text-to-speech System (Pisoni & Hunnicutt, 1980). As the ten year effort to build an unrestricted text-to-speech system drew to a close at MIT, it seemed appropriate to carry out a perceptual evaluation of the quality of the speech output produced by the system in order to establish an initial benchmark level of performance. In addition, the results of such an evaluation study would also prove valuable in uncovering any serious problems in the design and current

implementation of the system and in designing methodology and experimental procedures for evaluating the speech output of other voice response systems.

Although we wanted to obtain an objective measure the segmental intelligibility of the speech output produced by the system, we were also interested in finding out how well naive listeners could understand continuous speech produced by the system. This was thought to be an important aspect of the evaluation of total system performance since a version of the current system was to be implemented as a functional reading machine for the blind (Allen, 1973). In carrying out the evaluation, we patterned several aspects of the testing after earlier work done on the Haskins Laboratories reading machine project (Nye & Gaitenby, 1973, 1974). However, we also added several other tests to gain additional information about word recognition in meaningful sentences and listening comprehension for passages of fluent continuous text. These were designed to provide information about: (1) phoneme recognition, (2) word recognition in sentences and (3) listening comprehension. The results summarized in this section are based entirely on measures of performance accuracy, and, as such, they provide only relatively gross estimates of the differences in perception between natural and synthetic speech. In the next section, I will describe several more recent experiments that use more sensitive behavioral measures to examine processing differences in perception of synthetic and natural speech.

Phoneme Recognition

The Modified Rhyme Test (MRT) was used to measure segmental intelligibility. This perceptual test was chosen primarily because it is reliable, shows little effect of learning and is easy to administer to naive listeners. Seventy-two naive undergraduate students served as listeners. They listened to a single isolated monosyllabic English word on each test trial and then identified the test item from among six possible response alternatives. The test consisted of 50 trials. All stimuli were presented over high-quality headphones at a comfortable listening level.

Overall performance on the synthetic speech showed only 6.9% errors. Performance was somewhat better for consonants in initial position (4.6%) than final position (9.3%). Analysis of the response errors across manner classes showed systematic differences in the distributions of errors for consonants in intial and final position. While the overall level of performance for the synthetic speech is lower than that obtained using natural speech, a difference of 6.3%, the level of performance appears to be high enough for many text-to-speech applications where higher-level linguistic context would be expected to contribute to processes of word recognition and sentence comprehension.

It is also possible that performance for the synthetic speech was inflated by the use of the forced-choice testing format of the MRT test. The amount of uncertainty in a perceptual test is substantially reduced when the subject has to choose from among only six response alternatives that are presented to him before each test trial than when the number of response alternatives is unlimited. How well would subjects do with synthetic speech if the number of response alternatives was effectively the subjects' entire lexicon? To answer this question we ran two additional groups of listeners with an open, free-response

format. As before, subjects heard a single word on each trial. However, they were required to write down whatever word they heard and to the guess if necessary. The error rate for the natural speech increased only slightly from 0.6% to 2.8%. However, the error rate for the synthetic speech increased much more dramatically from 6.9% to 24.6%, a difference of 17.7%. Thus, although the segmental intelligibility for the synthetic speech initially appeared to be quite good when measured with a traditional closed-response six-alternative forced-choice testing procedure, subsequent tests using an open, free-response format revealed substantially worse performance.

In summary, the results of the Modified Rhyme Test revealed high levels of segmental intelligibility of the speech output from the MITalk text-to-speech system using naive listeners as subjects. Unfortunately, the contribution of other sources of knowledge to perception cannot be estimated from traditional measures of segmental intelligibility involving perception of isolated words. The role of prosody, speech timing and the systematic differences in segmental durations--what are referred to as "sentence-level effects" in perception can only be studied with sentential materials.

Word Recognition in Sentences

To evaluate word recognition performance in sentence contexts, we decided to obtain two different sets of data. One set was collected using 100 of the Harvard Psychoacoustic Sentences (Egan, 1948). These test sentences are all meaningful and contain a wide range of different syntactic constructions. In addition, the various segmental phonemes of English are represented in these sentences in accordance with their frequency of occurrence in the language (see also Ainsworth, 1974).

We also collected word recognition data with a second set of sentences. These were 100 syntactically normal but semantically anomalous sentences that Nye and Gaitenby (1974) developed at Haskins Laboratories for use in evaluating the intelligibility of their text-to-speech system. These test sentences permit a much finer assessment of the availability and quality of "bottom-up" acoustic-phonetic information and its potential contribution to word recognition. Since the materials are all meaningless sentences, the individual words cannot be identified or predicted from the listener's knowledge of the sentential context or his semantic interpretation. Thus, the results obtained with these sentences should provide an estimate of the upper bound on the contribution of acoustic-phonetic information to word recognition in sentence contexts.

Twenty-one naive adult subjects received the Harvard test sentences and an additional twenty-three subjects received the Haskins test sentences. The materials were presented one-at-a-time in a self-paced format. Subjects listened to each sentence and then attempted to transcribe it immediately. The sentences were output at a speaking rate in excess of 180 words per minute.

Correct word recognition for the Harvard sentences was quite good with an overall mean of 93.2% correct word recognition. Of the 6.7% errors observed, 30.3% were omissions of complete words while the remainder consisted of segmental errors involving substitutions, deletions and transpositions. In no case, however, did subjects generate permissible nonwords that could occur as potential lexical items in English.

As expected, word recognition performance for the Haskins anomalous sentences was substantially worse than the Harvard sentences, with a mean of 78.7% correct word recognition. Of the 21.3% errors recorded for the Haskins sentences, only 11% were omissions of complete words. The difference in error patterns, particularly in terms of the number of omissions, between the two types of sentence contexts suggests a substantial difference in the subjects' perceptual strategies in the two tests. Subjects used a much looser criterion for identifying words in the Haskins anomalous sentences since the number of permissible alternatives was substantially greater than those in the Harvard sentences.

The results of the two word-recognition tests demonstrate that the level of word recognition performance depends a great deal on the particular test format used and the type of information available to the subject. In one sense, the results of these two tests can be thought of as approximations to the upper and lower bounds on the accuracy of word recognition performance. On the one hand, the Harvard meaningful sentences provide an indication of how word recognition might proceed when both semantic and syntactic information is available to a listener under normal listening conditions. On the other hand, the Haskins anomalous sentences direct the subjects' attention specifically to the sensory-perceptual input and therefore provide an estimate of the reliance on the acoustic-phonetic information and sentence analysis routines available for word recognition in the absence of several important contextual constraints.

Listening Comprehension

One of the factors that obviously plays an important role in listening comprehension is the quality of the initial input signal--that is, the intelligibility of the speech signal itself. But intelligibility is only one factor that affects speech perception and spoken language understanding. As we have seen from our earlier results, attention must also be given to the contribution of higher-levels of knowledge to perception and comprehension (Pisoni, 1978). In this last part of our evaluation study, we wanted to obtain some preliminary estimate of how well listeners understand the linguistic content of continuous fluent speech produced by the MITalk text-to-speech system. Previous evaluations of synthetic speech generated by other text-to-speech systems have been concerned primarily with segmental intelligibility (i.e., phoneme intelligibility) or listener preferences of the quality or naturalness of the speech (McHugh, 1976).

To examine this problem, we selected fifteen narrative passages and an appropriate set of multiple-choice test questions from several standardized adult reading comprehension tests. The passages were quite diverse, covering a wide range of topics, writing styles and vocabulary.

We tested three groups of naive subjects with 20 subjects in each group. One group of subjects listened to synthetic versions of the passages, another listened to natural speech while a third group read the passages silently. All three groups answered the same set of test questions immediately after each passage.

The comprehension results for the three groups are summarized in Table 1. Averaged over the last thirteen test passages, the reading group showed a 7%

advantage over the synthetic speech group. However, the difference in performance between these two groups appeared only in the first half of the test. By the second half, performance for the synthetic group improved by over 10% whereas performance for the reading group remained the same. Although the scores for the natural speech group were slightly lower overall, no improvement was observed in their performance across both halves of the test.

Insert Table 1 about here

The finding of improved performance in the second half of the test for subjects in the listening group is consistent with the earlier results from the word recognition tests which showed that recognition performance improves for synthetic speech after only a short period of exposure. When the three comprehension groups were compared on the same passages in the last half of the test, their performance is equivalent. This result suggests that the overall difference between the groups is probably due to familiarity with the output of the synthesizer and not due to any inherent difference in the basic strategies used in comprehending or understanding the content of these passages (see also Carlson, Granstrom & Larsson, 1976). This conclusion is strengthened even further by the observation that the thirteen passages are highly correlated ($r = +.97$) across reading and listening conditions.

The multiple-choice comprehension test that we used in the initial evaluation study was probably too gross to distinguish between new knowledge acquired from listening to text and previous knowledge obtained from inferences drawn at the time of comprehension or later at the time of testing. To separate these issues, Luce (1982) recently completed a detailed study of the comprehension of fluent synthetic speech using more specific probe items in a question and answering format. As in the previous study, subjects listened to passages of continuous fluent synthetic speech. After presentation of each passage, they were required to answer four test questions about their understanding of the information expressed in the passage. Subjects were told to answer these YES-NO questions based on knowledge gained from listening to the passage and not from prior information brought to the testing situation. Measures of accuracy and response latency were obtained in the question-answering task. The results showed two interesting patterns. First, for accuracy in question-answering, performance on surface structure questions was better for synthetic passages than natural passages. However, performance on the other three types of questions which is based on more abstract information was better for the naturally produced passages than the synthetically produced ones. The question-answering latencies showed no consistent difference across synthetic and natural passages although the response times were related to the "depth of processing" required to answer the questions. Surface structure questions were responded to faster than high and low proposition questions, and, these in turn, were responded to faster than the inference questions. This pattern of results suggests that synthetic speech may force subjects to attend more closely to the physical or surface phonological properties of the signal. This reallocation of attention may therefore place somewhat greater demands on the cognitive processes involved in word recognition, lexical search and comprehension of meaningful

TABLE 1

Percent Correct Performance on Comprehension Test

	1st Half	2nd Half	Total
MITalk	64.1	74.8	70.3
Natural	65.6	68.5	67.8
Reading	76.1	77.2	77.2

connected speech. Luce's findings are consistent with other data showing a strong interaction between depth of processing and allocation of attention to different levels of linguistic information in the speech signal.

PROCESSING TIME STUDIES AND CAPACITY DEMANDS IN SHORT-TERM MEMORY

As noted earlier, our initial evaluation of the synthetic speech generated by the MITalk system relied entirely on performance measures involving response accuracy. Recently, we completed an experiment that was aimed at measuring the processing time required to recognize natural and synthetic words and permissible nonwords. In carrying out this study, we wanted to know how long it takes a human observer to recognize an isolated word and how the process of word recognition might be affected by the quality of the initial acoustic-phonetic information in the speech signal.

Lexical Decision Task

To measure how long it takes an observer to recognize isolated words, we used a lexical decision task (Pisoni, 1981). In this task, subjects are presented with either a word or a nonword stimulus item on each trial. The subject is required to classify the item as either a "word" or a "nonword" as fast as possible by pressing one of two buttons located on a response panel in front of him. Examples of the stimuli are shown in Table 2.

Insert Table 2 about here

The results for the first block of 100 trials in the experiment are shown in Figure 1. On the right, we have plotted the overall accuracy in terms of percent correct for the natural and synthetic items. Word and nonword responses are displayed as the parameter in this figure. Notice that performance is better overall for the natural speech items (98% correct) than the synthetic ones (79% correct). Moreover, these differences are present for both word and nonword responses alike.

Insert Figure 1 about here

The panel on the left of this figure displays the response times for correct responses to word and nonword stimuli. The response times for natural speech stimuli are consistently faster than the response times for the synthetic stimuli. The overall difference between the two types of speech signals was 145 milliseconds (msec). The differences between the word and nonword responses were comparable across the two types of speech signals; the overall difference was 140 msec.

TABLE 2

Examples of Test Stimuli Used in Lexical Decision Task

1. PROMINENT	.....	1. PRADAMENT
2. BAKED	.....	2. BEPT
3. TINY	.....	3. TADGY
4. GLASS	.....	4. GEEP
5. PARENTS	.....	5. PEEMERS
6. TOLD	.....	6. TAVED
7. BLACK	.....	7. BAEP
8. CONCERTS	.....	8. CAELIMPS
9. DARK	.....	9. DUT
10. BABBLE	.....	10. BURTLE
11. CRITIC	.....	11. CRAENICK
12. BOUGHT	.....	12. BUPPED
13. PAIN	.....	13. POON
14. GORGEOUS	.....	14. GAETLESS
15. COLORED	.....	15. COOBERKD

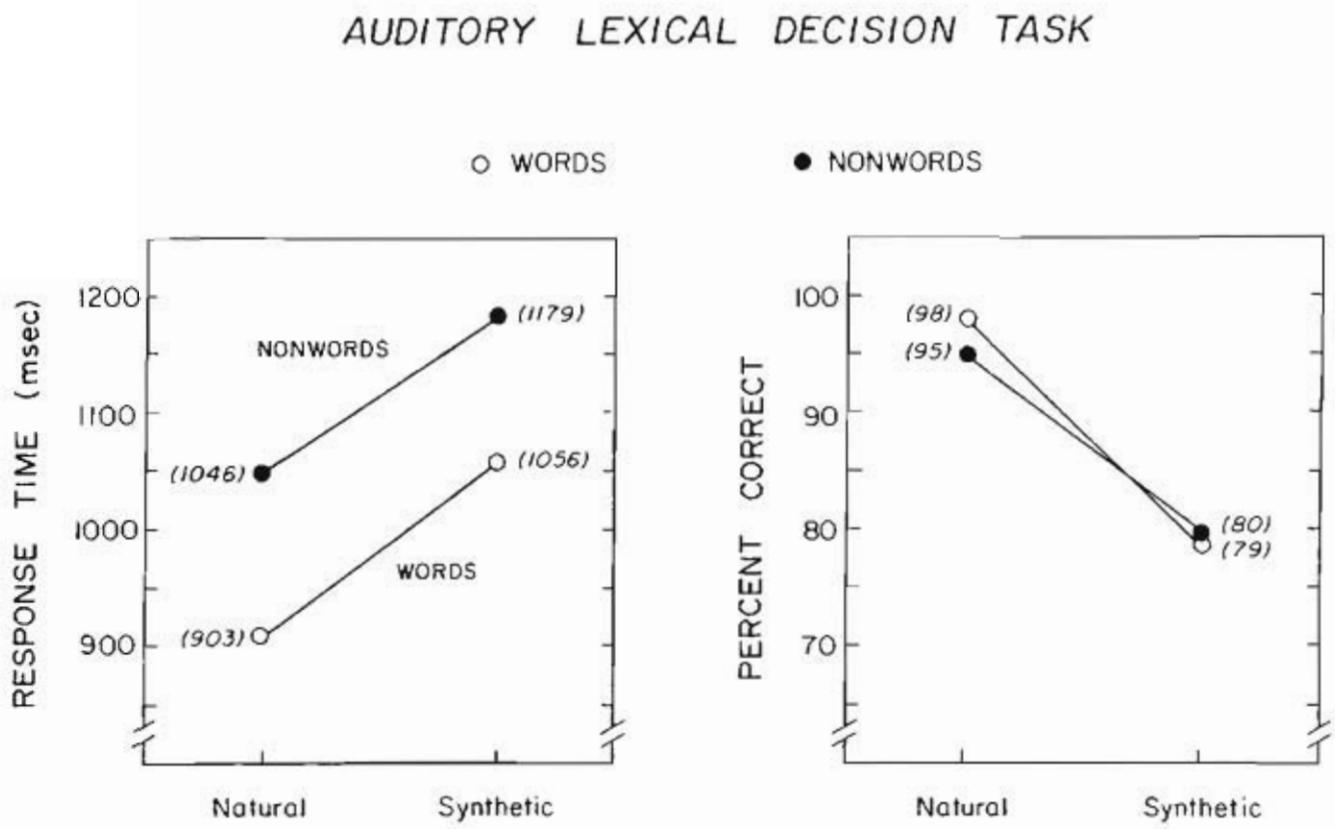


Figure 1. Results of the auditory lexical decision task for word and nonword stimuli. Reaction times are displayed in the left panel, error rates are shown on the right.

The results of this experiment indicate that synthetic speech requires more cognitive processing time to execute manual responses. We have found strong and consistent evidence that responses to both word and nonword stimuli are affected more-or-less equivalently when the test items are synthetic speech. The pattern of findings from this study suggests that the differences in processing time between natural and synthetic speech probably lie at processing stages involved in word recognition and the extraction of basic acoustic-phonetic information from the speech waveform -- that is, the early pattern recognition process itself rather than at the more cognitive levels involved in lexical search or retrieval of words from the mental lexicon.

Capacity Limitations

The findings we have discussed so far demonstrate the existence of both perceptual and cognitive difficulties in processing synthetic speech signals. Recently, Tim Feustel, Paul Luce and I carried out a series of experiments to pin down the specific locus of these difficulties in the information processing system. In these experiments, we were interested in determining whether the previously demonstrated differences in perception between natural and synthetic speech could be attributed to encoding and rehearsal processes in short-term memory (STM) or whether they were more perceptual in nature--that is, if the difficulties lie in the extraction of the important acoustic-phonetic information from the speech signal (see Feustel, Luce & Pisoni, 1981).

If the observed differences in processing synthetic speech are due to encoding difficulties at relatively early stages of perception, then there should be a measurable increase in the cognitive demands that synthetic speech signals place on the limited resources available in STM. These additional processing demands should then provide relatively less available capacity for any secondary task that the listeners might be asked to perform at the same time (Rabbitt, 1968). If this is true, then as the difficulty of the secondary task increases, performance on that task should decrease more rapidly for synthetic than for natural speech. To study this problem we examined recall of lists of words. We selected this experimental procedure because the difficulty of recall tasks can be easily and reliably manipulated and differences in these procedures are now well understood by cognitive psychologists.

In our first experiment, we presented six lists of 25 monosyllabic words to subjects over headphones for free recall. Three of the lists were synthetic and three were natural. The lists were presented at three different presentation rates: 1, 2 and 5 seconds (sec) per word. The subjects' task was simply to recall as many of the words from each list as possible during a 90 sec recall period immediately following presentation of each list.

The prediction for this experiment was quite straightforward, as the rate of presentation increased, the recall performance for the synthetic lists should decrease more rapidly than performance for the natural lists. This result was expected because any encoding difficulties in STM produced by the synthetic lists should influence the subjects' ability to rehearse and store the words.

The results are shown in Figure 2. The two graphs show the probability of word recall as a function of its position on the list -- what cognitive psychologists call serial position curves. The upper panel shows the natural

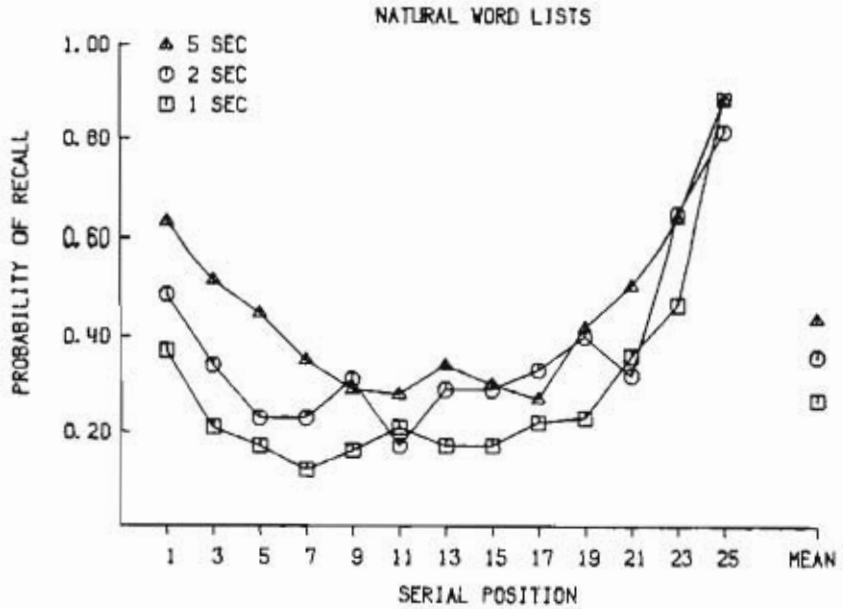
lists, the lower panel the synthetic lists. The three curves correspond to the three presentation rates. Note that the rate manipulation had the predicted effect of increasing the difficulty of the recall task. In both panels, there is an orderly decrease in the probability of recall as the presentation rate increases. The effect is largest in the early serial positions -- the "primacy" portion of the curves and smallest at the later positions -- the "recency" portion of the curve. This was expected because the last word in the list is still present and being rehearsed in STM as the recall period begins. Note that the synthetic words are recalled more poorly overall than the natural words and that the overall shape of the curves is quite similar for both the natural and synthetic lists. There is no clear indication in our analyses that the synthetic words required additional capacity or attention during encoding. However, the presentation rate of 1 sec per word may not have been fast enough to produce any reliable effects in this testing format. Moreover, it is possible that subjects were not being pushed to their capacity limitations since only one item was presented at a time.

Insert Figure 2 about here

In order to increase the stress on STM, we ran a second experiment in which we adapted a memory pre-load technique originally developed by Baddeley and Hitch (1974). This technique consists of loading STM with a short list of digits or letters which the subjects are asked to maintain in memory (i.e. actively rehearse) throughout presentation of the word lists. In this experiment, subjects listened to three synthetic and three natural lists. Prior to each word list the subjects saw a list of either 0, 3, or 6 digits presented visually, one at a time, on a CRT monitor. Subjects were told to remember the digits in the exact serial order in which they were presented. After the auditory items were presented, the subjects were first required to report the digit list and then to recall as many of the words from the list as they could remember.

Figure 3 shows the results for word recall. The average number of words recalled is plotted as a function of the digit pre-load. The results agree closely with those from the first experiment. Both pre-load and word type affect recall. The magnitude of the difference in recall performance between the natural and synthetic word lists is about the same as in the previous study. However, when the performance on recall of the pre-load items is examined, a different picture emerges. Figure 4 displays the number of subjects who correctly recalled all of the pre-load digits in the exact serial order in which they were presented. The expected interaction is present for recall of the pre-load digits. The number of subjects correctly recalling the digits decreased more rapidly for the synthetic lists as the load on STM increased from 3 to 6 items.

Insert Figures 3 & 4 about here



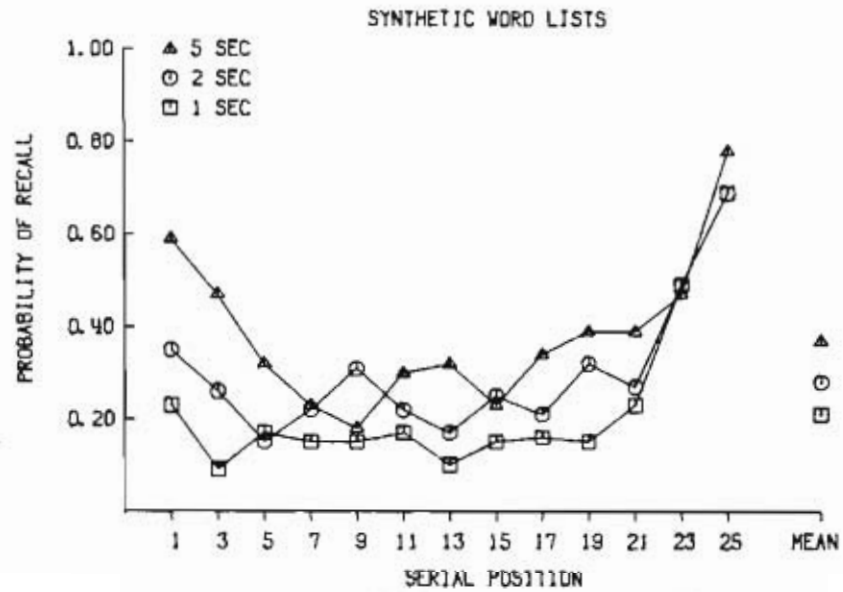


Figure 2. Serial position curves for recall of natural and synthetic word lists at three different presentation rates.

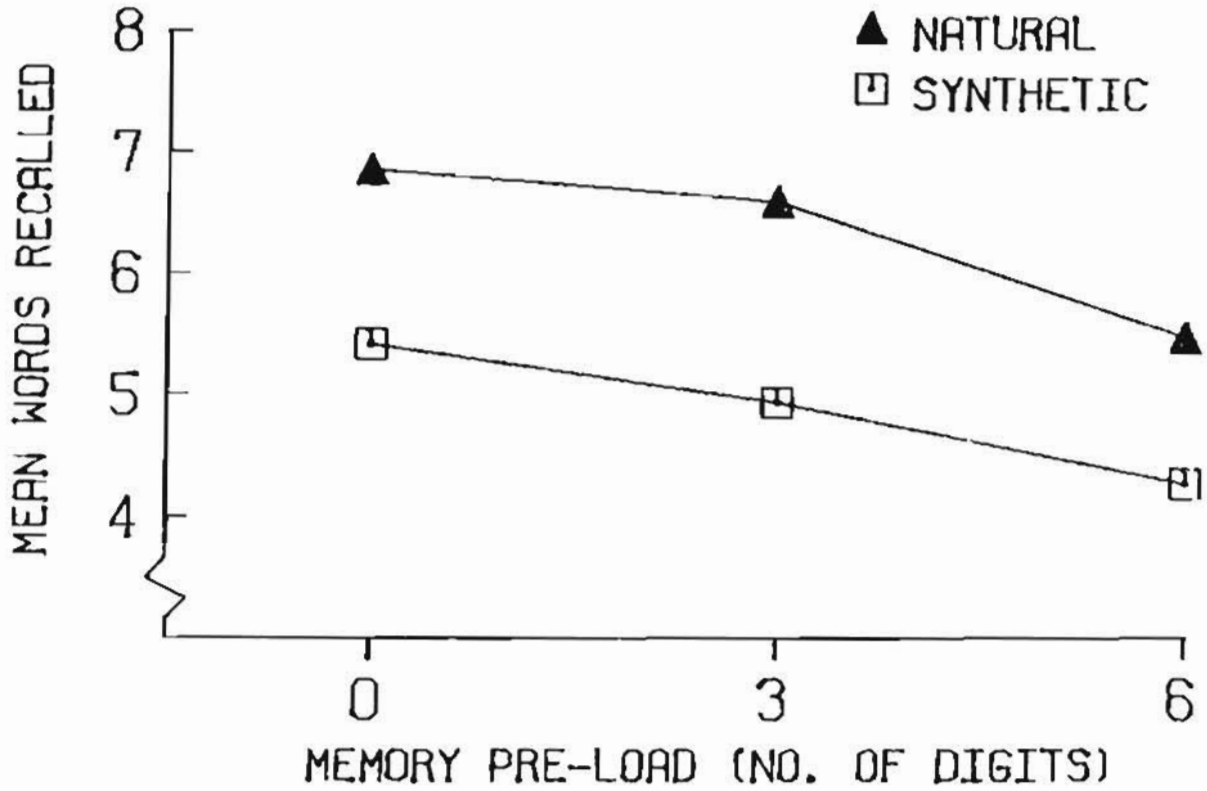


Figure 3. Mean number of natural and synthetic words recalled as a function of memory pre-load.

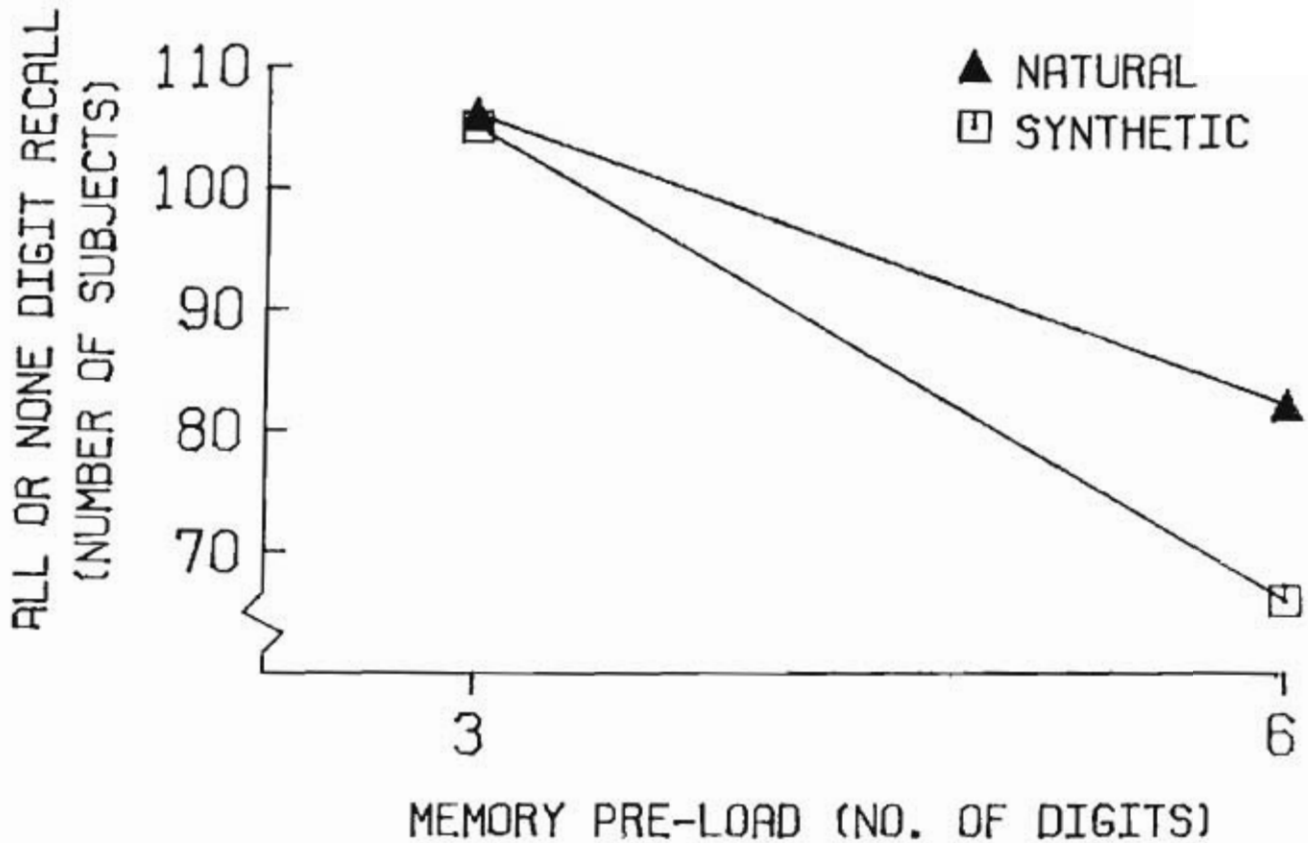


Figure 4. Number of subjects who correctly recalled all of the digits as a function of the memory pre-load condition.

The results from the second recall experiment using the pre-load technique indicate that synthetic speech is, in some manner, disrupting the subjects' ability to maintain information in STM. Moreover, this disruption appears to occur only under conditions of memory stress when the subject must actively maintain information concurrently in STM. The obvious interpretation of this effect, and the one we prefer, is that the subjects are "borrowing" from the limited capacity needed for maintenance rehearsal of the digits in STM in order to encode the synthetic word lists.

Two conclusions may be drawn from these recall experiments. First, the large constant decrement in recall performance observed in the first two experiments is probably due to a misencoding of some of the synthetic words in the lists at an early perceptual level. Second, and perhaps more important, as shown in our last experiment, the perceptual difficulties in the processing of synthetic speech also appear to be due to increased processing demands for these stimuli in STM. We believe these processing demands and the reallocation of resources may place important constraints on the use of synthetic speech in voice response devices that are used in high information load conditions such as aircraft cockpits, flight simulators, and computer assisted instructional devices where the human observer is carrying out more than one complex cognitive task at a time. In such applications, the use of synthetic speech may produce processing decrements that cannot be overcome by conscious real-location of resources and attention.

SUMMARY AND CONCLUSIONS

Our results on the perception of synthetic speech have important implications for the design, selection and use of voice-response systems that are to be used in various airborne applications such as cockpit design and interaction, air traffic control environment, training of pilots and other military personnel to operate complex equipment and manage control systems under hostile conditions. Moreover, our recent findings are quite relevant to several basic questions surrounding man-machine interaction using natural language. It should be obvious that additional research is clearly needed to gain a better understanding of the precise differences in perception between natural and synthetic speech. Well-motivated decisions concerning the choice and implementation of various voice response systems cannot be made until a number of important psychological problems are examined in greater detail. Basic research in voice technology over the next few years should be directed at questions such as: (1) the effects of noise and distortion on perception of synthetic speech; (2) perception of synthetic speech under various listening conditions requiring differential cognitive and attentional demands; (3) perceptual and cognitive processing time for recognition of synthetic speech; (4) effects of practice and familiarity; (5) comparative evaluations of various commercially available speech synthesizers and synthesis-by-rule systems; (6) questions surrounding the role of naturalness on intelligibility; and (7) relationships between traditional forced-choice measures of isolated word recognition and perception of words and comprehension of sentences in fluent speech where many sources of knowledge interact. Research on several of these problems is currently being carried out in our laboratory at Indiana University.

How do the various commercially available text-to-speech and voice-response systems perform? At the present time, we simply do not know. To our knowledge, no systematic comparative evaluations have ever been undertaken to assess the performance characteristics of these systems. From our informal observations, it is quite apparent that these products differ quite substantially from each other in level of speech quality and intelligibility and they may require many hours of practice and familiarity before the synthetic speech can be recognized and understood at levels comparable to natural speech. In some cases, performance levels may never reach those obtained with natural speech.

More seriously, however, is the fact that at the present time there simply are no uniform standards for evaluating the quality and intelligibility of the speech produced by speech synthesizers or voice response systems. This is unfortunate because after some thirty years of basic research on speech, the widespread use of text-to-speech and voice-response systems is now a realistic goal. The obstacles are no longer questions of basic research into the principles of speech production, perception and linguistic analysis; rather they are simply the practical matters of implementation and economics.

REFERENCES

Ainsworth, W. A. Performance of a speech synthesis system. International Journal of Man-Machine Studies, 1974, 6, 493-511.

Allen, J. Reading machines for the blind: The technical problems and the methods adopted for their solution. IEEE Transactions on Audio and Electroacoustics, 1973, Vol. AU-21, No. 3, 259-264.

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

Baddeley, A. D. and Hitch, G. Working memory. In G. H. Bower (Ed.), The psychology of learning and memory, Vol. 8, 1974, 47-90.

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

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

Feustel, T. C., Luce, P. A. and Pisoni, D. B. Capacity demands in short-term memory for synthetic and natural word lists. Journal of the Acoustical Society of America, 1981, 70, S98.

Lindsay, P. H. and Norman, D. A. Human Information Processing (2nd ed). NY: Academic Press, 1977.

Luce, P. A. Comprehension of synthetic and natural texts: Answering questions concerning various levels of textual representation. Paper to be presented at the 103rd Meeting of the Acoustical Society of America, Chicago, May 1982.

McHugh, A. Listener preference and comprehension tests of stress algorithms for a text-to-phonetic speech synthesis program. (NRL Report 8015). Washington, D.C.: Naval Research Laboratory, September 9, 1976.

Nye, P. W. and Gaitenby, J. Consonant intelligibility in synthetic speech and in a natural speech control (Modified Rhyme Test Results). Haskins Laboratories Status Report on Speech Research, SR-33, (1973). Pp. 77-91.

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

Pisoni, D. B. Speech Perception. In W. K. Estes (Ed.) Handbook of Learning and Cognitive Processes Volume 6. Hillsdale, NJ: Lawrence Erlbaum Associates, 1978, Pp. 167-233.

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

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

Rabbitt, P. Channel-capacity, intelligibility and immediate memory. Quarterly Journal of Experimental Psychology, 1968, 20, 241-248.

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

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

Visual Lexical Decision Times for Open- and Closed-Class Words and Nonwords

Paul A. Luce

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

This research was supported by grants from NIMH, Grant No. MH-24027, and NINCDS, Grant No. NS-12179. An earlier version of this paper was presented at the meeting of the Midwestern Psychological Association, Minneapolis, 1982. The author would like to thank David B. Pisoni for his advice and assistance at various stages of this project.

ABSTRACT

Bradley (1978) has proposed that the distinction between open- and closed-class words is a psychologically plausible notion. She has presented evidence that purports to show that closed-class words are retrieved from the mental lexicon by a computationally efficient mechanism that is independent of the mechanism responsible for retrieval of open-class words. In a visual lexical decision experiment using open- and closed-class words and nonwords, we found further evidence for the psychological plausibility of the open-/closed-class distinction. However, it is argued that the locus of this distinction as assessed by the visual lexical decision paradigm lies in either the search or comparison stages and that separate retrieval mechanisms for open- and closed-class words in the lexicon need not be postulated.

Visual Lexical Decision Times for Open- and Closed-Class Words and Nonwords

Bradley (1978; see also Bradley, Garrett, and Zurif, 1980) has recently shown that the traditional distinction between "content" and "function" words may have important implications for theories of language processing that attempt to account for lexical access and sentence comprehension. From a series of experiments on normal and aphasic subjects, Bradley obtained evidence suggesting that two separate retrieval mechanisms, one for function or "closed-class" items and one for content or "open-class" items, are operative in accessing the mental lexicon. The motivation for proposing two independent retrieval mechanisms comes from the finding that access to the set of closed-class words is computationally more efficient than access to the set of open-class words. If it is true that closed-class words are accessed at reduced computational cost, the burden of deriving structural analyses of sentences in fluent text or speech would be considerably reduced. That is, the parser would have immediate access to the closed-class vocabulary from which a structural analysis of a sentence could be quickly built. Establishing the "psychological plausibility" of the open-/closed-class distinction may therefore prove to be of great importance not only for models of lexical access but also for models that attempt to account for the parsing of fluent text and continuous speech.

Before reviewing the evidence for the psychological plausibility of the open-/closed-class distinction, it is perhaps best to digress here and discuss what is typically meant by open- and closed-class words. Fries (1952) is frequently cited as one of the first to explicitly separate words into the two classes of content and function words. He includes articles, prepositions, pronouns, numbers, conjunctions, and auxiliary verbs in the class of function words, and nouns, verbs, adjectives, and adverbs in the class of content words. According to Fries, function words act as grammatical operators, whereas content words carry semantic information. More explicitly, Bolinger (1975) states that the job of function words, or "grammatical morphemes," is to:

> "serve the main carriers of meaning, the lexical words: to relate them, refer back to them, combine them or separate them, augment or diminish them, substitute for them, and so on. Grammatical morphemes hover about the lexical words and groups of words, attaching themselves in front or behind and sometimes in the middle; they get less attention, are less clearly articulated and less frequently accented, and their second-class citizenship leads to . . . changes and reductions and losses of sound . . ." (pp. 119-120).

The dichotomy between function and content words is by no means clear-cut, but the above definitions should give the reader some feel for what does and does not constitute a function or content word. However, because of the arbitrariness of the distinction, considerable disagreement has arisen concerning what words

should be included in either of the two classes. To circumvent these disagreements, Bradley has adopted the more neutral terms "open-class" and "closed-class" to refer to the basic distinction between the two word-form classes, and we will comply in the remainder of this report with this convention. These newer labels cover roughly the same sets of words: Closed-class words are words from minor lexical categories that are relatively fixed in number and resist addition; open-class words are words from major lexical categories that may be added to freely.

A number of other characteristics of closed-class words serves to distinguish them from open-class words. First, the number of closed-class words is extremely small in comparison to the number of open-class words. Second, as Bolinger (1975) pointed out, closed-class words are typically unstressed in fluent speech (see Selkirk, 1972, 1980), although "stresslessness" of closed-class words is not a completely consistent phenomenon. For this reason, Flanagan, Coker, Rabiner, Schaefer, and Umeda (1970) have postulated a category of intermediate words containing less stressed open-class words and more stressed closed-class words. Nevertheless, the lack of stress speakers generally assign to closed-class words suggests that they may carry less "information" than open-class words (O'Shaughnessy, 1976). The line of reasoning supporting this claim is as follows: If closed-class words are unstressed and frequently reduced, the perceiver should be able to more easily predict closed-class than open-class words from context. And, because the degree of predictability of a given word is the inverse of the information load carried by that word, it follows that the closed-class words carry less information. This observation has important implications for the hypothesis that a separate retrieval mechanism is operative in accessing closed-class words, for the overall low information load of closed-class words may allow for computationally more efficient retrieval from the lexicon.

A third characteristic of closed-class words that distinguishes them from open-class words was demonstrated by Miller, Newman, and Friedman (1958). In their study, Miller et al. showed a stronger relationship between the length and frequency of closed-class words than between the length and frequency of open-class words. In particular, Miller et al. found that as the length of closed-class words increases, the frequency of these words decreases. The open-class words, on the other hand, show "a much more democratic division of responsibility" (p. 384), with frequency being a poor predictor of length. The implication from the Miller et al. findings is that constraints exist on the length of the closed-class words. These constraints are at least intuitively consistent with the notion of a separate retrieval mechanism that accesses the closed-class items at a considerable reduction in computational cost. [These constraints may be language specific, however. See Miller, Newman, and Friedman (1958).]

Apart from Bradley's study, which we will discuss below, empirical support for the open-/closed-class distinction has come from a variety of sources. In studies of speech production errors, Garrett (1976,1979) has shown a marked difference between the number and type of transpositions that occur with open-class words and those that occur with closed-class words. Language acquisition studies have shown that closed-class words are relatively late in

appearing in the productions of young children, resulting in telegraphic speech (Brown and Bellugi, 1964). However, Nelson (1973, see also Nelson, 1981) has observed that some children produce pronouns and other grammatical functors early in language acquisition, although, as she notes, "whether these terms could be considered 'vocabulary items' [is] problematic, since they [are] usually embedded in what [appear] to be unanalyzed formulas or routines. . ." (Nelson, 1981, p. 172). In addition to the research on speech production errors and language acquisition, Epstein (1961) has shown that closed-class items (including bound grammatical morphemes) facilitate memory for strings of nonwords.

In studies on proofreading errors, Haber and Schindler (1981) have shown that, all things being equal, subjects are less likely to detect misspellings in closed-class words than in open-class words. Healy (1976, 1980) has likewise shown that subjects make a large number of errors in letter detection tasks on one particular closed-class word, the word the. Although Drewnowski and Healy (1977) argue that all common words may be "unitized," or processed in units higher than the letter level, their results, in conjunction with the Haber and Schindler findings, suggest that visual processing of closed-class words may be, in some sense, more "superficial" than visual processing of open-class words. This conclusion is also supported by results from Ferres and Aaronson (1981) who found that in rapid serial visual presentation of continuous text, subjects spend less time reading closed-class words than open-class words (see also Ferres, 1981). Thus, there appears to be a corollary in the visual realm to the "stresslessness" characteristic of closed-class words in fluent speech. That is, subjects spend less time processing closed-class words presented visually in context because of high predictability and a concomitant low information load.

We now turn to the most compelling evidence for the open-/closed-class distinction. Using a visual lexical decision paradigm, Bradley (1978) found that normal subjects are not sensitive to the frequency of closed-class words, whereas subjects are sensitive to the frequency of open-class words, as has been well documented. In addition, she found that closed-class words embedded in initial portions of nonwords do not produce the interference effect well-documented for open-class words embedded in the initial portion of nonwords (Taft and Forster, 1975, 1976). Taken together, these two results suggest the operation of a retrieval mechanism for closed-class words that is separate from the retrieval mechanism for open-class words. The retrieval mechanism for closed-class words appears to be insensitive to word frequency (however, see the discussion of the Gordon and Caramazza (1982) study below) and is also not prone to nonword interference effects, presumably because of the manner in which this particular retrieval mechanism treats visually presented word forms.

This latter characteristic of the closed-class retrieval mechanism deserves further elaboration. Taft and Forster (1975, 1976) have shown that if the initial portion of a nonword contains an actual word in the language, reaction times are longer for deciding that the item is a nonword than if the item does not contain a real word in its initial portion. This result is interpreted as showing that the presence of a word in the initial portion of the nonword is sufficient to temporarily "trick" the retrieval mechanism into accepting the nonword as a word. Subsequent checks on the latter portion of the nonword causes a rejection of the item as a word.

Bradley (1978) tested the possibility that closed-class words embedded in the initial portions of nonwords would not produce the interference effect, and, as mentioned above, she found no interference effects for closed-class words. According to Bradley, two accounts of this lack of interference for closed-class words are possible: (1) Closed-class words are not accessed via the retrieval mechanism for open-class words which uses the initial portions of words to activate candidate representations in the lexicon; or (2) closed-class words are available to the same mechanism that retrieves open-class words, but when embedded in nonwords, closed-class nonword items are rejected at little or no time cost. Before considering which of these alternatives is the better one, we first turn to two additional experiments of Bradley's with aphasic subjects and then review a recent criticism of one of her findings.

Bradley (1978) performed the same two visual lexical decision experiments described above on subjects with Broca's aphasia. Some patients afflicted with Broca's aphasia are described as "agrammatic" because of their putative inability to process grammatical morphemes. Bradley reasoned that if agrammatism is caused by "damage" to the retrieval mechanism for closed-class words, aphasic subjects should not show differential treatment of open- and closed-class words. Consistent with her expectations, Bradley found that aphasics were sensitive to the frequency of closed-class words and moreover they showed the nonword interference effect for nonwords containing closed-class words. Apparently, Broca's aphasia is at least, in part, typified by an inability to treat open- and closed-class words differentially.

Although Bradley's results suggest that the open-/closed-class distinction is psychologically plausible and, more important, that there may be two separate retrieval mechanisms for the word-form classes, a recent study by Gordon and Caramazza (1982) has called into question one of Bradley's major findings. In particular, Gordon and Caramazza failed to replicate Bradley's experiment showing that normal subjects are insensitive to the frequency of closed-class words. They argue that both the frequency ranges compared by Bradley and the associated statistics were inappropriate. Once these factors were accounted for, Gordon and Caramazza found a definite sensitivity on the part of their subjects to the frequency of closed-class items.

Gordon and Caramazza's findings do not vitiate Bradley's claims, however, for the conclusions from the nonword interference effect still stand, as do the results from her experiments on lateral asymmetries and word class not discussed here. We therefore attempted to elaborate on Bradley's conclusions concerning nonword forms of closed-class words by conducting a visual lexical decision experiment in which nonwords constructed from open- and closed-class words were used. Instead of using nonwords constructed to contain real word forms, we compared nonwords systematically permuted from open- and closed-class words. More specifically, we replaced one vowel phoneme in each syllable of open- and closed-class words with a vowel phoneme that transformed the words into nonwords. By employing such a manipulation we were able to test (1) the possible existence of graphemic cues to word class and (2) the size of the search set of open- and closed-class words matched for frequency. We reasoned that if graphemic cues to word class exist, the reaction times for rejecting closed-class nonwords and open-class nonwords should reflect a near complete search of both vocabularies.

If the reaction times for the two sets of nonwords should differ, we would then be in a position to make some preliminary claims about a dissociation within the lexicon between open- and closed-class words.

METHOD

Subjects

The subjects were 24 undergraduates at Indiana University. All were native English speakers and received course credit for their participation.

Stimuli

ONE hundred and two words were used to construct two experimental lists. Each of these words fell into one of four categories: (1) high frequency closed-class words, (2) low frequency closed-class words, (3) high frequency open-class words, and (4) low frequency open-class words. All frequencies were obtained from Kucera and Francis (1967). The frequencies of the high frequency closed-class words were all above 7250 per million in the Kucera and Francis corpus. The frequencies of the low frequency closed-class and high frequency open-class words fell between 173 and 1984 per million. The frequencies of the low frequency open-class words were all one. The low frequency closed-class words and the high frequency open-class words were therefore matched for frequency, word length, and number of phonemes, and *t* tests showed that the two sets of words did not differ significantly on any account (for frequency, $t(80)=0.33$, $p>0.25$, for length, $t(80)=0.96$, $p>0.10$, and for number of phonemes, $t(80)=0.23$, $p>0.40$).

From each of these words, a nonword item was constructed by replacing one vowel phoneme in each syllable with a vowel phoneme that rendered the word a nonword form. This was done to preserve the consonantal structure (and most of the overall graphemic structure) of the word in its corresponding nonword form. All nonwords were pronounceable and, as best as could be determined, were composed of legal sequences of letters. The nonwords did not differ significantly in word length ($t(80)=0.90$, $p>0.10$) or number of phonemes ($t(80)=0.91$, $p>0.10$). Examples of the words and nonwords made from them are instead-anstud and program-pregrum.

The four classes of words were evenly divided and assigned to two separate lists. The nonwords were then divided and assigned to one of these two lists. No nonword was assigned to the list that contained the word from which it was derived. This procedure produced two lists of 51 words and 51 nonwords each.

Procedure

The words and nonwords were randomized and presented one at a time in the center of a CRT videodisplay monitor (GBC Model MV 10-A). Subjects received only one of the two experimental lists. All of the words were presented in capital letters to assess directly the effects of the graphemic structure of the stimuli with as little confounding of word shape as possible. Subjects responded "word"

or "nonword" to each item by pressing appropriately labelled buttons on response boxes in front of them. Subjects were instructed to respond "word" even if they did not know what the word meant but still believed it to be a real word in the English language. Responses and latencies were recorded and the entire experiment was run on-line under the control of a PDP 11/34 computer. Subjects were run in small groups of between four and six subjects each.

RESULTS

Figure 1 shows the percent correct for each of the four categories of words and nonwords.

Insert Figure 1 about here

Overall, there was no significant trend for nonwords to be responded to less accurately than words ($F(1,23)=0.07$, $p>0.70$). However, excluding the low frequency open-class items, words in general were responded to with 95 to 99 percent accuracy, whereas accuracy for nonwords dropped to 88 to 91 percent, which resulted in a significant word-nonword difference ($F(1,23)=12.47$, $p<.002$). Apparently, the low frequency open-class words, because they have a frequency of only one in the Kucera and Francis corpus, are so rare in the lexicon as to be relatively indistinguishable from nonwords for our subjects, whereas the nonwords made from them were easily identified as nonwords. Low frequency open-class items excluded, no other main effects ($F(2,46)=1.53$, $p>0.20$) or interactions ($F(2,46)=0.99$, $p>0.30$) were observed.

Figure 2 shows the reaction times for the words and nonwords for each of the four category types. Reaction times are shown for correct responses only.

Insert Figure 2 about here

The overall trend was for nonwords to be responded to more slowly than the words ($F(1,23)=17.64$, $p<.0003$), except for the low frequency open-class words. The explanation we offered for the accuracy data for the low frequency open-class items appears to hold here as well: Subjects apparently have a very difficult time making decisions regarding the uncommon open-class words which, as we mentioned above, were in fact very low frequency words. Subjects were able, however, to decide more quickly and accurately that the nonwords made from the low frequency open-class words were, in fact, nonwords.

Of special interest here, however, is the pattern of results obtained for the high frequency open-class words and the low frequency closed-class words. As

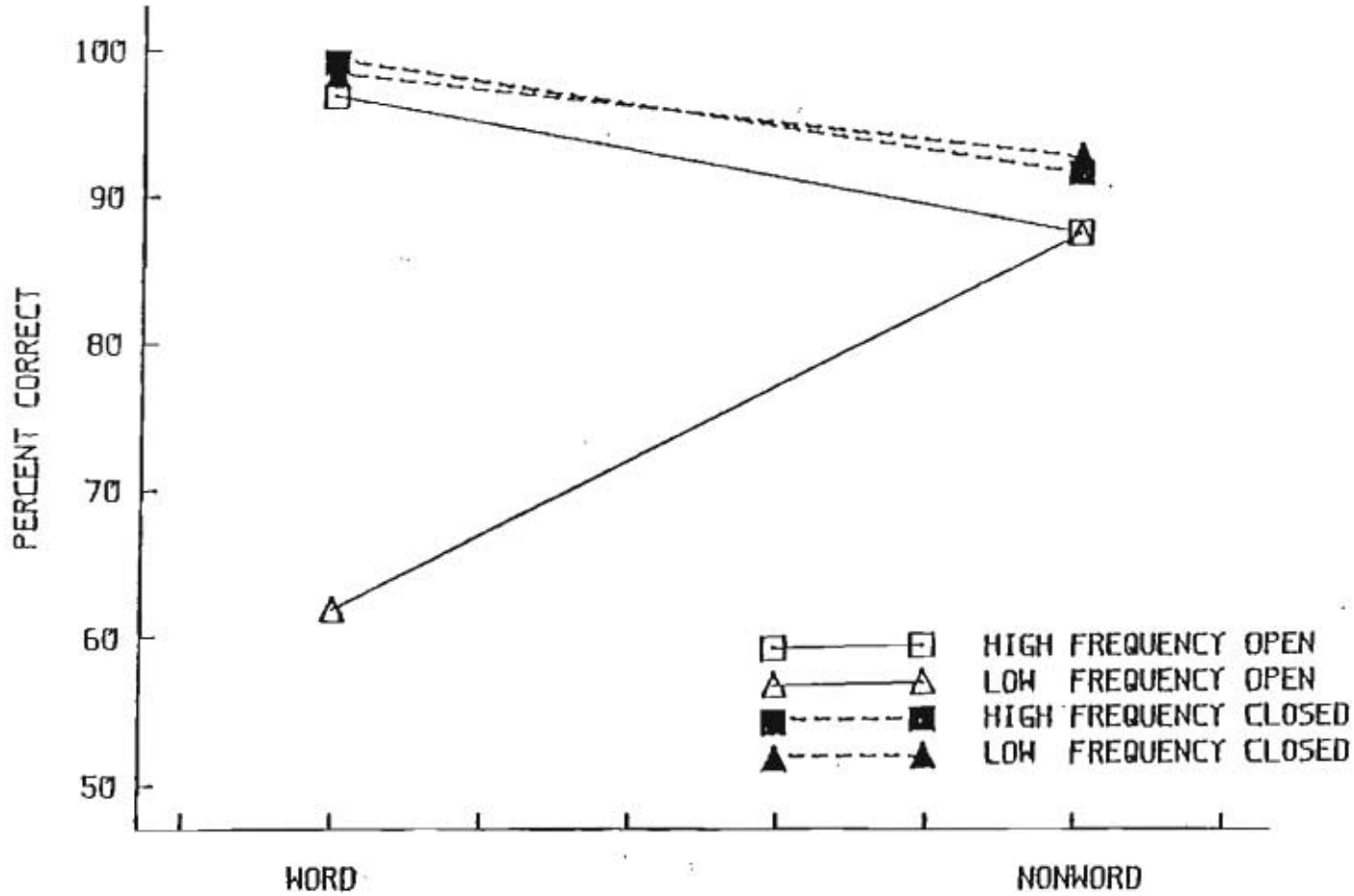


Figure 1. Percent correct for open- and closed-class words and nonwords.

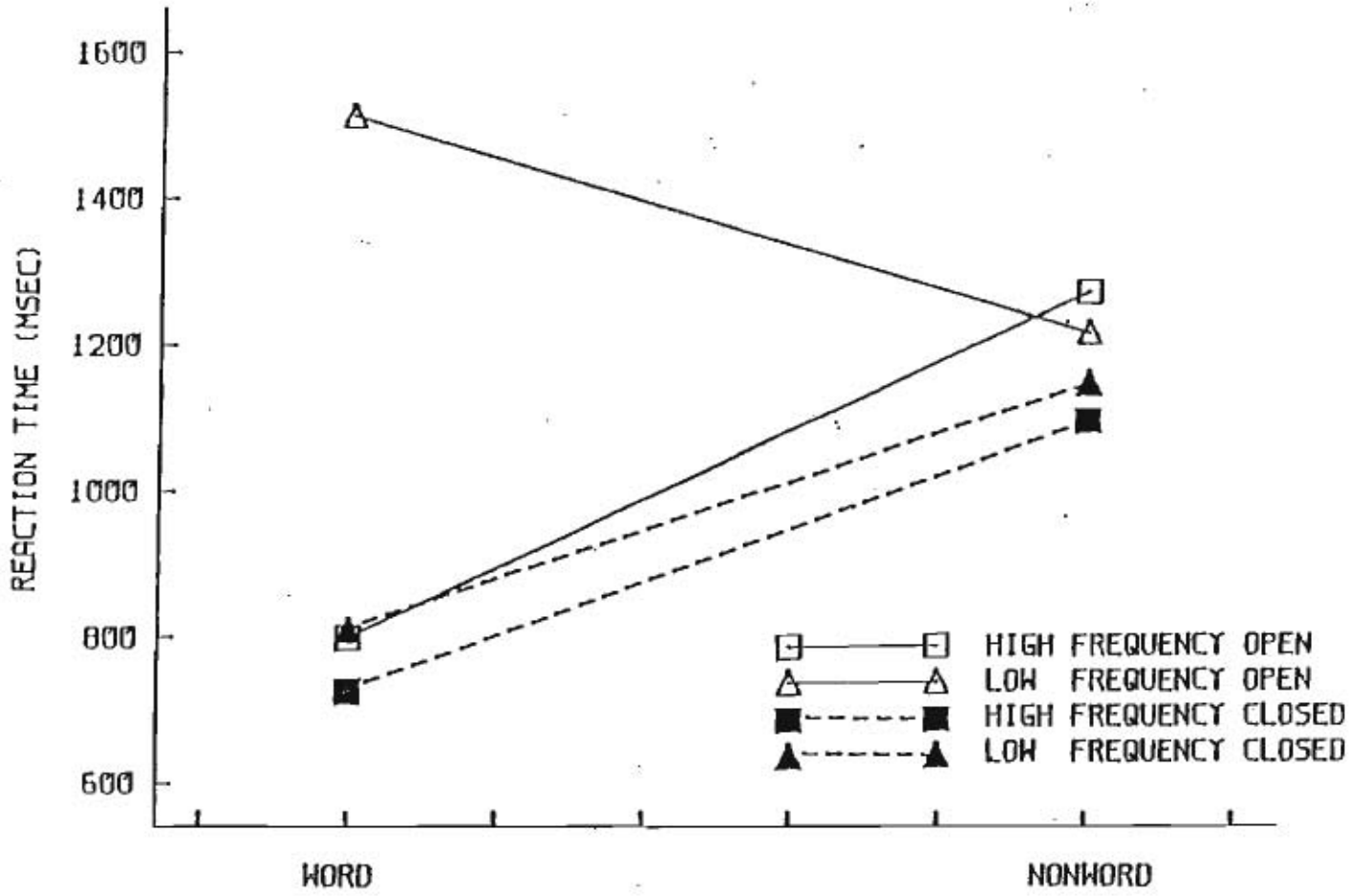


Figure 2. Reaction times for open- and closed-class words and nonwords.

we mentioned earlier, these two categories were matched for frequency and letter length. Note that for the words in both of these categories, there was essentially no difference in reaction times, thus experimentally validating the claim that the low frequency closed-class and high frequency open-class words were matched for frequency and length. However, the mean response times to decide on the lexical status of nonwords made from open-class words was approximately 150 msec longer than the times to decide on the lexical status of nonwords made from a closed-class words. This effect resulted in a significant difference between the low frequency closed-class items and the high frequency open-class items ($F(1,23)=7.49$, $p<0.02$), as well as a significant interaction ($F(1,23)=22.13$, $p<.0001$). Also note that the difference in reaction times between closed-class words and nonwords overall are approximately equal, thus producing no interaction ($F(1,23)=0.39$, $p>0.50$). That is, the slopes of the lines for both the high frequency closed-class and low frequency closed-class items are essentially identical, whereas a greater increase in reaction times from words to nonwords was obtained for the high frequency open-class words.

DISCUSSION

The large difference in mean reaction times between the high frequency open-class and low frequency closed-class nonwords suggests that there may be two important and interrelated properties of lexical access from visually presented word-forms: (1) Retrieval of items from the lexicon may be selectively directed to either the open- or closed-class vocabularies; and (2) the selection of which vocabulary to be searched is based, in part, on graphemic cues to word-form class. Given the demonstrations by Landauer and Streeter (1973) and Eukel (1980) that there are graphemic and phonotactic cues to word frequency, it is by no means surprising that closed-class items may be differentiated from open-class items on the basis of graphemic structure. However, it is not immediately clear at what stage of processing the graphemic structure of a word form becomes an important factor. In other words, one must ask if the graphemic structure of closed-class words makes them easier to encode, if graphemic structure serves as a cue to which vocabulary to search, or if graphemic structure speeds comparison of input items and lexical representations.

By graphemic structure we are referring to any number of regular sequences of letters and redundancies of consonant clusters that characterize the closed-class words. Although we have no stringent formal definition of the graphemic structure of closed-class items at this time, our results show that subjects are able to treat open- and closed-class nonwords differentially. Because these nonwords bear little or no information aside from what their graphemic structure relates, we believe our results support the hypothesis that graphemic cues to word-class membership exist.

Glanzer and Ehrenreich (1979) have recently addressed the issue of encoding with regard to the word frequency effect and have concluded that the word frequency effect is independent of encoding variables. They cite several studies (Stanners, Jastrzembski, and Westbrook, 1975; Landauer, Didner, and Fowlkes, Note 1; and Becker and Killion, 1977) that have employed Sternberg's (1969) additive factors analysis to determine if word frequency and various encoding variables

have interactive effects. These studies have all shown that word frequency and certain variables known to affect encoding have additive, not interactive, effects, thus supporting the notion that word frequency does not affect encoding processes. Because the proposal that word frequency may affect encoding is based, in part, on Landauer and Streeter's (1973) demonstration that high-frequency words are similar graphemically, it is no great inferential leap to generalise the argument of the independence of encoding and graphemic structure to the case of closed-class words.

Having dismissed on the basis of the available evidence that the reaction time differences observed in our experiment between high frequency open-class and low frequency closed-class nonwords did not arise in the encoding stage, what then accounts for the results obtained? At least two possibilities exist: (1) Search of the set of closed-class words can be terminated earlier than search of the set of open-class words because of the smaller set size of the closed-class vocabulary; or (2) comparison processes involving closed-class nonwords are carried out more quickly and/or efficiently than comparison processes for open-class nonwords. In short, the locus of our results may be attributed to either the search stage or the comparison stage in lexical access.

If we adopt a model of lexical access similar to that of Forster (1976, 1979), we can characterize the lexical access process in a visual lexical decision task as follows: When an item is presented, an orthographic code for that input item is built. This code is then used to search sequentially through a list of word-forms in the lexicon. If a match for the input item is found, search terminates. If no match is found, search is terminated either after an exhaustive search or after a response-time deadline is exceeded. Because the items we are attempting to account for are nonwords and are probably not represented in the lexicon at the time of first presentation, search of the lexicon will either be exhaustive or will continue until the subject terminates searching after some given time.

If, as we have argued, early decisions about the membership of an item in one of the two word-form classes can be made on the basis of graphemic cues, then search may be directed to one of two lists in the lexicon, one list containing the closed-class words and one containing the open-class words. If we assume that search of either vocabulary in the lexicon begins with higher frequency words and continues through successively lower frequency words, then open- and closed-class words matched for frequency will be accessed in an equivalent time span, which is precisely what we found. However, when nonwords are presented that are similar graphemically to either open- or closed-class words, an estimate of the time needed to complete an exhaustive or near exhaustive search of the open- and closed-class vocabularies is obtained from the time needed to decide that an input word-form is a nonword. Because the set of closed-class words is considerably smaller than the set of open-class words matched for frequency, a structural dissociation of the lexicon into open- and closed-class vocabularies should reveal itself in faster reaction times for rejecting nonwords made from closed-class words than for rejecting nonwords made from open-class words. This again is precisely what we found.

An explanation in terms of differences in search times is attractive in its simplicity and points directly to a structural dissociation in the lexicon between the open- and closed-class vocabularies. Also, given this interpretation of our findings, it is easy to incorporate the open-/closed-class distinction into many existing models of lexical access (although Morton's logogen model may prove more problematic, unless one is willing to postulate two sets of logogens, one for each word-form class). However, an alternative explanation of our results is possible.

As we noted in the introduction, the results reported by Healy (1978,1980), Drewnowski and Healy (1977), Haber and Schindler (1981), and Ferres and Aaronson (1981) suggest that closed-class words may be visually processed in perceptual units above the letter level. In other words, closed-class words may be highly automatized, in Schneider and Shiffrin's (1977; Shiffrin and Schneider, 1977) terms, and may be accessed with little computational effort. If this is the case, after preliminary graphemic cues have directed search toward the closed-class vocabulary, comparison processes may be almost effortless given the level of the unit at which the comparison is made. That is, the graphemic structure of closed-class words may be so cohesive that any deviation from this structure would be immediately identified. Thus, nonwords made from closed-class words could be compared to the representations for closed-class words and rejected at little or no time cost.

Notice that this explanation need not assume, as we have, that there is a real structural dissociation between the open- and closed-class vocabularies in the lexicon. The differences in reaction times we obtained between high frequency open-class and low frequency closed-class nonwords may simply be the result of faster comparison times for closed-class items. Thus, one need not propose a distinct structural dissociation of the two vocabularies in order to account for the present results. However, placing the locus of the reaction time differences in the comparison stage is tantamount to proposing a <u>functional</u> dissociation of the vocabularies.

An explanation in terms of differences in the comparison stage appears to be somewhat anomalous given the recent findings of Haber and Schindler (1981) who showed that closed-class words "conceal their misprints" (p. 573) better than open-class words of the same length. If misprints in closed-class words are likely to be missed, it may be that <u>more</u> rather than less time is spent in the comparison stage for correctly identifying that an item made from a closed-class word is a nonword. That is, any discrepancies noted in highly unitized closed-class items may require a time-consuming decomposition of the item prior to comparison. However, in the Haber and Schindler study, it is unclear whether syntactic predictability (or contextual constraint in general) caused subjects to overlook misprints in closed-class words or whether processing of the word-forms above the letter level (i.e., unitization) was the source of the errors in detecting misprints. In fact, as Haber and Schindler note, both predictability and unitization may have contributed to their results.

Because there was little, if any, influence of syntactic predictability in our study, we are left with the possibility that disturbances in highly cohesive graphemic structures are extremely salient to the reader and therefore reduce

comparison times. Moreover, given our finding that closed-class nonwords were actually responded to slightly more accurately than open-class nonwords matched for length and frequency, it is not likely that the closed-class items in our study were concealing their misprints. In fact, as we observed, the misprints become more rather than less salient. Obviously, further experiments will have to be performed in order to decide between the search and comparison hypotheses.

Under either hypothesis, though, we have found no convincing evidence to support Bradley's notion that there are separate retrieval mechanisms for the open- and closed-class vocabularies. If we take the search hypothesis to be an adequate account of our results, we may then propose a structural dissociation in the lexicon between the open- and closed-class vocabularies. If such a dichotomy should exist, the computational efficiency presumably afforded by a separate retrieval mechanism for closed-class words is easily accounted for by the reduced size of the search set. On the other hand, if we adopt the comparison-stage hypothesis, we need propose neither separate retrieval mechanisms nor a structural dissociation of the open- and closed-class vocabularies in the lexicon. However, we may still propose a functional distinction between the open- and closed-class vocabularies. That is, closed-class words on the whole may be more highly unitized or automatized than open-class words matched for frequency and length. Thus, the hypothesized computational efficiency involved in retrieving closed-class words may result from an extremely cohesive graphemic structure that allows fast and effortless retrieval of closed-class words from the lexicon (see Shiffrin, Dumais, and Schneider, 1981).

In summary, neither of the hypotheses we have considered favor the postulation of separate retrieval mechanisms for the open- and closed-class vocabularies. In the introduction, we mentioned two alternatives that Bradley suggested could account for her nonword interference results for closed-class items. Considering the evidence presented here, we prefer the second alternative: Nonwords constructed so that they contain closed-class words in their initial positions produce no interference effects because they can be rejected at little or no time cost, presumably because graphemic cues to word-form class are sufficient to allow fast rejection of nonwords containing closed-class words or because the graphemic structure of closed-class items is such that it allows fast and efficient comparison of items to representations residing in the lexicon. In spite of this conclusion, the psychological plausibility of the open-/closed-class distinction remains intact. However, identifying the source of the differences between the two word-form class has proven difficult. In short, our results suggest that the locus of the distinction may lie in a structural dissociation of the open- and closed-class vocabularies in the lexicon or in a functional dissociation involving differential comparison processes for closed-class items.

REFERENCES

Becker, C. A., & Killion, T. H. Interaction of visual and cognitive effects in word recognition. Journal of Experimental Psychology: Human Perception and Performance, 1976, 2, 556-566.

Bolinger, D. Aspects of language. New York: Harcourt, Brace, and Jovanovich, 1975.

Bradley, D. C. Computational distinctions of vocabulary type. Unpublished doctoral dissertation, MIT, 1978.

Bradley, D. C., Garrett, M. E., & Zurif, E. B. Syntactic deficits in Broca's aphasia. In D. Caplan (Ed.), Biological studies of mental processes. Cambridge, Mass.: MIT Press, 1980.

Brown, R., & Bellugi, U. Three processes in the child's acquisition of syntax. Harvard Educational Review, 1964, 34, 138-139.

Drewnowski, A., & Healy, A. f. Detection errors on the and and: Evidence for reading units larger than the word. Memory and Cognition, 1977, 5, 636-647.

Epstein, W. Influence of syntactical structure on learning. American Journal of Psychology, 1961, 74, 80-85.

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

Ferres, S. A word class encoding model for adults and children in comprehension and recall tasks. Unpublished doctoral dissertation, New York University, 1981.

Ferres, S., & Aaronson, D. A word class encoding model for reading times. Paper presented at the Psychonomics Society, Philadelphia, 1981.

Flanagan, J. L., Coker, C. H., Rabiner, L. R., Schaefer, R. W., & Umeda, N. Synthetic voices for computers. IEEE Spectrum, 1970, 7, 22-45.

Forster, K. I. Accessing the mental lexicon. In R. J. Wales and E. Walker (Eds.), New approaches to language mechanisms. Amsterdam: North Holland, 1976.

Forster, K. I. Levels of processing and the structure of the language processor. In W. Cooper & E. Walker (Eds.), Sentence processing. New York: Halstead, 1979.

Fries, C. C. The structure of English. New York: Harcourt, Brace, and World, 1952.

Garrett, M. Syntactic processes in sentence production. In R. J. Wales and E. Walker (Eds.), New approaches to language mechanisms. Amsterdam: North Holland, 1976.

Garrett, M. Levels of processing in sentence production. In E. Held, H.-L. Teuber, and E., Leibowitz (Eds.), Handbook of sensory physiology VIII. New York: Springer-Verlag, 1979.

Glanzer, M, & Ehrenreich, S. L. Structure and search of the internal lexicon. Journal of Verbal Learning and Verbal Behavior, 1979, 18, 381-398.

Gordon, B., & Caramazza, A. Lexical decision for open- and closed-class words: Failure to replicate differential frequency sensitivity. Brain and Language, 1982, 15, 143-160.

Haber, R. N, & Schindler, R. M. Errors in proofreading: Evidence of syntactic control of letter processing? Journal of Experimental Psychology: Human Perception and Performance, 1981, 7, 573-579.

Healy, A. F. Detection errors on the word the: Evidence for reading units larger than letters. Journal of Experimental Psychology: Human Perception and Performance, 1976, 2, 235-242.

Healy, A. F. Proofreading errors on the word the: New evidence on reading units. Journal of Experimental Psychology: Human Perception and Performance, 1980, 6, 45-57.

Holbrook, M. B. Effect of subjective interletter similarity, perceived word similarity and contextual variables on the recognition of letter substitution in a proofreading task. Perceptual and Motor Skills, 1978, 47, 243-250.

Kucera, H., & Francis, W. N. Computional analysis of present-day American English. Providence, R. I.: Brown University Press, 1967.

Landauer, T. K., & 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.

Miller, G. A., Newman, E. B., & Friedman, E. A. Length-frequency statistics for written English. Information and Control, 1958, 1, 370-389.

Morton, J. Interaction of information in word recognition. Psychological Review, 1968, 76, 165-178.

Nelson, K. Structure and strategy in learning to talk. Monographs of the Society for Research in Child Development, 1973, 38(1-2, Serial No. 149).

Nelson, K. Individual differences in language development: Implications for development and language. Developmental Psychology, 1981, 17, 170-187.

O'Shaughnessy, D. Modelling fundamental frequency, and its relationship to syntax, semantics, and phonetics. Unpublished doctoral dissertation, MIT, 1976.

Selkirk, E. The phrase phonology of English and French. Unpublished doctoral dissertation, MIT, 1972.

Selkirk, E. On prosodic structure and its relation to syntactic structure. Indiana University Linguistics Club, 1980.

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.

Schneider, W., & Shiffrin, R. M. Controlled and automatic human information processing: I. Detection, search, and attention. Psychological Review, 1977, 84, 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, 84, 127-190.

Shiffrin, R. M., Dumais, S. T., & Schneider, W. Characteristics of automatism. In J. Long & A. Baddeley (Eds.), Attention and Performance IX. Hillsdale, N. J.: Lawrence Erlbaum Associates, 1981.

Sternberg, S. The discovery of processing stages: Extensions of Donder's method. Acta Psychologica, 1969, 30, 276-315.

Taft, M., & Forster, K. I. Lexical storage and retrieval of prefixed words. Journal of Verbal Learning and Verbal Behavior, 1975, 14, 638-647.

Taft, M., & Forster, K. I. Lexical storage and retrieval of polymorphemic and polysyllabic words. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 607-620.

REFERENCE NOTES

1. Landauer, T. K., Didner, R. S., & Fowlkes, E. B. Processing stages in word naming: Reaction time effects of letter degradation and word-frequency. Unpublished paper, 1974.

II. SHORT REPORTS AND WORK-IN PROGRESS

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

Listening to Open and Closed Class Words in Fluent Speech*

Aita Salasoo and David B. Pisoni

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

*This research was supported by NIMH Research Grant MH-24027 and NINCDS Research Grant NS-12179.

Abstract

Listeners detected open class word targets (i.e. nouns, verbs and adjectives) more accurately than closed class targets (i.e. determiners, prepositions and conjunctions) in a dual-task paradigm that required simultaneous processing of fluent speech for comprehension. The results provide direct behavioral evidence that the open-closed class distinction reflects a fundamental dichotomy in the mental lexicon that is deployed in the real-time processing of fluent speech.

Damage to Broca's area of the brain is frequently associated with the omission of closed class words in speech production and corresponding abnormal deficits in responding to closed class words in listening, reading and writing (1,2). Open class words do not appear to be affected in the same way. Additional differences in processing between open and closed class words are evidenced by error patterns in speech production, by memory and parsing performance of sentences in the presence and absence of closed class words, by differential patterns of language development in the normal population and by results from proofreading and eye movement studies. For example, word exchange errors in speech production, e.g. "I left the briefcase in my cigar.", are restricted to transpositions of either closed class or open class items (3). Also, closed class words generally appear later than open class words in the speech and comprehension performance of children (4). Behavioral data from proofreading and eye movement studies also point to differences in reading open and closed class words: More visual attention is given to open class words when we normally read. Short, closed class words are rarely fixated during silent reading (5). And, more errors occur for closed class words in tasks requiring detection of particular letter targets or misspellings in written passages (5).

The present study examined the perception of open and closed class words in a task that required listeners to comprehend continuous speech. Using a dual-task paradigm, we found that normal listeners attended more to open class words than closed class words. This finding provides direct behavioral evidence for a word class distinction in processing fluent speech. The results suggest that the mental lexicon may be structured into two classes of words which play vastly different roles in the real-time analysis of spoken language.

In our study, open and closed class target words in spoken passages were selectively replaced by envelope-shaped noise. Listeners were required to monitor for the presence of these noise targets while they attempted to understand the passages. The assumption underlying this "dual-task" procedure is that the computational processes influencing the recognition of the original word in context also affect the accuracy of noise detection (6). The target words varied on three parameters: word class [open vs. closed], word length [short vs. long], and contextual constraint (CC) -- defined here as the predictability of the target word from the prior left-to-right context [low or high] (7).

Ten passages of connected text were chosen from scientific articles and speeches. Each passage contained 16 target words; half were open class and half were closed class items. The contextual constraint for each target word was determined independently using a variant of the Cloze procedure (8,9). Using this procedure, we calculated an index of contextual constraint for each target word, based on inter-subject prediction agreement: A highly constrained item corresponded to a word that was predictable from prior context; an item of "low" constraint corresponded to a word that was unpredictable.

The test passages were read at a normal rate by a male speaker, low-pass filtered and stored digitally via a 12-bit A/D converter. The beginning and end of each target word were located in the passages using a digitally controlled waveform editor; these segments of speech were then replaced entirely by new segments that consisted of envelope-shaped noise. The replacement noise was

matched to the duration and amplitude of the original segments of the waveform corresponding to each of the target words (10). Figure 1 shows examples of sound spectrograms of an original intact target word, right, (Panel A) and the target (Panel B) after it was replaced with envelope-shaped noise.

Insert Figure 1 about here

For each passage, five true-false comprehension questions were constructed. The primary comprehension task was designed to ensure that subjects actually attended to the content of each spoken passage and performed normal cognitive activities concurrently with the monitoring task. Two experimental conditions varied the amount of prior contextual knowledge available (presence or absence of typed transcripts) and the response mode.

In the transcript condition, 40 listeners had typed versions of each of the spoken passages available in front of them while they listened to the speech over headphones. Subjects were instructed to listen and try to understand the content and meaning of the passages. At the same time, they were also required to monitor for the presence of noise targets in the speech. Listeners responded to the noise targets by marking their location directly on the transcripts. Listeners in this condition had visual access to the identity of the actual target words in front of them. Thus, their responses were strongly related to both lexical and contextual properties of the target words in the passages.

In the second condition, another group of 58 subjects listened to the same spoken passages without transcripts. Thus, only information from the preceding spoken context, and from the noise target itself, could be used to detect the test items. Listeners responded to the noise targets by pressing a button that was interfaced to a minicomputer. In both conditions, the written comprehension questions were answered immediately after presentation of each passage.

The proportion of correct target detections to each test item was collected for each passage (11). The mean detection results summed over the ten passages are shown in Figure 2 as proportions of correct detections. The left-hand panel displays the data from the Transcript Condition in which subjects had the written passage in front of them; the right-hand panel shows the data from the No Transcript Condition, in which subjects listened only to the speech without any contextual support.

Insert Figure 2 about here

In both conditions, a significant word class effect was obtained: Open class targets were detected more often than closed class targets. [In the

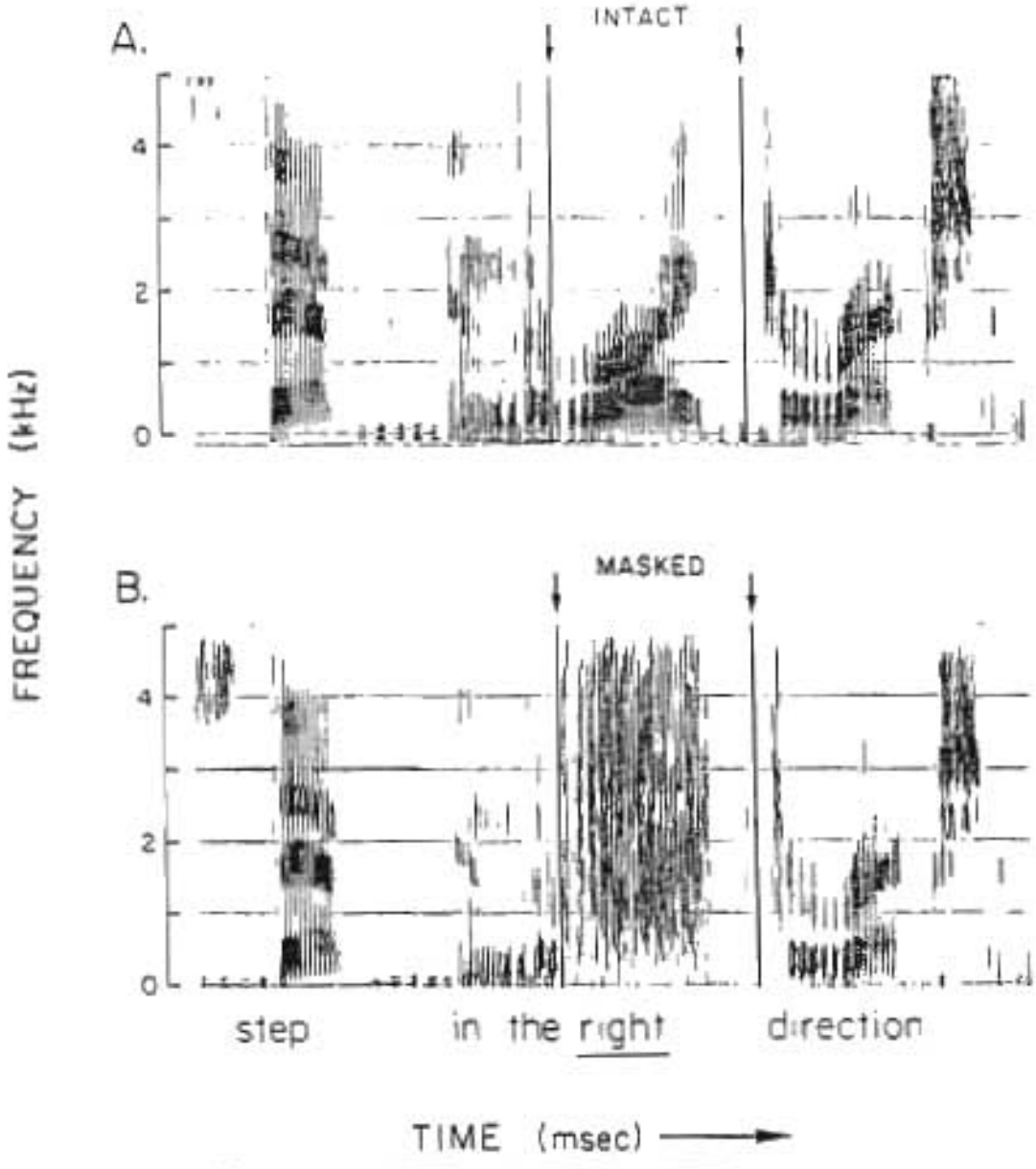


Fig. 1. (A) Spectrogram of the natural utterance "step in the right direction". The beginning and end of the original (intact) target word, right, are indicated by arrowheads. (B) Spectrogram of the same waveform after experimental manipulation of the target word, right. The original word was replaced entirely by envelope-shaped noise which preserved the overall duration and amplitude but obliterated the spectral information.

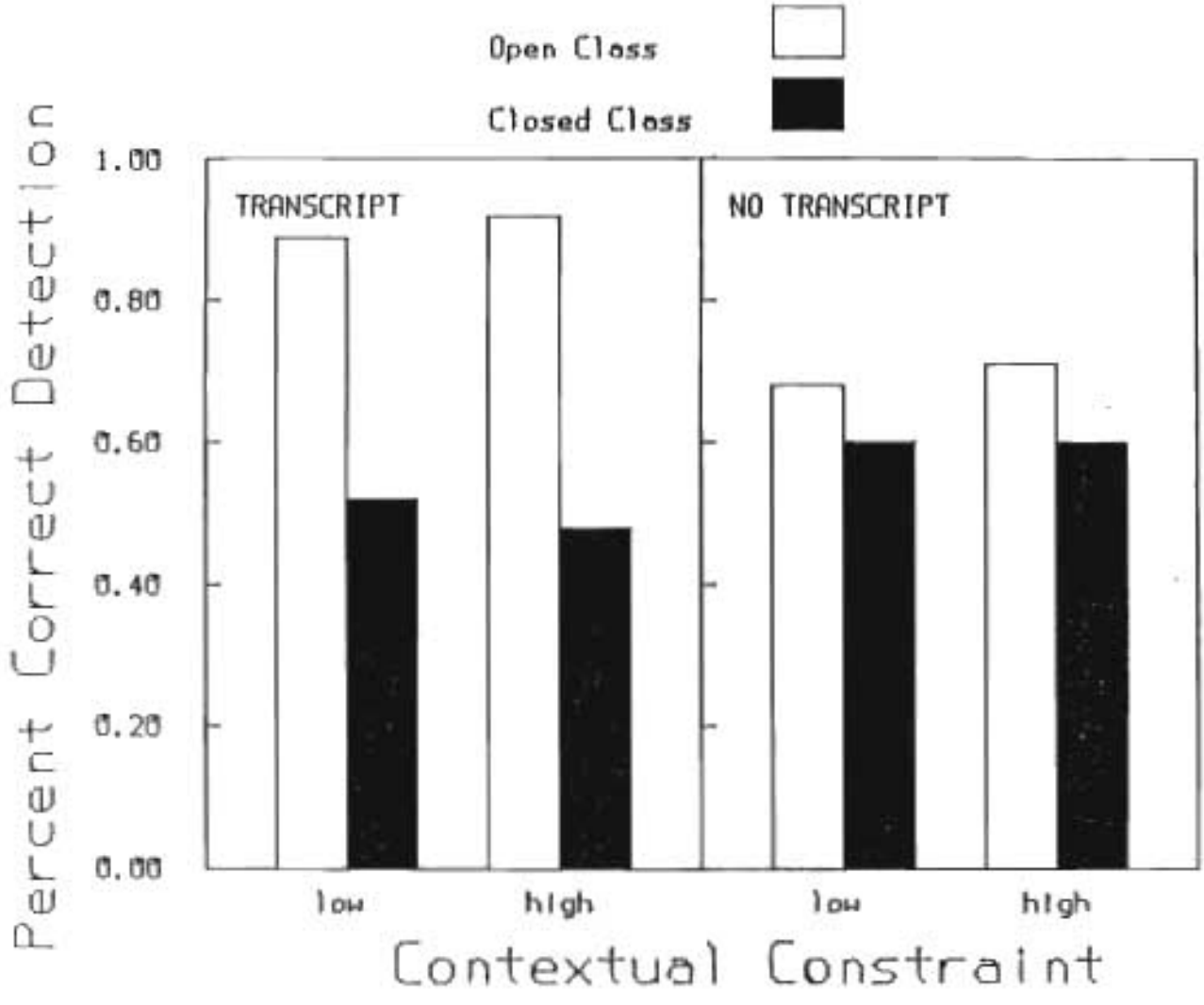


Fig. 2. Noise detection results for the Transcript Condition (visual and auditory context) and the No Transcript Condition (auditory context only). The figure displays the proportion of correct detections of open class word targets (open bars) and closed class word targets (solid bars) under both Low and High Contextual Constraint for each condition separately.

Transcript Condition, $F_{(1,36)}=173.17, p<.0001$; in the No Transcript Condition, $F_{(1,54)}=8.99$, $p<.005$.] No overall effects of contextual constraint of the target words were observed, nor were any significant correlations found between detection accuracy and comprehension scores for each passage. Open class targets were detected more accurately in the Transcript Condition, while closed class items were unaffected by this manipulation ($F_{(1,92)}=15.99$, $p<.001$). Thus, our major finding is that across both listening conditions, open class targets were detected with significantly greater accuracy than closed class targets, regardless of target word duration, knowledge of prior context and response mode.

The present monitoring data indicate that word class, in part, determines the level at which spoken words are processed by listeners during comprehension of spoken language. The word class effect occurs when the materials are presented auditorily, visually or in both sensory modalities: Both in reading text and in listening to fluent speech, more of the reader-listener's conscious attention is allocated to open class words than to closed class words. The linguistically motivated distinction between open and closed class words therefore has important behavioral implications for the on-line processing of fluent speech, since it may reflect a fundamental structural dichotomy in the organization of words in the mental lexicon. Such a structural dichotomy of the lexicon may facilitate on-line word recognition processes and subsequent syntactic analyses that enable rapid isolation and interpretation of meaning carried by open class words during listening comprehension (12).

In summary, the present results demonstrate that open and closed class words are processed differentially by adult listeners in perceiving passages of fluent speech. The finding that closed class noise targets are more difficult to detect than open class noise targets is consistent with the notion that closed class words form a special subset of the vocabulary of a language and serve as markers to the grammatical organization of sentences (1,12). Our behavioral data from the on-line perception of fluent speech add to the growing body of developmental, neurological and psycholinguistic research that supports the existence of important processing differences based on the structural distinction between open and closed class items in the mental lexicon. Taken together with these earlier findings, our results indicate that a functional dichotomy of the lexicon into two word classes may facilitate parsing and subsequent semantic interpretation that depends, in large part, on prior detailed structural analyses of the linguistic input. Our findings demonstrate that such a dichotomy is displayed in the behavior of listeners when asked to comprehend passages of fluent speech.

References and Notes

1. D. C. Bradley, Unpublished Ph.D. thesis, MIT, 1978; D. C. Bradley, M. F. Garrett & E. B. Zurif, In Biological Studies of Mental Processes, D. Caplan, Ed. (MIT, Cambridge, 1980); A. Caramazza & E.B. Zurif, Brain & Language, 3, 572 (1976); M.-L. Kean, Cog., 7, 69 (1979).

2. The open-closed class distinction has pervaded structural linguistics, e.g. C. C. Fries, The Structure of English. (Harcourt, Brace & World, New York, 1952). The contrast has enjoyed many labels and definitions. [D. C. Bradley, cf. Note 1; L. Gleitman & E. Wanner, Eds. Language Acquisition: The State of the Art (Cambridge U.P., Cambridge, 1982); M.-L. Kean, Cog., 5, 9 (1977); H. Kolk, Cog., 6, 349 (1979)] Briefly, open class words carry the semantic meaning in sentences, while closed class words are relational markers of syntactic structure. In English, the two classes of words differ in their vocabulary size, frequency of occurrence and length, as well as their phonological properties and stress assignment. Closed class words form a small subset of the lexicon, occur very frequently in the language and are typically very short in duration. They are also commonly acoustically reduced and occur in unstressed sentence positions.

3. V. A. Fromkin, Speech Errors as Linguistic Evidence (Mouton, The Hague, 1973); M. F. Garrett in New Approaches to Language Mechanisms, E. Walker & R. Wales, Eds. (North Holland, Amsterdam, 1976); M. F. Garrett in Language Production. Vol. I. Speech and Talk, B. Butterworth, Ed. (Academic, London, 1980) p.177.

4. Young children omit closed class items from their speech, have difficulty with tasks requiring repetitions and explicit attention to the internal structure of closed class items, and are typically not given explicit instruction about their usage until open class words have been acquired [L. Bloom, One Word at a Time (Mouton, The Hague, 1973); A. Drewnowski, J. Exp. Child Psych., 31, 154 (1981); M. H. Holden & W. H. MacGinitie, J. Ed. Psych., 63, 551 (1972); P. Rozin & L. Gleitman, in Toward a Psychology of Reading, A. S. Reiber & D. L. Scarborough, Eds. (LEA, Hillsdale, N.J., 1977).]

5. Proofreading and letter detection studies have shown that misspellings and target letters are less likely to be detected in some closed class words, e.g. "and", "the". [A. Drewnowski & A. F. Healy, J. Verbal Learning & Verbal Behav., 19, 247 (1977); A. Drewnowski & A. F. Healy, Mem. & Cog., 10, 145 (1982); A. F. Healy, J. Exp. Psych: HP & P, 6, 45 (1980)] These findings, however, are affected by the predictability of words from preceding context. [R. N. Haber & R. M. Schindler, J. Exp. Psych: HP & P, 7, 573 (1982)]. Word class effects on eye movements during reading also depend on preceding context [M. A. Just & P. A. Carpenter, Psych. Review, 87, 329 (1980); G. W. McConkie & D. Zola, in Interactive Processes in Reading, A. M. Lesgold & C. A. Perfetti, Eds. (Erlbaum, Hillsdale, N.J., 1981) p.155; K. O'Regan in Processing of Visible Language, 1, (Plenum, New York, 1979) p.49].

6. A. Cutler & D. Norris in Sentence Processing, W. Cooper & E. Walker, Eds. (Halsted, N.Y., 1979); J. A. Fodor, T. G. Bever & M. F. Garrett, The Psychology of Language (McGraw-Hill, N.Y., 1974).

7. Noting the confounding of sentence context effects and word class in proofreading and eye movement studies [c.f. Note 5], the two factors were manipulated separately in our study as the variables of contextual constraint and word class, respectively.

8. In this task, a separate group of 62 subjects was presented with the test passages on CRT screens up to the word immediately before the first target word and asked to guess the next word. When all subjects had written down their prediction of the deleted word, the original word and all the words in the passage up to the next target word were presented on the screen for subjects to read. In this way, only the preceding context was available to help readers predict each target word, presumably just as listeners have only the prior-spoken context to help them recognize each word when listening to fluent speech [W. L. Taylor, Journalism Quarterly, 30, 415 (1953)].

9. Subjects in the Cloze procedure and in all listening conditions were native English-speaking students enrolled at Indiana University, in Bloomington.

10. The envelope-shaped noise preserved the duration and amplitude of the original word while destroying the spectral information and formant structure used to identify the phonetic properties of words. This procedure was accomplished by randomizing the sign bit on output of the digital waveform through the D/A converter. [See Y. Horii, A. S. House & G. W. Hughes, JASA, 49, 1849 (1971)]

11. Regression-based analyses of variance were performed on arcsin transformations of correct detections collapsed over materials to partial out covariance due to word length. (cf. Note 1.) [F. N. Kerlinger & E. J. Pedhazur, Multiple Regression in Behavioral Research (Holt, Rinehart & Winston, New York, 1973)]

12. D. Aaronson, J. Exp. Psych.: HP & P, 2 42 (1976); A. Cutler & J. A. Fodor, Cog., 6 (1979); M. F. Garrett in Handbook of Sensory Physiology VIII, R. Held, H-L. Teuber & H. Leibowitz, Eds. (Spring Verlag, N.Y., 1979); J. Kimball, Cog., 2, 15 (1974).

13. We thank R. Bernacki, H. Brunner, J. Forshee and B. Greene for their assistance. This research was supported by NIMH research grant MH-24027 to Indiana University in Bloomington.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

Time-varying features in voiced and voiceless stops produced at different speaking rates*

Diane Kewley-Port and Paul A. Luce

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

*This work was supported by NINCDS Research Grant NS-12179 to Indiana University in Bloomington. We thank D. B. Pisoni for his interest and advice.

Abstract

Recent work has demonstrated that place of articulation in initial voiced stops can be accurately identified from visual displays of linear prediction smoothed spectra updated at 5 msec intervals (Kewley-Port, 1982). The present research extends this analysis in several directions. First, the visual displays were modified to incorporate auditory filtering and other frequency characteristics of the ear. Second, our earlier time-varying features were redefined to describe syllable initial stops in fluent speech. In the present experiment, we examined syllables beginning with voiced and voiceless stops paired with five vowels. Two males and two females read the syllables in a carrier phrase at fast, normal, and slow tempos. Employing a predefined set of time-varying features, judges first attempted to locate the burst and were then required to identify place of articulation in the visual displays. The results indicated that the new time-varying features were adequate for identifying place of articulation invariantly across voiced and voiceless stops, talkers, and speaking rates.

Kewley-Port (1982) has recently reported a study of the invariant time-varying acoustic properties of voiced stops. In this study, tokens of the stops /b/, /d/, and /g/ before eight vowels were obtained from two male talkers and one female talker. Visual displays of the linear prediction running spectra for the first 40 msec of each consonant-vowel syllable were prepared for inspection. From these displays, three time-varying features for identifying place of articulation were formally defined. Judges were then asked to determine, from the displays, place of articulation on the basis of the three features. Kewley-Port found that the judges were able to successfully identify place of articulation for 88% of the displays. Upon closer examination of the data, it was found that although place was identified invariantly across vowel contexts, it was not identified invariantly across speakers. In particular, identification of place of articulation was considerably worse for the female talker than for the two male talkers.

In the present study, we addressed the problem of specifying time-varying features that are invariant across talkers as well as vowel contexts. In addition, following suggestions made by Kewley-Port (1982), we extended the original research in a number of ways. First, we modified the analysis of the frequency characteristics of the waveforms in order to more closely model the auditory filtering properties of the ear. Second, we collected a new data base from two additional male and two additional female talkers producing both voiced and voiceless initial stop consonants. Finally, in a preliminary attempt to generalize Kewley-Port's findings to connected text, consonant-vowel syllables were spoken in a carrier phrase at three speaking rates: normal, fast, and slow.

In the previous study, running spectral displays for each consonant-vowel syllable were prepared. In this study, these displays were modified to include auditory filtering. For these auditory filtered displays, new time-varying features were also defined. In addition to identifying place of articulation on the basis of these new features, the judges in this study were asked to identify the onset of the burst and the onset of voicing. The purpose of the present experiment, therefore, was to determine if the new time-varying features derived from visual displays of the auditory filtered running spectral displays could be successfully employed by judges to identify the burst frame as well as place of articulation. Moreover, we were interested in determining whether these features were invariant across vowel contexts, talkers, and speaking rates.

The experiment consisted of two parts. In the first part, we examined only the voiceless stop consonants /p/, /t/, /k/. In the second part, we examined both voiced and voiceless stop consonants produced at three different speaking rates.

Method

Data collection. The stimuli for both parts of the experiment were collected in a single session for each talker. The talkers were recorded while they read 30 stop consonant-vowel syllables embedded in the carrier phrase "Teddy said ____." The syllables consisted of all combinations of /b,d,g,p,t,k/ paired with the vowels /i,æ,ɑ,ɔ,u/. Each list of 30 syllables in the carrier phrase was randomized and the stimuli were presented one at a time on a CRT videodisplay

monitor (GBC Model MV-10A) under computer control. The talkers were recorded in a sound treated booth (IAC Model 401A) using an Electro-Voice DO54 microphone and an Ampex AG-500 taperecorder. The talkers read five blocks of the 30 sentences. Each talker was instructed to read each sentence at a normal rate.

In addition, one male and one female talker read five blocks of the same stimuli at a fast rate followed by five blocks at a slow rate. The presentation of the stimuli in the fast and slow conditions was identical to the presentation of the stimuli in the normal rate condition.

Data analysis. Tokens from each talker from the third and fourth blocks of sentences were low-pass filtered at 4.8 kHz and digitized at a 10 kHz sampling rate via a twelve-bit analog-to-digital converter. When a token was deemed unacceptable because of a mispronunciation or because of noise in the signal, another token of the same utterance was taken from the fifth block of sentences. Mispronunciations were few and occurred mostly a the slow speaking rate.

After digitizing, the stop consonant-vowel syllables were spliced out of the carrier sentences using the waveform editing program, WAVES (see Luce and Carrell, 1982). The syllables were then edited so that at least 60 msec of the stop closure was included in the waveform. To produce the running spectral displays, linear prediction analysis was carried out using the program SPECTRUM (Kewley-Port, 1979). Twenty msec waveform segments were first pre-emphasized. Then, using a 25.6 msec Hamming window, fourteen autocorrelation coefficients were calculated. Each spectral section was calculated using a 256 point FFT, with the "frames" being updated every five msec. The first spectral section was positioned during the stop closure such that between one and six frames of closure randomly preceded the burst frame. A total of 15 frames (70 msec) were analyzed for each syllable.

A linear prediction running spectrum for the syllable /bu/ displayed on a standard linear frequency axis is shown in the top half of Fig. 1.

Insert Figure 1 about here

It was argued in the previous study (Kewley-Port, 1982) that representing energy in dB and updating the spectral frames every five msec is appropriate for modelling the processing of speech by the ear. However, this is not true for the frequency representation rendered by linear prediction analysis. The bandwidth of the frequency analysis performed by the ear is not equal across all frequencies, as it is represented in the top display in Fig. 1, but instead increases as frequency increases. Various estimates of auditory bandwidth have been made in the past and they range from at most one-half octave bandwidths to one-tenth octave bandwidths. The ranges most often reported are shown by the hatched area in Fig. 2.

/bu/

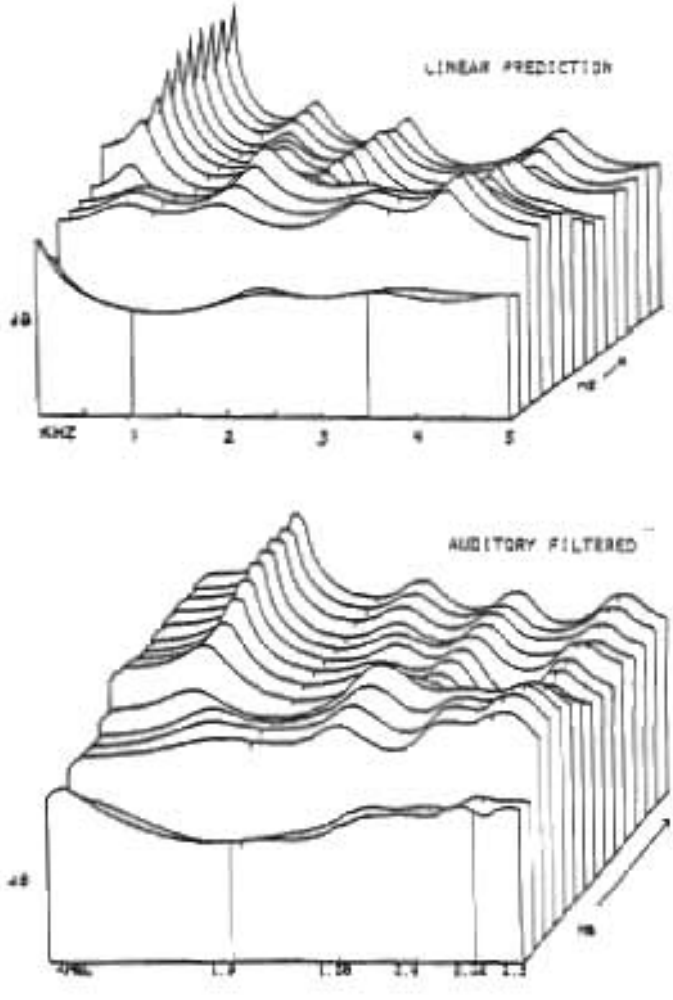


Figure 1. Linear prediction running spectral displays for /bu/ with (bottom panel) and without (top panel) auditory filtering.

Insert Figure 2 about here

The bark scale, proposed by Zwicker (1961), is shown as an upper limit, while one-sixth octave filtering similar to that proposed by Patterson (1976) is shown as the lower limit. Note how inappropriate the linear prediction bandwidths, labelled LP, are for modelling the frequency analysis performed by the ear.

To simulate auditory filtering in this study, one-sixth octave bandwidths were chosen following Patterson's results (Patterson, 1976; Patterson and Nimmo-Smith, 1980). The filter shape was trapezoidal and had a 75 dB/octave roll-off in the skirts. These filters were convolved with the linear prediction smoothed spectra to produce auditory filtered spectra. The frequency scale was changed to the technical Mel scale for displaying the running spectra, an example of which is shown in the lower half of Fig. 1. An auditory filtered running spectra was produced in this way for each syllable studied.

Four time-varying features were formally defined for the auditory filtered running spectral displays. The first feature--the occurrence of an abrupt increase in energy at high frequencies--was defined to allow identification of the burst frame. The second feature--the onset of a prominent, narrow, low-frequency peak that is continuous with succeeding frames--was defined to allow identification of onset of voicing. The third and fourth features were defined to allow assignment of place of articulation. These two features--the spectral tilt of the burst and succeeding voiceless frames and the presence or absence of mid-frequency peaks extending in time--were based on predictions made by the acoustic theory of speech production (Fant, 1960, 1973; Stevens and Blumstein, 1978). One feature used in the previous study was dropped from this analysis. This feature, called late onset of low frequency energy, is a measure of voice-onset time (VOT). Since three speaking rates were used in this study, it was presumed that this feature would not be invariant across the utterances examined here.

Judging. To illustrate how these features were used in judging the displays, we will refer to Fig. 3, where examples of the running spectral displays for /pi/, /da/, and /ku/ are shown.

Insert Figure 3 about here

Two graduate students in psychology with formal training in phonetics served as judges in this experiment. For each display, the judges were instructed to locate the burst frame and the onset of voicing, if present, and then to decide on the spectral tilt and the presence or absence of mid-frequency peaks. To locate the burst frame, the judges were instructed to find the first frame in

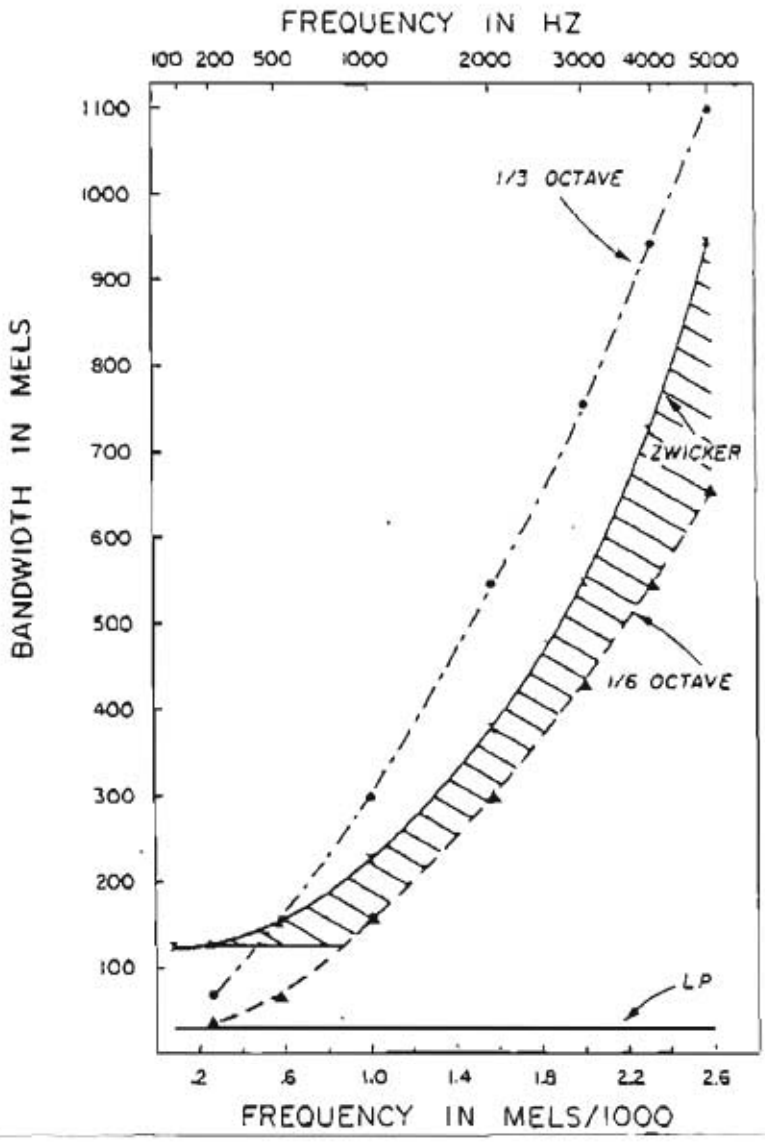


Figure 2. Four different filter bandwidths as a function of frequency are displayed using the technical Mel scale.

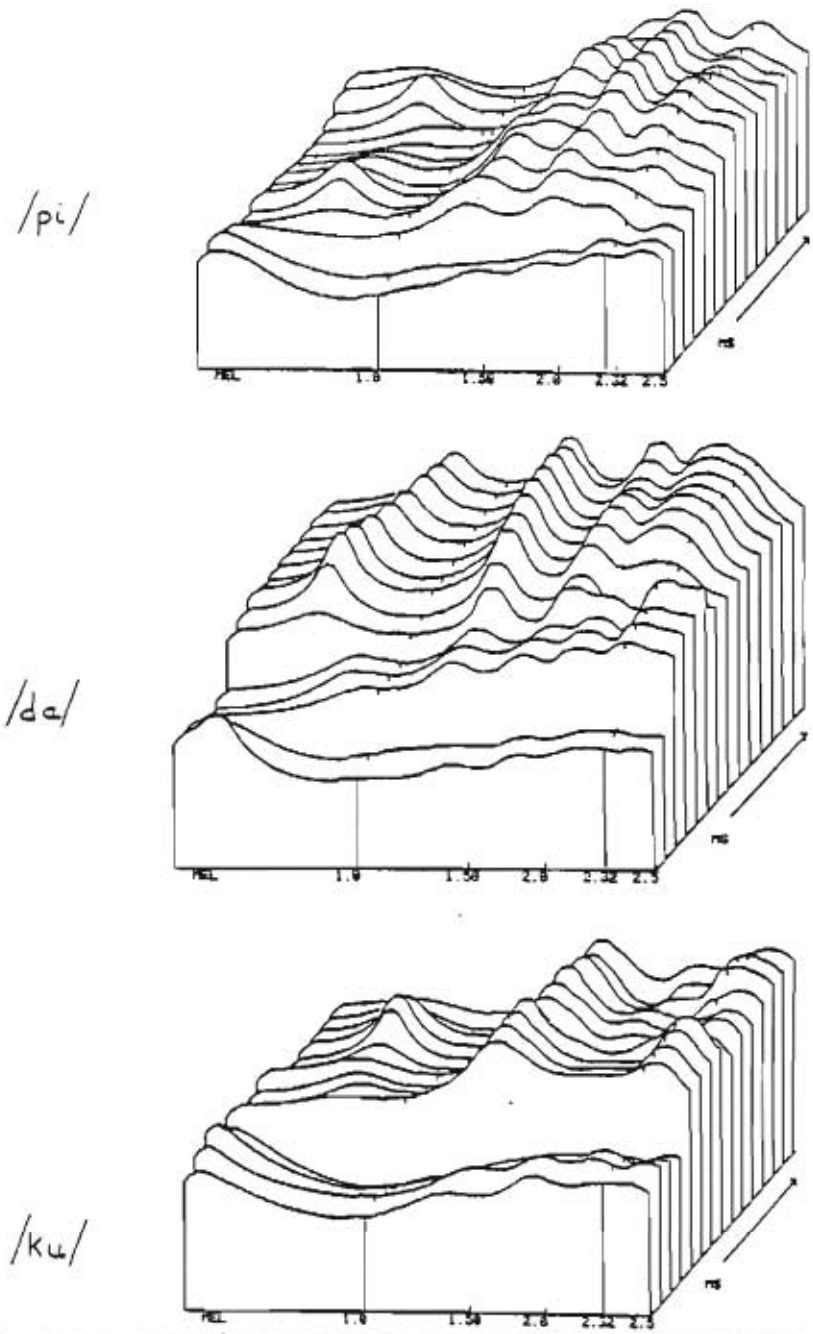


Figure 3. Auditory filtered running spectral displays for /pi/, /da/, and /ku/.

which there was an obvious increase in high-frequency energy and to record the number of that frame on their response sheets. In Fig. 3, the burst occurs in the third frame for /pi/ and /da/ and in the fifth frame for /ku/.

Next, the judges indicated the number of the first frame in which a low frequency peak became prominent and appeared continuous for the remaining frames. This frame defined the onset of voicing. In Fig. 3, the onset of voicing for /da/ begins in the seventh frame. Although /pi/ and /ku/ do show low frequency peaks, the peaks are not continuous with the remaining frames. Thus, no onset of voicing is present in these displays. In these cases, the judges recorded an N (for "not voiced"). When low frequency peaks began in the closure frames and continued for the remaining frames, the judges recorded a C (for "continuous voicing").

In this manner, the burst frame and up to six succeeding voiceless frames were identified. Judges then decided on the spectral tilt of these frames and recorded an F if the spectral tilt of the burst and voiceless frames was flat or slightly rising toward higher frequencies and an R if the frames were distinctly rising. Spectral tilt was subsequently used to assign place to the bilabials and dentals. Flat or slightly rising burst and voiceless frames are correlated with bilabials and distinctly rising frames are correlated with dentals (Fant, 1960; Stevens and Blumstein, 1978; see /pi/ and /da/ in Fig. 3).

Finally, the judges were required to identify the presence or absence of a prominent mid-frequency peak starting on the burst frame and extending for three or more frames. If mid-frequency peaks were present, they were located between the ticks on the display (see /ku/ in Fig. 3). The ticks were positioned on the display according to a vocal tract normalization rule proposed in the previous study (Kewley-Port, 1982). This rule is based on Fant's (1960) proposal that the spectral peak for velars is associated with F2 for low, back vowels and F3 or F4 for high, front vowels. Thus, the low frequency tick in these displays was positioned at the talker's F2 for /u/ and the high frequency tick was positioned at the F2 value of /i/ plus 500 Hz.

After judging the four features, the assignment matrix shown in Table 1 was used to determine place of articulation.

Insert Table 1 about here

If tilt was judged as flat or falling and no mid-frequency peaks were observed, the display was labelled as a bilabial. If the tilt was distinctly rising and no mid-frequency peaks were observed, the display was labelled as a dental. If mid-frequency peaks were present, the display was labelled as a velar regardless of spectral tilt.

Judging of the visual displays for both parts of the experiment was carried out in two sessions. In the first session, the two judges were instructed on the

Table 1. Assignment matrix used in judging place of articulation.

ASSIGNMENT MATRIX

TILT	MID-FREQUENCY PEAKS	PLACE
F	N	BILABIAL
R	N	DENTAL
*	Y	VELAR

(* INDICATES EITHER F OR R)

application of the definitions of the four time-varying features. Written versions of these definitions were available to the judges during both sessions. The judges then independently determined the location of the burst frame, the onset of voicing, and place of articulation. If there was disagreement between the judges on the assignment of place of articulation, the displays on which the judges disagreed were presented again in a second session in which the judges collaborated. If the judges were unable to agree on the assignment of any of the features in the collaborative judging session, they were instructed to mark on their response sheet that the feature was ambiguous.

Results

The overall results are shown in Tables 2 and 3.

Insert Tables 2 and 3 about here

Individual judging of the 288 displays was surprisingly poor, with place being correctly identified in only 49% of the displays. To evaluate the source of these errors, the feature assignments for each judge were compared, as shown in Table 3. It appeared that judges disagreed most often on assigning spectral tilt for the voiceless consonants. Often, the spectral tilt of voiceless stops changed from the burst frame over the first six frames (40 msec). Prior to the collaborative judging session, the judges were were reminded to examine the tilt over the first six voiceless frames and to weigh the later frames more heavily when in doubt. After further examination of the errors made in the individual judging session, we agreed with the judges that 6% of the displays clearly contained the inappropriate features for identifying place. Therefore, these displays were excluded from further analysis of the collaborative judging. Results from the 122 displays examined in the collaborative judging session are shown in Table 2. Place was incorrectly identified 6% of the time and 2% of the features were judged as ambiguous. Thus, after collaborative judging, overall correct place identification rose to 86%.

Table 4 shows the results for the first part of the experiment, which was aimed at evaluating the effectiveness of the new features for identifying place of articulation for voiceless stops.

Insert Table 4 about here

Two tokens from four talkers of all possible combinations of /p,t,k/ and /i,ae,ɔ, u/ resulted in 96 voiceless stop-vowel syllables. A three-way analysis of variance (talker X consonant X vowel) showed that place of articulation was identified invariantly across all talkers, $F(3,8)=1.0$, $p>.4$, although the percent

Table 2. Overall percent error in identification of place for individual and collaborative judging.

PLACE IDENTIFICATION RESULTS

	Percent Error
Individual Judging of 288 Displays:	
Place Not Identified Correctly by Both Judges	51
Collaborative Judging of 122 Displays:	
(Excluded 18 Displays Judged Incorrectly by Both Judges and Experimenters)	6
Features Judged as Ambiguous	2
Place Incorrectly Identified	6

Combined Results: 86% Correct Place Identification

Table 3. Agreement between the two judges in assigning the feature to the running spectra for individual judging expressed in percent.

FEATURE AGREEMENT RESULTS		
Burst Frame	Agreed	79%
	One Frame Difference	12%
	Other Differences	9%
Voicing Frame	Agreed	68%
	One Frame Difference	14%
	Other Differences	18%
Tilt of Burst	Agreed	65%
Mid-Frequency Peaks	Agreed	88%

Table 4. Percent correct place identification for voiceless stops across talker, consonant, and vowel.

VOICELESS STOP RESULTS

(N = 96)

VARIABLE		PERCENT CORRECT PLACE IDENTIFICATION	SIGNIFICANCE
TALKER:	MALE 1	92	N.S.
	MALE 2	79	
	FEMALE 1	92	
	FEMALE 2	92	
CONSONANT:	p	91	N.S.
	t	87	
	k	87	
VOWEL:	i	96	$p < .02$
	æ	96	
	ɔ	71	
	ɯ	92	
TOTAL:		89	

correct identification for the second male talker was lower than for the other three talkers. Place identification across consonants was likewise equivalent, $F(2,8)=.11$, $p>.8$. Vowel context, however, resulted in a significant difference, $F(3,8)=5.67$, $p<.02$., because of the poor identification performance on the vowel /ɔ/. In short, 89% of the voiceless stops were identified correctly.

The second part of the experiment was designed to approximate some of the conditions found in fluent speech, namely different talkers' speaking rates. The experimental variables and results are shown in Table 5.

Insert Table 5 about here

Running spectral displays were constructed from two tokens from one male and one female talker of all combinations of /p,t,k,b,d,g/ and /i,a,u/ spoken at three different rates (fast, normal, and slow) which resulted in 216 syllables. A five-way analysis of variance (talker X consonant X vowel X rate X voice) showed that the features were adequate in identifying place invariantly across talkers, $F(1,8)=.68$, $p>.4$, consonants, $F(2,8)=.51$, $p>.6$, vowel context, $F(2,8)=.89$, $p>.4$, and rate, $F(2,8)=.892$, $p>.4$. Although the normal rate produced better identification, it did not result in a statistically significant difference between speaking rates.

Voiced stops were, however, identified more poorly than voiceless stops, $F(1,8)=8.3$, $p<.01$, across all three rates. On closer examination of the data, we found this result to be caused by significantly poorer identification of the voiced stops at the fast and slow rates. Eighty-nine percent of the voiced stops were identified correctly at the normal rate, but 75% of the voiced stops were identified correctly at only the fast and slow rates. Rate had no effect, however, on the identification of place for the voiceless stops. Overall, 86% of the displays were identified correctly.

Discussion

The present study addressed several problems raised in Kewley-Port's (1982) study of place of articulation. Changing the running spectral displays from linear prediction spectra to auditory filtered spectra apparently made judgement of the tilt of the burst more difficult. Some of these difficulties were related to the indirect method used to produce the auditory filtered spectra. Further research should employ a more direct method of deriving the auditory filtered display, such as those proposed by Klatt (1976; 1979) and Flanagan and Christensen (1980).

Two phonetic variables examined in the previous study were also included here. As before, place was identified invariantly across vowel context, although the poorer performance for /ɔ/ merits further investigation. A new rule for locating a talker's mid-frequency range on the displays was implemented. Results showed that this rule was an improvement over the rule used in the previous study

Table 5. Percent correct place identification for the three speaking rates across talker, rate, consonant, voicing, and vowel.

RATE RESULTS

(N = 216)

Variable		Percent Correct Place Identification	Significance
Talker:	Male 1	84	N.S.
	Female 1	88	
Rate:	Normal	90	N.S.
	Fast	85	
	Slow	83	
Place:	Bilabial	89	N.S.
	Dental	86	
	Velar	83	
Voicing:	Voiced	80	$p < .02$
	Voiceless	93	
Vowel:	i	90	N.S.
	ɑ	85	
	ʉ	83	
Total:		86	

which resulted in identification of place invariantly across male and female talkers.

The present data base included both voiced and voiceless stop consonants produced at three different speaking rates. While identification of place was quite good for the voiceless stops using the time-varying features, the voiced stops were identified more poorly than in the earlier study. This result is probably related to the change from linear prediction to auditory filtered displays. One other possibility is that the poor performance on the voiced stops produced at the slow and fast rates may reflect poor identification of these stops auditorily. This is being investigated in a perception experiment that is currently underway.

The time-varying features employed in this study were used to locate the stop burst and the onset of voicing. Judges agreed on the choice of the burst frame within one frame 91% of the time (see Table 3). This feature of an abrupt change in high frequency energy was easy to observe in the visual displays and will hopefully serve to identify the classes of stop consonants from other phonetic class in fluent speech. The feature which located the onset of voicing served only to identify the voiceless frames for judging the tilt of burst. That is, the tilt of burst feature was not defined over a fixed temporal interval as in the Kewley-Port (1982) study, but rather varied according the number of voiceless frames present. Delgutte (1980) provides a good rationale for this approach. He suggests that the auditory system may process voiceless, low energy sounds differently than high energy voiced sounds. Furthermore, he suggests that the abrupt onset of either high frequency energy or low frequency voicing is a strongly marked event in the auditory system. Delgutte's hypotheses can be interpreted for the features employed in this study as follows: An abrupt change in high frequency energy signals the onset of a stop burst. The tilt of the spectral energy in succeeding voiceless frames is integrated over a time period of about 30 to 40 msec unless there is an abrupt onset of voicing. The onset of voicing terminates judgment of voiceless spectral tilt. If the burst and voicing onset occur simultaneously (as it normally would for /b/ and sometimes /d/), then the tilt of burst can be identified from the first five to ten msec of energy. It is clear from this description that the identification of place of articulation and voicing are interdependent in this analysis. While the judgment of the phonetic feature of voicing was not specifically made in this study, this is obviously an important next step for our research.

In summary, we believe we have successfully extended Kewley-Port's (1982) original findings in a number of important ways. Our results also further validate the use of time-varying features derived from running spectral displays to identify invariant cues to place of articulation. These features appear to be adequate for identification of the feature of voicing as well.

References

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

Fant, G. Acoustic theory of speech production. The Hague: Mouton, 1960.

Fant, G. Stops in CV-syllables. In G. Fant (Ed.), Speech Sounds and Features. Cambridge, Mass.: MIT, 1973, 110-139.

Flanagan, J. L., & Christensen, S. W. Computer studies on parametric coding of speech spectral. Journal of the Acoustical Society of America, 1980, 68, 420-430.

Kewley-Port, D. Spectrum: A program for analyzing the spectral properties of speech. Research on Speech Perception: Progress Report No. 5, Department of Psychology, Indiana University, 1979, 475-492.

Kewley-Port, D. Time-varying features as correlates of place of articulation in stop consonants. Journal of the Acoustical Society of America, 1980, in press.

Klatt, D. H. A digital filter bank for spectral matching. In C. Teacher (Ed.), Conference Record of the 1976 IEEE International Conference on Acoustics, Speech, and Signal Processing. Philadelphia, IEEE Catalog No. 76CH1067-8 ASSP, 1976, 537-540.

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

Luce, P. A., & Carrell, T. D. Waves: A program for creating and editing waveforms. Research on Speech Perception: Progress Report No. 7, Indiana University, 1982.

Patterson, R. D. Auditory filter shapes derived with noise stimuli. Journal of the Acoustical Society of America 1976, 59, 640-654.

Patterson, R. D., & Nimmo-Smith, I. Off frequency listening and auditory-filter symmetry. Journal of the Acoustical Society of America, 1980, 67, 229-245.

Stevens, K. N., & Blumstein, S. E. Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America, 1978, 64, 1358-1368.

Zwicker, E. Subdivision of audible frequency range into critical bandwidths (Frequenzgruppen). Journal of the Acoustical Society of America, 1961, 33, 248.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

In Defense of Segmental Representations in Speech Processing*

David B. Pisoni

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

*This is the text of an invited paper to be presented at the meeting of the Acoustical Society of America, Ottawa, Canada, May 19, 1981. Preparation of this paper was supported, in part, by NIH research grant NS-12179 and NIMH research grant MH-24027 to Indiana University. I am grateful to Dan Dinnsen, Becky Treiman and Louis Goldstein for comments and helpful suggestions regarding the arguments outlined in this paper. I am also indebted to T. D. Carrell, A. C. Walley, A. Salasso, Paul Luce, Jan Luce and C. Murphy for interesting discussions of these problems.

In Defense of Segmental Representations in Speech Processing

Introduction

When Dennis Klatt first proposed this special session on models of speech perception he asked me if I would be willing to present a paper defending the phoneme in speech processing. At first I was a little reluctant to do this because a number of people have been arguing strongly over the last few years against the use of segmental representations such as phonemes, including Dennis himself (see Klatt, 1980). However, I agreed to undertake this job mainly as an intellectual and, to some extent, as a theoretical exercise to see if I could marshall up enough evidence to convince myself that segmental representations such as phonemes were worth arguing about in public. Having now spent several weeks thinking and talking about the problems, I am convinced that it is worth raising some of the old issues again and worth bringing them out in the open for further consideration at a session such as this one. Before going into the background and arguments, I want to make it clear at the outset that I do not plan to present a new model or account of speech perception that claims to solve all of the old problems that people have been working on for thirty years (see Pisoni, 1978 for a review). Instead, my goal in this presentation is much more modest. In the time allocated I want to reexamine some of the historical reasons for supposing that segmental representations play an important role in speech processing. I plan to do this by first examining several of the obvious linguistic facts that are typically cited about the structure of human language, particularly as they relate to phonetics and phonology. Next, I will briefly summarize some of the psychological or processing data that is often cited as support for the existence of segmental representations. Finally, I will briefly consider several contemporary models of word recognition and lexical access in order to make several of their underlying assumptions more explicit in light of my previous observations about the need for segmental representations in speech processing.

For a number of years there has been a debate going on in journals, at meetings, and in backrooms concerning the use of segmental (i.e., phonemic) representations in on-going speech perception. Several theorists have totally abandoned reliance on computing an intermediate level of representation in favor of direct access models of speech perception (Klatt, 1979; Marslen-Wilson & Welsh, 1978). In models such as these, words are recognized without an analysis of their "internal structure" into computational elements like phonemes or morphemes. Proponents of this view have argued that their recognition models do not require the postulation of theoretical entities like phonemes or segments. I hope to show that while these theorists have attempted to ignore or even deny the existence of these processing units, such units nevertheless play an important -- if not critical role in these accounts of the word recognition process. Indeed, I hope to show that while these units are not explicitly acknowledged they are tacitly assumed by all theorists.

A. Linguistic Evidence for Segmental Representations

One of the fundamental assumptions of linguistic analysis is that the continuously varying speech signal can be broken down into a sequence of discrete units such as phones, phonetic segments or speech sounds. This assumption is central to our current conceptions of language as a system of rules which govern the sound patterns and sequences used to encode meanings and has been assumed since the days of the Sanskirt grammarians and as far back as Pannini. There are numerous reasons for postulating segmental representations in language. Among them are a few observations (see Kenstowicz & Kisseberth, 1977). These are shown in Table I.

Alternations. In most languages of the world, there are morphemes or words that have more than one phonetic realization. Various types of phonological rules have been formulated to select the appropriate phonetic alternative in different contexts. These rules operate on a specific aspect of the internal structure of words -- that is, they operate on segments, or the features of segments. We know that this aspect of linguistic behavior is rule governed because speakers can produce the correct inflection to nouns they never heard before. And, the pronunciation assigned to these novel forms follows a general principle. Similar observations can be made of languages that have vowel harmony. Presented with novel utterances, adults produce the correct phonetic forms by rule and these rules operate on segment-size of units.

Systematic Regularities. We know from phonological analyses of many languages that there are systematic regularities in the sound structure of the morphemes of the language. While some phonetic properties of a particular morpheme may be idiosyncratic or arbitrary, other properties reflect general manifestations of regularities in the sound patterns of the specific language. These regularities are controlled by rules which operate on segments or, more precisely, features of segments. The important point to emphasize here is that a description of the linguistic generalization requires analysis of words into their internal structure, a structure consisting of a linear sequence of sound types or phonemes.

In short, the mere existence of phonological processes and operations in spoken languages supports the assumption that speech is organized into a sequence of discrete units. Deletion and insertion rules apply to single segments in words rather than to whole words. The metathesis of adjacent sounds in an utterance indicates that the sounds are discrete entities that can be dissociated from the context they are encoded in. Moreover, rules which assign lexical stress to words require an analysis into discrete consonant and vowel segmental units.

Historical Sound Changes. Finally, additional support for the existence of segments comes from the detailed synchronic and diachronic analyses that have been done on sound changes in language. Differences in dialect as well as changes that have occurred in languages over time appear to be easily captured by rules that operate on the internal structure of words, specifically segments and features.

B. Psychological Evidence for Segmental Representations

The psychological evidence in support of segmental representations is very diverse and I will only summarize a few examples here. Each of these could be elaborated on at greater length. Table II provides several examples.

Orthography and Writing Systems. Another source of evidence comes from observations of speakers of unwritten languages who are attempting to develop writing systems. In his well-known article "The Psychological Reality of Phonemes," Sapir (1933, reprinted, 1963) cites several examples of where the orthographic choices of an illiterate speaker revealed a conscious awareness of phonological structure of his language. More recently, Read (1971, 1975) has described a number of examples of children who have invented their own orthographies spontaneously. The children's initial encounters with print show a systematic awareness of the segmental structure of language. Such observations would not be possible unless adults and eventually children had the abilities to analyze spoken language into representations consisting of discrete segments such as phonemes. Indeed, it has been suggested recently that young children's ability to learn to read an alphabetic writing system is highly dependent on the development of phonemic analysis skills -- that is, skills that permit the child to consciously analyze speech into segmental units such as phonemes (Gleitman & Rozin, 1977; Treiman & Baron, in press).

Language Games. The existence of language games based on insertion of a sound sequence at specifiable points in a word, the movement of a sound or sound sequence from one point to another or the deletion of a sound or sound sequence provides additional support for the existence of segmental representations of the internal structure of words. The existence of rhymes and the metrical structure of poetry all involve the awareness, in one way or another, that words have an internal structure and organization to them and that this structure can be represented as a sequence of discrete linear units distributed in time.

Speech Production Errors. An examination of errors in speech production has provided strong evidence that words and morphemes are represented in the lexicon in terms of some sort of segmental representation such as phonetic segments or phonemes. The high frequency of single segment speech errors such as substitutions and exchanges reveal evidence of the phonological structure of the language. It has been difficult, if not impossible, to explain these kinds of errors without assuming some kind of segmental representation in the organization of speech production (see Fromkin, 1980; Shattuck-Hufnagel & Klatt, 1979).

Speech Perception Studies. Over the years there have been many perceptual findings that can be interpreted as support for an analysis of speech into segmental representations (see Table III for examples of experimental procedures). Perhaps the most compelling data have come from numerous experiments involving an analysis of errors and confusions in short-term memory and of the errors produced in listening to words and nonsense syllables presented in noise (see Wickelgren, 1966; Miller & Nicely, 1955). While some of these findings were originally interpreted as support for various types of feature systems, they also provide strong evidence, in my view, for the idea that the listener carries out an analysis of the internal structure of the stimulus input into dimensions for encoding and storage in memory. However, these findings are not considered to be compelling to some investigators since they have also been

subject to alternative interpretations because of the specific task requirements. Apparently, the size of the perceptual unit changes as the level of analysis shifts according to the experimental task and instructions to subjects (see Savin & Bever, 1970; Foss & Swinney, 1973). If perceptual and short-term memory data were the only findings that could be cited in support of segmental representations, one might feel a little uneasy. However, there are other converging sources of evidence from perceptual studies which provide additional support for this view.

For example, there have been numerous reports describing the phoneme restoration effect, a phenomena that demonstrates the on-line synthesis of the segmental properties of fluent speech by the listener as it is heard (Warren, 1976). Numerous studies have been carried out by psycholinguists using the phoneme monitoring technique in which subjects are required to detect the presence of a specified target phoneme while listening to sentences or short utterances (Foss & Blank, 1980). Although some earlier findings suggested that listeners first recognized the word and then carried out an analysis of the segments within the word, other more recent findings indicate that subjects can detect phonemes in nonwords that are not present in the lexicon (Blank, 1979). Thus, subjects can detect phonemes based on two sources of knowledge, information from the sensory input and information developed from their knowledge of the phonological structure of the language.

A large body of data has also been collected on the detection of mispronunciations in fluent speech. While these findings have been interpreted recently by Cole as support for the primacy of word recognition in speech perception, the findings can, in my view, just as easily be used to support the idea that listeners can gain access to the internal structure of words in terms of their segmental representations, and they can do this while listening to continuous speech (Cole & Jakimik, 1980).

Finally, in terms of perceptual data, there is a small body of data on misperceptions of fluent speech. The errors collected in these studies suggest that a very large portion of the misperceptions involve segments rather than whole words (Bond & Garnes, 1980).

C. Models of Word Recognition and Lexical Access

At the present time there are about four or five different models of spoken word recognition and lexical access described in the published literature. These are outlined in Table IV. Unfortunately, it is very difficult to make comparisons among these models because they use quite different experimental tasks (i.e., recognition, shadowing, lexical decision and detection of mispronunciations) and attempt to account for a number of different phenomena (word frequency effect, context effects, word length effect, speed and earliness of word recognition). In some cases, the models make strong psychological claims about human speech perception as in the LAFS model of Klatt although no empirical data have ever been collected to test the basic assumptions (Klatt, 1979, 1980). In other cases, the models are based primarily on behavioral data collected with humans such as Morton's Logogen Model or Marslen-Wilson's Cohort Model (Morton, 1979; Marslen-Wilson & Tyler, 1980). These models make few claims about implementing their approach as a working computational device on a machine. What

is of interest in these current models is the degree to which they assume the existence of segmental representations in order to solve the primary recognition problem (see Table V for a list of major issues). Before going into the details of several of these models, it will be useful to provide some working definitions of precisely what is meant by "word recognition" and "lexical access" since these have often been confused in the literature. By word recognition, I mean those computational processes by which a listener identifies the phonetic and/or phonological form of a spoken utterance. That is, word recognition may be thought of as a form of pattern recognition. Thus, the psychological processes used in word recognition will be the same whether the input consists of words or nonwords. A good deal of the work carried out over the last thirty years has been concerned with what I call the "primary recognition problem." That is, how the form of a spoken utterance is recognized from an analysis of the acoustic waveform. Conscious identification of all of the phonetic segments is not necessary although it is certainly possible under special circumstances when a listener's attention is directed specifically to the sound structure of an utterance. Under normal listening conditions the listener does not have to identify the phonetic input to recognize the words.

By lexical access, I mean those computational processes that are involved in the activation of the meaning or meanings of words that currently exist in the listener's mental lexicon. I assume that the morphemes of a language are stored by the listener in his lexicon along with appropriate information about their syntactic and semantic attributes and their possible phonological forms. Thus, a word is accessed from the lexicon when its phonetic form or forms has been matched with some appropriate representation previously stored in the lexicon or with some representation generated by phonological and/or morphological rules. In the latter case, these rules provide the listener with the ability to recognize and understand novel words that he has never heard before.

Keeping in mind this distinction between word recognition and lexical access, let me now turn to a brief examination of four models of word recognition: Morton's LOGOGEN model, Klatt's LAFS model, Marslen-Wilson's COHORT model, and Forster's Autonomous Search Model. The first three models may be thought of as only models of word recognition, in the sense defined above, since they have very little to say about the structure of the lexicon or the types of representations that are used in accessing the meanings of words. In contrast, Forster's model is primarily concerned with lexical access. Nevertheless, we will consider each of them briefly in order to characterize what they have to say about segmental representations in speech perception and the process of word recognition.

Morton's Logogen Model. The Logogen model was formulated to account for several effects found in the visual word recognition literature, specifically, the word frequency effect and the effects of context on perceptual recognition (Morton, 1979). To deal with both of these effects, Morton postulated the existence of a hypothetical entity known as a Logogen -- a computational device that acts as a counter and accepts input from both the sensory input and context. Each word has a logogen associated with it that has a particular threshold for activation. To deal with the word frequency effect, Morton proposed that high frequency words have lower thresholds than low frequency words. And, to deal with the effects of context on perceptual recognition, he proposed that sentential context, for example, could selectively modify the threshold of a

logogen by lowering its sensitivity and therefore making the word more available for response. Morton's model has no lexicon, in the sense used above, nor is it very specific about the representation of information in a logogen. For the most part, Morton assumes that words are holistic entities and are recognized passively through template matching techniques. The logogen model says little about analyzing the internal structure of words in terms of phonemes or morphemes or the need for segmental representations.

Klatt's LAFS Model. Based on the apparent success of the HARPY speech understanding system (Lowerre & Reddy, 1979), Klatt has argued that it is possible, in principle, to carry out lexical access without any intermediate analysis of words into segments or phonemes (Klatt, 1979). By precompiling a network of context sensitive spectra for all word sequences in the language, Klatt believes that he will be able to recognize words directly without segmentation into phonemes or segments. The key to his proposal lies in two assumptions. First, that it is possible to precompile a network of spectral templates that will be able to deal with the context conditioned variation of segments in words produced by different speakers in different phonetic environments. Second, that low-level decisions are delayed pending additional information from the lexical level. The reason for this latter assumption, the assumption of "delayed binding," is Klatt's assertion that the recognition system needs to be able to recover from "errorful interpretation" which might occur if lower level acoustic-phonetic information in the signal were discarded after some initial form of phonetic categorization. Although not explicit, Klatt's model assumes that words do, in fact, have an internal structure that consists of a linear sequence of segments such as phonemes. To avoid dealing with abstract entities such as phonemes or segments, Klatt has simply substituted diphone-like computational units (i.e., context-sensitive spectral templates) and then precompiled this information into a passive network structure. Nevertheless, it should be pointed out here that his proposed system is crucially dependent on the recovery of the internal structure of words for primary recognition to take place. And, his proposed system is crucially dependent on the assumption that words consist of a linear sequence of segments with phoneme-like properties. On the surface it would appear that this system avoids the need to compute a distinct level of representation corresponding to phonemes or segments. However, on careful examination it can be shown that the information about the internal organization of words is actually encoded into the structural organization of the spectral network itself. To deal with new words and place their spectral properties in the network, Klatt had to develop another system, SCRIBER, which uses an analysis-by-synthesis scheme to enter the spectral properties of new words in the lexicon (Klatt, 1980).

Marslen-Wilson's COHORT Theory. Perhaps the most detailed account of word recognition to date is Marslen-Wilson's Cohort Theory (Marslen-Wilson & Tyler, 1980). The key to this theory of word recognition is the notion of a set of "word initial cohorts" or recognition candidates. These are defined by the acoustic-phonetic commonality of the initial sound sequences of words. A particular word is "recognized" at that point -- the "critical recognition point" where the word is uniquely distinguished from any other word in the language beginning with the same initial sound sequence. The theory accounts for the facilitatory effects of context in word recognition by assuming, as in the logogen model, that context acts to lower the threshold values for recognition of a particular word. However, context can also be used to deactivate noncandidate

words and therefore reduce the size of word initial cohort set that is active at any time. The interaction of context with the sensory input is assumed to occur at the level of word recognition (see Marslen-Wilson & Welsh, 1978). Processing at early sensory levels is assumed to occur automatically and is not influenced by other higher-order sources of knowledge. While specifically avoiding the problem of characterizing the representation of the acoustic-phonetic input by using euphemisms such as "sound sequences" or "sounds" or "sensory input," the cohort theory as well as LAFS is crucially dependent on the fundamental assumption that words have an internal structure to them and that structure must be recovered for primary recognition. Words are simply not undifferentiated wholistic processing entities. In cohort theory, the set of word initial cohorts that are activated is defined, in principle, by the internal segmental structure of the linear arrangement of speech sounds. To say, as Marslen-Wilson has done repeatedly, that his theory makes no claim about the structure of the input to the word recognition system in terms of phonemic representations is simply to deny the existence of a prima facie assumption that is central to the organization of his word initial cohort set. The theory, as currently formulated, would never work if the internal structure of words could not be described as a sequence of segment-like units.

Forster's Autonomous Search Model. Forster's Autonomous Search Model is primarily concerned with problems of lexical access although it is relevant to the present discussion because it specifically assumes that access to the lexicon can be carried out through various access files (Forster, 1979). Among these routes to the lexicon is a phonologically organized access file. This file contains a description of the phonetic composition of words in the language along with a pointer giving direct access to the corresponding entry in the master file. Within an access file, words are grouped in bins. And, within each bin words are arranged according to frequency. Locating a particular word requires a search through the bin. One of the nice features of Forster's model which is based almost entirely on data from visual perception of words is that access to the master file is possible through a number of different access files; semantic, syntactic and phonological information as well as orthographic information can be used to locate an entry in the lexicon.

Of the models considered thus far, Forster's is the only one that explicitly proposes lexical access via a segmental phonological code or representation (see Table IV). And, while the structure of Forster's lexicon is morphologically-based, search of the access file can only be initiated serially with bottom-up sensory information provided by the primary word recognition processes. Forster maintains autonomous levels in his model; as a consequence, higher-order semantic and syntactic information cannot be used to influence lower-level decisions involved in word recognition or lexical access. Thus, Forster argues that sentential context normally plays no role in the lexical access process. Sentential context and other sources of knowledge are only used post-lexically after an entry has been located in the master file. The primary search process through the phonological access file therefore functions autonomously and the search is dependent entirelyon the bottom-up acoustic-phonetic properties of the input signal.

While Forster's model has several desirable features associated with it such as the phonological access file and a morpheme-based lexicon, its major deficiency, in my view, lies in the assumption of autonomous levels of processing

and its inability to support interactive processing between and among various knowledge sources (Marslen-Wilson, 1975). There is strong evidence that words can be recognized very rapidly without a complete and detailed analysis of their sensory inputs. In several elegant experiments using a "gating procedure," Grosjean (1980) has shown very strong relations in word recognition between sentential context and the amount of stimulus information needed for identification of words. By eliminating the autonomy assumption and incorporating interaction between knowledge sources, Forster's model could easily account for the effects of higher-order sentence level constraints on word recognition while at the same time making use of segmental representations in primary recognition and lexical access.

D. Summary and Conclusions

The main point of this paper has been to argue for the importance of segmental representations in speech processing. Are words in spoken language simply undifferentiated holistic entities without dimensional organization or do they have a complex and rule-governed internal structure that needs to be recovered in accessing their meaning? After considering several sources of evidence, I am led to the position that some form of segmental representation--whether it be phones, segments, phonemes or speech sounds, is absolutely necessary to gain entry into the lexicon and recover the meaning of words. To deny that segmental representations exist or that they need not be "computed" during word recognition is simply, in my view, to ignore a fundamental property of language that words have an internal structure and a principled dimensional organization to them. This structure is rule-governed and is a distinctive property of spoken language. I would argue that this structure must be recovered either directly or indirectly for words to be recognized and their meanings accessed from the mental lexicon. The evidence that other theorists cite against mediation is certainly relevant to issues of word recognition and spoken language understanding but it is by no means conclusive. As I have tried to show, a number of the recent accounts that have explicitly denied the existence of segments have actually tacitly assumed them anyway in their analyses. For these analyses to work out correctly, segmental analysis must be assumed in precompiling the LAFS network or in computing the acoustic-phonetic input "on the fly" as in COHORT theory.

Some of you may wonder why I have had little to say in this paper about the very earliest stages of speech processing and why I have not mentioned Motor Theory, Analysis-by-Synthesis or even Feature Detectors, all of which suppose the computation of an autonomous level of phonetic representation. Many theorists no longer believe seriously that convincing evidence can be found for an autonomous level of processing that is independent of syntax, morphology and semantics; that is, an independent level corresponding to a linear sequence of phones. Moreover, and perhaps more importantly, the questions of interest in speech perception today are no longer ones dealing with the acoustic cues to nonsense syllables or the perception of phonetic features, phonemes or syllables in isolated contexts. Some of these are outlined in Table V. Instead, the major focus of recent work and interest has now shifted to questions concerning word recognition and to a more general concern for spoken language understanding, particularly the processing of meaningful connected fluent speech by human listeners. Investigators are much more interested in studying the perception of fluent

speech under more "naturalistic" conditions where various sources of knowledge can be used to support word recognition and language understanding. Hopefully this special session on word recognition will serve as a catalyst to encourage more research and discussion of these problems. There are many important basic issues that need to be resolved; the most critical of which must surely be the issue of representations in speech processing.

References

Blank, M. A. Dual-Mode Processing of Phonemes in Fluent Speech. Ph.D. thesis, University of Texas at Austin, 1979.

Bond, Z. S. & Garnes, S. Misperceptions of fluent speech. In R. A. Cole (Ed.), Perception and Production of Fluent Speech. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980. Pp. 115-132.

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

Cole, R. A. & Jakimik, J. A model of speech perception. In R. Cole (Ed.), Perception and Production of Fluent Speech. Hillsdale, NJ: Lawrence Erlbaum, 1980. Pp. 133-163.

Forster, K. I. Levels of processing and the structure of the language processor. In W. E. Cooper and E. C. T. Walker (Eds.), Sentence processing: Psycholinguistic Studies presented to Merrill Garrett. Hillsdale, NJ: Lawrence Erlbaum, 1979. Pp. 27-86.

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

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

Fromkin, V. A. (Ed.) Errors in linguistic performance. New York: Academic Press, Inc. 1980.

Gleitman, L. R. & Rozin, P. The structure and acquisition of reading I: Relations between orthographies and the structure of language. In A. Reiber & D. L. Scarbough (Eds.), Toward a psychology of reading. Hillsdale, NJ: Lawrence Erlbaum Associates, 1977. Pp. 1-54.

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

Jarvella, R. J. Immediate memory and discourse processing. In G. Bower (Ed.), The psychology of learning and motivation. Vol 13. NY: Academic Press, 1979. Pp. 379-421.

Kenstowicz, M. & Kisseberth, C. Topics in Phonological Theory. New York: Academic Press, 1977.

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

Klatt, D. H. Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. A. Cole (Ed.), Perception and Production of Fluent Speech. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

Lowerre, B. T. & Reddy, D. R. The HARPY Speech understanding system. In W. A. Lea (Ed.), Trends in Speech Recognition. Englewood Cliffs, NJ: Prentice-Hall, 1979.

Marslen-Wilson, W. D. Sentence perception as an interactive parallel process. Science, 1975, 189, 226-228.

Marslen-Wilson, W. D. & Tyler, L. K. The temporal structure of 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, 41, 329-335.

Miller, G. A. & Nicely, P. E. An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America, 1955, 27, 338-352.

Morton, J. Word recognition. In J. Morton and J. C. Marshall (Eds.), Psycholinguistics 2: Structures and processes, Cambridge: MIT Press, 1979. Pp. 107-156.

Nooteboom, S. G. Lexical retrieval from fragments of spoken words: Beginnings versus endings. Instituut voor Perceptie Onderzoek Den Dolech Z - Eindhoven, 1980, Manuscript No. 385/II. (To be submitted to J. of Phonetics).

Pisoni, D. B. Speech perception. In W. K. Estes (Ed.), Handbook of learning and cognitive processes (Vol. 6). Hillsdale, NJ: Lawrence Erlbaum Associates, 1978.

Read, C. Pre-school children's knowledge of English phonology. Harvard Educational Review, 1971, 41, 1-34.

Read, C. Children's categorization of speech sounds in English. Urbana, Illinois: National Council of Teachers of English, 1975.

Sapir, E. The psychological reality of phonemes. In D. G. Mandelbaum (Ed.), Selected writings of Edward Sapir in language, culture and personality. Berkeley: University of California Press, 1963. Pp. 46-60.

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

Shattuck-Hufnagel, S. & Klatt, D. H. The limited use of distinctive features and markedness in speech production: Evidence from speech error data. Journal of Verbal Learning and Verbal Behavior, 1979, 18, 41-55.

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

Warren, R. M. Auditory illusions and perceptual processes. In N. J. Lass (Ed.), Contemporary Issues in Experimental Phonetics. New York: Academic Press, 1976.

Wickelgren, W. A. Distinctive features and errors in short-term memory for English consonants. Journal of the Acoustical Society of America, 1966, 39, 388-398.

Wickelgren, W. A. Phonemic similarity and interference in short-term memory for single letters. Journal of Experimental Psychology, 1966, 71, 396-404.

Wickelgren, W. A. Short-term recognition memory for single letters and phonemic similarity of retroactive interference. Quarterly Journal of Experimental Psychology, 1966, 18, 55-62.

Wickelgren, W. A. Auditory or articulatory coding in verbal short-term memory. Psychological Review, 1969, 76(2), 232-235.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

Comprehension of fluent synthetic speech produced by rule*

Paul A. Luce

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

*This work was supported by grants from NIMH, Research Grant No. MH-24027, and NINCDS, Research Grant No. NS-12179. The author would like to thank David B. Pisoni and Hans Brunner for their assistance and advice on various stages of this project.

Abstract

Recent research presented to the Society by Feustel, Luce, and Pisoni (1981) showed that recall of synthetic word lists under conditions of increased load on short-term memory is consistently poorer than recall of natural word lists. In an attempt to extend these findings to the perception and comprehension of continuous synthetic speech, we carried out an experiment in which subjects were presented short meaningful passages of fluent synthetic or natural speech. After presentation of each passage, the subjects were required to answer questions keyed to various levels of linguistic information represented in the passages. The results showed that the synthetically produced passages forced subjects to attend more closely to the physical properties of the speech signal itself. These findings indicate that synthetic speech places increased demands on the cognitive processes involved in comprehension and understanding of meaningful connected speech.

At the last meeting of the Society, Feustel, Luce, and Pisoni (1981) presented the results from a series of experiments involving recall of lists of isolated synthetic and natural words. In these studies, we attempted to locate the source of previously demonstrated difficulties in the perception of synthetic speech. In particular, we were interested in determining whether the decreased intelligibility of synthetic speech could be attributed to encoding or rehearsal processes in short-term memory or whether these difficulties lie at some lower level of processing, such as the extraction of acoustic-phonetic information from the speech waveform.

To identify the factors that are responsible for the reduced intelligibility of synthetic speech, we required subjects to recall short lists of synthetic and natural words under various conditions of load on short-term memory. We reasoned that if processing of synthetic speech places increased demands on short-term memory, then reducing the available processing capacity should have greater effects on subjects' ability to recall synthetic word lists than natural word lists.

To assess the effect of reducing the capacity of short-term memory on the processing of the speech stimuli, we presented both the natural and synthetic word lists at three presentation rates: two, three, and five seconds per word. We predicted that at faster rates of presentation the reduced capacity to encode or rehearse the word lists would more severely reduce the number of synthetic words recalled relative to the number of natural words recalled. Our predictions were not borne out. Recall of the synthetic word lists was poorer overall, but increasing presentation rate did not differentially affect recall of the synthetic lists compared to the natural ones.

Because this manipulation of presentation rate may have placed too few demands on short-term memory, we performed a second experiment in which subjects were required to remember either zero, three, or five digits that were presented visually on a CRT display prior to the presentation of the word lists. Again, we reasoned that increasing the number of digits to be held in memory would reduce the capacity for encoding or rehearsing the words, thus differentially affecting recall of the synthetic word lists.

As in the first experiment, we found no differential decrease in recall of the synthetic words as the number of digits to be remembered increased.

Insert Figure 1 about here

However, as Figure 1 shows, we did find that the number of subjects who were able to correctly recall all of the digits decreased more rapidly for the synthetic lists than the natural lists as the load on short-term memory increased from three to six items. That is, performance on the digit task was differentially affected by having to recall a list of synthetic words.

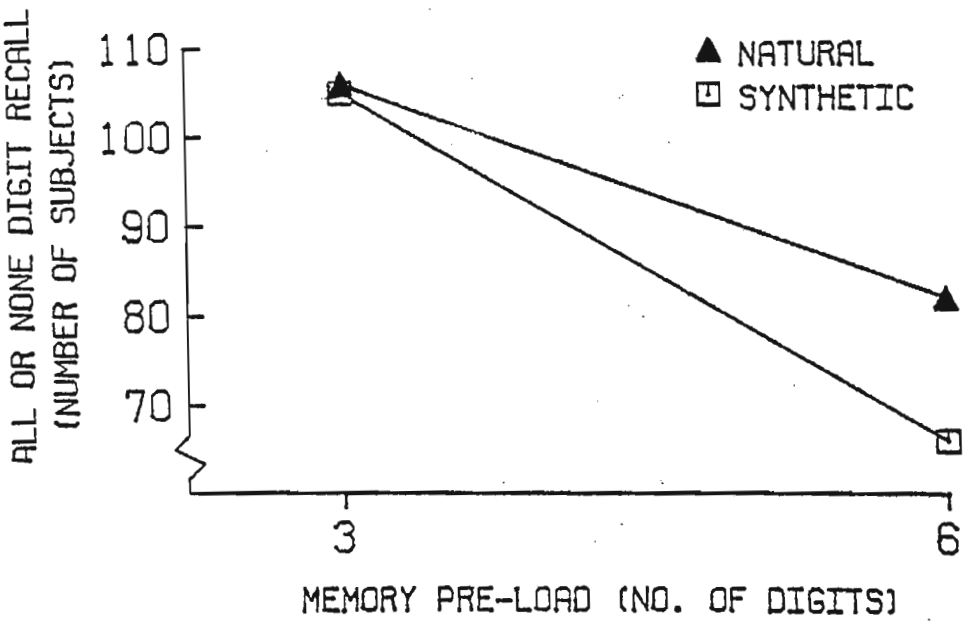

ALL OR NONE DIGIT RECALL
(NUMBER OF SUBJECTS)
110
100
90
80
70
NATURAL
SYNTHETIC
3
6
MEMORY PRE-LOAD (NO. OF DIGITS)

Figure 1.

In a third experiment, we were able to demonstrate more clearly the differential role of capacity demands in the perception of synthetic word lists. We presented subjects with short lists of synthetic and natural words and required them to recall the words in the exact order in which they were presented. (In the previous two experiments subjects were allowed to recall the words in any order.)

Insert Figure 2 about here

As Figure 2 shows, subjects' ability to recall the synthetic words from the beginning of the lists was significantly reduced relative to the recall of the natural words from the beginning of the lists. In other words, the primacy portion of the serial position curve for the synthetic word lists--seen on the left-hand side of this graph--was significantly lower than the primacy portion for the natural lists, whereas the recency portions of these two curves--seen on the right hand side of the graph--did not diplay such a difference. This difference in the primacy portion of the curves indicates that subjects are having relatively more difficulty encoding or rehearsing the earlier presented synthetic words. This finding, in conjunction with the earlier findings, strongly suggests that encoding and/or rehearsal processes in short-term memory are differentially affected in the processing of synthetic speech relative to natural speech.

Having identified one of the factors that affects intelligibility of synthetic speech, we were interested in assessing what effects, if any, these increased processing demands would have on the perception and comprehension of fluent synthetic speech. Recently, at a meeting of the Psychonomics Society, Jenkins and Franklin (1981) reported that recall of the gist of passages of synthetic speech is not demonstrably worse than recall of passages of natural speech. However, they used very simple grade school materials and their recall measure--memory for gist--was extremely insensitive. In light of our results with recall of word lists, however, we expected that a more sensitive measure of comprehension than free recall would reveal differences in the comprehension of synthetic and natural texts.

Method

Materials. To measure subjects' comprehension of fluent text, four types of questions were devised that were keyed to various levels of linguistic information within the text. Examples of these are shown in Figure 3.

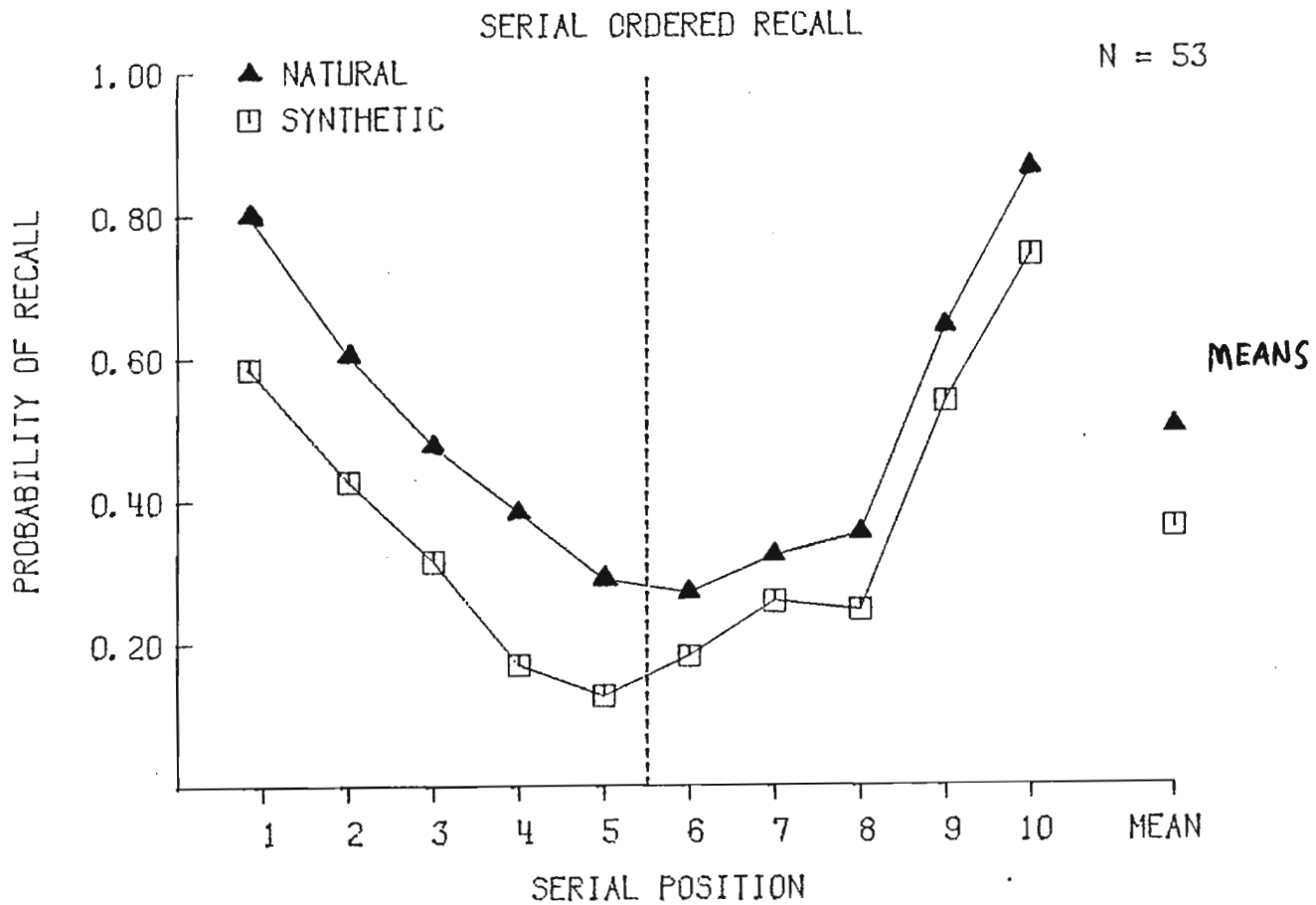

SERIAL ORDERED RECALL
N = 53
NATURAL
SYNTHETIC
1.00
0.80
0.60
0.40
0.20
PROBABILITY OF RECALL
1
2
3
4
5
6
7
8
9
10
MEAN
SERIAL POSITION
MEANS

Figure 2

Insert Figure 3 about here

The questions were based on those used in a recent experiment by Brunner and Pisoni (1982). The first type of question was aimed at determining the subjects' memory for the surface structural aspects of a text. Specifically, these questions involved asking the subjects if a particular word had occurred in the text. When the surface structure question was false, one of two types of foils were presented: Either a "synonomous" word or a "rhyming" word could replace the word that had actually occurred in the text.

The second type of question--the high proposition question--queried subjects' memory for a particular theme or larger message in the text. The third type of question--the low proposition question--examined subjects' memory for some detail presented in the text. The labels "high" and "low" refer to levels within the propositional hierarchy of the text (see Kintsch and van Dijk, 1978). Propositions high in this hierarchy are assumed to convey more general information in the text whereas low propositions are assumed to convey more detailed information. Finally, the inference question required that subjects draw some conclusion that was not explicitly stated in the text. These last three question types were constructed to examine subjects' comprehension of the semantic information or content of the passages. In contrast, the first type of question was designed to assess recall of structural information.

Procedure. Four groups of five subjects each were tested. Two of the groups heard short passages of fluent connected speech produced by the MITalk text-to-speech system and two groups heard passages read by a male speaker. Both the synthetic and natural passages were identical in content. All passages were digitized via a 12-bit analog-to-digital converter and presented to the subjects through matched and calibrated Telephonics (TDH-39) headphones. The passages were presented at 80 dB SPL against a background of wideband white noise at 50 dB SPL.

Twelve passages were presented to all four groups. The first two passages presented in each session were practice passages to familiarize the subjects with the experimental procedure and the speaker's voices. There were ten experimental texts in all.

After presentation of a single passage, subjects were visually presented one of each of the four types of questions on a CRT videodisplay monitor (GBC model MV-10A). The order of presentation of the questions was randomized for each text. Following each question subjects responded YES or NO to the surface structure questions and TRUE or FALSE to the remaining questions by pressing the appropriately labelled button on response boxes in front of them. Subjects were instructed to be as quick but as accurate as possible. Responses and latencies were recorded. The entire experiment was run under the control of a PDP 11/34 computer.

EXAMPLES OF COMPREHENSION QUESTIONS KEYED TO STRUCTURE OF TEXT

SURFACE STRUCTURE

DID THE WORD "FAIRLY" OCCUR IN THIS STORY?	TRUE
DID THE WORD "REASONABLY" OCCUR IN THIS STORY?	FALSE (SYNONYMOUS)
DID THE WORD "BARELY" OCCUR IN THIS STORY?	FALSE (RHYME)

HIGH PROPOSITION

THE LENS BUYER MUST BE CAUTIOUS.	TRUE
THE LENS BUYER MUST BE WEALTHY.	FALSE

LOW PROPOSITION

DEFICIENCIES OF LARGE-APERTURED LENSES ARE EVIDENT IN ENLARGED PRINTS.	TRUE
DEFICIENCIES OF LARGE-APERTURED LENSES ARE NOT EVIDENT IN ENLARGED PRINTS.	FALSE

INFERENCE QUESTIONS

THE BEST LENSES ARE NOT NECESSARILY EXPENSIVE.	TRUE
THE BEST LENSES ARE USUALLY EXPENSIVE.	FALSE

Figure 3.

Results and Discussion

Figure 4 shows the reaction times for each question type for both the synthetic and natural texts.

Insert Figure 4 about here

As expected from the earlier study by Brunner and Pisoni, reaction times were fastest to surface structure questions, followed by high proposition, low proposition, and inference questions, in that order. No differences were found between response times for the synthetic and natural passages nor were there any interactions.

Figure 5 shows the percent correct for each question type for the natural and synthetic passages.

Insert Figure 5 about here

For the high, low, and inference questions we found significantly lower performance for the synthetic versus natural passages, $F(1,18) = 4.43$, $p < .05$. Comprehension of the semantic content of the synthetic passages does appear to be affected relative to comprehension of the natural passages. Of special interest here, however, is the reversal of these results for the surface structure questions which resulted in a significant interaction, $F(3,54) = 3.76$, $p < .02$. Recall that the surface structure questions were aimed at evaluating the subjects' memory for particular words in the text. In this case, subjects' performance was actually better for the synthetic passages than for the natural passages, $F(1,18) = 4.33$, $p < .05$. This result is completely consistent with the view that synthetic speech requires greater processing demands. If subjects must spend more time encoding the surface phonological properties of synthetic speech, they should subsequently be better able to answer questions aimed at addressing memory for surface structure. Given that human observers are limited in processing capacity, if more time is spent processing the surface structure of a synthetic text, then comprehension will no doubt suffer as a consequence of reallocation of resources. This is precisely what we found in this study.

Conclusion

These results clearly indicate that with sensitive enough measures of comprehension and memory for fluent text, the increased processing demands demonstrated for isolated synthetic word lists can be shown to generalize to the processing of synthetic speech. We believe the line of research we have been

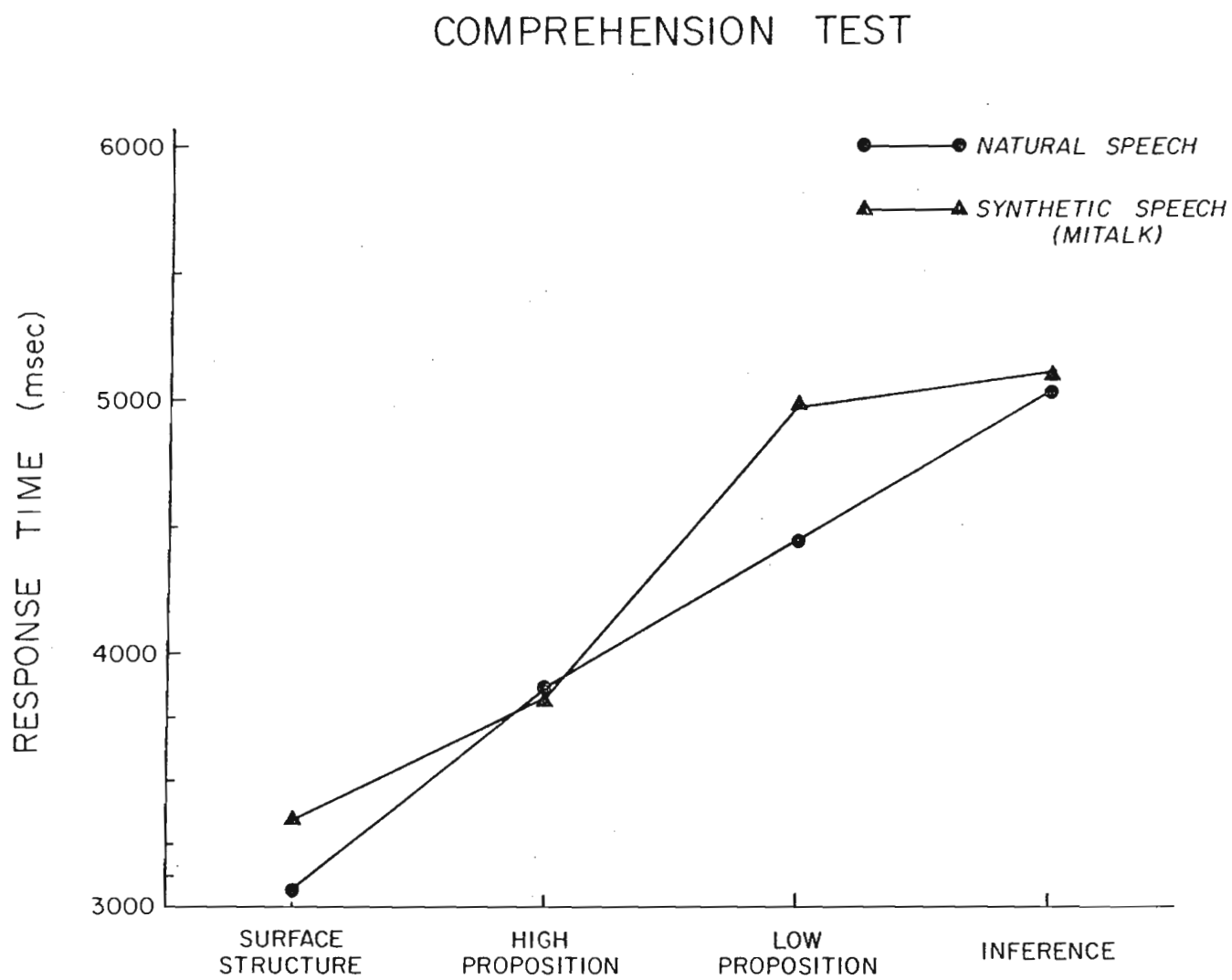

COMPREHENSION TEST
NATURAL SPEECH
SYNTHETIC SPEECH (MITALK)
RESPONSE TIME (msec)
6000
5000
4000
3000
SURFACE STRUCTURE
HIGH PROPOSITION
LOW PROPOSITION
INFERENCE

Figure 4.

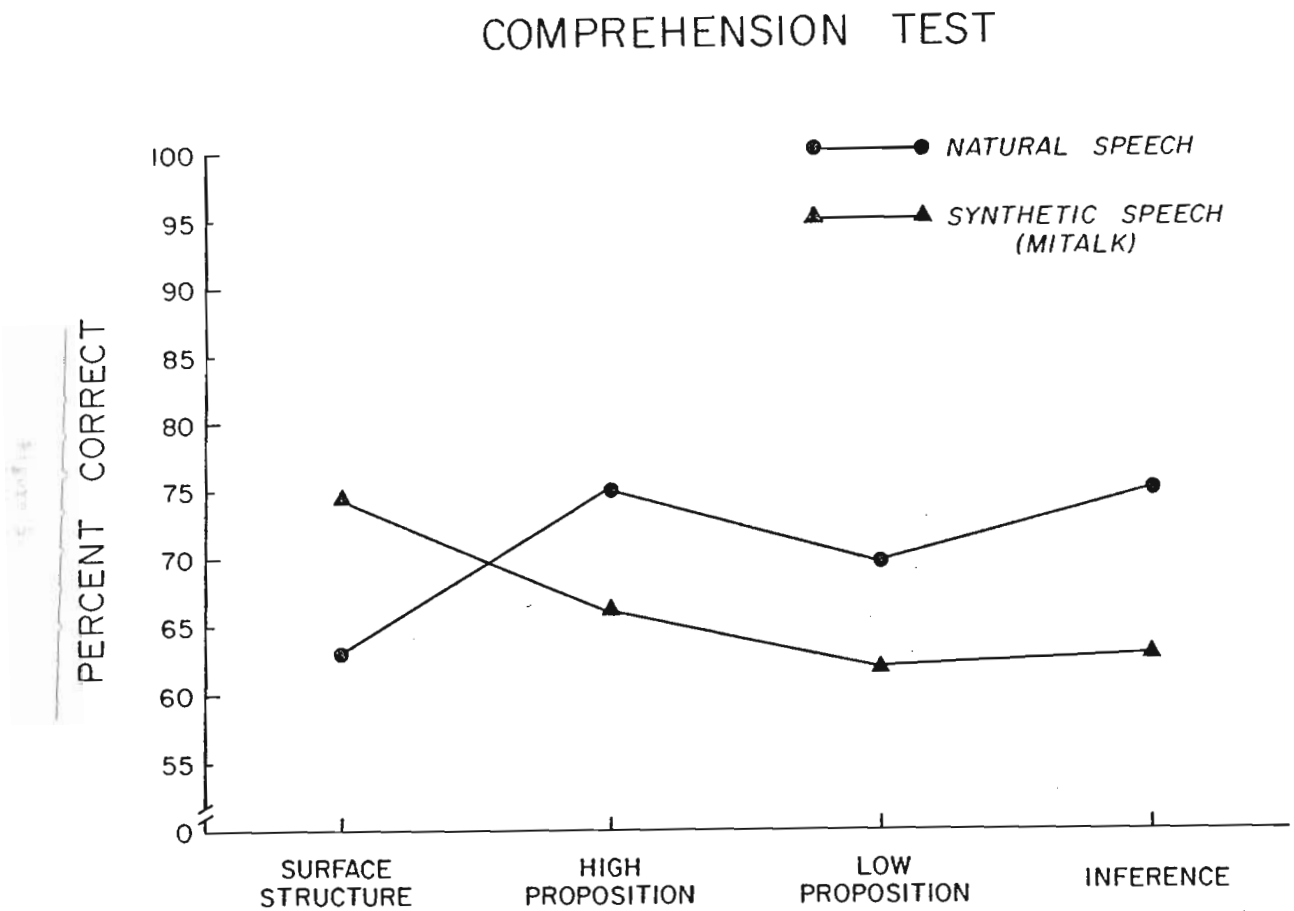

COMPREHENSION TEST
NATURAL SPEECH
SYNTHETIC SPEECH (MITALK)
PERCENT CORRECT
100
95
90
85
80
75
70
65
60
55
0
SURFACE STRUCTURE
HIGH PROPOSITION
LOW PROPOSITION
INFERENCE

Figure 5.

pursuing on the perception and comprehension of synthetic speech demonstrates important limitations on the processing of synthetic speech by human observers. Furthermore, these limitations have important implications for the implementation of synthetic speech systems in less that ideal environments where human observers are required to accurately and rapidly perceive and comprehend messages encoded with synthetic speech. Under conditions of increased demand on the cognitive mechanisms involved in speech processing, synthetic speech may detract from other tasks the listener is engaged in as well as increase the probability of errors in perception and comprehension.

References

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.

Feustel, T. C., Luce, P. A., & Pisoni, D. B. Capacity demands in short-term memory for synthetic and natural word lists. Paper presented at the meeting of the Acoustical Society of America, Miami, December, 1981.

Jenkins, J. J., & Franklin, L. D. Recall of passages of synthetic speech. Paper presented at the meeting of the Psychonomic Society, Philadelphia, November, 1981.

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

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

Some Comparisons of intelligibility of synthetic and natural speech at different speech-to-noise ratios*

David B. Pisoni and Esti Koen

Speech Research Laboratory
Psychology Department
Indiana University
Bloomington, Indiana 47405

*The research reported here was supported by NIH research grant NS-12179 to Indiana University. We thank T. D. Carrell for his help and assistance.

Abstract

Despite the extensive literature on the intelligibility of natural speech in noise, relatively little is currently known about the deleterious effects of noise on the perception (i.e., recognition, identification and understanding) of synthetic speech. This paper summarizes the results of a study that examined the effects of noise on the intelligibility of synthetically produced monosyllabic words using the traditional Modified Rhyme Test (MRT). Over a range of S/N ratios, we found that synthetic speech suffers a greater decrement in performance than naturally produced speech. Moreover, the decrement in performance was even greater when the MRT response format was changed from a closed six-alternative forced choice procedure to an open free response format. These findings have implications for voice response systems that employ synthetic speech in noisy listening environments where intelligibility of the linguistic message may be at a premium.

Insert Figure 1 about here

Within the next few years, there will be an extensive proliferation of various types of voice response devices in man-machine communication systems. Such systems will no doubt see application in situations such as: text-to-speech reading machines for the blind, speaking aids for the deaf, talking computer terminals, computer-aided instruction and sophisticated warning and informational systems in aircraft as well as a variety of consumer oriented products. The development of such automated voice response systems is no longer a matter of basic research and product development--the technology is currently available in the form of specialized speech processing devices which can be integrated into numerous systems requiring voice I/O.

Over the last three years we have been carrying out a program of basic research in our laboratory at Indiana University that is aimed, in part, at gaining a better understanding of how human observers perceive and understand synthetic speech that is generated automatically by rule. Our studies have looked at traditional measures of phoneme intelligibility in isolated words as well as perception of words in various types of sentential contexts where the amount of top-down linguistic knowledge has been manipulated. In addition, we have been exploring how human observers perceive, interpret and "understand" relatively long passages of meaningful connected speech that have been generated entirely by rule with a text-to-speech system. The results of these initial investigations were presented at the ICASSP meeting by Pisoni & Hunnicutt in 1980. More detailed studies on "processing time" and short-term recall for lists of natural and synthetic speech were presented at the Miami meeting of the Society. And, several papers will be presented at this meeting extending our initial work in a number of important directions.

Our earlier studies taken together with these more recent experiments using a variety of experimental techniques and a number of different types of perceptual and linguistic units suggest that important perceptual and cognitive limitations are present when synthetic speech is used in a variety of psychological tasks from phoneme recognition to word recognition to spoken language comprehension. Moreover, these differences in perception between natural and synthetic speech manifest themselves not only in terms of measures of response accuracy but also estimates of the psychological processing time required to execute manual and vocal responses to synthetic speech signals.

Insert Figure 2 about here

We believe that well-motivated decisions concerning the choice and implementation of various voice response systems cannot be made until a number of important psychological problems are examined. Basic research in voice

Sample Test Trials from the Modified Rhyme Test

1.	a)	bad	b)	back	c)	ban	d)	bass	e)	bat	f) bath
2.	a)	beam	b)	bead	c)	beach	d)	beat	e)	beak	f) bean
3.	a)	bus	b)	but	c)	bug	d)	buff	e)	bun	f) buck
4.	a)	case	b)	cave	c)	cape	d)	cane	e)	cake	f) came
5.	a)	cuff	b)	cut	c)	cuss	d)	cub	e)	cup	f) cud
6.	a)	dip	b)	din	c)	dill	d)	dig	e)	dim	f) did
7.	a)	dub	b)	dun	c)	dung	d)	dug	e)	duck	f) dud
8.	a)	fizz	b)	fin	c)	fill	d)	fig	e)	fib	f) fit
9.	a)	hear	b)	heath	c)	heal	d)	heave	e)	heat	f) heap
10.	a)	kid	b)	kit	c)	kill	d)	kin	e)	king	f) kick
11.	a)	lace	b)	lame	c)	lane	d)	lay	e)	lake	f) late
12.	a)	man	b)	math	c)	mad	d)	mat	e)	mass	f) map
13.	a)	pace	b)	pane	c)	pave	d)	page	e)	pay	f) pale
14.	a)	path	b)	pat	c)	pack	d)	pad	e)	pass	f) pan
15.	a)	peas	b)	peak	c)	peal	d)	peace	e)	peach	f) peat
16.	a)	pip	b)	pick	c)	pin	d)	pill	e)	pit	f) pig
17.	a)	puff	b)	pus	c)	pub	d)	pun	e)	puck	f) pup
18.	a)	rate	b)	race	c)	ray	d)	raze	e)	rave	f) rake
19.	a)	safe	b)	sake	c)	same	d)	sane	e)	save	f) sale
20.	a)	sat	b)	sag	c)	sack	d)	sap	e)	sass	f) sad
21.	a)	seed	b)	seek	c)	seen	d)	seep	e)	seem	f) seethe
22.	a)	sill	b)	sick	c)	sing	d)	sit	e)	sin	f) sip
23.	a)	sup	b)	sud	c)	sun	d)	sum	e)	sub	f) sung
24.	a)	tap	b)	tang	c)	tam	d)	tan	e)	tab	f) tack
25.	a)	tease	b)	tear	c)	teak	d)	teal	e)	team	f) teach

Fig. 1

SOME APPLICATIONS OF VOICE-RESPONSE SYSTEMS USING SYNTHETIC SPEECH:

1. MILITARY AND INDUSTRIAL WARNING SYSTEMS
2. TALKING CONSUMER PRODUCTS
3. SPEAKING AIDS FOR THE HANDICAPPED
4. TEXT-TO-SPEECH SYSTEMS
5. AUTOMATED INFORMATION RETRIEVAL SYSTEMS
6. COMPUTER-AIDED INSTRUCTION IN TEACHING AND TRAINING
7. FEEDBACK DEVICES IN AIRCRAFT AND MILITARY COMMAND/CONTROL APPLICATIONS

Fig. 2

technology is needed at this time on questions such as: (1) the effects of various kinds of noise on perception of synthetic speech; (2) perception of synthetic speech under various listening conditions requiring differential cognitive and attentional demands; (3) processing time studies of recognition and interpretation of synthetic speech; (4) effects of practice and familiarity; (5) comprehension of fluent continuous synthetic speech and the interaction of various knowledge sources in speech perception; (6) interaction of prosodic and segmental cues in perception of synthetic speech; (7) comparative evaluations of various commercially available speech synthesizers and synthesis-by-rule systems; (8) relations between size of message set and specific task requirements for use with synthetic speech; (9) questions surrounding naturalness and the effects of using synthetic speech on intelligibility; (10) relationships between traditional forced-choice measures of isolated word recognition and perception and comprehension of words and sentences in fluent speech where many different sources of knowledge interact in complex ways.

Research questions such as these should be examined under carefully controlled laboratory testing conditions in which comparisons between synthetically produced speech and natural speech can be undertaken. These studies will need to be done under various attentional and task specific testing conditions in order to map out the possible interactions between the observer, signal and task demands. Moreover, relevant information concerning the time-course of perceptual learning and adaptation to synthetic speech input should be examined as well in these studies with both naive and practiced observers since the amount of familiarity and experience with synthetic speech substantially affects its perception and comprehension, particularly under adverse listening conditions.

In the present paper we wish to report the results of a recent study that examined the effects of noise on the intelligibility of synthetically produced speech. To study intelligibility of isolated words, we used the Modified Rhyme Test. The testing format is illustrated in Figure 3 below.

Insert Figure 3 about here

The Modified Rhyme Test is a six-alternative forced-choice procedure that uses monosyllabic English words. We selected this test because it is reliable, shows little effect of learning and is easy to administer to naive listeners. Twelve groups of undergraduate students served as listeners. Six groups of subjects heard the MRT words in the traditional forced-choice format; the other six heard the same items in an open or "free-response" format in which they were required to write down what English word they heard on each trial.

The synthetic stimuli were generated automatically by rule on the MITalk text-to-speech system. The natural stimuli were produced by a male talker. The 300 test items from each set were digitized via a 12 bit A-D, stored on disk files and then output via a 12 bit D-A during the actual experimental sessions.

NEEDED RESEARCH ON THE PERCEPTION OF SYNTHETIC SPEECH:

1. PROCESSING TIME EXPERIMENTS
2. LISTENING TO SYNTHETIC SPEECH IN NOISE
3. PERCEPTION OF SYNTHETIC SPEECH UNDER DIFFERING ATTENTIONAL DEMANDS
4. EFFECTS OF SHORT AND LONG-TERM PRACTICE
5. COMPREHENSION AND UNDERSTANDING FLUENT SYNTHETIC SPEECH PRODUCED BY RULE
6. INTERACTION OF SEGMENTAL AND PROSODIC CUES
7. COMPARISONS OF DIFFERENT RULE SYSTEMS AND SPEECH SYNTHESIZERS
8. EFFECTS OF NATURALNESS ON INTELLIGIBILITY
9. GENERALIZATION OF SYNTHETIC SPEECH TO NOVEL UTTERANCES
10. EFFECTS OF MESSAGE SET SIZE ON SYNTHETIC VS. LOW-BIT RATE NATURAL SPEECH

Fig. 3

For each test format we examined performance for natural and synthetic words at three different speech-to-noise ratios: +30, +20 and 0 dB. The speech was always presented at an average of 80 dB SPL against a background of broadband white noise which was attenuated for the particular conditions. All signals were presented through TDH-39 headphones. The experimental sequences were controlled on-line with a PDP-11 computer.

Insert Figure 4 about here

Figure 4 shows the overall performance for our twelve groups of listeners at three S/N ratios. The closed response format is shown by the filled boxes, the open response format is shown by the open boxes. Natural speech is displayed as the triangles, synthetic speech is displayed as circles in this figure.

Notice that, in each case, performance for the synthetic speech is worse than the natural speech. And, note in particular, that for both natural and synthetic speech, performance is worse using an open free-response format than the standard forced-choice format. The drop in performance is quite apparent as we go from +20 to 0 dB for both natural and synthetic speech. We have not run the +10 dB S/N condition yet but we plan to do so within the next few months.

Insert Figure 5 about here

Figure 5 shows a table of difference scores for natural and synthetic speech as a function of testing format. The difference in performance at +20 and 0 S/N is shown in each box. Notice that for the MRT closed format on the left, the difference in performance is greater for the synthetic speech than the natural speech. And, going from left to right, the differences are consistently larger for the open format than the closed format. Turning to the open response format shown in the right hand column, we see that the differences between natural and synthetic speech are about the same, although the difference is slightly larger for natural speech. Notice also that performance for both natural and synthetic speech at 0 dB S/N ratio is extremely low and probably represents a floor effect for the synthetic speech.

We had expected the differences in performance to be even larger for the synthetic speech when the response format was changed from forced-choice to free-response but the floor effect may have prevented this from occurring in this experiment.

In summary, we have found that intelligibility of synthetic speech is affected more by masking noise than natural speech in the traditional forced-choice MRT format. Performance for both natural and synthetic speech decreased when the same items were presented in a free-response or open format. Our findings therefore suggest that synthetic speech may be affected to a greater

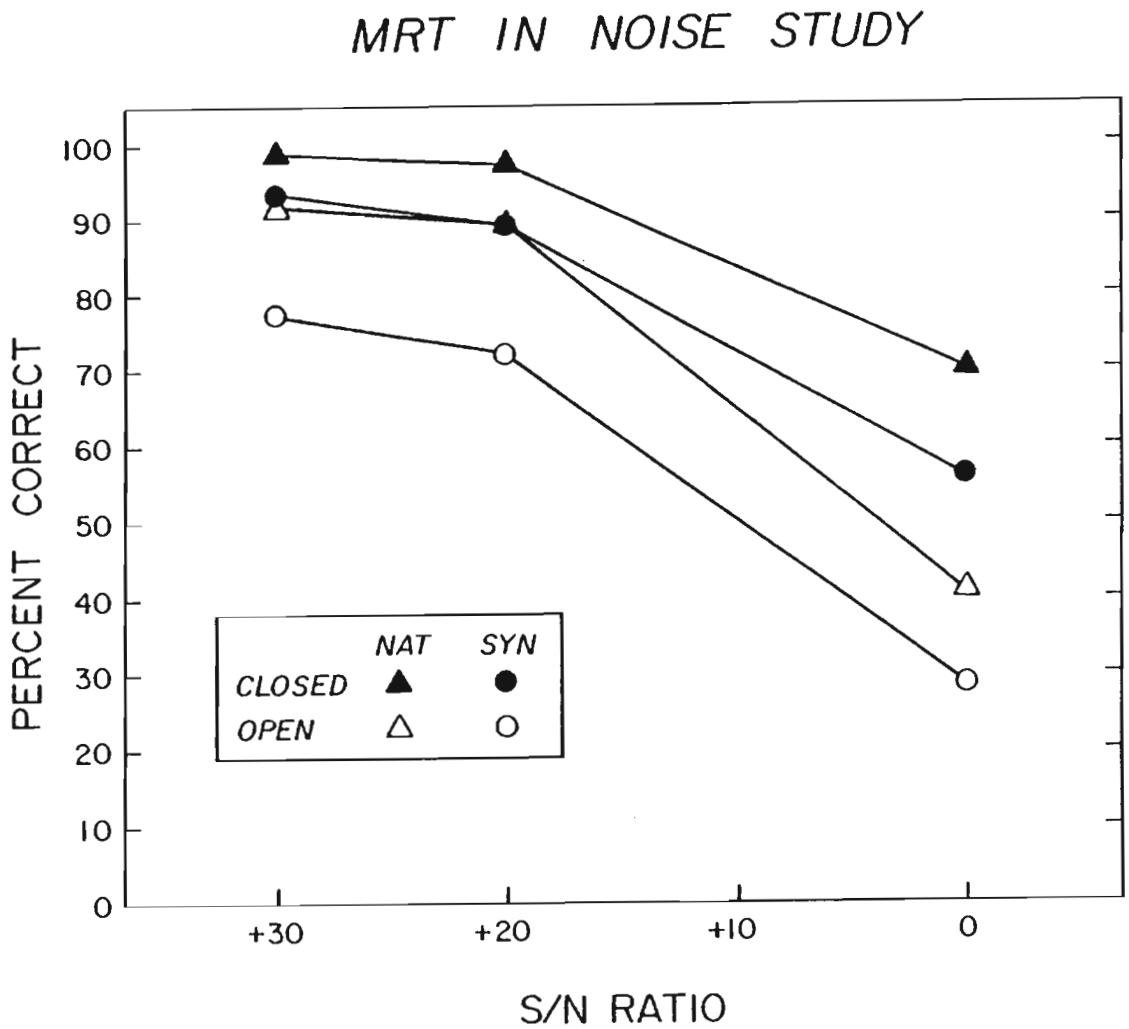


Figure 4. Percent correct word recognition in the MRT test as a function of S/N ratio for natural and synthetic items. The closed symbols display data collected in the closed forced-choice response format; the open symbols show data from a free-response open test format.

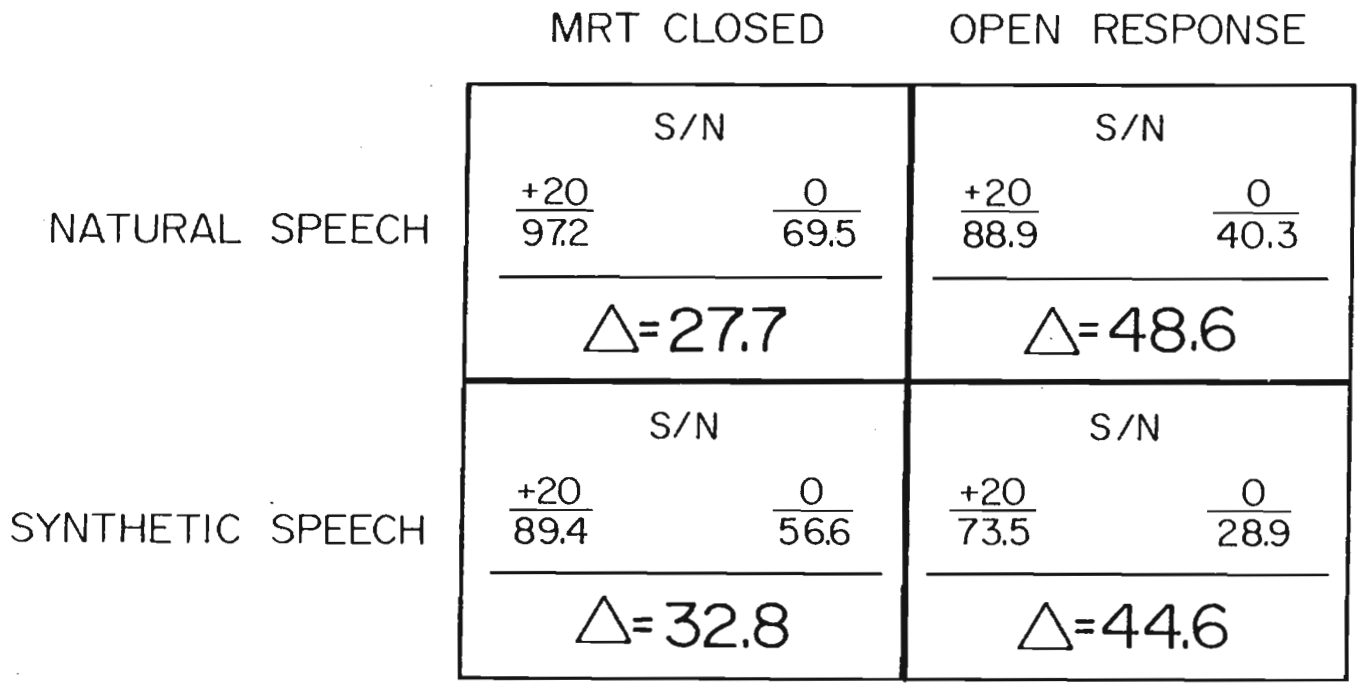


Figure 5. Difference scores for natural and synthetic speech across the two types of response formats in the MRT test.

degree than natural speech under various conditions of signal distortion. Moreover, such signal distortions may interact with the requirements of the processing task and the observer to produce deliterious effects on speech intelligibility that are unknown at the present time. Although there is an extensive literature on the intelligibility of natural speech in noise, relatively little is known about the effects of noise on the perception and comprehension of synthetic speech. Our study represents a first attempt to obtain empirical data on this issue. Other research dealing with the effects of noise on word recognition in sentences and the interaction of different knowledge sources is currently underway in our laboratory. We hope to report these findings at future meetings of the Society.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

Effects of practice on speeded classification of natural and synthetic speech*

Louisa M. Slowiaczek and David B. Pisoni

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

*The research reported here was supported, in part, by NINCDS Research Grant NS-12179 and NIMH Research Grant MH-24027 to Indiana University in Bloomington.

Abstract

At the last meeting of the Society in Miami, one of us (Pisoni) presented results from a speeded classification task demonstrating increased processing time for recognition of synthetic speech. In a lexical decision task, we showed an overall increase in response time of 145 msec for synthetically produced words and nonwords over naturally produced control items. Moreover, there was no interaction between signal type and classification response suggesting that the observed differences were due to early stages of perceptual analysis in which the segmental representation is developed from the acoustic-phonetic input. In the present investigation, we were interested in determining whether the observed differences in processing time between natural and synthetic speech could be attenuated or possibly eliminated by practice with the experimental materials and task over several days. We ran ten undergraduate subjects for five days in the lexical decision task. As expected, accuracy improved and latency decreased over the test sessions. However, the relative differences in response latency between natural and synthetic items remained roughly constant. Thus, the earlier differences observed in this speeded classification task do not appear to be a result of the subjects' unfamiliarity with synthetic speech. Rather they seem to reflect real differences in the perceptual and cognitive processes used to extract segmental information from the speech signal.

Over the last few years, the field of voice technology has advanced at an enormous rate. Within the next few years, there will be an extensive proliferation of various types of voice response devices in man-machine communication systems. Such systems will be applied in areas such as: text-to-speech reading machines for the blind, speaking aids for the deaf, computer aided instruction, feedback devices in aircraft, and other complex machinery. However, despite this progress, relatively little research has been directed at basic questions concerning how human observers process (i.e., perceive, encode and interpret) speech signals generated by these devices.

At the last meeting of the Society, Pisoni (1981) presented results from a speeded classification task demonstrating increased processing time for recognition of synthetic speech. Subjects performed a lexical decision task in which they had to classify items as either "words" or "nonwords." The results showed an overall increase of 145 msec in response time for synthetically produced items over naturally produced control items. Moreover, there was no interaction between signal type and classification response. These results suggest that the observed differences were due to early stages of perceptual analysis in which the segmental representation is developed or extracted from the acoustic-phonetic input. The experiment reported by Pisoni was run with naive observers who had little, if any, familiarity with synthetic speech. It is not clear how subjects might perform in the lexical decision task with practice and with greater familiarity with the synthetic materials.

The present investigation was therefore designed to determine whether the observed differences in processing time between natural and synthetic speech could be attentuated or possibly eliminated by practice with the experimental task and materials over several days. Ten undergraduate subjects performed a lexical decision task for one hour a day for five consecutive days. The subject was required to classify acoustic stimulus items as either a "word" or a "nonword" as fast as possible by pressing one of two buttons located on a response box in front of him.

Insert Figure 1 about here

Figure 1 shows some examples of the word and nonword stimuli used in the experiment. The nonword stimuli shown on the right were matched in number of syllables with the word stimuli on the left. Subjects were presented with two blocks of 100 trials on each day. Each block contained 50 words and 50 nonwords. Half of the items in a block were natural speech tokens produced by Dennis Klatt, and half were synthetic speech tokens produced by the MITalk text-to-speech system.

Insert Figure 2 about here

---

1. PROMINENT	1. PRADAMENT
2. BAKED	2. BEPT
3. TINY	3. TADGY
4. GLASS	4. GEEP
5. PARENTS	5. PEEMERS
6. TOLD	6. TAVED
7. BLACK	7. BAEP
8. CONCERTS	8. CAELIMPS
9. DARK	9. DUT
10. BABBLE	10. BURTLE
11. CRITIC	11. CRAENICK
12. BOUGHT	12. BUPPED
13. PAIN	13. POON
14. GORGEOUS	14. GAETLESS
15. COLORED	15. COOBERED

Figure 1. Examples of word and nonword stimuli used in the lexical decision task.

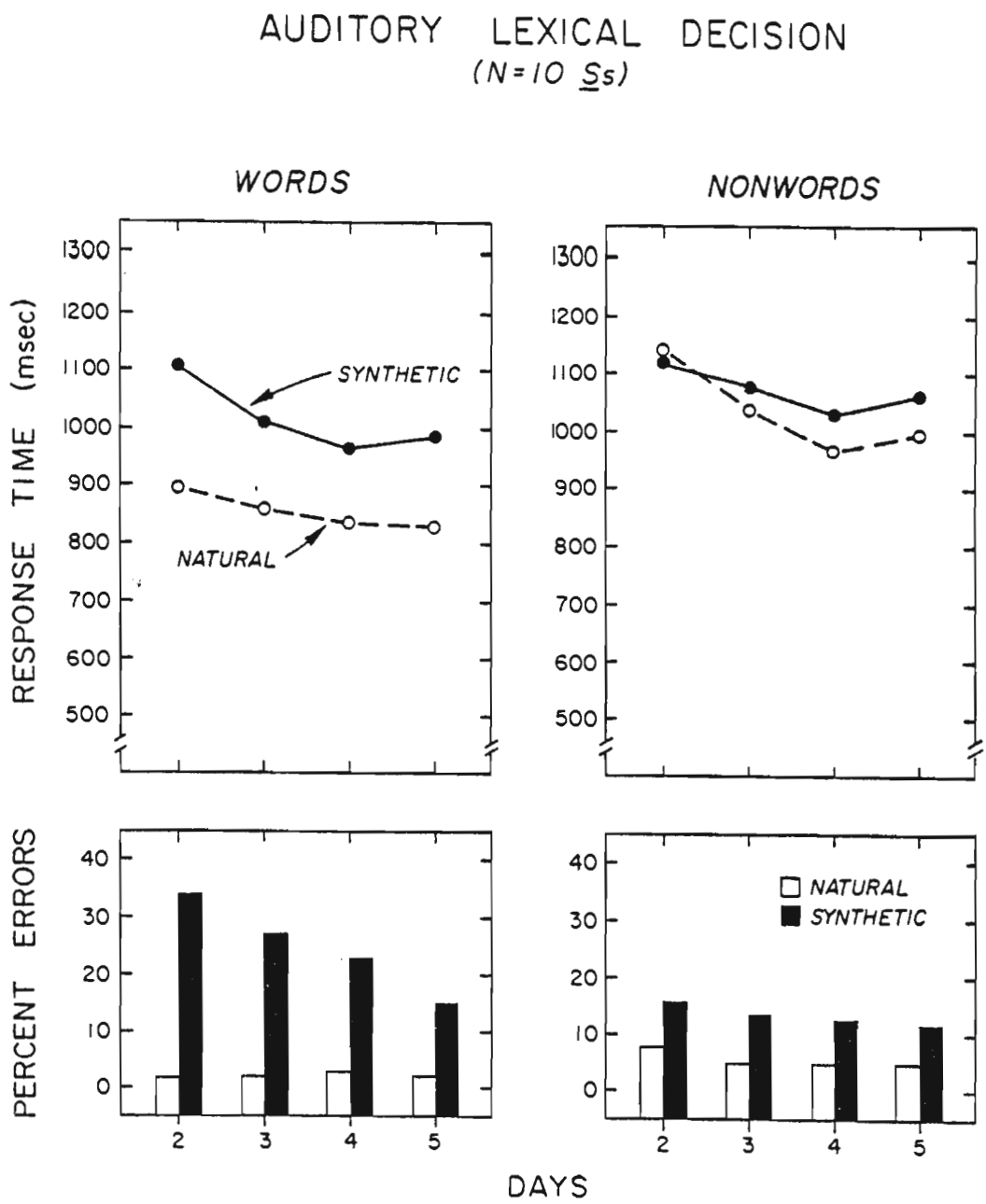

AUDITORY LEXICAL DECISION
(N=10 Ss)
WORDS
NONWORDS
RESPONSE TIME (msec)
1300
1200
1100
1000
900
800
700
600
500
SYNTHETIC
NATURAL
PERCENT ERRORS
40
30
20
10
0
NATURAL
SYNTHETIC
2
3
4
5
DAYS

Figure 2.

The results are shown in Figure 2. On the left are the results for word stimulus items and on the right are the results for nonword stimulus items. First, please direct your attention to the error rates that are shown at the bottom of the slide. Notice that the error rates for natural words and nonwords remain relatively constant for Days 2, 3, 4, and 5. The error rates for synthetic words and nonwords decline over days. However, the difference was only reliable for synthetic words.

The panels at the top of the slide display response times for correct responses to word and nonword stimuli. The response times for natural speech stimuli are consistently faster than the response times for the synthetic stimuli. The overall difference is 89 msec. As shown in the slide, the natural versus synthetic difference is greater for words than for nonwords. Notice that although there is a decrease in response time over days, the difference between natural and synthetic stimuli is maintained.

Since the present investigation was primarily concerned with a change in the observed differences in processing time between natural and synthetic speech with several days practice, a comparison of the results obtained by Pisoni (1981) and the fifth day of practice in the present investigation was of special interest to us.

Insert Figure 3 about here

The panel on the left of this figure displays the averaged results of Day 1 in the Pisoni (1981) study. On the right are the results of Day 5 for the present investigation. Notice that overall the results are very similar, except that the responses for the fifth day are faster.

Therefore, while overall accuracy increased, and response times decreased across the 5 day period of this study, the relative differences in response latency between natural and synthetic items remained roughly constant. It appears, therefore, that the earlier differences observed in the speeded classification task are not primarily a result of subjects being unfamiliar with synthetic speech. The present investigation demonstrates that practice over a five day period does not eliminate the observed differences in processing time between natural and synthetic speech. Thus, the differences in response latency reflect real differences in the perceptual and cognitive processes used to extract segmental information from the speech signal, particularly for synthetically produced speech signals.

Although the effects of short-term familiarity do not appear to improve performance in perception of synthetic speech, it is not clear what the effects would be for very long term practice in listening to synthetic stimuli (i.e., over a period of several weeks or months). Obviously, further studies will need to be conducted to determine the effects of long-term practice and the exact locus of these perceptual and cognitive differences in the human information processing system. For the present, however, our results and the ones presented earlier in Miami by Pisoni demonstrate reliable and potentially important

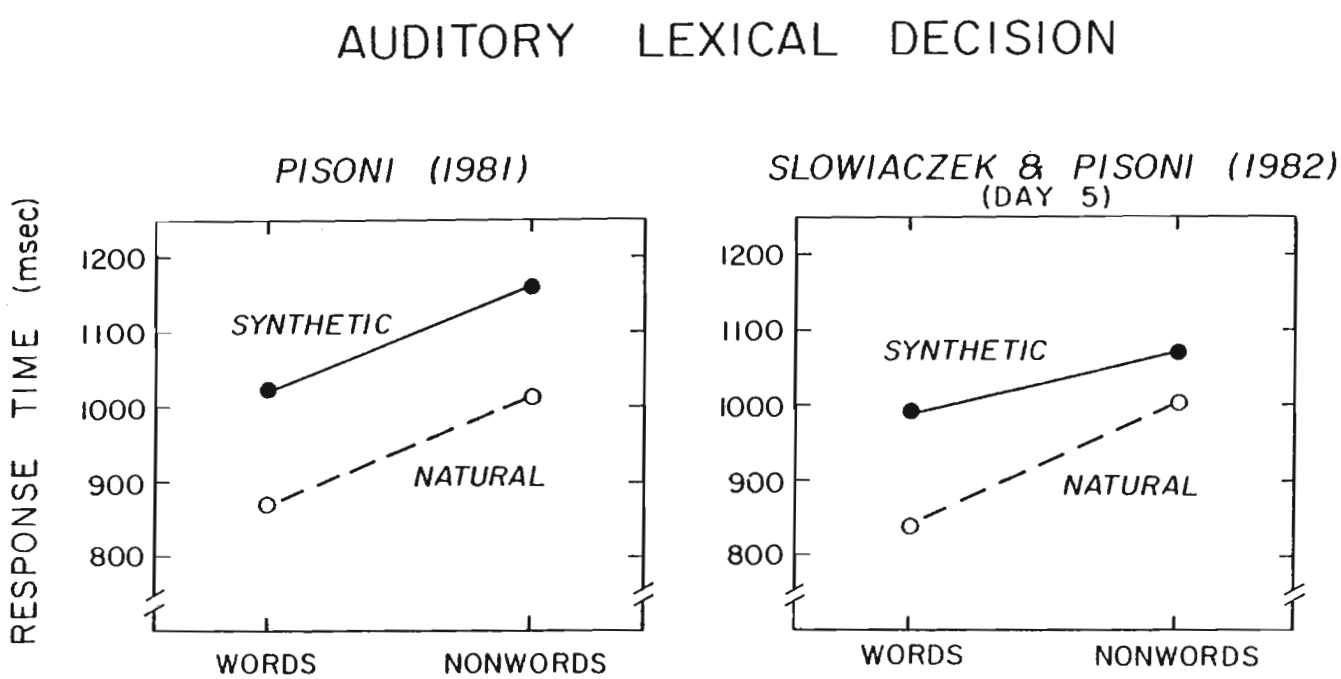

AUDITORY LEXICAL DECISION
PISONI (1981)
SLOWIACZEK & PISONI (1982)
(DAY 5)
RESPONSE TIME (msec)
1200
1100
1000
900
800
SYNTHETIC
NATURAL
WORDS
NONWORDS

Figure 3.

differences in processing time for recognition of synthetic words and nonwords. Such processing time differences in perception between natural and synthetic speech may well have important implications for applications where a human observer's attentional resources are severely limited. Such situations include having to process and respond to information from several sensory modalities at the same time, such as in the cockpit of an aircraft, or in complex command-control environments.

Reference

Pisoni, D. B. Speeded classification of natural and synthetic speech in a lexical decision task. Paper presented at the 102nd Meeting of the Acoustical Society of America, December 8, 1981, Miami Beach, Florida.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

Effects of linguistic context on the durations of lexical categories*

Jan Charles-Luce and Laurie Ann Walker

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

*This work was supported by grants from NIMH, Research Grant No. MH-24027, and NINCDS, Research Grant No. NS-12179.

Abstract

Findings from previous perception studies examining the effects of semantic and syntactic language constraints showed that subjects' performance was best when constraints were not violated and worst when both constraints were violated. The present study was an attempt to assess these findings for the production of speech. Subjects read grammatical, anomalous, and ungrammatical isolated sentences and grammatical, anomalous, and ungrammatical passages. Words from six lexical categories were digitally measured. The results show that speakers use nouns and adverbs as semantically governing units, or "islands of reliability". The results indicate that semantic constraints are more important in the production of speech than syntactic constraints. The differential effects of context may have implications for the development of data bases used in automatic speech recognition.

Several years ago a series of studies was undertaken to determine how perception and comprehension of sentences were affected when syntactic and semantic constraints were violated.

Insert Figure 1 about here

Miller and Isard (1963) asked subjects to shadow fifty grammatical, fifty anomalous, and fifty ungrammatical sentences. Examples of these are seen in the first figure. They found that percent correct shadowing increased as a function of sentence type and concluded that semantic and syntactic rules were involved in sentence perception. Marks and Miller (1964), again using grammatical, anomalous, and ungrammatical sentences, found in a free recall task that subjects were able to recall sentences better when fewer semantic and/or syntactic rules were violated. Wang (1970) found that subject comprehension ratings and recognition memory was best for grammatical sentences and worst for ungrammatical sentences; anomalous sentences consistently fell in the middle.

These three studies address only how a listener responds to violations of syntactic and semantic rules in perceptual and memory tasks. In the following study we examined how a speaker acts in producing these violations. Specifically, we were interested in what would happen to durations of individual words from six lexical categories when semantic and/or syntactic rules were violated. The six lexical categories examined were: nouns, verbs, adjectives, adverbs, prepositions, and articles. We expected that for all lexical categories duration would be influenced by the sentence type. That is, duration of a word would be longer in an anomalous sentence, with semantic violations, than a normal, grammatical sentence. Duration would be longest in an ungrammatical sentence, which has both semantic and syntactic violations.

Experiment 1

In the first experiment, the sentences generated by Miller and Isard were randomized by computer and presented individually on a CRT monitor to seven paid subjects from Indiana University. The subjects were asked to read each sentence aloud. Their productions were recorded on audio tape by one of the experimenters.

Five to nine tokens from each of the six lexical categories--nouns, verbs, adjectives, adverbs, prepositions, articles--had been selected before the recordings. The target words were measured from digital oscillograms of the recordings to the nearest .001 second.

EXAMPLE SENTENCES

(From Miller and Isard, 1963)

GRAMMATICAL:

The brightly colored toys pleased children.

The secret miracle ingredient worked wonders.

ANOMALOUS:

The nationally disguised toys covered customers.

The frenzied grey ingredient frayed autographs.

UNGRAMMATICAL:

Needed advertised cleverly the toys streets.

Autographs Latin mare worked sticky the.

Figure 1.

Insert Figure 2 about here

The results are shown in the second figure. The two panels show the mean durations for each lexical category across grammatical, anomalous, and ungrammatical sentences. There are significant differences in duration across the three types of sentence contexts for all function words in the left panel and for all content words in the right panel, except for nouns and adverbs. The significant differences at the $p < .05$ level are on the order of 70 msec for adjectives, 70 msec for verbs, 40 msec for prepositions, and 80 msec for articles. Where the durations are significantly different, they are shortest for the grammatical sentences and longest for the ungrammatical sentences, with anomalous between, as we expected.

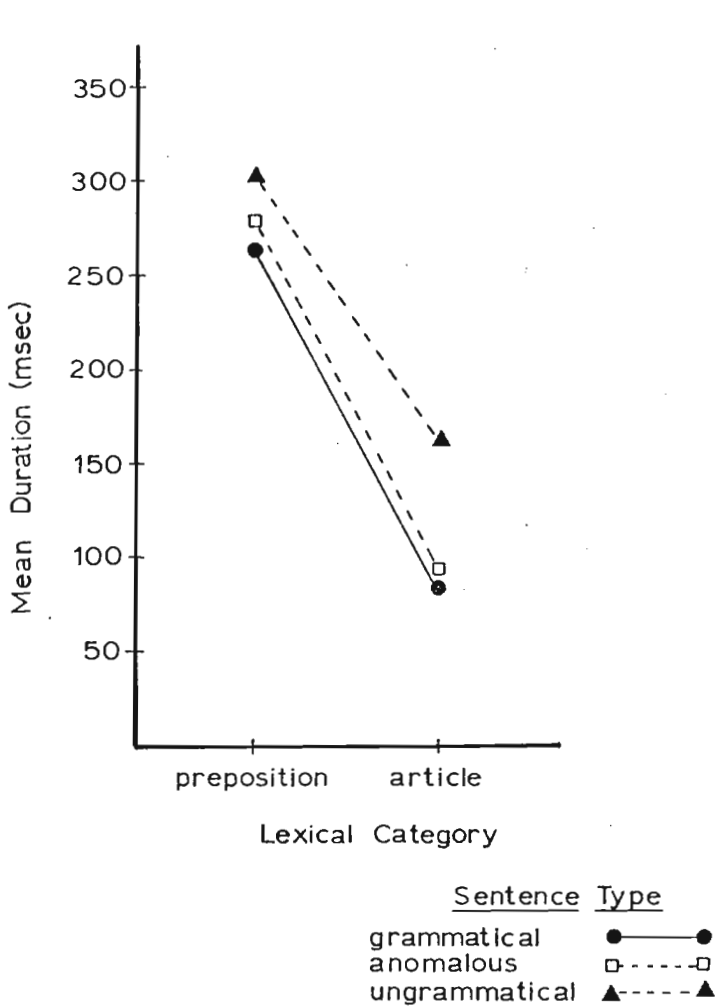
These results indicate that word duration for most syntactic categories is lengthened when semantic rules are violated and lengthened even more when both syntactic and semantic constraints are violated. However, nouns and adverbs appear to remain constant, suggesting that they may act as some informational unit for the speaker, perhaps as "islands of reliability". The observance of a lack of a uniform durational effect across all categories suggests that violation of semantic and syntactic rules in speech production is more complex, at least for content words, than we initially expected.

Chomsky (1965) has argued that verbs and adjectives are selectionally restricted by a relatively free noun. Nouns therefore semantically restrict or govern the choice of a verb and adjective. Adverbs, although much less studied, also act in a more unrestricted manner (Chomsky, 1965; Jackendoff, 1972; O'Shaugnessy, 1976). Adverbs are very rich in meaning and could provide for a stricter, more informative semantic interpretation of a sentence when they are present. A speaker may therefore treat nouns and adverbs as semantically governing units and then compensate in speech production for the other syntactic categories through increases in duration.

Experiment 2

A second experiment was carried out to see if we would obtain these same results when the same six lexical categories occurred in grammatical, anomalous, and ungrammatical passages of fluent text. In short, we wanted to see if greater context would affect the durational results found with isolated sentences; particularly if nouns and adverbs maintained constant durations across these contexts.

A reading passage about geology served as the grammatical passage. Anomalous and ungrammatical passages were constructed from this. Three Indiana University undergraduates were recorded while reading the passages aloud. Target words were then digitally measured as in the first experiment.


Mean Duration (msec)
350
300
250
200
150
100
50
preposition
article
Lexical Category
Sentence Type
grammatical
anomalous
ungrammatical

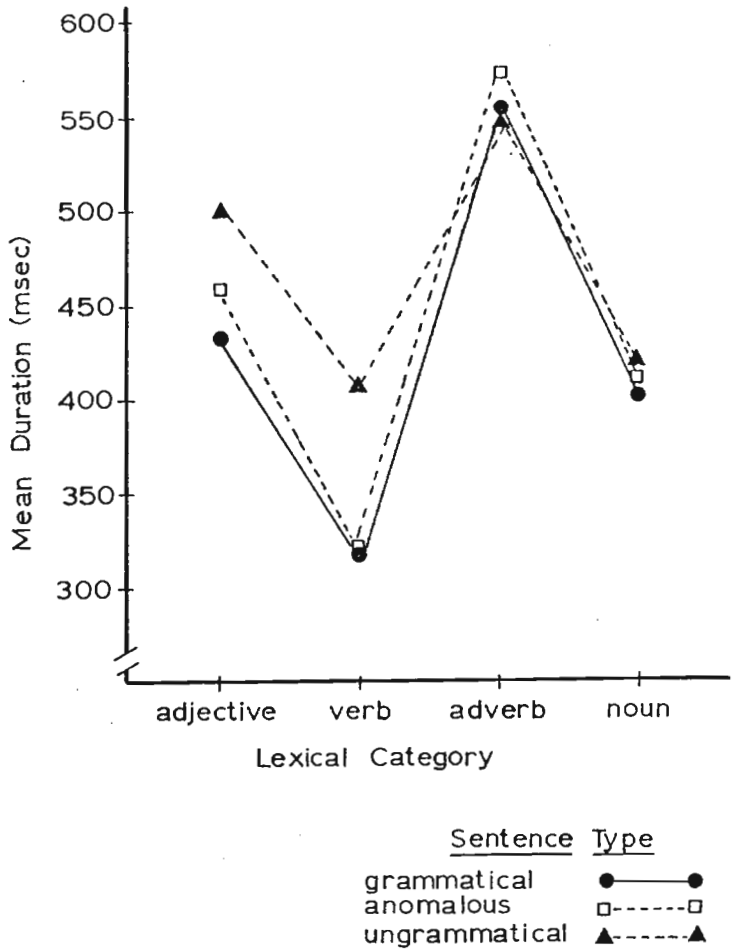

Mean Duration (msec)
600
550
500
450
400
350
300
adjective
verb
adverb
noun
Lexical Category
Sentence Type
grammatical
anomalous
ungrammatical

Figure 2.

Insert Figure 3 about here

The third figure shows the results of this experiment. As in the previous experiment, for all content words and function words the mean durations are shortest for the grammatical passages. However, only nouns and adverbs show significant durational changes across the three types of contexts. These significant differences are on the order of 80 msec for both nouns and adverbs. Context appears to reverse the findings of the first experiment.

Looking only at the content words in the right panel of the figure, we see that mean durations for adverbs and nouns in the anomalous and ungrammatical contexts are much longer in relation to their mean durations in the grammatical context, but there is little difference in mean duration between the anomalous and ungrammatical contexts. Again, anomalous and ungrammatical passages are similar because they both violate semantic rules. They are different, however, in that only ungrammatical passages violate syntactic rules as well as semantic rules. The minimal durational differences between anomalous and ungrammatical sentences suggest that syntactic violations do not produce lengthening in nouns and adverbs. Thus, meaning may be the most important constraint for lexical categories which selectionally restrict other categories in the grammar or which provide for a richer semantic interpretation of a sentence. Moreover, given more context, violation of semantic constraints affects the duration of those lexical categories which act as restrictive semantic units for the speaker.

Summary and Conclusion

This study suggests that speakers are sensitive to semantic and syntactic constraints in the production of speech. One way speakers compensate when these constraints are violated is by lengthening the durations of words in sentences. However, we did not find that all lexical categories acted alike nor that all categories necessarily varied in duration despite the same rule violations.

The most interesting result, in our view, involved the content words. Nouns and adverbs consistently opposed verbs and adjectives in the isolate sentences and in connected text. We noted that nouns and adverbs have traditionally been thought to be semantically more important categories in the grammar. As such, they serve as governing units for the speaker when semantic violations occurred. In the isolated sentences, where the context was limited, speakers used these categories to stabilize the violations, and did not need to compensate with longer durations. However, the other categories were lengthened in order to compensate for their violation of the governing units. On the other hand, when anomalous and ungrammatical sentences were embedded in longer passages, the speaker over-compensated for the violations by lengthening the units--nouns and adverbs--which semantically govern in an attempt to rectify the violations by emphasizing their function in the grammar. Verbs and adjectives were, then, not the speakers' focus; these categories were produced at their inherent durations

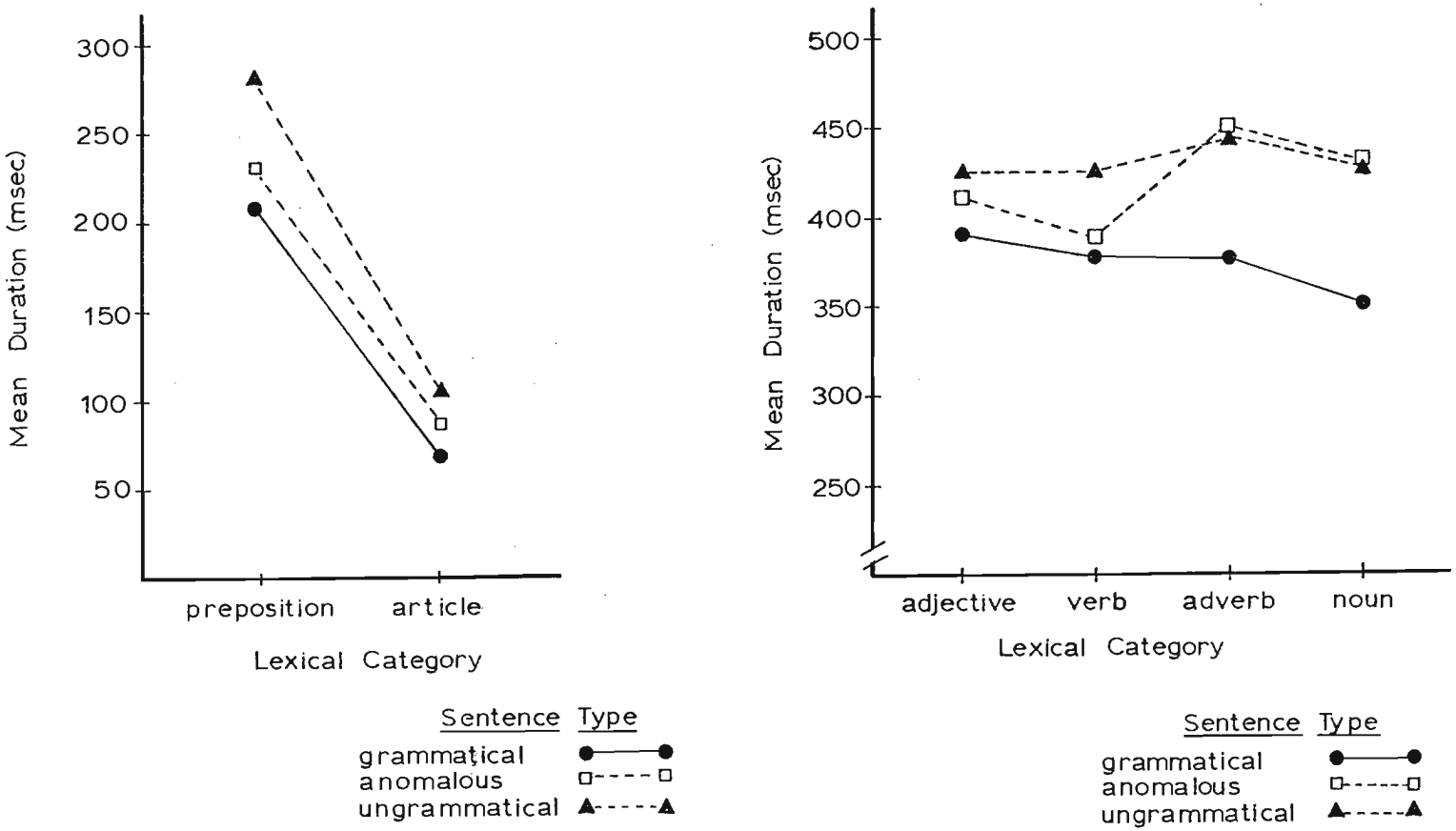

Mean Duration (msec)
300
250
200
150
100
50
preposition
article
Lexical Category
Sentence Type
grammatical
anomalous
ungrammatical
Mean Duration (msec)
500
450
400
350
300
250
adjective
verb
adverb
noun
Lexical Category
Sentence Type
grammatical
anomalous
ungrammatical

Figure 3.

for the speakers regardless of any violations. We suggest that a larger context depends more heavily on semantic constraints for individual sentences within it. Moreover, we suggest that semantic constraints may be more important for the speaker than syntactic constraints and, in turn, these appear to influence the speaker's control over the durations of words in speech. We believe the results of this study may have implications for the construction of large-scale data bases which have been used recently in speech recognition systems. If the durations of words in sentences are affected by semantic and syntactic constraints, such factors will obviously have to be incorporated in some way into the algorithms currently used to recognize the acoustic-phonetic properties of words in fluent speech.

References

Chomsky, N. Aspects of the theory of syntax. Cambridge, Mass.: MIT Press, 1965.

Jackendoff, R. S. Semantic interpretation in generative grammar. Cambridge, Mass.: MIT Press, 1972.

Marks, L. E. & Miller, G. A. The role of semantic and syntactic constraints the memorization of English sentences. Journal of Verbal Learning and Verbal Behavior, 1964, 3, 1-5.

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

O'Shaugnessy, D. Modelling fundamental frequency, and its relationship to syntax, semantics, and phonetics. Unpublished doctoral dissertation, MIT, 1976.

Wang, M. D. Influence of linguistic structure on comprehensibility and recognition. Journal of Experimental Psychology, 1970, 85, 83-89.

III. INSTRUMENTATION AND SOFTWARE DEVELOPMENT

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

WAVMOD: A Program to Modify Digital Waveforms*

Bob Bernacki

Speech Research Laboratory
Psychology Department
Indiana University
Bloomington, IN 47405

*This work was supported, in part, by NINCDS Research Grant No. NS-12179 and NIMH Research Grant No. MH-24027 to Indiana University.

WAVMOD was developed to provide a comprehensive integrated package for manipulating natural and synthetic digitized acoustic waveforms. The WAVMOD program consists of a support structure and various signal processing routines. The support structure is composed of stimulus file handlers, a level control system, a segment file management system, and an operation logging system. The signal processing routines provide the means to modify or replace digitized waveforms.

WAVMOD Functions

The fundamental modifications are signal replacement operations. These replacements consist of tone, envelope noise, and white noise. A fourth modification, level adjustment, is available as an independent command, as well as within the other modification commands.

An important feature of WAVMOD is that it provides the user with the capability to modify a digitized speech stimulus file entirely, or by selected segments. Another program on the SRL system, WAVES, provides the facility to view, edit, and preview digitized signals (see Luce and Carrell, 1982). Waves may be used in conjunction with WAVMOD to define segment times. Since modification proceeds on a sample-by-sample basis, the resolution of the segment time is .1 msec for a 10KHz sampling rate.

WAVMOD was designed to accommodate novice and experienced users without sacrifice of speed or clarity. An independent monitor screen is utilized as an adjunct to the terminal screen for display of menu information that would otherwise cause loss of program continuity if presented on the terminal screen. The adjunct screen displays include the WAVMOD session log, SEGMENT file contents, and options within the commands. Furthermore, the user can invoke a display of menus or options by simply typing a carriage return in response to a prompt. These two features provide important information to the novice, yet the experienced user need not be burdened with unnecessary detail. The parameter prompts allow default values to further simplify the user interaction. At various points in WAVMOD, the user may abort the current command and return to the WAVMOD command entry.

WAVMOD uses special terminology for the input and output stimulus files to emphasize their functions in this system. The input file is refered to as the SOURCE file, and the output file is the RESULTANT file. In addition to the stimulus files, LOG files and SEGMENT files are created and maintained by WAVMOD. The LOG is a text record of all WAVMOD operations, and is compatible with sequential access text files. A LOG file is kept with a stimulus file as a "lab notebook" to ensure that detailed documentation of the stimulus generation is available.

Additions to the Original Package

The menu of WAVMOD modifications has been expanded to include glottal-like modulation of the SOURCE waveform (generally time varying sinusoids), and a signal and noise mixer. To accommodate the needs of new experimental paradigms, a user command was installed to provide a mechanism for expansion without the need for alteration of the program's core structure. With the US subroutine, a knowledgeable programmer can develop software, in a cookbook fashion, to implement specialized replacement or signal processing modifications. The current SRL system version of WAVMOD has a user extension configured with five new commands. The first two invoke a second digital signal mixer which is used to combine two signals together with selectable level ratios. The second pair of commands form a versatile digital filter system to meet specialized requirements not available with analog equipment, and the last command performs infinite peak clipping. The WAVMOD menu is listed in Table 1.

Insert Table 1 about here

WAVMOD Hardware Requirements

The basic version of WAVMOD (not including the user extensions) can operate with an RK05 disk drive, EIS arithmetic and 28k of memory with the DEC RT-11 SJ operating system. The SRL system configuration that supports the full version of WAVMOD is considerably larger. The disc drive used for the digitized file is an 80mbyte CDC 9762 unit. A VRM-11 video monitor and a VT-100 terminal provide the user interface facility. In addition, an FP-11A floating point hardware unit and 80k of extended memory support the advanced signal processing features. (see Forshee, 1979)

WAVMOD File Commands

The OW command invokes the disc file management system. Stimulus file names can be any six character name, but the device name and the extension are added by WAVMOD. These are (DK:) and (.STM) respectively. Stimulus files are composed of a header block and one or more blocks of digitized waveform. The header block contains the name, date and time of creation, a one line user description, and the length of the file in minutes and seconds. There are two basic modes of file handling. The first is an input copy to an output file. The second is a single file mode, where the file is quite long and more efficiently copied in the monitor. In addition to the file open command, there are two associated commands. The first is the abort command, AB, which terminates all WAVMOD activity and returns control immediately to the monitor. The second command, EX, completes the transfer of the remaining portion of the SOURCE file to the RESULTANT file and exits WAVMOD. The copy process in the exit command can represent a substantial period of time for a long file. The AB command skips this final transfer and deletes the RESULTANT file to avoid consuming excess time.

Table 1

WAVMOD COMMANDS

AB	ABORT AND RETURN TO MONITOR IMMEDIATELY
EC	EXPERIMENT CONTROL FLAG ENTRY
EN	ENVELOPE SHAPED NOISE
EX	EXIT (CLOSE FILES AND RETURN TO MONITOR)
GL	GLOTTAL MODULATION
LT	LISTEN TO RESULTANT
LV	SEGMENT LEVEL ADJUST
ML	MEASURE AND RECORD SEGMENT LEVEL
OW	OPEN NEW FILES (CLOSES OLD FILES)
QF	QUERY FILE NAMES AND TIME POINTERS
SF	SEGMENT FILE CREATION
SN	SIGNAL TO NOISE MIX
TO	TONE REPLACEMENT
US	USER EXTENSION
VL	VIEW CURRENT SESSION LOG
WN	WHITE NOISE REPLACEMENT

The scope of WAVMOD operations within a stimulus file is controlled with a SEGMENT list. The SEGMENT files command, SF, is used to create, SEGMENT lists, and to save them on disc as SEGMENT files. The SEGMENT file has the same six letter name as the stimulus file, but uses the .WSF extension. Segments are identified by number from one to a maximum of sixty four. The SF command gives the user considerable flexibility in creating, modifying and saving SEGMENT lists. Table 2 provides a list of the SF command options.

Insert Table 2 about here

Segment control within the commands permits the user to establish one of three modes. Option 1 starts the command operation at the current file pointer time, and stops at the end of the file. This option is selected when modifying a complete stimulus file. Option 2 allows a list of segment times to be entered from the keyboard to define one or more segments for the command. Option 3 copies the SEGMENT file list established by the SF command into the segment control buffer. Two additional options are available to view the SEGMENT list, and to abort the command in the case of an error.

Level Control

Waveform energy may be adjusted with command LV and in replacement modification commands. The level of the segment may be a level relative to the input waveform, or an absolute level. The user is asked to establish level control at the start of each segment in a command. When segment control options 2 or 3 are in effect, two or more segments will be modified in a WAVMOD command. Group mode duplicates the user's responses from the first segment of the group for all subsequent segments within a command. Group control provides automation to the WAVMOD system, substantially reducing time in generating large numbers of new stimuli. All level control options are valid in the group mode.

Two options are available for replacement modifications. The equivalent RMS level adjustment, E, provides a means of control with reference to the level of the input waveform. A measurement of the input segment is added to an offset specified by the user. The resulting value then controls the replacement signal level. Peak level, P, is used to set an absolute peak level for replacement signals. Since replacement signals have a defined ratio of peak to RMS (tone RMS is 3 dB below peak, and white noise RMS is 6 dB below the peak), the absolute RMS can also be set with the P option.

Level adjustment of the SOURCE can be controlled in three ways. Control relative to the RMS level of the input signal, R, allows a new level to be specified as an offset. Peakclips are calculated prior to modification and the user is given a warning message. The offset may then be changed to provide the proper maximum. Scaling, S, applies a direct ratio to the SOURCE to produce a new level. The ratio can be considered in terms of either peak or RMS since for

Table 2

The SEGMENT File Command Options.

OPTION	DESCRIPTION
B	Modify the current segment file list by adding an offset to the beginning time of all segments.
C	Change the begin and end times of a segment. The range of the new values starts at the previous segment end time, and extends to the following segment begin time.
D	Display the contents of the current SEGMENT file on the VR monitor screen.
E	Modify the current SEGMENT file list by subtracting an offset from the end time of all segments.
I	Insert a segment after the segment specified.
K	Keyboard entry of a SEGMENT list is the first step in creating a SEGMENT file. Up to 64 nonoverlapping sequential SEGMENT time pairs may be entered.
L	Load the SEGMENT file from a disc SEGMENT file.
P	Print (record) the current SEGMENT list in the WAVMOD log.
R	Remove the specified segment from the list.
S	Save the current SEGMENT file list as disc file.
X	Exit the SF command and return to WAVMOD command entry.

scaling, the results are the same. The scaling control is useful for reduction of a signal when the absolute value or the result is not of concern. Scaling provides a warning for peak clips prior to segment modification. The target level control adjusts the signal to an absolute RMS value. The dB value provided by the user and the measured RMS level of the SOURCE are used to calculate an offset ratio. This ratio is then used to adjust the SOURCE level to the target level. As in the relative control, the user is given a warning about clipping prior to the modification.

In addition to the level control options, there are two other functions of the level control system. The first is the ramp control option K. Attack and decay rate (ramps) of a signal are important in suppressing onset and offset clicks. The tone replacement operation enables the ramping option automatically. The default values for the ramp durations have been set to 20 msec (determined with listening tests on the SRL headphone system). The last function is the measurement option, M. This option measures three aspects of the segment signal level. These are the RMS level, the peak voltage, and the peak voltage expressed in dB. The RMS level is a good indication of relative loudness, while the peak value must be considered in order to avoid clipping distortion. The SRL system uses 12 bit A/D and D/A converters which have a maximum peak level of 10.24 volts (66.2 dB above .005 volts).

WAVMOD Modification Commands

WAVMOD signal modifications are the heart of the system. There are six operations in the core version of the program. Level adjustment of the input waveform may be performed with the level control system. Additionally, attack and decay ramps can be imposed on the waveform with this command.

Tone replacement generates a pure sinewave tone which may be used to replace the signal in the specified segment. The frequency range for the tone may be from 50 to 4800 Hz. Envelope shaped noise follows a procedure developed by O'Malley and Peterson (O'Malley, 1966). This procedure preservesthe instantaneous envelope and exact RMS energy of the waveform while obliterating formant and related spectral cues. The SRL digital version is implemented with a large normally distributed pseudo random table that provides a number sequence. Values from the sequence are compared to a switch point value which, when exceeded, reverses the signal's polarity. The switch point is set to a level that produces a mean reversal rate of 3000 times per second.

In the Signal-to-Noise mix operation, the SOURCE signal is mixed with a background signal of Tone, Envelope noise, or White Noise. White noise replacement replaces the SOURCE waveform with White Gaussian noise. The second standard deviation of the noise (Z=2) is used as a reference for the level control by arbitrarily making it the peak voltage reference. The average peak level will be 6dB above the RMS, but 4% of the peaks will clip.

The function of the glottal modulation command is to impose a triangular envelope on time varying sinusoids. The effect of this manipulation is to give sidebands to the component tones of the SOURCE. Spectrograms of resulting

waveforms contain patterns similar to those of speech formants. In addition, vertical striations of the glottal pulses are present in wideband spectrograms.

Utility Commands

Several utility commands are available to give the user access to various types of information. SOURCE file level can be measured with the ML command. This is useful for verification of stimulus levels. A listening test command allows the user to review the results of a modification with the audio system. QF will display the current file time pointer, and the file names of the SOURCE and RESULTANT, and VL provides a display of the LOG file on the video monitor.

The experiment control command inserts control flags into the stimulus file. The digitized audio data requires 12 bits of the 16 bits in each PDP-11 data word. The remaining four bits will be utilized as real-time flags to signal one of 15 possible conditions. These can be hardware controls, or program flags. The flags will be accurate to .1 msec for precise sychronization to the waveform. (The DAC hardware to intercept these flags will be implemented very shortly.)

The User Command system

The US command invokes the user extension system. This is a system of program expansion to enable the addition of new modifications in a structured manner. USER commands comprising three special modification systems are presently available in WAVMOD. The most complex is the digital filtering command. The filter is of the direct form type, and can implement 62 poles and 63 zeros. The filter coefficients are obtained from a common array that is loaded with the LC command. Filters are designed using routines from the IEEE signal processing package. (The Digital Signal Processing Committee of the IEEE, 1979). The load coefficient command can input a filter parameter file or select one of several immediate mode standard filters. These standard filters are of a general purpose nature and may be used with little knowledge of the complexities of filter design.

The signal mixer combines the SOURCE file segments with the signal in the mix buffer in a sequential fashion. The buffer pointer may be cycled, so that the contents can be used repeatedly. The load mixer buffer command takes a signal from a SOURCE file that is to be a background signal and loads it into the mixer buffer. This buffer is 80k words of extended memory.

The infinite peak clipper amplifies the SOURCE signal to remove all level information. The zero crossings are preserved, and the peak level is 66 dB. The output can be obtained as either the clipped signal, or as a series of narrow spikes that indicate polarity and axis crossings.

Session Example

A WAVMOD session is presented to provide an example of the user environment. A 1000 Hz tone is to be inserted into a stimulus consisting of a 2000 Hz tone. A SEGMENT file will be created to save the segment times for later use in a perceptual experiment.

The first step is the RUN WAVMOD command in the RT-11 monitor. File option 1 is selected to open the files. The SOURCE name specified is TONE, and the RESULTANT name is TUTOR. An identifier line for the file header is also entered. WAVMOD now requests a command entry. For the first operation, the SF command is selected. The K option provides for the entry of a SEGMENT list. The segments are: (.1,.15) (.2,.25). The segfile list is now displayed on the video screen. The S option saves the file on disc, and the X option returns control to the WAVMOD command entry. Figure 1 shows the LOG record that was generated by this sequence of WAVMOD commands.

The TO command is used to insert a tone into the the file. Segment control option 3 is used to specify the SEGMENT list that was created in the SF command. A frequency value of 1000 Hz is entered for the tone. The group level control mode is used, and option P selects peak level. The default peak dB level is used. WAVMOD begins modification with segment time messages on the terminal screen to provide an indication of progress. If any peakclips occur, the user is informed of those immediately. When tone is finished, the the EX command will end the WAVMOD session. Figure 2 is a WAVES display of the first segment of the waveform after modification. Note the ramps at the start and end of the inserted tone.

Insert Figures 1 and 2 about here

Summary

WAVMOD has been in use for over one year in our laboratory. It has been successful in meeting it's initial goals of signal modification, and has demonstrated flexibility in supporting expansion in a variety of directions. The modular design of WAVMOD will be able to support future signal processing needs with various additions of custom FORTRAN modules to the current program version.

```
WAVMOD SESSION LOG      20-MAY-82 @ 14:37:17      PAGE  1     (NL=  22)
>TONE.STM    (SOURCE) WAS CREATED WITH WAVMOD ON 28-APR-82 @ 18:18:34
TEST TONE OF 2000 HZ AT 60DB RMS.
>TUTOR.STM  (RESULTANT)                 DURATION=  0.3216 SECONDS
REPLACE SEGMENTS OF THE SOURCE WITH TONES.

#SEGMENT FILE ENTERED
   1=  0.1000,  0.1500   2=  0.2000,  0.2500
#SAVE SEGMENT FILE
  SEGMENT FILE: TUTOR.WSF            CREATED ON: 20-MAY-82
2 SEGS FOR TUTOR
  1=  0.1000,  0.1500   2=  0.2000,  0.2500

#TONE REPLACEMENT
 0.1000,  0.1500  53.00 DB RMS(P @ 56.00 DB) AT= 20.0MS, DT= 20.0MS
                   0. PEAK CLIPS             FREQUENCY=1000.HZ
 0.2000,  0.2500  53.00 DB RMS(P @ 56.00 DB) AT= 20.0MS, DT= 20.0MS
                   0. PEAK CLIPS            FREQUENCY=1000.HZ
** SOURCE FILE TONE.STM AND RESULTANT FILE TUTOR.STM ARE CLOSED **
                  ------------------------------
```

Figure 1. WAVMOD session log for the tutorial

Notes:
1. > designates stimulus file names.
2. # marks the WAVMOD operation to follow.
3. 0.1000, 0.1500 are the first segment times
4. 53 dB is the level for the segment.
5. P@ 56 dB is the level given in the Peak option.
6. AT, DT are the ramp times.

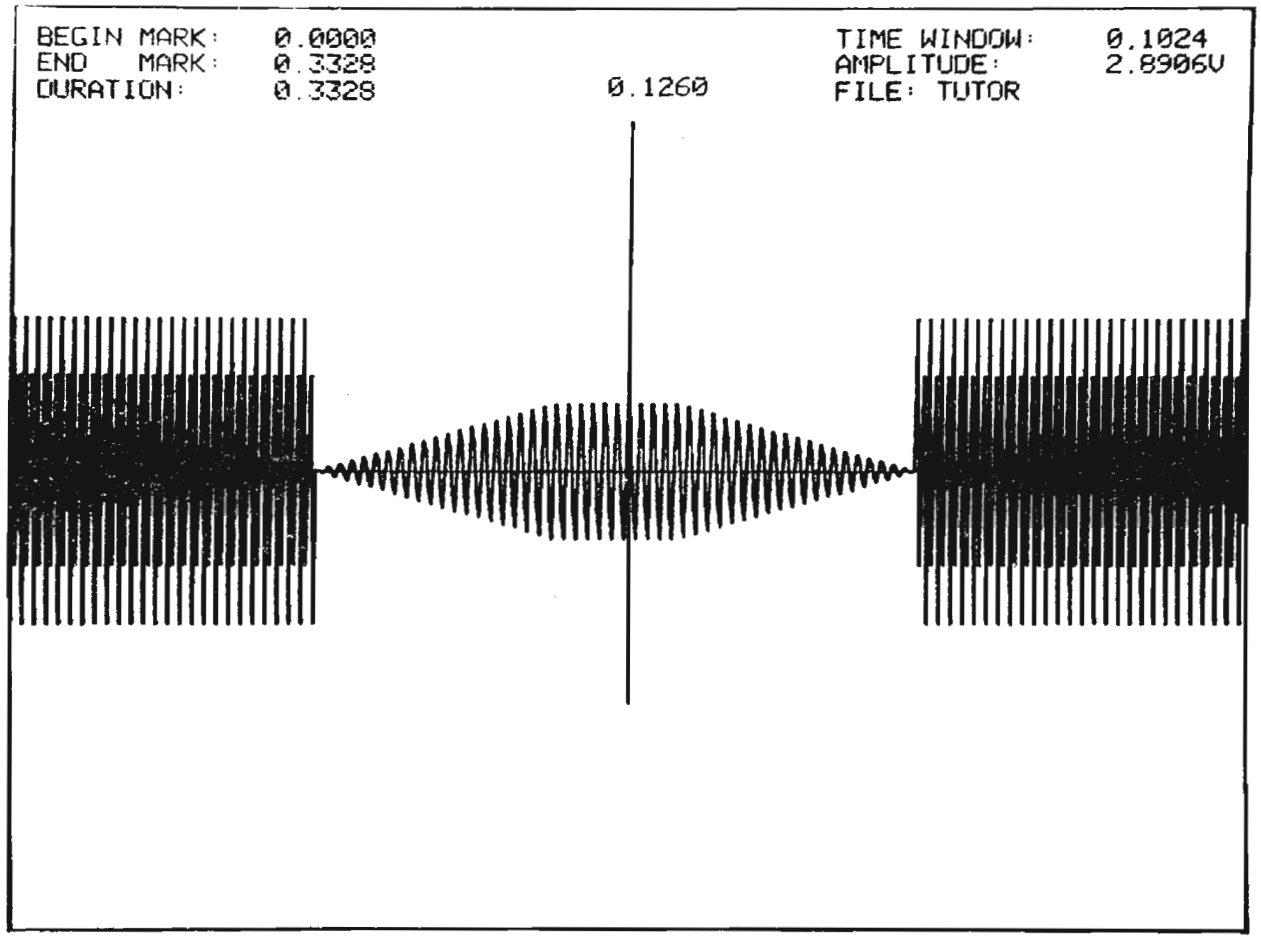

BEGIN MARK: 0.0000
END MARK: 0.3328
DURATION: 0.3328
0.1260
TIME WINDOW: 0.1024
AMPLITUDE: 2.8906V
FILE: TUTOR

Figure 2

References

The Digital Signal processing committee of the IEEE
Programs for Digital Signal Processing. New York: Institute of Electrical and Electronics Engineers, 1979.

Forshee, J. Computer facilities in the Speech Perception Laboratory. Research on Speech Perception: Progress Report No. 5, 1979, 449-473.

Luce, P. and Carrell, T. Creating and editing waveforms using WAVES. Research on Speech Perception: Progress Report No. 7, 1982.

O'Malley, M. and Peterson, G. An Experimental Method for Prosodic Analysis. Phonetica, 1966, 15, 1-13.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 7 (1981) Indiana University]

Creating and Editing Waveforms Using WAVES*

Paul A. Luce and Thomas D. Carrell

Speech Research Laboratory
Psychology Department
Indiana University
Bloomington, IN 47405

*This work was supported by NIH Research Grant NS-12179 and NIMH Research Grant MH-24027 to Indiana University in Bloomington. We thank Jerry C. Forshee for his help and assistance on various phases of this work.

Introduction

WAVES is an interactive program for creating and editing waveforms on the PDP-11/34 computer in the Speech Research Laboratory. The version of WAVES currently implemented on our system was adapted by Tim Smith from an earlier version of the program written for the PDP-11/40 computer in the Psychoacoustics Research Laboratory at Northwestern University (Smith, 1981). WAVES was designed to provide users with a powerful and flexible means of measuring and editing digitized waveforms through the use of interactive computer graphics. Since its implementation, the WAVES program has proved to be an invaluable research tool in our laboratory.

WAVES was written in Macro-11 and runs as a background job under the RT-11 foreground/background monitor. In addition to the PDP-11/34 computer, the hardware used by WAVES includes an ADDS Regent 100 console terminal, two 80 megabyte CDC disks, a DEC VT-11 graphics display system, 12-bit analog-to-digital and 12-bit digital-to-analog converters, and floating-point processor hardware. The two 80 megabyte disks are used for creating and temporarily storing waveform files. In principle, at a sampling rate of 10 kHz a maximum of 133 minutes of digitized waveforms can be stored on each disk, although in actual practice some disk space is used for storage of the operating system and other system programs, including WAVES. Hardcopies of the graphics displays created by WAVES using the VT-11 display system can be obtained by copying the display to a Tektronix 4010 terminal and using a Tektronix 4631 hardcopy unit (see Forshee, 1979).

WAVES Command Modes

The WAVES program operates under two command modes: an edit mode and an immediate mode. Each command in the edit mode consists of two characters, the first of which is always an E (for "Edit"). The edit mode commands are given in Table 1.

Insert Table 1 about here

The edit mode commands are primarily used for inputting and outputting waveform files; that is, for sampling from the analog-to-digital converter and for playing waveform files over the digital-to-analog converter. These commands are also used for establishing and deleting begin and end marks within a waveform file, making measurements, and performing various kinds of utility functions.

Immediate mode commands are single characters used to control the display of the waveform file on the screen of the VT-11 display system. WAVES immediate mode commands are given in Table 2.

Table 1.

WAVES Edit Mode Commands
(Adapted from Smith, 1981)

Command	Action
EA	deletes begin marker
EB	marks begin point of edit segment
EC	closes output file
ED	deletes begin and end markers
EE	marks end point of edit segment
EH	dislpays immediate mode keypad configuration (help command)
EI	copies VT-11 display to the Tektronix 4010
EJ	jumps to any point in the waveform file
EK	kills output file
EM	types the menu of WAVES options on the terminal
EN	deletes end marker
EO	writes edit segment to output file
EP	reproduces waveform over the D/A converter
ER	establishes input file
ES	samples from the A/D converter
ET	changes sample rate
EW	establishes output file
EX	exits from WAVES
EZ	writes specified number of zeros to output file

Insert Table 2 about here

Sampling and Reading Waveform Files

Upon entering the WAVES program, the user must give one of two edit mode commands: ES or ER. The ES ("Edit Sample") command causes WAVES to sample a waveform from the 12-bit analog-to-digital converter. After issuing the ES command, WAVES prompts the user for a file name and for the length in seconds of the waveform to be digitized. Waveforms may be digitized from either a previously recorded audio tape or on-line from a microphone at a remote user's station. After sampling, WAVES asks the user if he/she wishes to open the waveform file. If the user responds YES, the waveform is displayed on the VT-11 display screen. If the user responds NO, WAVES is ready to digitize or edit another waveform or to return to the monitor.

The second edit mode command that may be issued upon entering WAVES is ER ("Edit Read"). This command allows the user to open an already digitized waveform file. If the file exists, it will be displayed on the VT-11. If the file is not found, an appropriate error message is given to the user.

Waveform Display and Immediate Mode Commands

An example of a waveform display is shown in Figure 1.

Insert Figure 1 about here

The display shown in Figure 1 covers an effective 1024 points of the digital waveform, although only every other point is displayed in order to reduce flicker. Time is represented on the horizontal axis and amplitude in volts is represented on the vertical axis. Because this waveform was digitized at a sampling rate of 10 kHz, the time window of this display is 102.4 msec, as indicated in the upper right-hand corner of the display. Also shown in the upper right-hand corner is the amplitude of the waveform at the center-line cursor (the vertical line in the center of the display) and the name of the waveform file. In the left-hand corner of the display, the locations in seconds of the begin and end marks are shown. The duration of the waveform segment between these two marks is also shown. In this display, no begin and end marks have been designated, so the location of these marks has defaulted to the beginning and ending of the waveform file.

Table 2.

WAVES Immediate Mode Commands
(Adapted from Smith, 1981)

Key Label	Action
STOP SCROLL	stops scrolling of waveform display on VT-11
SCROLL <	scrolls waveform display toward beginning of file
SCROLL >	scrolls waveform display toward end of file
POINT <	steps waveform display one point toward beginning of file
POINT >	steps waveform display one point toward end of file
VIEW MODE	changes size of display window on VT-11
READOUTS	toggles numeric readouts on VT-11 on and off
RESET TIME	resets time 0.0 sec to the point in the waveform file currently displayed at the center line cursor

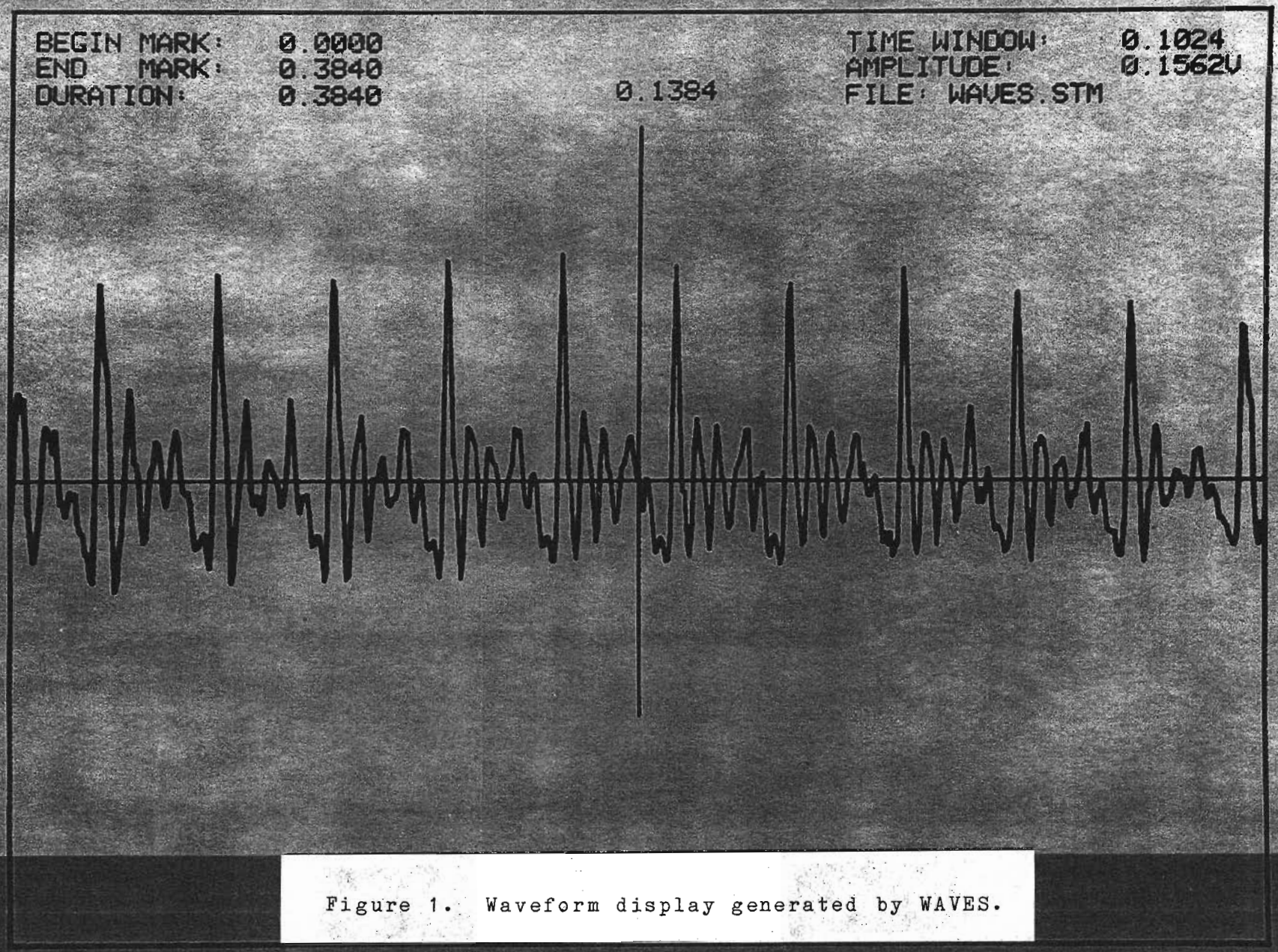


Figure 1. Waveform display generated by WAVES.

Once displayed, the user has a number of options in viewing the waveform using the immediate mode commands. The immediate mode commands are assigned to the 0 through 7 keys and thus can be conveniently invoked using the numeric keypad on the ADDS terminal. Return is not required after an immediate mode command. Striking the 1 key causes the waveform to scroll off to the left of the VT-11 screen toward the beginning of the waveform file. Striking the 2 key causes the waveform to scroll in the opposite direction. Repeatedly striking either of these keys causes the scroll rate to increase. When the display is scrolling in one direction, striking the key to scroll in the opposite direction causes the scroll rate to slow; with repeated striking, the direction of the scroll will reverse. Scrolling may be stopped at any time by striking the zero or "stop scroll" key. Keys 4 and 5 may be used to scroll the waveform display one point at a time for more precise control when editing waveforms.

Three additional immediate mode commands are "view mode", assigned to the 3 key, "reset time", assigned to the 6 key, and "read outs", assigned to the 7 key. The "view mode" command causes the normal waveform display of 1024 points to be reduced to 256 points, thus reducing the time window to 25.6 msec and thereby expanding the waveform display. Figure 2 shows an expanded portion of the waveform shown in Figure 1.

Insert Figure 2 about here

Striking the "view mode" key again causes the display to revert to the normal time window. The "reset time" command causes the zero time reference point--originally at the beginning of the file--to be set to the location of the center-line cursor, which can be defined by the user. Finally, the "read outs" command allows the user to erase the read outs from the top corners of the display. Striking the 7 key again redisplays the read outs.

Measuring and Editing a Waveform File

WAVES allows fast and precise measurement of segments within a waveform file. When an input file has been activated, the beginning of the segment to be measured can be located using the immediate mode commands. Once the center-line cursor is positioned exactly at the beginning of the segment to be measured, the EB ("Edit Begin") command can be used to establish the begin mark. The center-line cursor may then be moved to the end of the segment being measured and the EE ("Edit End") command may be used to establish an end mark. Upon establishing the end mark, the duration of the segment will be shown automatically in the upper left-hand corner of the display. Figure 3 shows a portion of a waveform with the begin and end marks established.

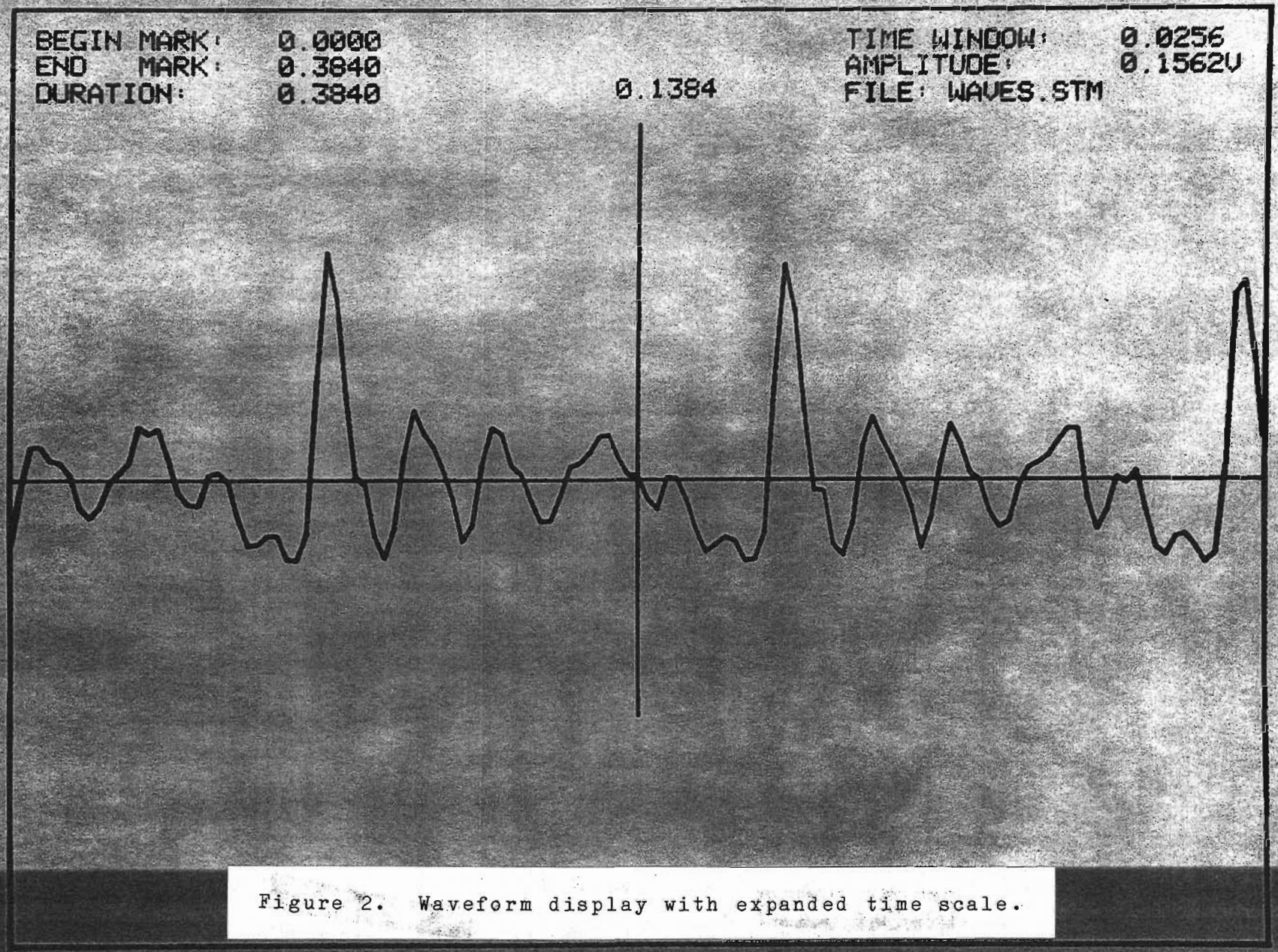


Figure 2. Waveform display with expanded time scale.

Insert Figure 3 about here

By employing the immediate mode commands to position the begin and end marks, editing of waveform files becomes a simple process that is conceptually similar to manual splicing of audio tape. Once the cursors are positioned at the desired points within the waveform file, the portion of the file between the begin and end marks can be written to an output file. Writing to the output file can continue until the file is closed, with each new segment being concatenated onto the end of the previously written segment or segments.

To illustrate a typical sequence of commands used in editing a waveform, consider the problem of "splicing off" the vowel /a/ from the consonant-vowel syllable /ba/ and adding to the /b/ an /i/ from the syllable /ki/. First, the input file containing the /ba/ is opened and the center-line cursor is located at the onset of the vowel /a/. The center-line cursor is then designated as the end mark using the EE command. (The begin mark, when unspecified, defaults to the beginning of the file. We will assume that the onset of the /b/ marks the beginning of this waveform file.) To assure that the desired segment has been located, the EP ("Edit Play") command can be given in order to listen to the segment between the begin and end marks. The EW ("Edit Write") command is now used to create an output file, at which time the file is named. The EO ("Edit Output") command is then used to write the portion of the input waveform file up to the end mark to the previously established output file. That is, the segment /b/ is written to the output file. Although only one output file may be open at a time, any number of input files may be activated. Thus, the waveform file containing /ki/ can now be opened and the center-line cursor placed at the onset of the vowel /i/. Because we are interested in writing the latter portion of the file--the segment /i/--to the output file, the center-line cursor is designated as the begin mark using the EB command. The EO command is then used to write the /i/ to the output file. The output file contains the waveforms from the /b/ in the /ba/ syllable and the /i/ in the /ki/ syllable. Because our editing session is complete, the EC ("Edit Close") command is used to close the output file.

This example illustrates the ease afforded by the WAVES program in digitally splicing and recombining waveform files. A variety of editing tasks can be quickly and almost effortlessly handled by WAVES, such as truncating waveform files, increasing or decreasing the duration of fricatives, vowels, bursts, and closures, and inserting periods of silence at any point in the file. This last task can be accomplished by using the EZ ("Edit Zero") command to write a specified number of zeros to the output file.

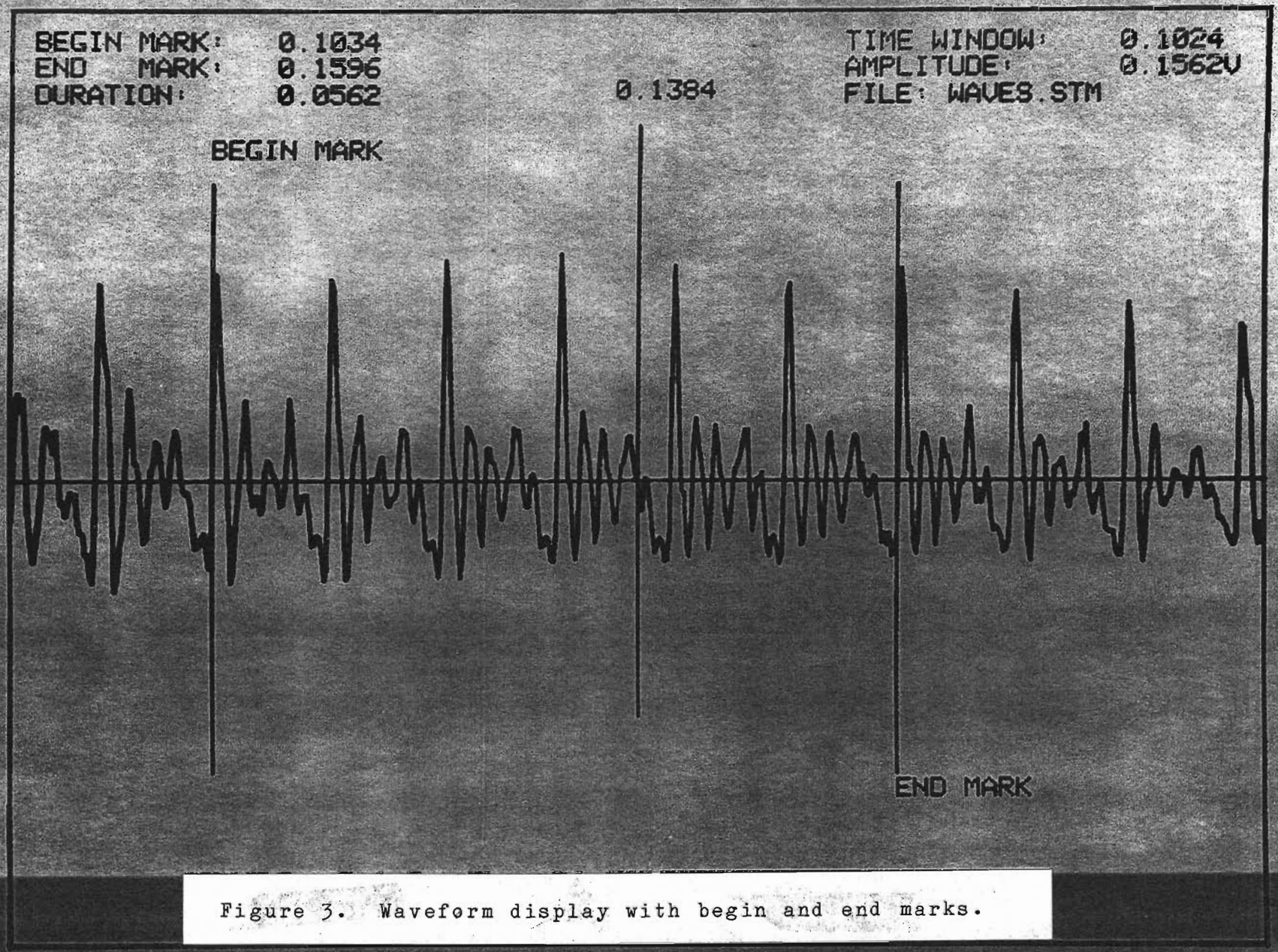


Figure 3. Waveform display with begin and end marks.

Summary

WAVES provides the user with a powerful and efficient means of digitally measuring and editing waveform files in a precise but conceptually simple manner. Through the use of the immediate mode commands to manipulate the waveform display, WAVES allows easy inspection of the waveform. In addition, WAVES provides a user-oriented means of digitally splicing and recombining waveform files to create new files to be used in perceptual research.

References

Forshee, J. Computer facilities in the Speech Perception Laboratory. Research in Speech Perception: Progress Report No. 5, 1979, 449-473.

Smith, T. WAVES: A program for creating and editing waveforms on the PDP-11 computer. Speech Research Laboratory Software Documentation, 1981.

IV. Publications

Pisoni, D. B. Variability of vowel formant frequencies and the Quantal Theory of Speech: A first report. Phonetica, 1980, 37, 285-305.

Carrell, T. D., Smith, L. B. and Pisoni, D. B. Some Perceptual Dependencies in Speeded Classification of Vowel Color and Pitch. Perception & Psychophysics, 1981, 29, (1), 1-10.

Remez, R. E., Rubin, P. E., Pisoni, D. B. and Carrell, T. D. Speech Perception without Traditional Speech Cues. Science, 1981, 212, No. 4497, 947-950.

Walley, A. C., Pisoni, D. B. and Aslin, R. N. The Role of Early Experience in the Development of Speech Perception. In R. N. Aslin, J. Alberts and M. R. Petersen (Eds.), The Development of Perception: Psychobiological Perspectives. New York: Academic Press, 1981, Pp. 219-255.

Blank, M. A., Pisoni, D. B. and McClasky, C. Effects of Target Monitoring on Comprehension of Fluent Speech. Perception & Psychophysics, 1981, 29, 4, 383-388.

Aslin, R. N., Pisoni, D. B., Hennessy, B. L. and Perey, A. J. Discrimination of Voice-onset Time by Human Infants: New Findings Concerning Phonetic Development. Child Development, 1981, 52, 1135-1145.

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

Technical Reports:

Kewley-Port, D. Representation of Spectral Change as Cues to Place of Articulation in Stop Consonants. Technical Report No. 3, December 18, 1980.

Manuscripts to be published:

Sinnott, J. M. and Pisoni, D. B. Pure Tone Thresholds in the Human Infant and Adult. Infant Behavior and Development, 1982 (In Press).

Grunke, M. E. and Pisoni, D. B. Some Experiments on Perceptual Learning of Mirror-Image Acoustic Patterns. Perception & Psychophysics, 1982 (In Press).

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 (In Press).

Brunner, H. and Pisoni, D. B. Some effects of perceptual load on spoken text comprehension. Journal of Verbal Learning and Verbal Behavior, 1982 (In Press).

Walley, A. C. and Pisoni, D. B. Review of J. Morton & J Marshall (Ed.), "Psycholinguistics." Journal of Communication Disorders, 1982 (In Press).

Green, B. G., Craig, J. C., Wilson, A. M., Pisoni, D. B. and Rhodes, R. P. Vibrotactile identification of vowel spectra. Journal of the Acoustical Society of America, 1982 (In Press).

Pisoni, D. B. Perception of Speech: The Human Listener as a Cognitive Interface. Speech Technology, 1982 (In Press).

Kewley-Port, D. Measurement of formant transitions in naturally produced stop consonant-vowel syllables. Journal of the Acoustical Society of America, 1982 (In Press).

V. Speech Research Laboratory Staff, Associated Faculty and Technical Personnel:

Research Personnel:

David B. Pisoni, Ph.D. -------- Professor of Psychology and Director
Richard N. Aslin, Ph.D. ------- Associate Professor of Psychology
Beth G. Greene, Ph.D. --------- Research Associate
Diane Kewley-Port, Ph.D. ------ Research Associate
Michael R. Petersen, Ph.D. ---- Assistant Professor of Psychology
Eileen E. Schwab, Ph.D. ------- Visiting Assistant Professor
Joan M. Sinnott, Ph.D. -------- Research Associate
Rebecca Treiman, Ph.D. -------- Assistant Professor of Psychology

Hans Brunner, 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