NIH Public Access
Author Manuscript
Hum Factors. Author manuscript; available in PMC 2012 December 04.

Published in final edited form as:
Hum Factors. 1983 February ; 25(1): 17–32.

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

PAUL A. LUCE[1], **TIMOTHY C. FEUSTEL**, and **DAVID B. PISONI**
Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana

Abstract

Three experiments were performed that compared recall for synthetic and natural lists of monosyllabic words. In the first experiment, presentation intervals of 1, 2, and 5 s per word were used. 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. In a second experiment, zero, three, and six digits were presented visually for retention prior to free recall of each spoken word list in a preload paradigm. Fewer subjects were able to correctly recall all of the digits for the six-digit list than the three-digit list when the following word lists were synthetic. The third experiment required ordered recall of lists of natural and synthetic words. 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. These 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.

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 unpracticed listeners, 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 research, we investigated a number of possible explanations for the difficulties typically observed in the perception and comprehension of synthetic speech.

Allen (1976) and Nickerson (1975) have suggested that prosodic differences between synthetic and natural speech present the major difficulty to 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

© 1983, The Human Factors Society, Inc. All rights reserved.

[1]Requests for reprints should be sent to Paul A. Luce, Speech Research Laboratory, Department of Psychology, Indiana University. Bloomington, IN 47405.

$watermark-text
$watermark-text
$watermark-text

$watermark-text

$watermark-text

$watermark-text

high-quality speech synthesis, it appears that, if perception and comprehension are to proceed normally, these natural variations must also be incorporated into the speech.

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 the initial acoustic-to-phonetic encoding of speech 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 does 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 representations 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 the encoding of speech perceived in noise produce subsequent difficulties in rehearsal processes in short-term memory and, therefore, the 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 than for natural speech (see Pisoni, 1982). 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, and 1981, 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 with 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 synthetic speech was about 6% lower than that for natural speech. For the anomalous sentences, however, recall performance for synthetic speech was about 19% lower than it was for natural speech.

In addition to the Pisoni and Hunnicutt findings, Jenkins and Franklin (1981), using both the VOTRAX and FOVE 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 the finding of Pisoni and Hunnicutt 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

$watermark-text

$watermark-text

$watermark-text

in anomalous sentences can only be recognized from a 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 ms slower than response times for natural words and nonwords. This study demonstrates that for isolated words, significant decrements in response latency can be shown for synthetic speech relative to natural speech even if the error rates are comparable.

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 synthetic speech is presented, When meaningful sentences or simple passages are used, intelligibility and comprehension appear, at first glance, to suffer relatively little when compared with 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 acoustic-phonetic signal itself. It is apparent, then, that there are several important differences in the perception and comprehension of synthetically generated speech. Our goal in the present paper has been to attempt to isolate the locus of these differences 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 was suggested by several earlier studies on the effects of noise on retention of spoken word lists. Two studies in the literature have shown that both perception and memory are affected insofar as both require or share limited processing capacity. In one study, 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. In another study, Rabbitt (1968), employing a similar paradigm, also found that recognition errors and capacity limitations in short-term memory were responsible for decreased recall of digits presented in noise. In a different 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 synthetic speech places 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 and 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 speech than for natural speech. To examine this problem, we selected free recall of lists of synthetic and natural words as the experimental task. This paradigm was chosen because the difficulty of a free recall task can be easily and reliably manipulated by a number of well-understood experimental variables.

$watermark-text

$watermark-text

$watermark-text

EXPERIMENT 1

In Experiment 1 the difficulty of the free recall task was manipulated by varying the presentation rate of the individual words in the lists. We predicted that, as the presentation rate increased, recall performance for the synthetic lists would decrease more rapidly than would recall performance for the natural lists. This outcome was expected 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—Seventy-two undergraduates were drawn from a paid subject pool. 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 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 et al., 1965). The lists were constructed so that the words on successive lists differed only by either the initial or final phoneme. Both natural and synthetic recordings of each of the lists were generated. 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 10 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 that was interfaced to matched and calibrated headphones. The words were presented at a comfortable listening level of 80 dB SPL against a background of wide-band Gaussian white noise at 50 dB SPL. Presentation of the stimuli was controlled in real time by a minicomputer.

Procedure—Twelve groups of six subjects were tested 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, and no subject heard a given natural or synthetic list more than once. 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 the natural lists was presented at three interword intervals: 1, 2, and 5 s. The order of the lists and the presentation rates were counterbalanced across groups according to a Latin square design.

At the beginning of the 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 procedure. The practice lists consisted of 10 words presented at a presentation interval of 2 s.

Immediately preceding the presentation of each list, the subjects heard a 500-ms 1000-Hz warning tone. Following presentation of the test list, another tone signaled the end of the list and the beginning of a 2-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 signaled the end of the recall period. There was a short break between the third and fourth lists in the session.

$watermark-text

$watermark-text

$watermark-text

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 interword interval. 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 presentation interval increased from 1 to 2 to 5 s. No interaction was observed between voice type (natural versus synthetic) and presentation rate, $F(1, 142) < 1.00$.

A similar pattern of results was observed when recall intrusions were scored. Any word in the subjects' response protocols that 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.

Again, there were significant main effects for synthetic versus natural voice, $F(1, 71) = 40.11$, $p < 0.01$, and presentation rate, $F(2, 142) = 4.01$, $p < 0.02$. The interaction of these two variables was not significant, $F(2, 142) = 1.75$, $p > 0.18$. Although a slight increase in intrusions is apparent at the 5-s interval for the synthetic lists, a Scheffé 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. This conclusion is supported by the finding that recall of the synthetic items was *not* differentially affected by increasing presentation rate. Thus, there was 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 word per second was not fast enough to reveal any encoding and/or rehearsal effects that 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, a second experiment was performed in which a secondary task was added to the primary recall task.

EXPERIMENT 2

Experiment 2 employed a memory preload 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 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 preload 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.

$watermark-text

$watermark-text

$watermark-text

Method

Subjects—The subjects were 120 undergraduates. 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 did 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 15 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. 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 2 s. The interval between presentation of the digits was 1 s. The presentation interval for the words was fixed at 2 s.

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 s. Counterbalancing was the same as in Experiment 1, with the digit preload manipulation substituted for presentation rate.

The subjects were instructed to remember the preload 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 10 words long and were preceded by a preload of three digits. As in the experimental lists, the presentation interval was set at 2 s.

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 preload condition. As in Experiment 1, the natural word lists were recalled better overall than were the synthetic word lists across all three preload conditions, $F(1, 119) = 106.93$, $p < 0.01$. The main effect of preload condition was also statistically significant, $F(2, 238) = 37.36$, $p < 0.01$, with mean word recall for both lists decreasing with increasing preload. No interaction between voice type and preload 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 preload condition. Again, intrusions were significantly greater for the synthetic than for the natural word lists, $F(1, 119) = 75.57$, $p < 0.01$. However, no main effect of preload condition was observed, $F(2, 238) = 2.59$, $p > 0.07$, nor was there an interaction between these two variables, $F(2, 238) < 1.00$.

$watermark-text

$watermark-text

$watermark-text

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

Digit recall—Three different analyses of the preload 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.

The interaction that we expected to find for word recall is clearly present for recall of the preload 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 preload items in short-term memory increased from three to six items, $z = 1.63$, $p < 0.05$, one-tailed.

Further analyses performed on the digit recall results are shown in Figure 6. In the upper panel, the average percentage of the digits recalled is plotted as a function of load condition without regard for the order in which the preload digits were recalled. In the lower panel of Figure 6, the digits were scored as correct only if they were recalled in the exact 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 preload, $F(1, 119) = 6.31$, $p < 0.02$, and $F(1, 119) = 163.03$, $p < 0.01$, respectively. The predicted interaction, although in the right direction, was not significant, $F(1, 119) = 2.47$, $p > 0.11$. Thus, the data reveal a trend for the digit recall to be poorer under the high preload condition for the synthetic relative to the natural word lists, although the effect is not statistically significant.

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

In summary, the analyses for the digit recall show a tendency for performance in the six-digit preload condition relative to the three-digit preload 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 and Rossman, 1965). Moreover, the greater effects of the synthetic words lists relative to the natural word lists appear to occur only under conditions in which memory stress is present. The obvious interpretation of these effects is that the subjects "borrow" from the limited 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 preload 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. Moreover, the result was just barely significant at $p \leq 0.05$. To obtain stronger evidence for the increased capacity demands of

synthetic speech, a third experiment was conducted 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 the 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. Informal analyses of the serial position curves from Experiments 1 and 2 did not reveal differential effects of voice type on the primacy and recency portions of the curves. However, because subjects in the first two experiments were not constrained as to the order of recall, they need not have encoded order information with the items. We reasoned that requiring ordered recall would lower the primacy portion of the serial position curve for synthetic speech as compared with natural speech because of the increased demands placed on the processing system when encoding of order information is required. This prediction was based on the hypothesis that increased demands on encoding and/or rehearsal processes arising from synthetic speech would cause fewer items presented early in the synthetic lists to be transferred to long-term memory (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 in comparison 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. They received credit for an introductory psychology course for their participation. All of the subjects met the same criteria for participation as did 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 10 words per list for the six lists. The order of the words within each list was again 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 10 words each were presented prior to the presentation of the six experimental lists. The words in both the practice and experimental lists were presented at a presentation interval of 2 s, and the recall periods were 90 s 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.

The serial position curves were obtained by scoring an item as correct only if it was recalled in the same position in which it was presented. The number of correct responses for each

$watermark-text $watermark-text $watermark-text

$watermark-text

$watermark-text

$watermark-text

serial position for each voice was then summed across all subjects, and an overall percentage-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 the presence of 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 as compared with 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 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 extensive 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 the earlier digit preload 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, arise from 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 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 observation that recall intrusions were, on the average, between one and two words more frequent for the synthetic than for the 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 recall of the preload items.

$watermark-text

$watermark-text

$watermark-text

One account of the failure to find differential effects of synthetic speech on word recall in the preload experiment (Experiment 2) comes from Rabbitt (1968). Rehearsal of the preload items that 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 the presence of 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 more 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" or degraded items by placing increased capacity 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 therefore 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, perceptual confusions, or poor synthesis of the distinctive cues. Second, and perhaps more relevant for our purposes, items that are poorly encoded may then 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 (1982) 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. Luce's 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 also Aaronson, 1976).

In another study, Pisoni and Koen (1982) have recently found differences in intelligibility of natural and synthetic words presented in noise at several different speech-to-noise ratios. Intelligibility of synthetic words is affected more by noise than are the same naturally produced words. Thus, the effects of noise produce a greater decrement in 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 more 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

$watermark-text

$watermark-text

$watermark-text

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 situations, particularly under 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.

In addition, our results demonstrate that the use of synthetic speech in various applied settings may place important constraints on rate of learning as well as the degree to which contextual materials may be comprehended rapidly and effectively. The use of synthetic speech in aids for the handicapped, such as a reading machine for the blind, must be considered in light of the acoustic-phonetic quality of the speech and its interaction with the cognitive processes involved in comprehension. Even with highly intelligible speech, such as that produced by the MITalk system, we were able to demonstrate that resources that would ordinarily be available for comprehension are depleted by the demands synthetic speech places on the perceptual system. The depletion of these processing resources may have important ramifications not only for aids for the handicapped, but also for computer-aided instruction using synthetic speech. Employing voice output devices using synthetic speech in educational settings may produce slower and less complete learning as well as increased frustration on the part of the student.

Our results suggest, furthermore, that the use of synthetic speech systems in noisy environments, such as in aircraft cockpits, must be carefully considered when messages crucial to the safety of the aircraft and its occupants are encoded in synthetic speech. In military applications, where aircraft personnel are required to monitor several sources of visual input simultaneously, and where noise levels may be extremely high, the demands synthetic speech places on attentional and memory mechanisms should be of great concern. The present findings demonstrate that even under ideal listening conditions, where the perceiver's cognitive load is relatively small, synthetic speech produces significant decrements in performance. We would expect that implementing voice output systems using synthetic speech in real-world applications would exaggerate the differences we observed in a controlled laboratory setting. In conclusion, then, we believe our findings have demonstrated that both consumers and vendors of products using synthetic speech output should become aware of the demands this type of speech places on the human perceiver and of the potential interactions among the listener, task demands, and speech quality.

Acknowledgments

This research was supported by NIH Research Grant NS-12179 to Indiana University in Bloomington. The authors would like to thank Dr. Howard C. Nusbaum for his comments and suggestions.

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. Linguistic-based algorithms offer practical text-to-speech systems. Speech Technology. 1981; 1:12–16.

Allen, J.; Hunnicutt, S.; Carlson, S.; Granstrom, B. MITalk-79: The 1979 MIT text-to-speech system. In: Wolf, JJ.; Klatt, DH., editors. Speech communication papers presented at the 97th meeting of the Acoustical Society of America. New York: Acoustical Society of America; 1979. p. 507-510.

$watermark-text

$watermark-text

$watermark-text

Baddeley, AD.; Hitch, G. Working memory. In: Bower, GH., editor. The psychology of learning and memory. Vol. 8. New York: Academic Press; 1974.

Dallett KM. Intelligibility and short-term memory in the repetition of digit strings. Journal of Speech and Hearing Research. 1964; 7:362–368. [PubMed: 14239015]

House AS, Williams CE, Hecker MHL, Kryter KD. Articulation-testing methods: Consonantal differentiation with a closed-response set. Journal of the Acoustical Society of America. 1965; 37:158–166. [PubMed: 14265103]

Jenkins, JJ.; Franklin, LD. Recall of passages of synthetic speech. Paper presented at the meeting of the Psychonomic Society; Philadelphia. November, 1981;

Klatt DH. Linguistic use of segmental duration in English: Acoustic perceptual evidence. Journal of the Acoustical Society of America. 1976; 59:1208–1221. [PubMed: 956516]

Klatt DH. Software for a cascade/parallel format synthesizer. Journal of the Acoustical Society of America. 1980; 67:971–995.

Luce PA. Comprehension of fluent synthetic speech produced by rule. Journal of the Acoustical Society of America. 1982; 71:UU11.

Nickerson RS. Characteristics of the speech of deaf persons. The Volta Review. 1975; 77:342–362.

Pisoni DB. Speeded classification of natural and synthetic speech in a lexical decision task. Journal of the Acoustical Society of America. 1981; 70:S98.

Pisoni DB. Perception of speech: The human listener as a cognitive interface. Speech Technology. 1982; 1:10–23.

Pisoni, DB.; Hunnicutt, S. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. 1980; IEEE International Conference Record on Acoustics, Speech and Signal Processing; New York: IEEE; 1980.

Pisoni DB, 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.

Posner MI, Rossman E. Effect of size and location of informational transforms upon short-term retention. Journal of Experimental Psychology. 1965; 67:496–505. [PubMed: 5833671]

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. [PubMed: 5683763]

Shiffrin, RM. Capacity limitations in information processing, attention, and memory. In: Estes, WK., editor. Handbook of learning and cognitive processes. Vol. 4. Hillsdale, NJ: Erlbaum; 1976.

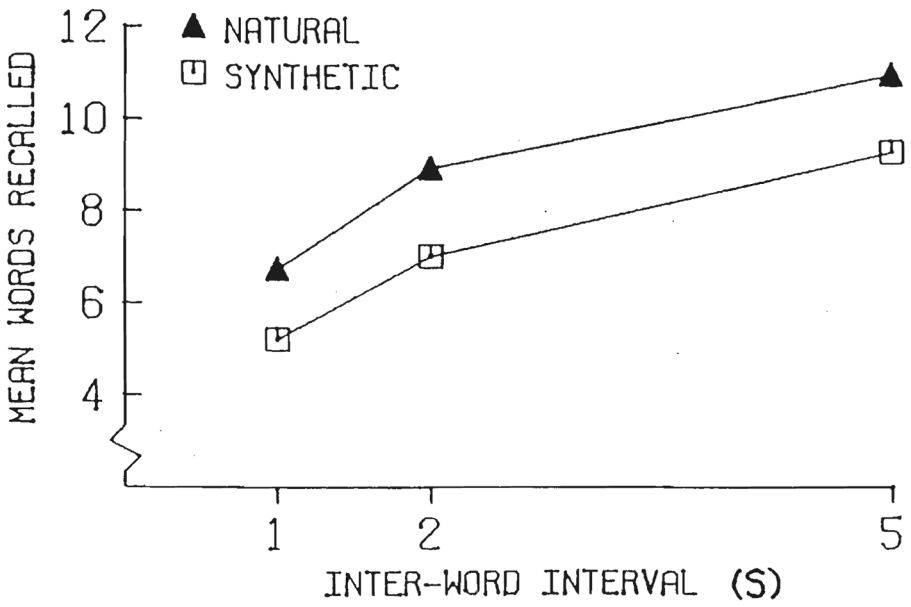

Figure 1.
Mean number of natural and synthetic words recalled as a function of interword interval.

$watermark-text
$watermark-text
$watermark-text

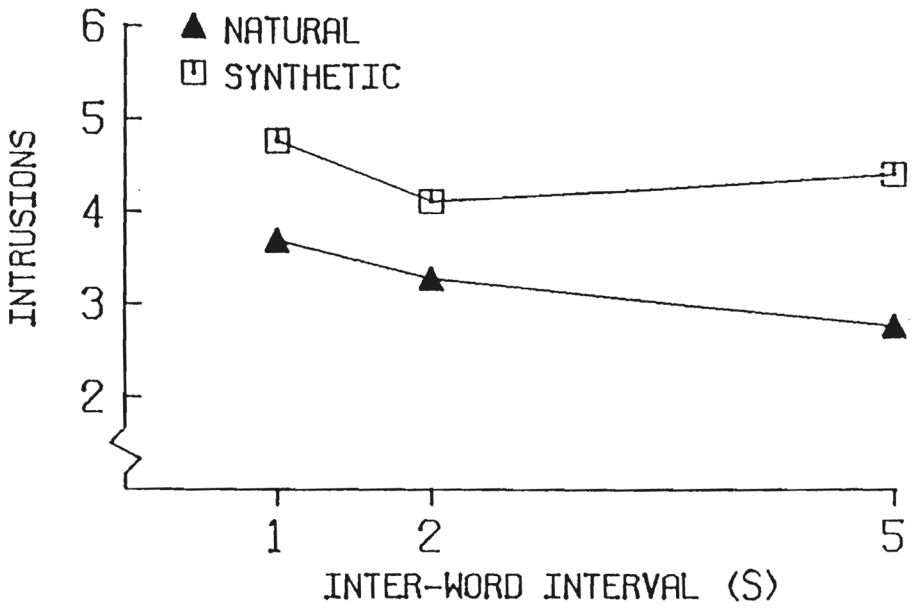

Figure 2.
Mean number of intrusions for both natural and synthetic word lists as a function of interword interval.

$watermark-text
$watermark-text
$watermark-text

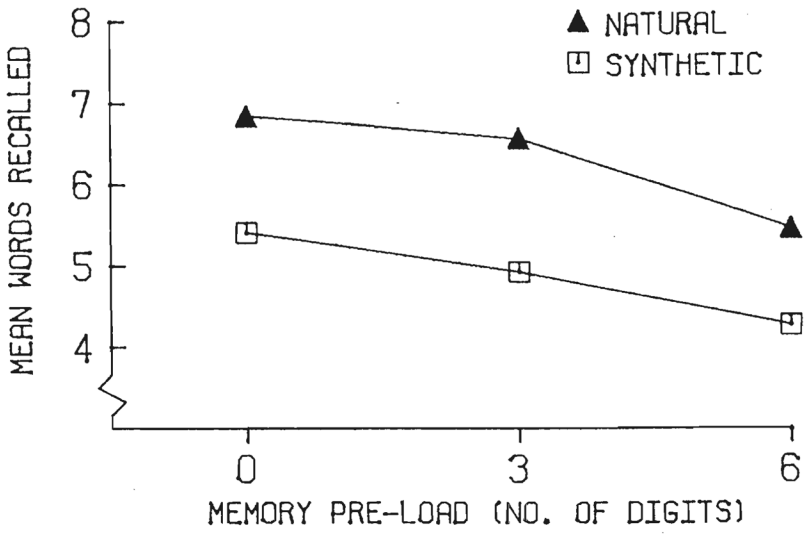

Figure 3.
Mean number of natural and synthetic words recalled as a function of memory preload.

$watermark-text
$watermark-text
$watermark-text

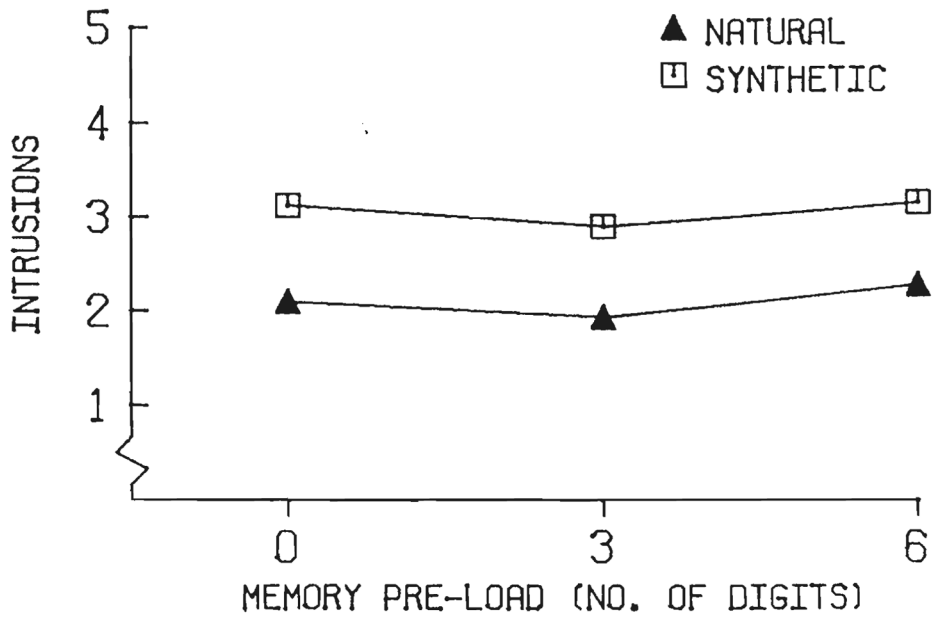

Figure 4.
Mean number of intrusions for both natural and synthetic word lists as a function of memory preload.

$watermark-text
$watermark-text
$watermark-text

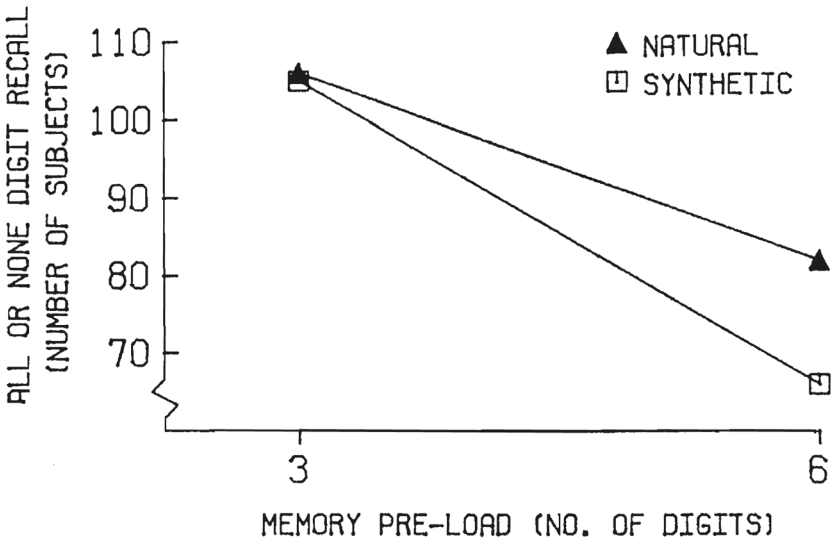

Figure 5.
Number of subjects correctly recalling all of the digits as a function of memory preload.

$watermark-text
$watermark-text
$watermark-text

$watermark-text
$watermark-text
$watermark-text

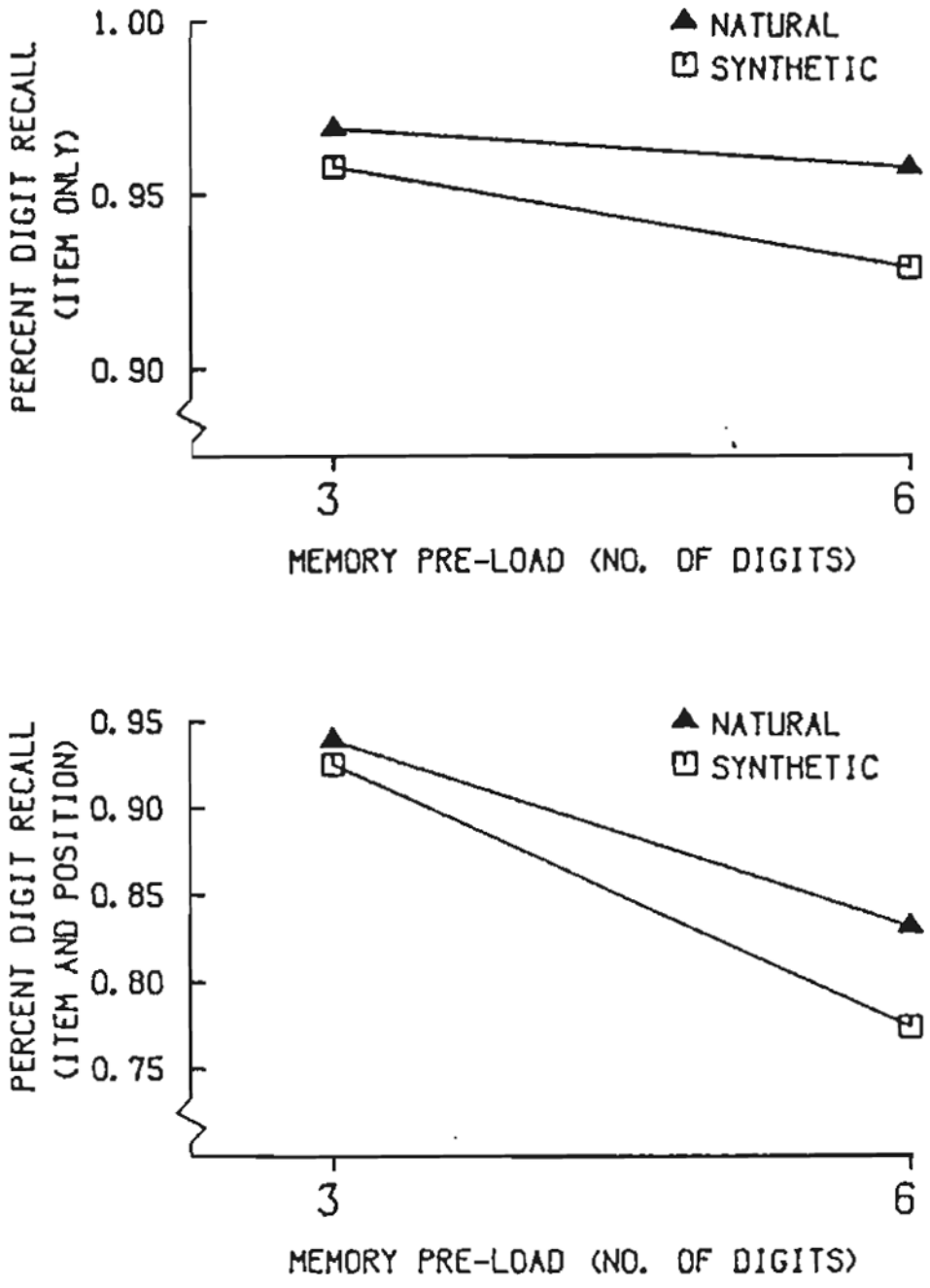

Figure 6.
Percentage of digits recalled as a function of memory preload. The upper panel shows percentage correct when the digits were scored without respect to the positions in which they were recalled. The lower panel shows the percentage correct for the digits scored according to position.

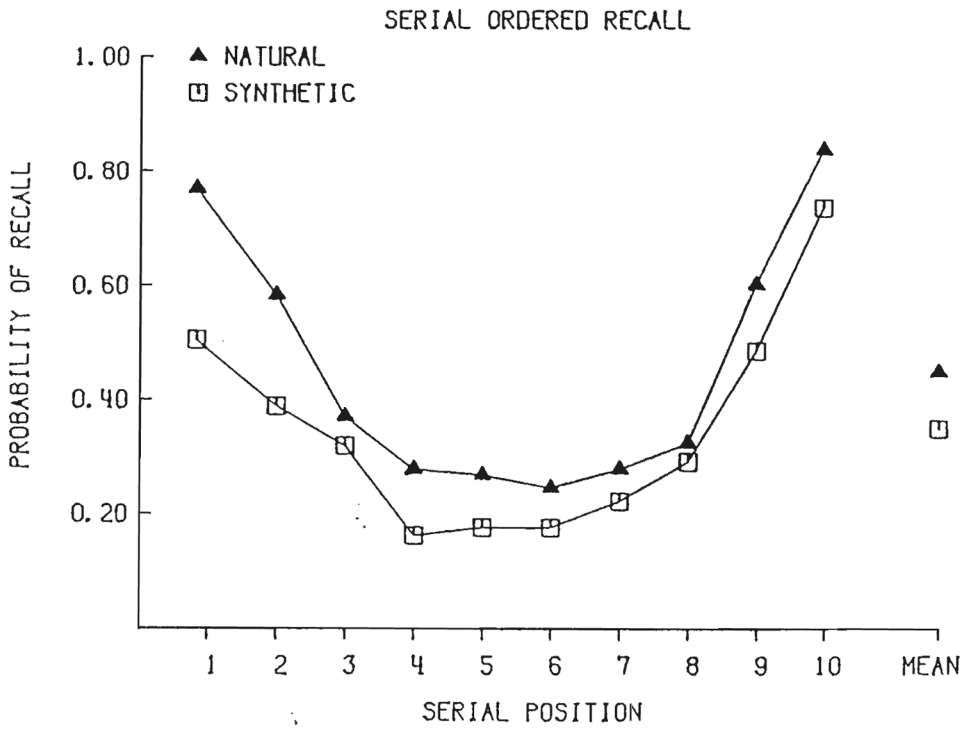

Figure 7.
Overall serial position curves for the natural and synthetic word lists.

$watermark-text
$watermark-text
$watermark-text