Published in final edited form as:

Hum Factors. 1991 August ; 33(4): 471–491.

Comprehension of Synthetic Speech Produced by Rule: Word Monitoring and Sentence-by-Sentence Listening Times

JAMES V. RALSTON, DAVID B. PISONI¹, SCOTT E. LIVELY, BETH G. GREENE, and JOHN W. MULLENNIX²

Indiana University, Bloomington, Indiana

Abstract

Previous comprehension studies using postperceptual memory tests have often reported negligible differences in performance between natural speech and several kinds of synthetic speech produced by rule, despite large differences in segmental intelligibility. The present experiments investigated the comprehension of natural and synthetic speech using two different on-line tasks: word monitoring and sentence-by-sentence listening. On-line task performance was slower and less accurate for passages of synthetic speech than for passages of natural speech. Recognition memory performance in both experiments was less accurate following passages of synthetic speech than of natural speech. Monitoring performance, sentence listening times, and recognition memory accuracy all showed moderate correlations with intelligibility scores obtained using the Modified Rhyme Test. The results suggest that poorer comprehension of passages of synthetic speech is attributable in part to the greater encoding demands of synthetic speech. In contrast to earlier studies, the present results demonstrate that on-line tasks can be used to measure differences in comprehension performance between natural and synthetic speech.

INTRODUCTION

The recent proliferation of voice output technology has extended its practical range of applications as a useful human-machine interface. Limited-vocabulary applications, such as warning and feedback systems, make only modest demands on computer storage capacity by using prestored utterances (McCauley, 1984). However, other important applications, such as prostheses for the language impaired, reading machines for the blind, computer-aided instruction, and information retrieval systems, all require the generation of a potentially unlimited vocabulary of unrestricted text. In these situations it is desirable to make use of speech synthesizers controlled by rule-based systems (McCauley, 1984). Devices that automatically convert text into fluent speech are referred to as *text-to-speech* (TTS) systems (Klatt, 1987). A wide variety of TTS systems are currently available on the commercial market. These systems range from inexpensive, hobbyist products that typically produce speech of poor quality to relatively expensive systems that produce highly intelligible and natural-sounding speech (see Logan, Greene, and Pisoni, 1989).

Because of the increasing use of voice response systems in many applications, it has become important to study how synthetic speech is perceived and understood by listeners (Pisoni, Nusbaum, and Greene, 1985). The bulk of research on this problem has been concerned primarily with assessing the intelligibility of isolated syllables or words (Hoover, Reichle, van Tasell, and Cole, 1987; Logan et al., 1989; Nye and Gaitenby, 1973). For example,

Logan et al. (1989) recently reported intelligibility scores obtained using the Modified Rhyme Test (MRT) with synthetic speech produced by eight different TTS devices and a natural speech control. They observed wide variation in error rates across the various types of speech, from 0.53% for natural speech to 35.56% for the Echo TTS system. Thus at the phoneme and word levels, the quality of speech synthesis varies over a wide range. Low intelligibility scores on the MRT for a particular synthesis device suggest the possibility that a listener has serious problems in understanding longer messages, particularly if the materials are novel and/or linguistically complex (Pisoni et al., 1985).

Although intelligibility tests provide important information about speech perception using simplified stimuli in highly constrained tasks, it is not likely that one will be able to accurately predict listening performance when the stimulus materials are longer passages of fluent, connected speech. Subjects in forced-choice intelligibility tests such as the MRT are not required to access the meanings of isolated words but only to recognize different sound patterns. In more natural communication situations, in which fluent speech is used, listeners must not only recognize words but also must derive meaning from successive strings of words in sentences. In this report we are interested in the extent to which comprehension measures can be predicted from intelligibility scores obtained with the MRT. Given this goal, it is useful to review some of the previous work on the comprehension of synthetic speech.

Several studies using passages of natural and synthetic speech have found either no differences in comprehension or differences that disappeared with moderate exposure or training (Jenkins and Franklin, 1981; McHugh, 1976; Pisoni and Hunnicutt, 1980). Pisoni and Hunnicutt (1980) compared the comprehension of passages produced by a human talker and by the MITalk-79 TTS system (Allen, Klatt, and Hunnicutt, 1987) with a reading control. After each passage was presented, subjects answered a series of multiple-choice questions. Comprehension accuracy increased from the first half to the second half of the test session for the passages of synthetic speech. Accuracy for the synthetic speech group was significantly worse than that of the reading control during the first half of testing but was equivalent to that of the reading and natural speech controls during the second half of testing.

In contrast, other comprehension studies have reported reliable differences in comprehension accuracy between natural and high-quality synthetic speech (Hersch and Tartaglia, 1983; Luce, 1981; Moody and Joost, 1987). In one study Luce (1981) used a recognition memory task to compare the comprehension of natural speech and synthetic speech produced by the MITalk system. Subjects judged the truth of short sentences with respect to information presented in the preceding passage. The sentences tested memory for information that ranged from the occurrence of specific words, to factual information not related to exact wording (*gist*), to information that was available only from semantic inferences based on understanding textual information. Recognition accuracy was uniformly higher for passages of natural speech. However, Luce also found an interesting crossover interaction between voice and sentence type. Recognition accuracy for words was higher for subjects listening to synthetic speech, but recognition accuracy for meaningful information was higher for subjects listening to natural speech.

Luce suggested that the interaction provided support for a resource-sharing model of spoken language comprehension (Kintsch and van Dijk, 1978; LaBerge and Samuels, 1974). According to this model, the various comprehension processes—from analyses of phonology to analyses of semantic information—share a limited pool of processing resources. Luce argued that subjects listening to synthetic speech allocate a greater proportion of their cognitive resources to analyzing the initial acoustic-phonetic structure of

the signal, leaving fewer resources available for comprehending the semantic content of the passages. He also proposed that this difference in encoding strategies could explain the observed differences in recognition memory performance for natural and synthetic speech. Other experiments have also found evidence suggesting that perceptual encoding of synthetic speech requires greater processing resources (Hersch and Tartaglia, 1983; Lee and Nusbaum, 1989; Luce, Feustel, and Pisoni, 1983).

Taken together, results of previous intelligibility and comprehension studies of synthetic speech provide a mixed picture. Studies of intelligibility have consistently demonstrated reliable differences in performance between natural and synthetic speech. In contrast, comprehension studies using passages of connected speech have produced conflicting evidence, some reporting sizable differences between natural and synthetic speech (Hersch and Tartaglia, 1983; Luce, 1981; Moody and Joost, 1987) and others reporting no differences at all (Jenkins and Franklin, 1981; McHugh, 1976; Pisoni and Hunnicutt, 1980; Schwab, Nusbaum, and Pisoni, 1985). How can these findings be accounted for in a principled way? We believe that the discrepancies observed across these various comprehension studies may be the result of the specific experimental methods used to measure comprehension and the selection of stimulus materials.

In the present study we investigated both of these issues. We hypothesized that previous research using long passages of connected speech may have used relatively crude and insensitive experimental techniques to measure comprehension. In particular, all of the previous studies that have reported negative results used *successive measures* (Levelt, 1978). That is, comprehension was assessed by measuring responses to questions presented *after* listeners heard the test passages. Although successive measures are useful because they examine the contents of memory and the products of the comprehension process, they are known to be highly vulnerable to subject strategies such as inferencing and reconstruction, which may selectively modify the memory representation for the stimulus materials (Bartlett, 1932; Kintsch and van Dijk, 1978; Levelt, 1978).

In contrast, *simultaneous* or *on-line* measures provide an experimental methodology that can be used to assess comprehension as it proceeds in real time (Levelt, 1978). Because on-line measures are taken during the process of comprehension, they are assumed to be less susceptible to postperceptual cognitive processes (Levelt, 1978). In the present set of experiments we used two on-line tasks—a word monitoring task and a novel sentence-by-sentence listening time task—combined with a postperceptual recognition memory task to examine the comprehension of passages of connected speech. Using these two measures, we hoped to address a number of unresolved issues surrounding the comprehension of passages of synthetic speech.

First, we attempted to determine whether more subtle differences can be detected using a simultaneous measure of comprehension, even if successive measures fail to reveal differences. Previous studies have demonstrated that the word monitoring task is sensitive to a number of linguistic variables that influence cognitive workload during comprehension (Cutler, 1976; Foss, 1969). Interpretations of monitoring data assume that both comprehension processes and the secondary monitoring task place competing demands on a common pool of processing resources. Increases in secondary-task latency or decreases in secondary-task accuracy are taken as evidence that the difficulty or processing load of the primary comprehension task has increased (Posner and Boies, 1971; Wickens, 1987).

Although the listening time task has been used only once before in an exploratory study (Mimmack, 1982), numerous experiments with an analogous self-paced reading task have revealed that it, too, is sensitive to a number of linguistic variables (Haberlandt, 1984).

However, instead of indexing cognitive workload, the reading time task assesses the speed of comprehension.

Second, we attempted to determine whether reliable differences in comprehension performance between natural and synthetic speech could be detected using a recognition memory verification test with more carefully controlled items. The test questions used in several of the earlier studies could have been answered without listening to the passages (Jenkins and Franklin, 1981; Moody and Joost, 1987). Therefore we prescreened all test sentences used in the verification task to eliminate items that could be answered without knowledge of the contents of the passage.

Third, we attempted to determine whether the increased demands associated with the recognition of synthetic speech interact with comprehension and/or memory load. To accomplish this, difficulty of comprehension was manipulated by employing two sets of passages that differed in text difficulty (see Britton, Westbrook, and Holdredge, 1978, for a similar manipulation). Memory load was manipulated in the word monitoring experiment by requiring subjects to remember and monitor for zero, two, or four target words.

Finally, we attempted to assess the relationship between intelligibility and comprehension measures. Manous, Pisoni, Dedina, and Nusbaum (1985) found a strong relationship between transcription accuracy and verification performance using isolated sentences in a sentence verification task ($r = -0.86$). However, the relationship between intelligibility and comprehension of fluent speech has never been studied, so far as we know. Because discourse processing relies heavily on linguistic knowledge, we would expect that the correlations between intelligibility and comprehension would be lower than those observed by Manous et al. (1985).

EXPERIMENT 1

Method

Subjects—Subjects were 112 volunteers enrolled in introductory psychology classes at Indiana University in Bloomington. They either received course credit or were paid \$4.00 for their participation. All subjects were native speakers of English and reported no history of a speech or hearing disorder at the time of testing.

Materials—A subset of 50 words from the MRT (House, Williams, Hecker, and Kryter, 1965) was selected to assess intelligibility. The MRT measures the intelligibility of syllable-initial and syllable-final consonants using a six-alternative forced-choice procedure.

Passages—Tables 1 and 2 list the passages used in the different text difficulty conditions. The passages listed in Table 1 were taken from Pauk's six-way paragraphs (Pauk, 1983). Each was designed for fourth-grade readers. The passages listed in Table 2 were selected from a variety of written comprehension tests designed for college-level readers. Tables 1 and 2 also include the results of analyses conducted on the written form of the texts, including standard readability indices. Results of the analyses indicate that the passages in Table 1 are more readable, or less difficult, than the passages in Table 2. The passages were written in narrative or expository styles and generally concerned scientific and cultural topics. We refer to the difference between these sets of passages as *text difficulty*, reflecting differences in their readability (Haberlandt, 1984; Kintsch and Vipond, 1979).

Word targets—Target words for each passage were selected on the basis of several criteria, including word-initial phoneme, number of syllables, and grammatical class. Because we used existing texts, we were unable to completely control the properties of the

target words. Target words were selected according to a set of priorities. The most common target words were monosyllabic nouns with word-initial stop consonants. However, if all words of this type had been selected from a passage, other words, such as monosyllabic nouns with word-initial fricatives or affricates, were used. The target words were balanced across different experimental conditions.

Recognition memory test—Eight sentences were created for each passage to assess comprehension. These sentences were presented visually on a CRT monitor immediately following the presentation of each passage. The sentences ranged from five to ten words in length and were stated in a declarative form. Four of the sentences (*word recognition* sentences) assessed listeners' memory for specific words that occurred in the passage, and four of the sentences (*proposition-recognition* sentences) assessed memory for propositions of the passage. Based on information in the preceding passage, half of all sentences were true and half were false. Word-recognition sentences all had the same frame, "The word XXX appeared in the passage." None of the target words in the word recognition sentences appeared as word-monitoring targets in the same passage. True proposition recognition sentences described relationships between concepts that occurred in the passage. False proposition recognition sentences contained concepts and relationships from the passage that were recombined in semantically meaningful ways.

A large set of sentences was initially generated and then tested in written form with a separate group of introductory psychology students. The pretest identified proposition recognition sentences that could be correctly answered without knowledge of the passage and proposition recognition sentences that could not be correctly answered even after reading the passage. The pretest also identified word recognition sentences for which verification accuracy was unacceptably low even after subjects had read the passages. In the pretest half of the subjects were given copies of the passages with the full set of verification sentences and the other half were given only proposition recognition sentences, without the passages. Proposition recognition sentences were discarded if accuracy was lower than 80% correct for both test groups of subjects or if accuracy was greater than 80% correct for both test groups. This left a set of sentences for which accuracy was relatively low (50–75% correct) for subjects who had not read the passages and relatively high (75–100%) for subjects who had read the passages. Word recognition sentences that were recognized with 85% or better accuracy after subjects had read the passages were selected from the pretest.

Recording and playback techniques—A male native speaker of American English with a midwestern accent read both the MRT list and the eleven passages. The materials were recorded in an IAC sound-attenuated booth with an Electro-Voice D054 microphone. An identical set of MRT words and passages was produced by a Votrax Type-N-Talk speech synthesizer controlled by a VAX 11/750 computer. No special stress or pronunciation correction was applied to the synthetic speech. Both the natural and the synthetic speech were recorded onto one channel of an audiotape at 15 in/s using an Ampex AG500 tape recorder.

All stimuli were then low-pass-filtered at 4.8 kHz, sampled, and digitized using a 12-bit A/D converter running at 10 kHz. With a digital waveform editing program, flags were set at the onset of each target word using the upper 4 bits of each 16-bit computer word (Luce and Carrell, 1981). All stimuli were equated for RMS amplitude. The speech stimuli were then recorded back onto one channel of an audiotape using a D/A converter. At this time the bit flags were used to trigger a tone burst generator that output a timing tone onto the second channel of the audiotape. Each audiotape contained a calibration vowel, 50 MRT words, and 11 comprehension passages. Target-set size and voice of the passages were between-subjects variables. Thus an experimental tape contained either natural or synthetic speech

and had zero, two, or four word targets. Speech signals were mixed with 55 dB broadband noise to mask tape hiss and ambient room noise and were presented to listeners binaurally over TDH-39 matched and calibrated headphones at 80 dB SPL.

Procedure—Subjects were run in groups of five or fewer in a quiet room. Each subject sat at a sound-treated booth equipped with a set of headphones, a video display monitor (GBC Standard CRT Model MV-10A), and a two-button response box that was connected to the computer. Each subject was given a booklet containing instructions and response forms for the MRT test, instructions for the comprehension task, and a postexperimental questionnaire that was designed to record subjective reactions to the experimental stimuli and tasks. After reading the MRT instructions, subjects were presented with a block of 50 trials. Subjects were required to circle one of six response alternatives in their answer booklets after each trial.

After completing the MRT, subjects were asked to read the instructions for the comprehension portion of the experiment. These instructions were also thoroughly explained by the research assistant. Subjects were told that they would be required to simultaneously detect all target words and comprehend the passage, and that they would be required to respond to a series of test sentences at the end of each passage. The first passage served as practice to acquaint the subjects with the experimental procedures and the quality of the voice used in each condition.

Each trial of the comprehension test consisted of three parts. In the first part zero, two, or four word targets were presented visually on the CRT display for a 30-s memorization period. When the memorization period ended, the target words disappeared from the screen and the passage was presented over headphones. Subjects were instructed to press a button on the response box each time they detected one of the target words in the memory set. At the end of each passage, eight sentences were presented visually on the CRT monitor for 4 s each. Subjects were instructed to determine whether the sentence was true or false and to record their responses by pressing one of two buttons on the response box while the sentences were presented. No feedback was provided to the subjects about the accuracy of their responses. Button position (left or right) was counterbalanced. Subjects received a 5-min break after the fifth passage. The comprehension portion of the experiment was controlled in real time by a PDP 11/34 minicomputer that presented the target words and test sentences and recorded subjects' responses. An experimental session lasted approximately one hour.

Results and Discussion

In the original experiment data from an eight target-set size monitoring condition were also collected. However, preliminary analysis indicated that the subjects in this condition were unable to monitor reliably for the target items and simultaneously comprehend the test passages. Thus this condition was eliminated from further analyses. Two criteria were used to screen the data of the remaining subjects. First, subjects' data were excluded from further analyses if they failed to make any word-monitoring responses on at least five of the ten test passages. Second, subjects' data were excluded if they failed to respond to at least 8 of the 80 comprehension sentences that were presented after the test passages. The data of 21 subjects failed to meet one or both of these criteria and were eliminated from further analysis. The absolute rate of rejection per condition was relatively low; the largest number of subjects was rejected in the Votrax/4 target condition (4.5% of all subjects). The final analyses (to be reported) were based on data provided by 91 subjects. A chi-square test confirmed that there was no differential rejection rate as a function of experimental condition, $\chi^2 = 0.013$, $p > 0.05$.

The results will be presented in three major sections corresponding to the three tasks: intelligibility, word monitoring, and recognition memory. These data were subjected to analyses of variance, treating voice (natural or Votrax) and target-set size (0, 2, or 4) as between-subjects variables and text difficulty (fourth-grade or college) and recognition sentence type (word or proposition recognition) as within-subjects variables. All post hoc statistical analyses were done using Newman-Keuls tests.

MRT—MRT accuracy refers to the percentage of correct responses out of 50 trials. Analysis of variance revealed a significant main effect of voice, $F(1,85) = 743.65$, $p < 0.001$. None of the other main effects or interactions were significant. Mean accuracy in the MRT was 94% correct for natural speech and 60% correct for synthetic speech. This overall level of performance is consistent with previously reported data based on more extensive testing with the full set of MRT words (see Logan et al., 1989).

Word monitoring—Mean accuracy and latency data were examined in separate analyses of variance. Monitoring hits were defined as those responses that occurred from 100 ms to 1200 ms after the onset of a target word. Responses outside these limits were presumed to reflect either anticipation or inattention by the subjects. Because the probability of false alarms was exceedingly low (the mean $p[FA]$ was less than 0.01), accuracy is presented in terms of the probability of a correct detection (i.e., a hit).

Accuracy—Word-monitoring accuracy data are displayed in Figure 1. Word-monitoring accuracy was higher for natural speech than for synthetic speech, $F(1,54) = 17.39$, $p < 0.001$. Monitoring accuracy was also higher for two targets than for four targets, $F(1,54) = 55.87$, $p < 0.001$. Although the interaction between text difficulty and target-set size was significant in the ANOVA, $F(1,54) = 4.32$, $p = 0.042$, subsequent post hoc analyses failed to reveal the precise locus of the significant difference between the means underlying the interaction.

Latency—Word-monitoring latencies are shown in Figure 2. Overall, monitoring responses were faster for target words in natural speech than in synthetic speech, $F(1,54) = 11.44$, $p = 0.001$. Responses were faster for the two-target word condition than for the four-target word condition, $F(1,54) = 8.76$, $p = 0.004$. Monitoring responses were also slower for the college-level passages than for the fourth-grade passages, $F(1,54) = 75.36$, $p < 0.001$. Finally, the interaction between voice and text difficulty was significant, $F(1,54) = 10.53$, $p = 0.002$. Post hoc analyses revealed that the increase in response latency from fourth grade to college level was greater for the passages of synthetic speech than for those of natural speech.

Recognition memory—Recognition accuracy data were analyzed with an ANOVA, treating voice and target-set size as between-subjects variables and treating sentence type and text difficulty as within-subjects variables. The accuracy data are displayed in Figure 3 as a function of target-set size. The panels on the left present data for word recognition sentences, and the panels on the right present data for proposition recognition sentences. The top panels present data for fourth-grade passages, and the bottom panels present data for college-level passages.

Subjects recognized sentences more accurately after listening to passages of natural speech than after listening to synthetic speech, $F(1,85) = 37.21$, $p < 0.001$. Accuracy was also higher for word recognition sentences than for proposition recognition sentences, $F(1,85) = 9.94$, $p = 0.002$. Apparently natural speech provides a more robust representation of information in the passages. This finding is consistent with the previous results of Luce (1981) and of Moody and Joost (1987), who reported that passages of natural speech were comprehended better than passages of synthetic speech.

The interaction between voice and sentence type was also highly significant, $F(1,85) = 10.95$, $p = 0.001$. Post hoc analyses revealed that although no difference was observed as a function of sentence type in the natural speech condition, accuracy was significantly lower for the proposition recognition sentences in the synthetic speech condition. This result is similar to findings reported by Luce (1981). In the present experiment, however, accuracy for word recognition items was not better for synthetic speech than for natural speech. Instead, we found that synthetic speech was uniformly poorer than natural speech in all conditions.

We suggest that the interaction between voice and sentence type may arise from the combination of two mechanisms. One possible explanation, the *misidentified word* hypothesis, suggests that phonological encoding influences comprehension in a fairly direct fashion. According to this hypothesis, the meaning of a passage of synthetic speech is comprehended poorly because the constituent words are identified incorrectly. A second explanation, proposed by Luce (1981), suggests that phonological encoding affects comprehension in a more indirect fashion. According to this hypothesis, listeners shift their attentional resources toward phonological encoding when intelligibility decreases. As a result, resources are drawn away from higher-level comprehension processes.

Overall, accuracy decreased with increasing target-set size, $F(2,85) = 3.19$, $p = 0.046$. Post hoc analysis revealed that accuracy for the four-target condition was significantly lower than was accuracy for the zero- and two-target conditions. However, the latter two conditions were not significantly different from each other. This result replicates earlier findings reported by Brunner and Pisoni (1982), who found that recognition performance after monitoring for a single word target was the same as verification performance after listening to a passage without monitoring for any word targets.

Recognition accuracy was also significantly lower for sentences following college-level passages than for those following fourth-grade-level passages, $F(1,85) = 4.18$, $p = 0.044$. In addition, the interaction between sentence type and target-set size was significant, $F(2,85) = 6.79$, $p = 0.002$. Post hoc analyses showed that word recognition accuracy did not vary as a function of target-set size, but that proposition recognition accuracy was significantly lower for the four-target condition than for any of the other conditions. This result suggests that comprehension and memory for the target words compete for the same limited pool of processing resources.

The three-way interaction between text difficulty, voice, and target-set size was significant, $F(2,85) = 3.13$, $p = 0.048$. Post hoc analyses revealed an unexpected effect: accuracy was greater for fourth-grade texts than for college-level texts for listeners in the natural speech condition who were not required to monitor for targets. However, no differences in performance between fourth-grade and college-level texts were observed for the other voice and target-size combinations.

Regression analyses—Regression analyses were carried out with intelligibility as the predictor variable and word monitoring and recognition accuracy data as predicted variables. Data were collapsed across the voice, text difficulty, and target-set size variables. The results are presented in Table 3. The left column displays the correlation coefficients (Pearson's r), and the right column displays the probability values associated with F -statistics that were used to test whether the observed coefficients were different from zero.

The magnitudes of the correlation coefficients, though significant, are moderate in their predictive power. For example, the correlation between speech intelligibility as measured by the MRT and proposition recognition accuracy is +0.47, indicating that variation in

intelligibility accounts for only 21% of the observed variation in the proposition recognition accuracy scores. Although direct statistical comparisons are not possible, these correlations are, on the whole, smaller than those reported by Manous et al. (1985). Because higher-level cognitive contributions are very important for the comprehension of complex linguistic materials, such as passages of connected speech, recognition memory can be predicted only moderately well from intelligibility scores alone.

Summary

The major findings from this experiment can be summarized in terms of the results obtained in each of the three tasks. First, results from the MRT were consistent with previous findings (Logan et al., 1989). Intelligibility was much poorer for Votrax speech than for natural speech. Second, results from the word-monitoring task demonstrated that this on-line measure of comprehension was highly sensitive to several experimental variables. Monitoring responses were more accurate and faster for detecting words in passages of natural speech than in passages of synthetic speech. Accuracy and response speed in the word-monitoring task were also related to target-set size. As the number of word targets in the memory set increased, accuracy of detection decreased and the latency of the monitoring responses increased. Monitoring responses were also slower when the word targets were embedded in more difficult passages.

Finally, a significant interaction was observed between voice and text difficulty in the latency data. The increase observed in monitoring latency for difficult texts was larger for synthetic speech than for natural speech. The results indicate that the comprehension of passages of connected synthetic speech is poorer than comprehension of passages of natural speech. Furthermore, the observed interaction suggests that stimulus encoding and comprehension processes share a common pool of processing resources (Kahneman, 1973; Posner and Boies, 1971; Wickens, 1987). This point is important: comprehension deficits for synthetic speech become larger when task difficulty increases (see also Logan et al., 1989),

Several differences in comprehension between natural and synthetic speech were also observed in the recognition memory data. First, recognition responses were more accurate following passages of natural speech than following synthetic speech. Second, a significant interaction between voice and sentence type was observed. Although no difference was observed between sentence types following natural speech, accuracy was significantly poorer for proposition recognition sentences following passages of synthetic speech.

Regression analyses revealed moderate but significant relationships between intelligibility and comprehension accuracy. These findings are not surprising given that comprehension—especially for more complex texts—involves much more than simply perceiving sound patterns of isolated words. Comprehension of spoken language involves accessing word meanings, parsing information within sentences, integrating information across sentences, and drawing inferences from a semantic propositional base (Kintsch and van Dijk, 1978). These resource-intensive processes may be more variable across listeners than processes underlying the perception and recognition of words. Thus it is not surprising that comprehension performance for long passages of connected speech is not predicted very well simply by considering intelligibility of isolated words in a highly constrained task such as the MRT.

EXPERIMENT 2

Introduction

A second experiment was conducted to obtain converging evidence about differences in the time course of comprehension for passages of natural and synthetic speech. The experiment

utilized a sentence-by-sentence listening task that was similar to procedures developed in reading research (Kintsch and Keenan, 1973; Miller and McKean, 1964). In the reading studies, passages of continuous text were presented visually one sentence at a time. Subjects controlled the presentation of successive sentences by pressing a button. The dependent variable, reading time, was measured as the difference between the onset of visual presentation and the subsequent button press.

Using multiple regression techniques, several studies have demonstrated that sentence-by-sentence reading times are sensitive to a number of linguistic variables that affect reading comprehension (see Haberlandt, 1984, for an exhaustive listing). Generally, sentence-by-sentence reading times increase with increasing text difficulty. The reading times are assumed to be composites of several cognitive operations that underlie comprehension (Haberlandt, 1984).

The use of the sentence-by-sentence listening time task (SBSLTT) was motivated by two factors. First, we wanted to develop another procedure that would complement existing on-line techniques that have been used to measure comprehension, such as phoneme and word-monitoring tasks. Both phoneme and word monitoring encourage listeners to allocate their attention to the surface structure of a passage rather than to its abstract propositional content (Blank, Pisoni, and McClaskey, 1981; Brunner and Pisoni, 1982). Because the SBSLTT made use of the same passages and comprehension task, the results could be compared directly with the results obtained in Experiment 1. This comparison provided a baseline to assess the effects of the SBSLTT task on comprehension performance.

The second motivation for using the SBSLTT was to provide converging evidence about the effects of synthetic speech on comprehension. Because the SBSLTT measures a listener's performance during comprehension, it can provide a second on-line index of processing load. As such, it can be used to assess the generality of the effects observed in Experiment 1 using the word-monitoring procedure.

Method

Subjects—The subjects were 30 adult native speakers of English. They were paid \$5.00 each for their participation. All subjects were living in the Bloomington, Indiana, area at the time of testing, and none reported any speech or hearing difficulties. None of the subjects was in the previous experiment.

Materials—The MRT word lists, comprehension passages, and recognition sentences were the same as those used in Experiment 1.

Stimulus preparation—The MRT preparation and presentation methods were identical to those used in Experiment 1.

Passages—Digitized versions of the ten passages used in Experiment 1 were segmented into individual sentences using a digital waveform editing program.

Procedure

After completing the MRT, subjects were given instructions for the sentence-by-sentence listening time procedure. Instructions were presented one sentence at a time, both visually on the CRT monitor in front of the subjects and auditorily over the headphones. At the end of each sentence, a cue light on the response box was illuminated. After the subject pressed a button labeled "Continue," the next sentence from the instructions was presented. Thus subjects controlled the rate of presentation—that is, the temporal interval between sentences

in each passage—by pressing the Continue button on the response box. Two hundred fifty milliseconds after the subject pressed the Continue button, the next sentence began. After the instructions, a practice passage and the ten test passages were presented using the same self-paced, sentence-by-sentence format. Sentences from the comprehension passages were presented only over headphones. At the end of each passage, the same set of recognition memory sentences used in Experiment 1 was presented on the CRT display to assess comprehension.

Latencies between the offset of each sentence and the subsequent button press in the SBSLTT, as well as the comprehension responses, were recorded by a PDP 11/34 minicomputer. Subjects were given a 5-min break after the fifth passage. The order of the passages and the recognition sentences was identical to that used in Experiment 1. Each experimental session lasted approximately one hour.

Results and Discussion

The results will be presented separately for the MRT, the SBSLTT, and the recognition memory test. Correlations were also computed between the MRT accuracy scores and several different comprehension measures. Voice (natural or synthetic) was treated as a between-subjects variable; text difficulty (fourth grade or college) and sentence type (word recognition or proposition recognition) were treated as within-subjects variables. Newman-Keuls tests were used as post hoc probes to assess statistical effects revealed by ANOVAs.

MRT—Performance on the MRT was similar to that observed in the first experiment. Subjects who listened to natural speech averaged 96% correct, whereas subjects who listened to synthetic speech averaged 63% correct. This difference was highly significant, $t(14) = -17.14, p < 0.001$.

SBSLTT—In order to reduce variability, response latencies greater or less than two standard deviations from the mean for each voice group were replaced with the mean value for the same group. Using this procedure, 2.1% of the responses from the natural speech condition and 1.7% of the responses from the synthetic speech condition were replaced. Data for each subject were then averaged across the different sentences within a passage as well as across the fourth-grade and college passages. Figure 4 shows the overall sentence-by-sentence listening times.

Sentence listening times were significantly longer for synthetic speech than for natural speech, $F(1,28) = 6.74, p = 0.015$. Listening times were also longer for the college passages than for the fourth-grade passages, $F(1,28) = 38.93, p < 0.001$. The interaction between voice and text difficulty was not significant. These results demonstrate the validity of the sentence-by-sentence listening time procedure for measuring on-line processing activities. The methodology indexed differences in comprehension between natural and synthetic speech and also was sensitive to the text difficulty variable.

Further evidence of these differences was revealed when the sentence listening times were analyzed separately for each passage. Figure 5 displays the sentence listening times for each passage. The left panel shows data for the five fourth-grade passages, and the right panel shows data for the five college-level passages. Within each panel are the passage name and the results of *t*-tests that were conducted for each passage. The differences between natural and synthetic speech were highly significant for four of the five fourth-grade passages. The difference just missed significance for the fifth passage, $t(28) = 1.93, p = 0.06$. The differences between natural and synthetic Speech were also significant for four of the five college-level passages. Although not significant, the difference between voices for the fifth passage was in the same direction as in the other passages, $t(28) = 1.70, p = .10$. In short, the

differences in listening times for natural and synthetic speech were reliable across the individual passages used in the experiment.

Figure 6 shows the sentence listening times for two representative passages as a function of voice and serial position of each sentence within a passage. The upper panel shows the listening times for a fourth-grade passage composed of 14 sentences, and the lower panel shows the listening times for a college-level passage with 16 sentences. These two examples reveal that differences in sentence listening times are highly consistent across individual sentences within as well as across passages.

In short, the sentence-by-sentence listening time data reveal on-line differences in the speed of comprehension processes. These processes include components involved in word recognition, lexical access and parsing for individual sentences, and the integration of information across different sentences (Just and Carpenter, 1980; Kintsch and van Dijk, 1978). Thus the SBSLTT provides a converging procedure that can be used to index online spoken language comprehension processes. Both the word-monitoring data from the first experiment and the SBSLTT data from the present experiment suggest that comprehension proceeds more slowly for passages of synthetic speech than for passages of natural speech, and that comprehension proceeds more slowly for difficult passages than for easy passages. The present results are consistent with those obtained in Experiment 1 using a word-monitoring task.

Recognition memory—Figure 7 displays the recognition accuracy data. The upper panel shows data for the word recognition sentences, and the lower panel shows data for the proposition recognition sentences. Accuracy was greater for sentences following passages of natural speech than for those following passages of synthetic speech, $F(1,28) = 27.54$, $p < 0.001$. This finding demonstrates not only that real-time comprehension proceeds more slowly for passages of synthetic speech but also that the products of comprehension appear to be more fragile than those obtained from natural speech. These results are also consistent with the findings obtained in the first experiment. Although the main effects of sentence type and text difficulty were not significant, both of these factors entered into significant interactions.

As in the first experiment, the interaction between voice and sentence type was significant, $F(1,28) = 4.91$, $p = 0.035$. This interaction is shown in Figure 8. When comparing accuracy for word and proposition recognition sentences, a large increase in performance was observed for sentences following passages of natural speech, whereas a small decrease was observed for sentences following passages of synthetic speech. Post hoc analyses revealed that although the increase from word recognition to proposition recognition sentences for natural speech was significant, the decrease for synthetic speech was not significant. This effect is similar to the crossover interaction observed earlier by Luce (1981), and it replicates the interaction between voice and sentence type found in Experiment 1. No other interactions involving voice were significant.

An analysis was also carried out to compare the recognition accuracy results with the findings obtained in the zero-target conditions of Experiment 1. Except for two effects, the patterns of results in the two experiments were identical. The first difference was found in the overall level of accuracy across the two experiments. Subjects were more accurate at recognizing sentences in Experiment 2 than in Experiment 1, $F(1,59) = 26.22$, $p < 0.001$. This result is not surprising because the SBSLTT encourages subjects to seek optimal levels of comprehension performance.

The second difference was revealed in a three-way interaction between experiment, voice, and difficulty, $F(1,59) = 10.47$, $p = 0.002$. The interaction was attributable entirely to the lowered performance in Experiment 1 on the college-level passages for natural speech. The decrease was not present in Experiment 2, again suggesting that subjects attempted to reach optimal levels of performance in the SBSLTT. Despite these differences, the recognition memory results obtained in Experiment 2 closely resemble those observed in Experiment 1. This pattern suggests that the SBSLTT does not adversely affect the final products of comprehension. Indeed, subjects in Experiment 2 appear to have used the self-paced nature of the procedure to enhance their overall comprehension performance. In fact, the increased verification accuracy obtained in Experiment 2 may be partly attributable to a speed/accuracy trade-off introduced by the unlimited response interval in the second experiment (Reed, 1976).

Regression analyses—Table 4 displays the correlations between intelligibility and several of the comprehension measures. Data were collapsed across voice and text difficulty. The left column shows correlation coefficients (Pearson's r), and the right column shows the probability values associated with F -statistics that were used to test whether the observed correlation coefficients are different from zero. Overall, the correlations of the MRT with the sentence-by-sentence listening times and recognition memory accuracy were moderate, though statistically significant. The correlation of the MRT accuracy with sentence listening time was negative, indicating that as intelligibility increased, response latencies decreased. In contrast, the correlations of MRT accuracy with recognition accuracy were positive, indicating that as intelligibility increased, recognition memory accuracy increased. Although the results show a moderate relationship between intelligibility and comprehension, the magnitude of the correlation coefficients ($r = +0.40$ to $+0.63$) suggests that comprehension of fluent speech involves more than merely phoneme and word recognition. Not surprisingly, some proportion of the variance in comprehension is produced by processes that go beyond the acoustic-phonetic information present in the speech waveform.

Summary

The results of Experiment 2 corroborate and extend the major findings obtained in Experiment 1. First, the results demonstrate that SBSLTT is a useful tool for studying spoken language comprehension, particularly on-line processes employed during the initial stages of comprehension. Although the paradigm has been used profitably for several years with printed text (Cirilo, 1981; Cirilo and Foss, 1980; Graesser, Hoffman, and Clark, 1980; Haviland and Clark, 1974; Kieras, 1974; Lesgold, Roth, and Curtis, 1979; Miller and McKean, 1964), it has never been used to study spoken language comprehension. The finding that the observed latency measures varied with text difficulty demonstrates that the measurement technique is sensitive to several components of the comprehension process (see Haberlandt, 1984, for comparable arguments for the visual analog of the paradigm).

The SBSLTT data also demonstrate that comprehension proceeds more slowly when the sentences are produced by a low-quality speech synthesizer such as Votrax than when sentences are produced by a natural talker. This finding is consistent with the results of the word-monitoring data, in which detection latencies were reliably slower for words in passages of Votrax synthetic speech than for words in natural speech.

The recognition memory data largely replicate results obtained in the word-monitoring experiment. In fact, an analysis comparing data from this experiment with data from the zero-target condition of the first experiment revealed only small differences in recognition memory accuracy. Instead of interfering with comprehension, the self-paced nature of the

sentence listening task appears to enhance comprehension by permitting subjects to seek optimal presentation rates.

GENERAL CONCLUSIONS

The results obtained in the present set of experiments are similar to findings reported in previous comprehension studies. In experiments by Luce (1981), Hersch and Tartaglia (1983), and Moody and Joost (1987), reliable differences between natural and synthetic speech were observed in several comprehension measures, using postperceptual tests. The present findings suggest that passages of synthetic speech do indeed provide a degraded representation of textual information in memory.

As noted earlier, however, other studies have found relatively small differences in comprehension between natural and synthetic speech when postperceptual measures were used (Jenkins and Franklin, 1981; McHugh, 1976; Pisoni and Hunnicutt, 1980). We believe these studies may have used procedures that were simply too insensitive to measure differences in the comprehension processes. For example, in one of their experiments, Jenkins and Franklin (1981) used a free recall task that encourages subjects to adopt a wide variety of postperceptual strategies. In the studies conducted by McHugh (1976) and Pisoni and Hunnicutt (1980), the test items may have been insensitive because of baseline floor-and-ceiling effects.

The two on-line measures used in the present experiments are largely complementary. Word-monitoring performance measures the mental workload of comprehension, whereas sentence-by-sentence listening times provide a composite measure of the speed and efficiency of the comprehension process. In the present studies both procedures have demonstrated that long passages of poor-quality synthetic speech are more difficult to comprehend than long passages of natural speech.

Several implications for applications of synthetic speech are suggested by the present results. First, because of slower comprehension of synthetic speech, poorer-quality systems should not be used when rapid response to the linguistic content of a text is critical. Second, in light of the relatively high cognitive workload associated with synthetic speech, poorer-quality systems should not be used with linguistically complex or demanding secondary tasks that may require the operator to share attentional capacity differentially among several competing inputs.

In summary, several conclusions can be drawn from the results obtained in the present investigation. First, on-line comprehension of synthetic speech is poorer than comprehension of natural speech. Second, comprehension of synthetic speech places greater workload demands on listeners than does natural speech. Third, the sentence-by-sentence listening time task appears to be a sensitive measure that can be employed successfully to study the real-time comprehension of spoken language. Fourth, comprehension accuracy for passages of connected speech can be predicted only moderately well from standardized tests of phoneme intelligibility for isolated words. Finally, the present results suggest that poor-quality synthetic speech may be inappropriate for applications in which there are difficult texts, concurrent tasks requiring differential attention, or severe constraints on response accuracy or speed.

Acknowledgments

The authors thank Susan Rivera for help in running subjects and collecting data in Experiment 2. The authors also thank Paul Luce, Bob Peterson, and Steve Goldinger for their comments on an earlier draft. This research was supported by National Science Foundation Research Grant IRI-86-17847 to Indiana University.

References

- Allen, J.; Klatt, DH.; Hunnicutt, S. From text to speech: The MITalk system. Cambridge, England: Cambridge University Press; 1987.
- Bartlett, FC. Remembering. Cambridge, England: Cambridge University Press; 1932.
- Blank MA, Pisoni DB, McClaskey CL. Effects of target monitoring on understanding fluent speech. *Perception and Psychophysics*. 1981; 29:383–388. [PubMed: 7279562]
- Britton BK, Westbrook RD, Holdredge T. Reading and cognitive capacity: Effects of text difficulty. *Journal of Experimental Psychology: Human Learning and Memory*. 1978; 4:582–591.
- Brunner H, Pisoni DB. Some effects of perceptual load on spoken text comprehension. *Journal of Verbal Learning and Verbal Behavior*. 1982; 21:186–195.
- Cirilo RK. Referential coherence and text structure in story comprehension. *Journal of Verbal Learning and Verbal Behavior*. 1981; 20:358–367.
- Cirilo RK, Foss DJ. Text structure and reading time for sentences. *Journal of Verbal Learning and Verbal Behavior*. 1980; 19:96–109.
- Cutler A. Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics*. 1976; 20:55–60.
- Educational Testing Service. Cooperative English tests: Reading comprehension, Form 1B. Princeton, NJ: Author; 1960.
- Flesch, R. The art of readable writing. New York: Harper; 1949.
- Foss DJ. Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behavior*. 1969; 8:457–462.
- Graesser AC, Hoffman NL, Clark LF. Structural components in reading times. *Journal of Verbal Learning and Verbal Behavior*. 1980; 19:135–151.
- Gunning, R. The technique of clear writing. New York: McGraw-Hill; 1968.
- Haberlandt, K. Components of sentence and reading times. In: Kieras, DE.; Just, MA., editors. *New methods in reading comprehension research*. Hillsdale, NJ: Erlbaum; 1984. p. 219–251.
- Harcourt Brace Jovanovich. Iowa Silent Reading Tests: Level 3, Form E. New York: Author; 1972a.
- Harcourt Brace Jovanovich. Stanford Test of Academic Skills: Reading, College Level, II-A. New York: Author; 1972b.
- Haviland SE, Clark HH. What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*. 1974; 13:512–521.
- Hersch, HM.; Tartaglia, L. Understanding synthetic speech. Maynard, MA: Digital Equipment Corporation, User Research Group, Corporate Research and Architecture; 1983.
- Hoover J, Reichle J, van Tasell D, Cole D. The intelligibility of synthesized speech: Echo II versus Votrax. *Journal of Speech and Hearing Research*. 1987; 30:425–431. [PubMed: 2959817]
- Houghton-Mifflin. The Nelson-Denny Reading Test: Form D. Boston: Author; 1973.
- House AS, Williams CE, Hecker MH, 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 Psychonomics Society Meeting; Philadelphia, PA. 1981 Nov.
- Just MA, Carpenter PA. A theory of reading: From eye fixations to comprehension. *Psychological Review*. 1980; 87:329–354. [PubMed: 7413885]
- Kahneman, D. Attention and effort. Englewood Cliffs, NJ: Prentice-Hall; 1973.
- Kieras, DE. Unpublished doctoral dissertation. University of Michigan; Ann Arbor, MI: 1974. Analysis of the effect of word properties and limited reading time in a sentence comprehension and verification task.
- Kintsch W, Keenan JM. Reading rate as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*. 1973; 5:257–274.
- Kintsch W, van Dijk TA. Toward a model of text comprehension and production. *Psychological Review*. 1978; 85:363–394.

- Klatt DH. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*. 1987; 82:737–793. [PubMed: 2958525]
- LaBerge D, Samuels SL. Toward a theory of automatic information processing in reading. *Cognitive Psychology*. 1974; 6:293–323.
- Lee L, Nusbaum HC. The effects of perceptual learning on capacity demands for recognizing synthetic speech. *Journal of the Acoustical Society of America*. 1989; 85(Suppl 1):S25.
- Lesgold AM, Roth SF, Curtis ME. Foregrounding effects in discourse comprehension. *Journal of Verbal Learning and Verbal Behavior*. 1979; 18:291–308.
- Levelt, WJM. A survey of studies in sentence perception. In: Levelt, WJM.; Flores d'Arcais, GB., editors. *Studies in the perception of language*. New York: Wiley; 1978. p. 1-74.
- Logan JS, Greene BG, Pisoni DB. Intelligibility of eight text-to-speech systems. *Journal of the Acoustical Society of America*. 1989; 86:566–581. [PubMed: 2527884]
- Luce, PA. Research on Speech Perception, Progress Report 7. Bloomington: Indiana University, Speech Research Laboratory; 1981. Comprehension of fluent synthetic speech produced by rule.
- Luce, PA.; Carrell, TD. Research on Speech Perception, Progress Report 7. Bloomington: Indiana University, Speech Research Laboratory; 1981. Creating and editing waveforms using WAVES.
- Luce PA, Feustel TC, Pisoni DB. Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*. 1983; 25:17–32. [PubMed: 6840769]
- Manous, LM.; Pisoni, DB.; Dedina, MJ.; Nusbaum, HC. Research on Speech Perception, Progress Report 11. Bloomington: Indiana University, Speech Research Laboratory; 1985. Comprehension of natural and synthetic speech using a sentence verification task.
- McCauley, ME. Human factors in voice technology. In: Muckler, FA., editor. *Human factors review*. Santa Monica, CA: Human Factors Society; 1984. p. 131-166.
- McHugh, A. Naval Research Laboratory Report 8015. Washington, DC: Naval Research Laboratory; 1976. Listener preference and comprehension tests of stress algorithms for a text-to-phonetic speech synthesis program.
- Miller GA, McKean KO. A chronometric study of some relations between sentences. *Quarterly Journal of Experimental Psychology*. 1964; 16:297–308.
- Mimmack, PC. Unpublished master's thesis. Indiana University; Bloomington, IN: 1982. Sentence-by-sentence listening times for spoken passages: Text structure and listeners' goals.
- Moody, TS.; Joost, MG. The effects of various types of speech output on comprehension rates. In: Bullinger, HJ.; Shackel, B., editors. *Human computer interactions*. Amsterdam: Elsevier; 1987. p. 573-579.
- Nye, PW.; Gaitenby, J. Haskins Laboratory Status Report on Speech Research, SR-33. New Haven, CT: Haskins Laboratories; 1973. Consonant intelligibility in synthetic speech and in a natural speech control (Modified Rhyme Test results).
- Pauk, W. Six-way paragraphs: Middle level. Providence, RI: Jamestown; 1983.
- 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. p. 572-575.
- Pisoni DB, Nusbaum HC, Greene BG. Perception of synthetic speech generated by rule. *Proceedings of the IEEE*. 1985; 11:1665–1676.
- Posner MI, Boies SJ. Components of attention. *Psychological Review*. 1971; 78:391–408.
- Reed AV. List length and the time-course of recognition in immediate memory. *Memory and Cognition*. 1976; 4:16–30.
- Schwab EC, Nusbaum HC, Pisoni DB. Some effects of training on the perception of synthetic speech. *Human Factors*. 1985; 27:395–408. [PubMed: 2936671]
- Wickens, CD. Information processing, decision making, and cognition. In: Salvendy, G., editor. *Handbook of human factors*. New York: Wiley; 1987. p. 72-107.

Figure 1.

Word-monitoring accuracy (probability of a correct detection) as a function of target-set size in Experiment 1. The upper panel shows data for fourth-grade passages, and the lower panel shows data for college-level passages. Open bars represent accuracy for passages of natural speech, and striped bars represent accuracy for passages of Votrax synthetic speech. Error bars represent one standard error of the sample means.

Figure 2. Word-monitoring response latency (in milliseconds) as a function of target-set size in Experiment 1. The upper panel shows data for fourth-grade passages, the lower panel for college-level passages. Open bars represent latencies for passages of natural speech, and striped bars represent latencies for passages of Votrax synthetic speech. Error bars represent one standard error of the sample means.

Figure 3.

Recognition memory accuracy (probability correct) as a function of target-set size in Experiment 1. The panels on the left show data for word recognition sentences, and the panels on the right show data for proposition recognition sentences. The upper panels show data for fourth-grade passages, the lower panels for college-level passages. Open bars represent accuracy for sentences after passages of natural speech, striped bars for sentences after passages of synthetic speech. Error bars represent one standard error of the sample means.

Figure 4. Sentence-by-sentence listening times as a function of voice and text difficulty in Experiment 2. Open bars represent response latencies for passages of natural speech, and striped bars represent response latencies for passages of synthetic speech. Error bars represent one standard error of the sample means.

Figure 5.

Sentence-by-sentence listening times as a function of voice in Experiment 2. The panels on the left display response latencies for the five fourth-grade passages; panels on the right display response latencies for the five college-level passages. Open bars represent response latencies for passages of natural speech, and striped bars represent response latencies for passages of synthetic speech. Error bars represent one standard error of the sample means.

Figure 6.

Sentence-by-sentence listening times as a function of the position of a sentence within a passage in Experiment 2. The upper panel displays data for a representative fourth-grade passage, the lower panel for a representative college-level passage. Open circles and connecting lines represent response latencies for passages of natural speech. Solid circles and connecting lines represent response latencies for passages of synthetic speech.

Figure 7. Recognition memory accuracy (probability correct) as a function of text difficulty in Experiment 2. The upper panel shows data for word recognition sentences, the lower panel for proposition recognition sentences. Open bars represent accuracy for sentences after passages of natural speech, and striped bars represent accuracy for sentences after passages of synthetic speech. Error bars represent one standard error of the sample means.

Figure 8. Recognition memory accuracy (probability correct) as a function of sentence type in Experiment 2. Open bars represent accuracy for sentences after passages of natural speech, striped bars for sentences after passages of synthetic speech. Error bars represent one standard error of the sample means.

TABLE 1

Text Statistics for Fourth-Grade Passages

Measure	Passage				
	Bears	Bees	Birds	Firsts	Turtle
Readability indices					
Flesch ^a	84	84	89	88	85
Gunning ^b	7	7	5	7	7
Flesch-Kincaid ^c	5	5	4	3	5
Paragraph level					
No. paragraphs	2	3	3	3	5
Average no. sentences	12.5	6.6	4.6	7.6	4.2
Sentence level					
No. sentences	25	20	14	23	21
Average no. words	11.9	11.3	11.2	9.5	12
Word level					
Total number	298	226	157	220	253
Average no. syllables	1.3	1.3	1.3	1.3	1.3

^aFlesch Reading Ease (Flesch, 1949). Varies from 0 (*very difficult to read*) to 100 (*very easy to read*).

^bGunning's Fog Index (Gunning, 1968). Corresponds roughly to the grade level needed to understand the text.

^cFlesch-Kincaid (Flesch, 1949). Value represents the average number of school years needed to understand the text.

\$watermark-text

\$watermark-text

\$watermark-text

TABLE 2

Text Statistics and Sources for College-Level Passages

Measure	Passage				
	Iberia ^a	Build ^b	Lens ^c	Radio ^d	Speech ^d
Readability indices					
Flesch	61	68	40	32	43
Gunning	14	11	19	22	16
Flesch-Kincaid	10	9	15	17	13
Paragraph level					
No. paragraphs	2	2	1	2	3
Average no. sentences	6	6.5	7	3	5
Sentence level					
No. sentences	12	13	7	6	15
Average no. words	21	18.4	29.5	33	21.7
Word level					
Total number	252	240	207	198	326
Average no. syllables	1.5	1.4	1.6	1.7	1.7
Average					1.6

Note: See Table 1 for explanation of reliability indices.

Source:

- ^a *Cooperative English Tests: Reading Comprehension* (Educational Testing Service, 1960).
- ^b *Iowa Silent Reading Tests* (Harcourt Brace Jovanovich, 1972a).
- ^c *The Nelson-Denny Reading Test* (Houghton-Mifflin, 1973).
- ^d *Stanford Test of Academic Skills: Reading* (Harcourt Brace Jovanovich, 1972b).

TABLE 3

Correlations of Modified Rhyme Test Accuracy and Comprehension Measures in Experiment 1

Predicted Variable	Correlation	Probability
Word monitoring		
Probability of correct detection	0.31	0.01
Response latency	−0.34	<0.01
Recognition accuracy		
Word recognition	0.46	<0.01
Proposition recognition	0.47	<0.01

TABLE 4

Correlations of Modified Rhyme Test Accuracy and Comprehension Measures in Experiment 2

Predicted Variable	Correlation	Probability
Sentence listening time		
Response latency	−0.40	0.03
Recognition accuracy		
Word recognition	0.55	<0.01
Proposition recognition	0.63	<0.01