Published in final edited form as:

Hum Factors. 1985 August ; 27(4): 395–408.

Some Effects of Training on the Perception of Synthetic Speech

Eileen C. Schwab¹, Howard C. Nusbaum, and David B. Pisoni

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

Abstract

The present study was conducted to determine the effects of training on the perception of synthetic speech. Three groups of subjects were tested with synthetic speech using the same tasks before and after training. One group was trained with synthetic speech. A second group went through the identical training procedures using natural speech. The third group received no training. Although performance of the three groups was the same prior to training, significant differences on the post-test measures of word recognition were observed: the group trained with synthetic speech performed much better than the other two groups. A six-month follow-up indicated that the group trained with synthetic speech displayed long-term retention of the knowledge and experience gained with prior exposure to synthetic speech generated by a text-to-speech system.

INTRODUCTION

In the next few years we can expect to see an increase in the use of voice response systems in a variety of consumer, military, and industrial applications. Many of these systems will use synthetically produced speech that is generated by rule from a text-to-speech system. Despite the potential applications for this technology, several recent studies have demonstrated that synthetic speech, even high-quality synthetic speech, is more difficult to perceive and understand than natural speech (Pisoni, 1982; Pisoni and Hunnicutt, 1980).

The differences in perception between natural and synthetic speech could be a major obstacle to the widespread use of synthetic speech in voice-output systems, particularly under conditions of high information load, such as in the cockpit of an aircraft or in a command-control application. To overcome this problem, it is necessary to maximize intelligibility so that performance with synthetic speech approximates performance with natural speech as closely as possible. There are two ways to achieve this goal. One way to improve intelligibility is to increase the quality and naturalness of the synthetic speech so that it contains more of the critical acoustic cues and redundancy of natural speech. A number of efforts are currently being devoted to improving synthesis via greater attention to the acoustic-phonetic detail and the effects of context and prosody on perception (e.g., Allen, 1981). Unfortunately, the results of these efforts may not be incorporated into commercially available products for quite some time. The second possibility is to train listeners to perceive synthetic speech more accurately. Because humans are very flexible and are capable of learning new ways of processing information, perceptual training and exposure could change the strategies humans use in perceiving synthetic speech. Thus, it might be possible to improve intelligibility and hence the usability of this new technology through short-term selective exposure and training.

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

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

METHOD

Subjects

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

Materials

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

PB lists—The first set of stimuli consisted of 12 lists of 50 monosyllabic phonetically balanced (PB) words. These lists were a subset of 20 lists originally designed for testing speech intelligibility (Acoustical Society of America, 1960; Egan, 1948).

MRT lists—The second set of stimuli consisted of four sets of 50 monosyllabic words taken from the Modified Rhyme Test (House, Williams, Hecker, and Kryter, 1965). This test was designed primarily to assess the intelligibility of consonants in natural speech.

Harvard sentences—The third set of materials consisted of 100 Harvard psychoacoustic sentences (Egan, 1948; IEEE, 1969). These are normal English sentences containing five key words plus a variable number of function words, arranged in a variety of syntactic structures.

Haskins sentences—The fourth set of materials consisted of 100 syntactically normal but semantically anomalous sentences developed at Haskins Laboratories (Nye and Gaitenby, 1974). These sentences contained high-frequency monosyllabic words. Each stimulus had four key words, which were presented in the following syntactic structure: The (*adjective*) (*noun*) (*verb, past tense*) the (*noun*). The sentences were semantically anomalous in order to reduce word identification cues based on the meaning of the utterance.

Prose passages—Finally, the fifth set of materials consisted of 39 prose passages taken from a variety of sources including popular magazines and reading comprehension tests. Eight of the passages were used as training stories and were taken from the Nelson-Denny Reading Comprehension Test (Nelson and Denny, 1930) along with the original multiple-choice comprehension questions. Another nine passages were selected from other adult reading tests (Brown, 1973; Derrick, Harris, and Walker, 1960; Farr, 1972; Gardener, Callis, Merwin, and Madden, 1972). The remaining 21 passages were modified versions of texts that have been published in magazines such as *Time* and *Scientific American*. For these 30 passages, we used yes/no comprehension questions constructed to examine information at different levels of text comprehension. These levels ranged from a relatively low level, in which subjects decided if a particular word occurred in a text, to a high level, in which subjects evaluated the theme of the text. In addition, several questions were designed to examine inferences that could be drawn from the passages but were not explicitly stated in the passage.

Procedure

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

Testing—The entire experiment was conducted during a two-week period. Each experimental session lasted approximately one hour. Since the experiment was run during the business week, a two-day break intervened between Days 5 and 6 of the experiment.

All three groups of subjects were tested on the first (Day 1) and last (Day 10) day of the experiment using the same procedures and order of tasks (see Table 1). On testing days, all subjects listened to speech produced by the Votrax Type-’N-Talk text-to-speech system. The Votrax system was chosen primarily because of the relatively poor quality of its segmental (i.e., consonant and vowel) synthesis. Thus, ceiling effects could not obscure any effects of training. The pretest and posttest design allowed us to assess the improvements, if any, in

recognition performance for the synthetic speech. During the testing sessions, no feedback was provided to the subjects about their performance.

On each testing day, the first task always consisted of identifying isolated words. Subjects were presented with two 50-item PB lists, one word at a time. Subjects were asked to write down an English word on every trial and to guess if they were not sure. After writing a response, each subject pushed a button on a computer-controlled response box to indicate completion of the current trial. When all subjects had responded, the next word was presented.

For the second task, subjects listened to and identified 100 isolated words from the Modified Rhyme Test (MRT lists). After hearing a stimulus, subjects were presented with six alternative words centered on a CRT screen. The subjects were instructed to choose one of six buttons on a response box in front of them to indicate the word they had heard. The subjects were instructed to respond as quickly and as accurately as possible. Response latency and accuracy were recorded by the computer and stored on disk for later analysis.

For the third test, subjects were instructed to transcribe 10 Harvard sentences. After each sentence was presented, subjects wrote down what they heard. Subjects were encouraged to guess if they were not sure of the words in the Utterance. After transcribing the sentence, subjects pushed a response button to indicate they were ready for the next sentence.

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

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

Training—Although all three groups of subjects participated in the experiment on Day 1 (pretest) and Day 10 (posttest), only the Votrax and natural groups actually received any training between the pretest and the posttest. Both groups heard the same utterances and went through the identical training procedures on Day 2 through Day 9 of the experiment. The only difference in treatment was whether they received synthetic or natural speech.

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

subjects scored their response by pressing one of two buttons (labeled *correct* or *incorrect*) to indicate their accuracy in transcribing the stimulus.

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

RESULTS

Five subjects did not complete the experimental sessions, and their data were excluded from the statistical analyses. The reason given by the subjects for not completing all sessions ranged from oversleeping to simply forgetting the appointment. At the beginning of the experiment, subjects were warned that if they missed one session, they would be dropped from the experiment. Of the remaining 26 subjects who did complete the experiment, 9 were in the group trained with synthetic speech, 7 were in the group trained with natural speech, and 10 were in the group that received no training.

Pretest and Posttest Data

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

A three-way analysis of variance (Groups $\times$ Days $\times$ Lists) was conducted on the pretest and posttest scores. A significant main effect of lists was obtained, $F(1,23) = 83.88$, $p < 0.0001$. Performance on the second PB list on each day was higher than performance on the first PB list. A significant Groups $\times$ Days interaction was also obtained, $F(2,23) = 134.93$, $p < 0.0001$. This two-way interaction can be seen in Figure 1, which shows the mean accuracy for the three groups of subjects (averaged over the two PB Lists presented) for Day 1 and Day 10. A simple-effects analysis of this interaction showed that on the first day of the experiment, all three groups of subjects had equivalent levels of performance, $F(2,35) = 0.73$, $p > 0.45$ —subjects correctly identified approximately 25% of the stimuli. On Day 10, the groups showed very different levels of performance, $F(1,35) = 122.11$, $p < 0.0001$. A Newman-Keuls analysis revealed that the Votrax group performed better than either the control or natural group ($p < 0.05$). No difference was found in performance between the control and natural groups on Day 10 ($p > 0.05$). Simple-effects analyses also indicated that the Votrax group showed a significant improvement from Day 1 to Day 10, $F(1,23) = 705.55$, $p < 0.0001$. Performance improved from approximately 25% correct identification on Day 1 to approximately 70% correct identification. The control and natural groups also had a significant improvement from approximately 25% correct identification on Day 1 to approximately 35% correct word identification on Day 10, $F(1,23) = 27.99$, $p < 0.0001$ and $F(1,23) = 56.39$, $p < 0.0001$, respectively.

MRT lists—The pretest and posttest performance scores for the Modified Rhyme Test are shown in Table 2. Day 1 accuracy on the MRT was much higher than the PB words (31% versus 61%), a result that is due primarily to the limited response set used in the MRT test (i.e., six alternatives versus an open response set). A three-way analysis of variance revealed

a significant Days $\times$ Groups interaction, $F(2,23) = 12.89, p < 0.0005$. Examination of this interaction using simple-effects tests indicated that all three groups had equivalent levels of accuracy on Day 1, $F(2,46) = 0.07, p > 0.9$. However, on Day 10, the groups differed significantly, $F(2,46) = 28.9, p < 0.0001$. A Newman-Keuls analysis indicated that the Votrax group performed better ($p < 0.05$) than both the control and natural groups, which were not different from each other. In addition, a simple-effects analysis of the Votrax group data showed a significant improvement in performance on Day 10 for this test, $F(1,23) = 34.65, p < 0.0001$. However, the performance of the control and natural groups did not change from Day 1 to Day 10, $F(1,23) = 1.05, p > 0.3$ and $F(1,23) = 0.70, p > 0.4$, respectively.

Table 2 also shows the mean response latencies for the three groups of subjects for the MRT stimuli on Days 1 and 10. A three-way analysis of variance on the response latencies from the MRT test revealed a significant Days $\times$ Groups interaction, $F(2,23) = 3.69, p < 0.05$. Examination of this interaction indicated that the response latencies for the three groups did not differ on Day 1, $F(2,33) = 0.20, p > 0.8$. However, on Day 10 the three groups showed significantly different mean latencies, $F(2,33) = 3.37, p < 0.05$. Post-hoc analyses indicated that neither the control group nor the natural group showed any significant change in response latency from Day 1 to Day 10, $F(1,23) = 1.14, p > 0.25$ and $F(1,23) = 3.4, p = 0.078$, respectively. However, the Votrax group responded more rapidly on Day 10 as compared with Day 1, $F(1,23) = 22.03, p < 0.0002$. Thus, the Votrax group not only responded more accurately on the MRT on the posttest (Day 10), but they also responded much faster than did the other two groups.

Harvard sentences—Each test sentence contained five key words that were scored for accuracy. In order for a word to be scored as correct, subjects had to transcribe the exact word with no additions or deletions. Table 2 lists the mean percentage of correct word identifications for key words in the Harvard sentences. A two-way analysis of variance (Groups $\times$ Days) revealed a significant two-way interaction, $F(2,23) = 39.47, p < 0.0001$. Post-hoc analysis of this interaction indicated that, on Day 1, all three groups had equivalent levels of performance, $F(2,35) = 1.53, p > 0.2$. By Day 10, however, a significant difference in performance was observed, $F(2,35) = 44.21, p < 0.0001$. A Newman-Keuls analysis indicated that the Votrax group performed better on Day 10 than either the control group ($p < 0.05$) or the natural group ($p < 0.05$), whereas the control and natural groups did not differ from each other. Post-hoc analyses also revealed that the Votrax group improved in performance from Day 1 to Day 10, $F(1,23) = 55.54, p < 0.0001$. By contrast, the performance of the control and natural groups did not change from Day 1 to Day 10, $F(1,23) < 1.00, p > 0.9$, and $F(1,23) = 0.40, p > 0.5$, respectively.

Haskins sentences—Each anomalous sentence contained four key words that were scored for accuracy. Table 2 shows the mean percentage of correct word identifications for the three groups of subjects. A two-way analysis of variance (Groups $\times$ Days) revealed a significant two-way interaction, $F(2,23) = 14.32, p < 0.0005$. A post-hoc analysis of the Groups $\times$ Days interaction showed that on Day 1, all three groups obtained equivalent levels of performance, $F(2,45) = 1.16, p > 0.3$. In contrast, a difference in accuracy for the three groups was observed on Day 10, $F(2,45) = 32.26, p < 0.0001$. A Newman-Keuls analysis indicated that the Votrax group performed better than the control ($p < 0.05$) and natural ($p < 0.05$) groups. As before, performance for the control and natural groups on the posttest (Day 10) was not different. However, post hoc analyses of the scores on Day 1 and Day 10 also indicated that on Day 10 all three groups showed improved word identification relative to Day 1: control, $F(1,23) = 25.35, p < 0.0001$; natural, $F(1,23) = 24.91, p < 0.0001$, and Votrax, $F(1,23) = 137.27, p < 0.0001$. Thus, all groups displayed some improvement.

However, this improvement was greatest for the subjects in the Votrax group, who received specific training with and exposure to synthetic speech during Days 2 through 9.

Prose passages—The mean accuracy for the comprehension questions is shown in Table 2. A two-way analysis of variance was conducted on the mean percentage of correct responses to the comprehension questions. A significant effect of days was observed, $F(1,23) = 5.65, p < 0.05$. All subjects were more accurate on Day 1 than on Day 10. No significant effect of groups was observed, $F(2,23) = 1.60, p > 0.2$ and the Groups $\times$ Days interaction was not significant, $F(2,23) = 0.76, p > 0.4$. A two-way analysis of variance was also conducted on the response latencies (shown in Table 2) and no effect of groups was observed, $F(2,23) = 1.39, p > 0.25$. The Groups $\times$ Days interaction was not significant, $F(2,23) = 2.93, p = 0.073$. However, a significant effect of days was observed, $F(1,23) = 4.53, p < 0.05$, indicating that subjects were faster to respond on Day 10 than on Day 1. An examination of the data demonstrated that the effect of days was due primarily to the Votrax group. A simple-effects test showed a significant decrease in latencies for the Votrax group, $F(1,23) = 9.63, p < 0.001$, but no significant change in latency for either the control group, $F(1,23) = 0.04, p > 0.8$, or the natural group $F(1,23) = 0.61, p > 0.4$.

Training Data

All training data were averaged over two-day periods in order to reduce variability due to the small number of subjects in each group. Because the control group received no training, the results are presented only for the Votrax and natural groups. It is important to reemphasize that these two groups received exactly the same training procedures with the same words, sentences, and passages. The only difference was that the Votrax group listened to synthetic speech whereas the natural group listened to natural speech. We expected performance for the natural group to be consistently higher than performance for the Votrax group throughout training because natural speech is more intelligible and easier to recognize.

Word identification—The mean accuracy for the transcription of isolated PB words, words in Harvard sentences, and words in Haskins sentences for the Votrax and natural groups on each of the training days is given in Table 3. As expected, the natural group showed ceiling levels of performance in recognizing isolated words and words in sentences throughout the course of training. These data contrast sharply with the performance shown by the Votrax group. Subjects in this group showed a consistent increase in their performance over the course of training. A two-way analysis of variance (Groups $\times$ Days) was conducted on the identification scores for each of the word identification tasks. The different patterns of performance for the two groups produced a significant Groups $\times$ Days interaction for PB lists, $F(3,42) = 74.90, p < 0.0001$; for Harvard sentences, $F(3,42) = 55.66, p < 0.0001$, and for Haskins sentences, $F(3,42) = 25.81, p < 0.0001$. As expected, post hoc analysis of this interaction showed no changes in performance for the natural group for any of the tasks over the course of the training, $F(3,42) = 1.76, p > 0.15$, for PB lists; $F(3,42) = 0.06, p > 0.9$, for Harvard sentences; and $F(3,42) = 0.16, p > 0.9$, for Haskins sentences. However, significant changes in accuracy were found for these tasks for the Votrax group, $F(3,42) = 210.67, p < 0.0001$, for PB lists; $F(3,42) = 125.76, p < 0.0001$, for Harvard sentences; $F(3,42) = 64.56, p < 0.0001$, for Haskins sentences. Newman-Keuls analyses indicated that consistent improvements were obtained by the Votrax group for each successive two-day period for word recognition in all three tasks ($p < 0.05$).

Prose passages—Because of a computer malfunction, the prose data for the Votrax group on Day 7 were accidentally deleted and could not be recovered. To keep the design balanced, the comprehension data on Day 7 for the natural group were excluded from all

subsequent analyses. Table 3 displays the mean accuracy for comprehension questions for the three test passages presented on each training day. As shown here, a different pattern of results was obtained compared with the data from the previous tasks. No clear separation was observed in performance for the two groups. A Groups $\times$ Days analysis of variance was conducted and a significant two-way interaction was observed, $F(3,42) = 5.37$, $p < 0.005$. A simple-effects analysis of this interaction revealed that the natural group performed better than the Votrax group for Days 2–3, $F(1,39) = 9.07$, $p < 0.005$, and Day 6, $F(1,39) = 7.53$, $p < 0.01$ of the experiment. For the remaining days, no difference in accuracy between the two groups was observed. Also, no consistent pattern of differences in response latencies was observed.

DISCUSSION

With the exception of the data on comprehension for the prose passages, the results from the pretest and posttest demonstrate that exposure and training with synthetic speech produce substantial improvements in performance. Moreover, the experimental design revealed that this improvement in performance is primarily a function of specific training with and exposure to the synthetic speech rather than familiarity with the experimental procedures and tasks. Note that similar training with natural speech did not yield improved performance with synthetic speech. Thus, the observed improvement in performance appears to be due to changes in the perceptual processing of the synthetic speech rather than to task familiarity.

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

A similar pattern was observed with the Harvard sentences, which constrain word recognition to only a few plausible alternatives. Normal semantic and syntactic constraints substantially reduce the number of words that are possible at any point in the sentence (cf. Grosjean, 1980; Salasoo and Pisoni, 1985). For example, if a subject hears a word in a sentence such as *The funny clown made them laugh*, that could not be identified, relatively few plausible substitutions for the word *laugh* exist. Thus, it is possible that subjects adopted a recognition strategy that exploited the inherent task constraints in the MRT and the Harvard sentences.

However, the PB lists and the Haskins sentences are much more difficult tasks due to the lack of constraints on either the response set or the set of plausible word candidates, and a different pattern of results was obtained for these tasks. All groups improved their performance from Day 1 to Day 10. However, the Votrax group performed much better than either the control or natural groups on Day 10. Given the reduced top-down constraints, these tasks appear to be more sensitive to improvements in the perception of synthetic speech. Indeed, the improvements shown on these tasks by the natural and control groups may have resulted from the single exposure to synthetic speech on the pretest. Since the natural group received eight additional days of practice with the tasks, the natural group should have been even better than the control group if the improvement were due to mastery

of the procedures. However, the lack of a difference between the natural and control groups suggests that the slight improvement in performance shown by these groups may be attributable to exposure to the synthetic speech that occurred on the first day of the experiment—the pretest. It should be emphasized, however, that even though all groups improved on these two tasks, the Votrax group showed the greatest improvement.

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

The one set of materials for which subjects did not show a posttest change in performance after training was the prose passages. For these materials, no increase in performance from the pretest to the posttest was obtained for any group of subjects. Analysis of the data indicated that the variability in performance was more a function of the particular passages heard and the particular test questions than the nature of the signal. Because of the poor sensitivity of this task to differences in perception between natural and synthetic speech, a more detailed analysis of the results was not conducted.

In summary, our results demonstrate that it is possible to produce dramatic changes in the perception of synthetic words and sentences with only moderate amounts of training. Moreover, these improvements were obtained with synthetic speech produced by a low-cost, commercially available text-to-speech system. Measuring performance of a text-to-speech system with naive or unpracticed subjects may underestimate the performance characteristics of the system for certain applications. If a particular text-to-speech system is going to be used routinely on a daily basis by the same person, it may be difficult to predict performance from group recognition scores obtained with inexperienced subjects. Our results show that even moderate amounts of exposure to synthetic speech, using poor-quality synthetic speech, can produce significant improvements in performance. However, in some applications there may be little or no time to familiarize the listener with the synthetic output. For these applications, more natural-sounding synthetic speech of much higher quality will be required. In other applications, providing a listener with some exposure to synthetic speech will permit the use of lower-cost voice response systems.

SIX-MONTH FOLLOW-UP EVALUATION

In addition to examining the effectiveness of training, we were also interested in the persistence of the training effects over some period of time. To investigate this problem, we retested subjects from the natural and Votrax groups six months after training was completed. At that time, tests were conducted to assess performance and detail the possible changes that may have occurred.

Method

Subjects—Only five of the original subjects from each group were willing or able to participate in this follow-up evaluation. The subjects were paid for their participation as in the earlier study. All 10 subjects reported that they had had no exposure to synthetic speech during the six-month interval.

Materials and procedure: In the six-month follow-up, subjects were again presented with synthetic speech produced by the Votrax Type-'N-Talk and were tested with the same procedures used for the pretest and the post-test. The subjects transcribed words from two 50-item PB lists, two sets of 10 Harvard sentences, and two sets of 10 Haskins sentences. Subjects were also given two sets of 50 words from the Modified Rhyme Test. Prose passages were not used because of the failure to observe a consistent pattern of performance in the earlier experiment. All of the words and sentences were different tokens from the materials used six months earlier.

Results and Discussion—All responses were scored for accuracy in the same manner as described previously. The pretest (Day 1) and posttest (Day 10) data presented in this section are only from the 10 subjects who participated in the six-month follow-up. Table 4 summarizes the data for each of the different tests (i.e., MRT, PB lists, and Harvard and Haskins sentences).

The pattern of results shown in Table 4 is quite consistent across tasks. In general, the performance of the subjects trained on synthetic speech was faster and more accurate six months after training than on the pretest on Day 1 of the experiment. In addition, after six months, the performance of the subjects trained on synthetic speech was better than the performance of the subjects trained on natural speech. Finally, after six months, subjects trained on synthetic speech were less accurate than on the posttest on Day 10 of the experiment. This pattern of results is exemplified by performance on the PB lists. More detailed analyses of the six-month follow-up data are reported by Schwab, Nusbaum, and Pisoni (1983).

A two-way analysis of variance conducted on the mean accuracy scores for the PB lists revealed a significant Groups $\times$ Days interaction, $F(2,16) = 98.42, p < 0.0001$. The post hoc analyses indicated that, for the subjects who participated in the follow-up, both the natural and the Votrax groups showed significant changes in performance over the three test periods, $F(2,16) = 54.76, p < 0.0001$, and $F(2,16) = 454.16, p < 0.0001$, respectively. Newman-Keuls analyses indicated that both groups showed significant improvements in accuracy from Day 1 to Day 10 and that both groups showed a decrease in accuracy from Day 10 to six months later ($p < 0.05$). However, performance for the Votrax group was still significantly better after six months than it was on Day 1 ($p < 0.05$), indicating that performance did not drop back to baseline. Furthermore, simple-effects tests also showed that the Votrax group was better in transcribing isolated words than was the natural group on both Day 10, $F(1,11) = 92.08, p < 0.0001$, and six months later, $F(1,11) = 61.18, p < 0.0001$. This pattern of results suggests that although the effects of training decreased somewhat over six months, the members of the Votrax group retained a considerable amount of their experience with the synthetic output from the text-to-speech system.

The results of the six-month follow-up study demonstrate that, in general, subjects maintained the gains in performance that they had obtained as a result of training with the synthetic speech. The enhanced levels of performance six months after the completion of training relative to pretest performance demonstrates that subjects learned and retained fairly detailed information about the acoustic-phonetic properties of the Votrax text-to-speech system. These results are even more interesting considering that the subjects had no exposure to synthetic speech during the intervening six-month period. Thus, no opportunity occurred for the subjects to actively use the knowledge gained from listening to synthetic speech during the original training.

CONCLUSIONS

The results of the present study suggest several important conclusions about the effects of training. First, the effect of training appears to be localized to a specific improvement in the encoding of synthetic speech produced by the Votrax Type-'N-Talk. Clearly, subjects did not simply learn to perform the various experimental tasks better, because the subjects trained on natural speech showed little or no improvement in performance. Moreover, since different stimuli were presented on each day, subjects could not have memorized particular tokens from a fixed ensemble of stimuli. In addition, the training affected performance similarly with isolated words and words in sentences, and for closed and open response sets. The overall pattern of results indicates that subjects in the group trained on synthetic speech did not just learn special strategies for sophisticated guessing; that is, they did not simply learn to use linguistic knowledge or task constraints to improve recognition. Rather, subjects abstracted detailed information about the structural characteristics of the synthetic speech as represented in the rule system of this particular text-to-speech system.

The six month follow-up study demonstrated that subjects were able to retain many of the effects of training even after a relatively long intervening period in which they had no further contact with the synthetic speech. Thus, it appears that the training produced a relatively stable and long-term change in the perceptual encoding processes. Furthermore, it is likely that more extensive training would have produced even greater persistence of the training effects. If subjects had been trained to asymptotic levels of performance, the long-term effects of training might have been even more sizable than those observed here.

These findings suggest that low-cost, poor-quality, commercially available text-to-speech systems can be used to provide speech output in certain applications. If the cost of training is minimal compared with the cost of voice response systems, then for some well-defined applications a low-cost and relatively poor-quality text-to-speech system may provide acceptable performance even when the system is not used on a regular basis.

It is important to note here that although we have demonstrated that training can improve the perception of synthetic speech, we did not attempt to measure changes in the capacity demands imposed by processing the synthetic speech. Whether training of the sort we provided here would reduce the attentional limitations that Luce, Feustel, and Pisoni (1983) recently found for perception of synthetic speech remains to be examined in future research. In this connection, Hakkinen and Williges (1984) recently claimed that training and familiarity will reduce the capacity demands incurred by perception of synthetic speech. Indeed, they asserted that if Luce et al. (1983) had familiarized their subjects with the test vocabulary, no capacity limitations would have been found. We believe their claim is unwarranted because in the study by Luce et al., the synthetic speech that was used was, in fact, of very high quality and was highly intelligible compared with the synthetic speech used in the present investigation. Pisoni and Hunnicutt (1980) reported intelligibility scores of 93.1% for the synthetic words used by Luce et al. Thus, the inference that Hakkinen and Williges (1984) attempted to draw about the results of Luce et al. is based on the assumption that the subjects in the Luce et al. study may have had difficulty perceiving the synthetic speech due to a lack of familiarity. The subjects did not have any difficulty perceiving the synthetic speech. The assertion about the effects of familiarity is therefore not only misleading, it is factually incorrect given the earlier data reporting the intelligibility of these stimulus items. Furthermore, Luce and Pisoni (1983) have recently used synthetic words carefully matched in intelligibility to natural speech and have found capacity limitations similar to those reported earlier by Luce et al.

Given these results, it is clear that low intelligibility per se is not the primary source of the capacity limitations imposed by perception of synthetic speech. It is still an open question whether training will reduce the capacity demands of synthetic speech. The present training study examined changes in perception of synthetic speech in relatively easy tasks imposing little, if any, cognitive load on the listener. Obviously, future research should be directed towards the problems inherent in high-information-load tasks where the observer may be required to carry out several attention-demanding activities simultaneously. Work on these problems is currently under way in Our laboratory, and reports of these studies should be forthcoming.

Acknowledgments

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

REFERENCES

- Acoustical Society of America. Method for measurement of monosyllabic word intelligibility. New York: American National Standards Institute; 1960.
- Allen J. Linguistic-based algorithms offer practical text-to-speech systems. *Speech Technology*. 1981; 1:12–16.
- Brown, JI. Nelson-Denny Reading Test, Form D. Boston: Houghton-Mifflin; 1973.
- Carlson, R.; Granstrom, B.; Larsson, K. Evaluation of a text-to-speech system as a reading machine for the blind. (QPSR 2–3). Stockholm: Royal Institute of Technology; 1976. p. 9-13.
- Derrick, C.; Harris, DP.; Walker, B. Cooperative English Test: Reading Comprehension, Form 1B. Princeton, NJ: Educational Testing Service; 1960.
- Egan JP. Articulation testing methods. *Laryngoscope*. 1948; 58:955–991. [PubMed: 18887435]
- Farr, R., editor. Iowa Silent Reading Tests, Level 3, Form E. New York: Harcourt Brace Jovanovich; 1972.
- Gardener, EF.; Callis, R.; Merwin, JC.; Madden, R. Stanford Test of Academic Skills: Reading, College Level II-A. New York: Harcourt Brace Jovanovich; 1972.
- Grosjean F. Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*. 1980; 28:267–283. [PubMed: 7465310]
- Hakkinen MT, Williges BH. Synthesized warning messages: Effects of an alerting cue in single- and multiple-function voice synthesis systems. *Human Factors*. 1984; 26:185–195.
- 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]
- IEEE. IEEE recommended practice for speech quality measurements. (IEEE No. 297). New York: Author; 1969.
- Luce PA, Fuestel TC, Pisoni DB. Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*. 1983; 25:17–32. [PubMed: 6840769]
- Luce, PA.; Pisoni, DB. Capacity-demanding encoding of synthetic speech in serial-ordered recall. (Research on Speech Perception, Progress Report 9). Bloomington, IN: Indiana University; 1983. p. 295-309.
- Nelson, MJ.; Denny, EC. Nelson-Denny Reading Test for Colleges and High Schools, Forms A and B. New York: Houghton-Mifflin; 1930.
- Nye, PW.; Gaitenby, J. Haskins Laboratories Status Report on Speech Research. Vol. 38. New Haven, CT: Haskins Laboratories; 1974. The intelligibility of synthetic monosyllable words in short, syntactically normal sentences; p. 169-190.
- 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. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. New York: IEEE Press; 1980. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system; p. 572-575.
- Salasoo A, Pisoni DB. Sources of knowledge in spoken word identification. *Journal of Verbal Learning and Verbal Behavior*. 1985; 24:210–231.
- Schwab, EC.; Nusbaum, HC.; Pisoni, DB. Some effects of training on the perception of synthetic speech. (Research on Speech Perception, Progress Report 9). Bloomington, IN: Indiana University; 1983. p. 39-77.
- Slowiaczek L, Pisoni DB. Effects of practice on speeded classification of natural and synthetic speech. *Journal of the Acoustical Society of America*. 1982; 71:S95.
- Walker, LA. Ears: A simple auditory screening test. (Research on Speech Perception, Progress Report 8). Bloomington, IN: Indiana University; 1982. p. 319-323.

Figure 1.
Mean accuracy for Day 1 (pretest) and Day 10 (posttest) for transcribing isolated synthetic words (PB Lists).

TABLE 1

Tasks Administered During the Pretest (Day 1) and the Posttest (Day 10)

Stimuli	Response Type	Total Responses	Dependent Measures
2 PB lists	open class	100	exact correct word
2 MRT lists	6 AFC	100	accuracy and RT
1 Harvard list (10 sentences)	open class	50 content words	exact correct word
1 Haskins list (10 sentences)	open class	40 content words	exact correct word
3 prose passages	Y/N questions	30	accuracy and RT

TABLE 2

Pretest (Day 1) and Posttest (Day 10) Measures of Intelligibility and Comprehension of Synthetic Speech for the Three Groups.

		Votrax	Natural	Control
MRT—% correct	Day 1	63.4	62.4	62.5
	Day 10	80.3	65.1	59.7
MRT—RT (ms)	Day 1	1966	1902	2007
	Day 10	1508	1696	1908
Harvard—% correct	Day 1	42.0	42.0	34.4
	Day 10	77.8	38.6	34.4
Haskins—% correct	Day 1	26.7	26.1	20.3
	Day 10	76.9	50.4	40.8
Passages—% correct	Day 1	65.3	74.1	72.5
	Day 10	62.5	65.5	60.2
Passages—RT (ms)	Day 1	3342	3249	3601
	Day 10	2912	3126	3628

TABLE 3

Measures of Intelligibility and Comprehension for the Training Days (Days 2–9) for Natural and Synthetic Speech

		Votrax	Natural
PB—% correct	Days 2–3	35.6	94.6
	Days 4–5	48.4	95.4
	Days 6–7	56.3	97.1
	Days 8–9	66.9	97.4
Harvard—% correct	Days 2–3	59.7	99.9
	Days 4–5	70.0	99.1
	Days 6–7	80.1	99.6
	Days 8–9	87.0	99.6
Haskins—% correct	Days 2–3	53.7	98.0
	Days 4–5	61.6	98.7
	Days 6–7	72.2	99.6
	Days 8–9	81.2	99.1
Passages—% correct	Days 2–3	70.4	82.5
	Days 4–5	80.1	76.6
	Day 6	72.7	83.6
	Days 8–9	71.6	74.7
Passages—RT (ms)	Days 2–3	3029	2993
	Days 4–5	2882	2640
	Day 6	2620	2354
	Days 8–9	2827	2691

TABLE 4

Pretest (Day 1), Posttest (Day 10), and Six-Month Follow-Up Measures of Intelligibility of Synthetic Speech for the Two Groups Receiving Training

		Votrax	Natural
PB—% correct	Day 1	26.4	22.6
	Day 10	70.8	38.4
	6 months	58.0	31.8
MRT—% correct	Day 1	65.2	60.2
	Day 10	80.0	66.0
	6 months	76.0	64.8
MRT—RT (ms)	Day 1	1950	1916
	Day 10	1506	1812
	6 months	1483	1558
Harvard—% correct	Day 1	42.4	41.2
	Day 10	80.0	39.6
	6 months	76.6	63.2
Haskins—% correct	Day 1	30.5	24.5
	Day 10	78.0	46.5
	6 months	65.5	50.0