Constraints on the perception of synthetic speech generated by rule

HOWARD C. NUSBAUM and DAVID B. PISONI
Speech Research Laboratory, Indiana University, Bloomington, Indiana

Within the next few years, there will be an extensive proliferation of various types of voice response devices in human-machine communication systems. Unfortunately, at present, relatively little basic or applied research has been carried out on the intelligibility, comprehension, and perceptual processing of synthetic speech produced by these devices. On the basis of our research, we identify five factors that must be considered in studying the perception of synthetic speech: (1) the specific demands imposed by a particular task, (2) the inherent limitations of the human information processing system, (3) the experience and training of the human listener, (4) the linguistic structure of the message set, and (5) the structure and quality of the speech signal.

We are beginning to see the introduction of practical, commercially available speech synthesis and speech recognition devices. Within the next few years, these systems will be utilized for a variety of applications to facilitate human-machine communication and as sensory aids for the handicapped. Soon we will converse with vending machines, cash registers, elevators, cars, clocks, and computers. Pilots will request and receive information by talking and listening to flight instruments. In short, speech technology will provide the ability to interact rapidly with machines. However, although there has been a great deal of attention paid to the development of the hardware and systems, there has been almost no effort made to understand how humans will utilize this technology. To date, there has been very little research concerned with the impact of speech technology on the human user. The prevailing assumption seems to be that simply providing automated voice response and voice data entry will solve most of the human factors problems inherent in the user-system interface. At present, this assumption is untested. In some cases, the introduction of voice response and voice data entry systems may create a new set of human factors problems.

To understand how the user will interact with these new speech processing devices, it is necessary to understand much more about the human observer. In other words, we must understand how the human processes information. More specifically, we must know how the human perceives, encodes, stores, and retrieves speech and how

these operations interact with the specific tasks the observer must perform.

In the Speech Research Laboratory at Indiana University, research projects have been directed at investigating various aspects of perception of synthetic speech generated automatically by rule using several text-to-speech systems (see Nusbaum & Pisoni, 1984, Nusbaum, Schwab, & Pisoni, 1983, and Pisoni, 1982). Strictly speaking, this work is not human factors research; that is, it is not designed to answer specific questions regarding the development and use of specific products. Rather, the goal is to provide more basic knowledge about the perception of synthetic speech. This research can then serve as a foundation for subsequent human factors studies that may be motivated by specific problems. In general, this research is concerned with the ability of human listeners to perceive synthetic speech under various task demands and conditions. In addition, we have carried out several basic comparisons of the performance of human listeners on standardized tasks with synthetic speech generated by various text-to-speech systems.

CONSTRAINTS ON HUMAN PERFORMANCE

To interpret the results of evaluation studies, it is necessary to consider some of the basic factors that may interact to affect an observer's performance: (1) the specific demands imposed by a particular task, (2) the inherent limitations of the human information processing system, (3) the experience and training of the human listener, (4) the linguistic structure of the message set, and (5) the structure and quality of the speech signal.

Task Complexity

The first factor that constrains performance concerns the complexity of the tasks that engage an observer during the perception of speech. In some tasks, the response demands are relatively simple, such as deciding which of two known words was spoken. Other tasks are extremely

This research was supported in part by NIH Grant NS-12179, in part by Contract No. AF-F 33615-83-K-0501 with the Air Force Systems Command, AFOSR through the Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, OH, and in part by a contract from Digital Equipment Corporation with Indiana University. We thank Beth Greene for her assistance in preparing this paper. Requests for reprints should be sent to Howard C. Nusbaum, Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405.

complex, such as trying to recognize an unknown utterance from a virtually unlimited number of response alternatives, while engaging in an activity that already requires attention. There is a substantial amount of research in cognitive psychology and human factors that demonstrates the powerful effects of perceptual set, instructions, subjective expectancies, cognitive load, and response set on performance in a variety of perceptual and cognitive tasks. The amount of context and the degree of uncertainty in the task also strongly affect an observer's performance in substantial ways.

Limitations on the Observer

The second factor influencing recognition of synthetic speech concerns the substantial limitations on the human information processing system's ability to perceive, encode, store, and retrieve information. Because the nervous system cannot maintain all aspects of sensory stimulation (and therefore must integrate acoustic energy over time), very severe processing limitations have been found in the capacity to encode and store raw sensory data in the human memory system. To overcome these capacity limitations, the listener must rapidly transform sensory input into more abstract neural codes for more stable storage in memory and subsequent processing operations. The bulk of the research on cognitive processes over the last 25 years has identified human short-term memory (STM) as a major limitation on processing sensory input (Shiffrin, 1976). The amount of information that can be processed in and out of STM is severely limited by the listener's attentional state, past experience, and the quality of the sensory input.

Experience and Training

The third factor concerns the ability of human observers to quickly learn effective cognitive and perceptual strategies to improve performance in almost any sort of task. When given appropriate feedback and training, subjects can learn to classify novel stimuli, remember complex pattern sequences, and respond to rapidly changing stimulus patterns in different sensory modalities. Clearly, the flexibility of subjects in adapting to the specific demands of a task is an important constraint that must be evaluated, or at least controlled in any attempt to evaluate synthetic speech.

Message Set

The fourth factor relates to the structure of the message set, that is, to the constraints on the number of possible messages and the organization and linguistic properties of the message set. This linguistic constraint depends on the listener's knowledge of language.

Signal Characteristics

In comparison, the fifth factor refers to the acoustic-phonetic and prosodic structure of a synthetic utterance. This constraint refers to the veridicality of the acoustic

properties of the synthetic speech signal compared with naturally produced speech.

Speech signals may be thought of as the physical consequence of a complex and hierarchically organized system of linguistic rules that map sounds onto meanings and meanings back onto sounds. At the lowest level in the system, the distinctive properties of the speech signal are constrained in substantial ways by vocal tract acoustics and articulation. The choice and arrangement of speech sounds into words is constrained by the phonological rules of language; the arrangement of words in sentences is constrained by syntax; and finally, the meaning of individual words and the overall meaning of sentences in a text are constrained by semantics and pragmatics. The contribution of these various levels of linguistic structure to perception will vary substantially from isolated words, to sentences, to passages of fluent continuous speech. In addition to linguistic structure, the ambient noise level and the spectrotemporal properties of noise in the environment in which the speech signal occurs will also affect recognition.

PERCEPTUAL EVALUATION OF SYNTHETIC SPEECH

There are basically three areas in which a text-to-speech system could be deficient that would impact the overall intelligibility of the speech: (1) the spelling-to-sound rules, (2) the computation and production of suprasegmental information, and (3) the phonetic implementation rules that convert the internal representation of phonemes and/or allophones into a speech waveform. In previous research, we found that phonetic implementation rules are a major factor in determining the segmental intelligibility of a voice response system (Nusbaum & Pisoni, 1982). The task that is generally used as a standard measure of the segmental intelligibility of speech is the Modified Rhyme Test (MRT), in which subjects are asked to identify a single word by choosing one of six alternative response words differing by a single phoneme in either initial or final position (House, Williams, Hecker, & Kryter, 1965). All the stimuli in the MRT are consonant-vowel-consonant (CVC) words; on half the trials, the responses share the VC of the stimulus and on the other half, the responses share the CV. Thus, the MRT provides a measure of the ability of listeners to identify either the initial or final phoneme of a set of spoken words. To date, we have evaluated natural speech and speech produced by four different text-to-speech systems: the Votrax Type-'n-Talk, the Speech Plus Prose-2000, the MITalk-79 research system, and DECTalk (Greene, Manous, & Pisoni, 1984). Word identification performance for natural speech was the best at 99.4% correct. For DECTalk, we evaluated speech produced by Paul and Betty (two of DECTalk's nine voices) and found different levels of performance—96.7% of the words spoken by the Paul voice were identified correctly, whereas only 94.4% of Betty's words were iden-

tified correctly. However, the level of performance for the Paul voice comes quite close to natural speech and is considerably higher than performance for any other text-to-speech system we have studied to date. Performance on MITalk-produced speech was somewhat lower than that on either of the DECTalk voices—93.1% correct word identification. The prototype of the Prose-2000 produced speech that was identified at 87.6% correct, although the current working version of the Prose-2000 is slightly improved, with performance at 91.1% correct. Finally, the least intelligible synthetic speech was produced by the Votrax Type-'n-Talk—67.2% correct word identification. These results, obtained under closely matched testing conditions, show a wide range of variation among currently available text-to-speech systems that seems to reflect the amount of basic research that was carried out to develop the phonetic implementation rules of these different voice response systems.

In addition to these tasks, we have used an open-response format version of the MRT, in which listeners are instructed simply to write the word that was heard on each trial. This open-response format provides a measure of performance when constraints on the response set are minimized (compared with the six-alternative forced-choice version), and it also provides information about the intelligibility of the vowels that is not available in the closed-response-set version of the MRT. A comparison of the closed- and open-response versions of the MRT for speech produced by different text-to-speech systems with natural speech indicates the degree to which listeners rely on response-set constraints. Performance on the open-response-set MRT for natural speech was at 97.2% correct exact word identification, compared with 99.4% correct in the closed-response-set task. Even when there are no strong constraints on the number of alternative responses for natural speech, performance is better than for any text-to-speech system with a constrained set of responses. For the MITalk-79 research system, performance in the open-set task was considerably worse at 75.4% correct. Similarly, DECTalk's Paul voice produced words that were identified at the 86.7% level. These results show a large and reliable interaction between intelligibility measured in the closed-response format MRT and the open-response format MRT. Even though the rank ordering of intelligibility stays the same across the two forms of the MRT, it is clear that as speech becomes less intelligible, listeners rely more heavily on response-set constraints to aid performance.

To examine the contribution of linguistic constraints on performance, we compared word recognition in two types of sentence contexts. The first type of sentence context was syntactically correct and meaningful—the Harvard psychoacoustic sentences (Egan, 1948). An example is: Add salt before you fry the egg.

The second type of sentence context was syntactically correct, but these sentences were semantically anomalous—the Haskins syntactic sentences (Nye & Gaitenby,

Table 1
Percent Correct Word Identification for Meaningful and Semantically Anomalous Sentence Contexts

Type of Speech	Type of Sentence Context	
	Meaningful	Anomalous
Natural	99.2%	97.7%
MITalk-79	93.3%	78.7%
Prose-2000	83.7%	64.5%
DEC Paul	95.3%	86.8%
DEC Betty	90.5%	75.1%

1974). These test sentences had the syntactic form of normal sentences, but they were nonsense. An example of this type of nonsense sentence is: The old farm cost the blood.

By comparing word recognition performance for these two classes of sentences, it was possible to determine the influence of sentence meaning and linguistic constraints on word perception (Greene et al., 1984). Table 1 shows percent correct word identification for meaningful and semantically anomalous sentences for natural speech and synthetic speech produced by MITalk-79, the Speech Plus Prose-2000 prototype, and for DECTalk's Paul and Betty voices. For natural and synthetic speech, word recognition was much better in meaningful sentences than in the semantically anomalous sentences. Furthermore, a comparison of correct word identification in these sentences reveals an interaction in performance such that semantic constraints are relied on by listeners much more for less intelligible speech.

CAPACITY LIMITATIONS AND PERCEPTUAL ENCODING

The results of the MRT and word identification studies of natural and synthetic speech clearly indicate that synthetic speech is less intelligible than natural speech. In addition, these studies demonstrate that as synthetic speech becomes less intelligible, listeners rely more on linguistic and response-set constraints to aid word identification. However, these studies do not account for why this difference in perception of natural and synthetic speech exists. In order to address this issue, we carried out a series of experiments that were aimed at measuring the time required to recognize natural and synthetic words and permissible nonwords. In carrying out these studies, we wanted to know how long it takes a human listener to recognize an isolated word and how the process of word recognition might be affected by the quality of the acoustic-phonetic information in the signal. To measure how long it takes an observer to recognize isolated words, Pisoni (1981; Slowiaczek & Pisoni, 1982) used a lexical decision task. Subjects were presented with a word or a nonword stimulus item on each trial. The listener was required to classify the item as either a "word" or a "nonword" as fast as possible by pressing one of two buttons located on a response box.

The mean response times for correct responses showed significant differences between synthetic and natural test items. Subjects responded significantly faster to natural words (903 msec) and nonwords (1,046 msec) than to synthetic words (1,056 msec) and nonwords (1,179 msec). On the average, response times to the synthetic speech were 145 msec longer than response times to the natural speech. These findings demonstrate that the perception of synthetic speech requires more cognitive "effort" than the perception of natural speech. This difference was observed for both words and nonwords alike, suggesting that the extra processing does not depend on the lexical status of the test item. Thus, the phonological encoding of synthetic speech appears to require more "effort" or resources than the encoding of natural speech.

Similar results were obtained by Pisoni (1982) for naming latencies for natural and synthetic words and nonwords. As in the lexical decision task, subjects were much slower to name synthetic test items than natural test items, for both words and nonwords. These results demonstrate that the extra processing time needed for synthetic speech does not depend on the type of response made by the listener since the results were comparable for both manual and vocal responses. Early stages of encoding of synthetic speech require more processing time than encoding of natural speech.

This conclusion receives further support from Luce, Feustel, and Pisoni (1983), whose experiments were designed to study the effects of processing synthetic speech on the capacity of short-term memory. In one study, subjects were given a visual digit string to remember followed by a list of 10 natural or 10 synthetic words. The most important finding was an interaction for recall performance between the type of speech presented (synthetic vs. natural) and the number of visual digits presented (three vs. six). Synthetic speech impaired recall of the visually presented digits more with increasing digit list size than did natural speech. These results demonstrate that synthetic speech required more short-term memory capacity than natural speech.

In another experiment, Luce et al. (1983) presented subjects with lists of 10 natural words or 10 synthetic words to be memorized and recalled in serial order. Overall, natural words were recalled better than synthetic words. However, an interaction was obtained such that there was a significantly decreased primacy effect for recall of synthetic words compared with natural words. This result suggests that, in the synthetic lists, the words presented later in each list interfered with active maintenance of the words presented earlier. This is precisely the result that would be expected if the perceptual encoding of the synthetic words placed an additional load on short-term memory, thus impairing the rehearsal of words presented in the first half of the list.

These studies suggest that the problems in perception of synthetic speech are tied largely to the processes that encode and recognize the acoustic-phonetic structure of words. Recently, Slowiaczek and Nusbaum (1983) found

that the contribution of suprasegmental structure to the intelligibility of synthetic speech was quite small compared with the effects of degrading acoustic-phonetic structure. Therefore, it appears that much of the difference in intelligibility of natural and synthetic speech is probably the result of the phonetic implementation rules that convert symbolic phonetic strings into the time-varying speech waveform.

ACOUSTIC-PHONETIC STRUCTURE AND PERCEPTUAL ENCODING

Several hypotheses can account for the greater difficulty of encoding synthetic utterances. One hypothesis is that synthetic speech is simply equivalent to "noisy" natural speech. That is, the acoustic-phonetic structure of synthetic speech is hard to encode for the same reasons that natural speech presented in noise is hard to perceive—the basic cues are obscured, masked, or physically degraded in some way. According to this view, synthetic speech is on the same continuum as natural speech, but is degraded in comparison with natural speech. In contrast, an alternative hypothesis is that synthetic speech is not "noisy" or degraded speech, but instead may be thought of as "perceptually impoverished" relative to natural speech. By this account, synthetic speech is different from natural speech in both degree and kind.

Spoken language is structurally rich and redundant at all levels of linguistic analysis, and it is clear that listeners will make use of the linguistic redundancy that can be provided by semantics and syntax to aid in the perception of speech (see Pisoni, 1982). In addition, natural speech is highly redundant at the level of acoustic-phonetic structure. In natural speech, there are many acoustic cues that change as a function of context, speaking rate, and talker. However, in synthetic speech, only a small subset of the possible cues are implemented as phonetic production rules. As a result, some phonetic distinctions may be minimally cued, perhaps by only a single acoustic attribute. If all cues do not have equal importance in different phonetic contexts, a single cue may not be perceptually sufficient to convey a particular phonetic distinction in all utterances. Moreover, the reliance of synthetic speech on a minimal cue set could be disastrous if a particular cue is incorrectly synthesized or masked by environmental noise.

These two hypotheses about the encoding problems encountered in perceiving synthetic speech make different predictions about the types of errors and the distribution of perceptual confusions that should be obtained with synthetic speech. According to the "noisy speech" hypothesis, synthetic speech is similar to natural speech that has been degraded by the addition of noise. Therefore, the perceptual confusions that occur with synthetic speech should be very similar to those obtained with natural speech heard in noise. The "impoverished speech" hypothesis, however, claims that the acoustic-phonetic structure of synthetic speech is not as rich in segmental

cues as natural speech. According to this hypothesis, two types of confusion errors should occur in the perception of synthetic speech. When the acoustic cues used to specify a phonetic segment are not sufficiently distinctive, confusions should occur between minimally cued segments that are phonetically similar. This type of error should be similar to the errors predicted by the noisy speech hypothesis, since perceptual confusions of natural speech in noise also depend on the acoustic-phonetic similarity of the segments (Miller & Nicely, 1955; Wang & Bilger, 1973). However, the two hypotheses may be distinguished by the second type of error that is predicted only by the impoverished speech hypothesis. If the minimal acoustic cues used to signal phonetic segments are incorrect or contradictory, then confusions that are not based on the nominal acoustic-phonetic similarity of the confused segments should occur. Instead, these confusions should be entirely determined by the perceptual interpretation of the misleading cues and therefore should result in confusions of segments that are phonetically quite different from the intended ones.

In order to investigate the predictions made by these two hypotheses, we carried out an experiment to directly measure the confusions that arise within a set of natural and synthetic phonetic segments (Nusbaum, Dedina, & Pisoni, 1984). We used 48 CV syllables as stimuli constructed from the vowels /i, a, u/ and the consonants /b, d, g, p, t, k, n, m, r, l, w, j, s, f, z, v/. These syllables were produced by a male talker and by three text-to-speech systems—the Votrax Type-*n*-Talk, the Speech Plus Prose-2000, and the Digital Equipment Corporation DEC-Talk. To assess the type of perceptual confusions that occur for natural speech, the natural syllables were presented at four signal-to-noise ratios of +28, 0, -5, and -10 dB.

When averaged over vowel contexts, the results showed that natural speech at +28 dB S/N was the most intelligible (96.6% correct), followed by DECTalk (92.2% correct), followed by the Prose-2000 (62.8% correct), with the Type-*n*-Talk finishing last (27.0% correct). Of special interest were the results of more detailed error analyses, which revealed that the distributions of perceptual confusions obtained for natural and synthetic speech were often quite different. For example, in the case of DECTalk, 100% of the errors made in identifying the segment /r/ were due to confusions with /b/ even though this type of error never occurred for natural speech at +28 dB S/N. Even at the worst S/N (-10 dB), for which the intelligibility of natural speech (29.1% correct) was actually worse than DECTalk (92.2% correct), this type of error accounted for only 3% of the total errors made on this segment.

In order to compare the segmental confusions that occurred for natural and synthetic speech, we examined the confusion matrices for synthetic speech and natural speech presented at a signal-to-noise ratio that resulted in comparable overall levels of identification performance. The confusion matrices for the Prose-2000 were compared with natural speech presented at 0 dB S/N and the con-

fusion matrices for Votrax with natural speech presented at -10 dB S/N. An examination of the proportion of the total errors contributed by each response class (stop, nasal, liquid/glide, fricative, other) indicated that, for natural speech, most of the errors in identifying stops were due to responses that were other stop consonants. In contrast, the errors found with the Prose-2000 appeared to be more evenly distributed between stop, liquid/glide, and fricative responses. In other words, more intrusions appeared from other manner classes in the errors observed with the Prose-2000 synthetic speech.

The comparison between natural speech at -10 dB S/N and Votrax speech indicated that the pattern of errors in identifying stops was more similar for these conditions. Indeed, the comparison of identification errors for natural speech at 0 and -10 dB S/N was quite similar to the comparison between Votrax and natural speech. At least for the perception of stop consonants, the confusions of Votrax speech seem to be based on the acoustic-phonetic similarity of the confused segments, as in noisy speech. Moreover, the overall performance level for Votrax speech was low to begin with. However, the different pattern of errors obtained for Prose-2000 and natural speech suggests that the errors produced by the Prose-2000 may be phonetic miscues rather than true phonetic confusions.

A very different pattern of results was obtained for the errors that occurred in the perception of liquids and glides. The distribution of errors for Prose-2000 speech and natural speech revealed that similar confusions were made for liquids and glides for both types of speech. However, the results were quite different for a comparison of Votrax speech and natural speech for these phonemes. For liquids and glides, the largest number of errors for Votrax speech resulted from confusions with stop consonants, whereas, for natural speech, relatively few stop responses were observed. Thus, for liquids and glides, errors in perception of Prose-2000 speech seem to be based on acoustic-phonetic similarity, whereas the errors for Votrax speech seem to be phonetic miscues.

On the basis of these confusion analyses, the predictions made by the noisy speech hypothesis are incorrect. Two different types of errors have been observed in the perception of synthetic speech. Some consonant identification errors were based on the acoustic-phonetic similarity of the confused segments. Others followed a pattern that can only be explained as the result of phonetic miscues in which the acoustic structure specified the wrong segment in a particular context.

These results support the conclusion that the differences in perception of natural and synthetic speech are largely the result of differences in the acoustic-phonetic structure of the signals. More recently, we have found further support for this hypothesis using the gating paradigm (Grosjean, 1980; Salasoo & Pisoni, 1985) to investigate perception of the acoustic-phonetic structure of natural and synthetic words. Listeners were presented with short segments of spoken words for identification. On the first trial with a particular word, the first 50 msec of the word was

presented. On subsequent trials, the amount of stimulus was increased in 50-msec steps so that, on the next trial, 100 msec of the word was presented, and on the next trial 150 msec of the word was heard, and so on, until the entire word had been presented. We found that, on the average, natural words could be identified after 67% of a word had been heard, whereas, for synthetic words, it was necessary for listeners to hear 75% of a word for correct word identification. These results demonstrate that the acoustic-phonetic structure of synthetic words conveys less information (per unit of time) than the acoustic-phonetic structure of natural speech (see Manous & Pisoni, 1984).

IMPROVING INTELLIGIBILITY OF SYNTHETIC SPEECH

The human observer is a very flexible processor of information. With sufficient experience, practice, and specialized training, observers may be able to overcome some of the limitations on performance observed in our previous studies. Indeed, several researchers (e.g., Pisoni & Hunnicutt, 1980) reported a rapid improvement in recognition of synthetic speech during the course of their experiments. These improvements appear to have been the result of subjects' learning to process the acoustic-phonetic structure of synthetic speech more effectively. However, it is also possible that the reported improvements in intelligibility of synthetic speech were actually due to an increased familiarity with the experimental procedures rather than to an increased familiarity with the synthetic speech. In order to test these alternatives, we conducted an experiment to separate the effects of training on task performance from improvements in the recognition of synthetic speech (Schwab, Nusbaum, & Pisoni, *in press*).

Three groups of subjects were given a pretest with synthetic speech on Day 1 of the experiment and a posttest with synthetic speech on Day 10. The pretest determined baseline performance for the Votrax Type-'n-Talk text-to-speech system, and the posttest on Day 10 was given to determine whether any improvements had occurred in recognition for the synthetic speech. We selected the low-cost Votrax system primarily because of the poor quality of its segmental synthesis. Thus, ceiling effects would not obscure any effects of training and there would be room for improvement.

The three groups of subjects were treated differently on Days 2-9. One group received training with Votrax synthetic speech. One group was trained with natural speech using the same words, sentences, and paragraphs as the group trained on synthetic speech. This second group served to control for familiarity with the specific experimental tasks. Finally, a third group received no training at all on Days 2-9.

On the pretest and posttest days, the subjects were given the MRT, isolated phonetically balanced (PB) words, and sentences for transcription. The word lists were taken

from PB lists; the sentences were both meaningful and semantically anomalous sentences. Subjects were given different materials on every day. During all the training sessions, subjects were presented with spoken words and sentences, and received feedback indicating the correct response on each trial.

The results showed that performance improved dramatically for only one group—the subjects that were trained with the Votrax synthetic speech during Days 2-9. At the end of training, the Votrax-trained group showed significantly higher levels of performance than the other two groups with these stimuli. For example, performance on identifying isolated PB words improved for the Votrax-trained group from about 25% correct to almost 70% correct word recognition. Similar improvements were found for all the word identification tasks.

The results of our training study suggest several important conclusions. First, the effect of training is apparently that of improving the encoding of synthetic words produced by the Votrax Type-'n-Talk. Clearly, subjects were not learning simply to perform the various tasks better, since the subjects trained on natural speech showed little or no improvement in performance. Moreover, training affected performance similarly with isolated words and words in sentences, and for closed and open response sets. This pattern of results indicates that subjects in the group trained on synthetic speech were not learning special strategies; that is, they were not learning to use linguistic knowledge or task constraints to improve recognition. Rather, subjects seem to have learned something about the structural characteristics of this particular synthetic speech system that enabled them to perform better regardless of the task. This conclusion is further supported by the design of the training study. Improvements in performance were obtained on novel materials even though the subjects never heard the same words or sentences twice. In order to show improvements in performance, subjects must have learned something about the detailed acoustic-phonetic properties of the synthetic speech produced by the system.

In addition, we found that subjects retained the training even after 6 months with no further contact with the synthetic speech. Thus, it appears that training produced a relatively stable and long-term change in the perceptual encoding processes used by subjects. Furthermore, it is likely that more extensive training would have produced 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 stable.

SUMMARY AND CONCLUSIONS

Our research has demonstrated that listeners rely more on the constraints of response-set size and linguistic context as the quality of synthetic speech becomes worse. Furthermore, this research has indicated that the segmental

intelligibility of synthetic speech is a major factor in word perception. The segmental structure of synthetic speech can be viewed as impoverished by comparison with the structural redundancy that is inherent in natural speech. Also, it appears that it is the impoverished acoustic-phonetic structure of synthetic speech that is responsible for the increased capacity demands that occur during perceptual encoding of synthetic speech.

For low-cost speech synthesis systems in which the quality of segmental synthesis may be poor, the best performance will be achieved when the set of possible messages is very small and the user is highly familiar with the message set. It may also be important for the different messages in the set to be maximally distinctive, such as the military alphabet (alpha, bravo, etc.). In this regard, the human user should be regarded in somewhat the same way as an isolated-word speech recognition system. Of course, this consideration becomes less important if the spoken messages are accompanied by a visual display of the same information. When the user can see a copy of the spoken message, any voice response system will seem, at first glance, to be quite intelligible. Although providing visual feedback may reduce the utility of a voice response device, a low-cost text-to-speech system could be used in this way to provide adequate spoken confirmation of data-base entries. In situations in which visual feedback cannot be provided and the messages are not restricted to a small predetermined set, extensive training or a more sophisticated text-to-speech system would be advisable.

Assessing the intelligibility of a voice response unit is an important part of evaluating any system for applications. But it is equally important to understand how the use of synthetic speech may interact with other cognitive operations carried out by the human observer. If the use of speech input/output interferes with other cognitive processes, performance of other tasks might be impaired if carried out concurrently with other speech processing activities. For example, a pilot who is listening to talking flight instruments using synthetic speech might miss a warning light, forget important flight information, or misunderstand the flight controller. Therefore, it is important to understand the capacity limitations imposed on human information processing by the use of synthetic speech.

Furthermore, it should be recognized that the ability to respond to synthetic speech in very demanding applications cannot be predicted from the results of the traditional forced-choice MRT. In the forced-choice MRT, the listener can utilize the response constraints inherent in the task, provided by the restricted set of alternatives. However, outside the laboratory, the observer is seldom provided with these constraints. There is no simple or direct method of estimating performance in less constrained situations from the results of the forced-choice MRT. Instead, evaluation of voice response systems should be carried out under the same task requirements that are imposed in the intended application.

From our research on the perception of synthetic speech, we have been able to specify some of the constraints on the use of voice response systems. However, there is still a great deal of research to be done. Basic research is needed to understand the effects of noise and distortion on the processing of synthetic speech, how perception is influenced by practice and prior experience, and how naturalness interacts with intelligibility. Now that the technology has been developed, research on these problems and other related issues will allow us to realize both the potential and the capabilities of voice response systems and to understand their limitations in various applications.

REFERENCES

- EGAN, J. P. (1948). Articulation testing methods. *Laryngoscope*, **58**, 955-991.
- GREENE, B. G., MANOUS, L. M., & PISONI, D. B. (1984). *Preliminary evaluation of DECTalk* (Tech. Note No. 84-03). Bloomington: Indiana University, Speech Research Laboratory.
- GROSJEAN, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, **19**, 267-283.
- HOUSE, A. S., WILLIAMS, C. E., HECKER, M. H. L., & KRYTER, K. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, **37**, 158-166.
- LUCE, P. A., FEUSTEL, T. C., & PISONI, D. B. (1983). Capacity demands in short-term memory for synthetic and natural word lists. *Human Factors*, **25**, 17-32.
- MANOUS, L. M., & PISONI, D. B. (1984). *Gating natural and synthetic words* (Research on Speech Perception Progress Report No. 10). Bloomington: Indiana University, Department of Psychology, Speech Research Laboratory.
- MILLER, G. A., & NICELY, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 338-352.
- NUSBAUM, H. C., DEDINA, M. J., & PISONI, D. B. (1984). *Perceptual confusions of consonants in natural and synthetic CV syllables* (Tech. Note No. 84-02). Bloomington: Indiana University, Speech Research Laboratory.
- NUSBAUM, H. C., & PISONI, D. B. (1982). Perceptual and cognitive constraints on the use of voice response systems. In *Proceedings of the 2nd Voice Data Entry Systems Applications Conference*. Sunnyvale, CA: Lockheed.
- NUSBAUM, H. C., & PISONI, D. B. (1984). Perceptual evaluation of synthetic speech generated by rule. In *Proceedings of the 4th Voice Data Entry Systems Applications Conference*. Sunnyvale, CA: Lockheed.
- NUSBAUM, H. C., SCHWAB, E. C., & PISONI, D. B. (1983). Perceptual evaluation of synthetic speech: Some constraints on the use of voice response systems. In *Proceedings of the 3rd Voice Data Entry Systems Applications Conference*. Sunnyvale, CA: Lockheed.
- NYE, P. W., & GAITENBY, J. (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. *Haskins Laboratories Status Report on Speech Research*, **38**, 169-190.
- PISONI, D. B. (1981). Speeded classification of natural and synthetic speech in a lexical decision task. *Journal of the Acoustical Society of America*, **70**, S98.
- PISONI, D. B. (1982). Perception of speech: The human listener as a cognitive interface. *Speech Technology*, **1**, 10-23.
- PISONI, D. B., & HUNNICUTT, S. (1980). Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. In *1980 IEEE International Conference Record on Acoustics, Speech, and Signal Processing* (pp. 572-575). New York: IEEE Press.
- SALASOO, A., & PISONI, D. B. (1985). Interaction of knowledge sources in spoken word identification. *Journal of Memory and Language*, **24**, 210-231.

- SCHWAB, E. C., NUSBAUM, H. C., & PISONI, D. B. (in press). Effects of training on the perception of synthetic speech. *Human Factors*.
- SHIFFRIN, R. M. (1976). Capacity limitations in information processing, attention, and memory. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (Vol. 4). Hillsdale, NJ: Erlbaum.
- SLOWIACZEK, L. M., & NUSBAUM, H. C. (1983). Intelligibility of fluent synthetic sentences: Effects of speech rate, pitch contour, and meaning. *Journal of the Acoustical Society of America*, **73**, S103.
- SLOWIACZEK, L. M., & PISONI, D. B. (1982). Effects of practice on speeded classification of natural and synthetic speech. *Journal of the Acoustical Society of America*, **71**, S95.
- WANG, M. D., & BILGER, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America*, **54**, 1248-1266.