

Appendix: Audio Demos

The sound example demonstrates our system.

43

Perception of Synthetic Speech

David B. Pisoni

ABSTRACT This chapter summarizes the results we obtained over the last 15 years at Indiana University on the perception of synthetic speech produced by rule. A wide variety of behavioral studies have been carried out on phoneme intelligibility, word recognition, and comprehension to learn more about how human listeners perceive and understand synthetic speech. Some of this research, particularly the earlier studies on segmental intelligibility, was directed toward applied issues dealing with perceptual evaluation and assessment of different synthesis systems. Other aspects of the research program have been more theoretically motivated and were designed to learn more about speech perception and spoken language comprehension. Our findings have shown that the perception of synthetic speech depends on several general factors including the acoustic-phonetic properties of the speech signal, the specific cognitive demands of the information-processing task the listener is asked to perform, and the previous background and experience of the listener. Suggestions for future research on improving naturalness, intelligibility, and comprehension are offered in light of several recent findings on the role of stimulus variability and the contribution of indexical factors to speech perception and spoken word recognition. Our perceptual findings have shown the importance of behavioral testing with human listeners as an integral component of evaluation and assessment techniques in synthesis research and development.

43.1 Introduction

My interest in the perception of synthetic speech dates back to 1979 when the MITalk text-to-speech system was nearing completion [AHK87]. At that time, a number of people, including Dennis Klatt, Sheri Hunnicutt, Rolf Carlson, and Bjorn Granstrom, were working in the Speech Group at MIT on various aspects of this system. Given my earlier research on speech perception, it seemed quite appropriate to carry out a series of perceptual studies with human listeners to assess how good the MITalk synthetic speech actually was and what differences, if any, would be found in perception between natural speech and synthetic speech produced by rule. Since that time, my research group at Indiana has conducted a large number of experiments to learn more about the perception and comprehension of synthetic speech produced by rule. This chapter provides a summary and interpretation of the major findings obtained over the last 15 years and some suggestions for future research directions.

TABLE 43.1. Needed research on the perception of synthetic speech [Pis81a].

- | | |
|-----|--|
| 1. | Processing time experiments |
| 2. | Listening to synthetic speech in noise |
| 3. | Perception under differing attentional demands |
| 4. | Effects of short- and long-term practice |
| 5. | Comprehension of fluent synthetic speech |
| 6. | Interaction of segmental and prosodic cues |
| 7. | Comparisons of different rule systems and synthesizers |
| 8. | Effects of naturalness on intelligibility |
| 9. | Generalization to novel utterances |
| 10. | Effects of message set size |

In placing our earlier work in context, it is important at the outset to draw a distinction between basic research on the perception of synthetic speech and more applied or practical work dealing with questions concerning assessment and the development of evaluation techniques. Although we have been involved with both kinds of activities, most of our research has been oriented toward basic research issues that deal with the perceptual analysis of speech. In particular, we have been concerned with trying to understand some of the important differences in perception between natural speech and several kinds of synthetic speech. By studying the perception of synthetic speech produced by rule, we hoped to learn more about the mechanisms and processes used to perceive and understand spoken language more generally [Kla87, PNG85].

A few years after this research program began, I generated a list of about a dozen basic issues that seemed at the time to be important topics for future research [Pis81a]. Table 43.1 lists these research issues. Many of these questions have been studied over the years since I constructed this list, but some of the topics still remain to be explored in future research.

A number of researchers have taken our initial set of findings and developed much more detailed assessment and evaluation techniques to test various types of voice output devices [Pol89a, Pol92, van94]. The goal of this recent work has been to develop reliable methods of evaluation and assessment so that standards can be formulated for use in a variety of languages across a range of applications [BB92, FHBH89]. One approach to assessment was proposed recently by Pols [Pol89b]. He suggests that assessment techniques be categorized into four broad classes: (1) *global*, including acceptability, preference, naturalness, and usefulness; (2) *diagnostic*, including segmentals, intelligibility, and prosody; (3) *objective*, including the speech transmission index (STI) and articulation index (AI); and (4) *application-specific*, including newspapers, reading machines, telephone information services, and weather briefings. Much of our early research on the perception of synthetic speech was concerned with global and diagnostic issues, topics that continue to occupy most researchers even today.

Although some aspects of our research have been concerned with evaluation and assessment, such as the intelligibility studies using the modified rhyme test (MRT),

many other studies over the years have focused on why some types of synthetic speech are difficult to perceive and understand and how listeners compensate for the generally poor-quality acoustic-phonetic information in the signal. In the sections below, I provide a brief summary of the major findings and conclusions from our research program. Both the perceptual evaluation studies and the experimental work have suggested a number of general conclusions about the factors that affect the perception of synthetic speech. Finally, I offer several suggestions for future research.

43.2 Intelligibility of Synthetic Speech

A text-to-speech system can generate three different kinds of errors that may affect the overall intelligibility of the speech. These errors include incorrect spelling-to-sound rules, the computation and production of incorrect or inappropriate suprasegmental information, and the use of error-prone phonetic implementation rules that are used to convert the internal representation of allophones into a speech waveform [AHK87, PNG85]. In the studies described below, my collaborators and I have focused much of our attention on measures of segmental intelligibility, assuming that the letter-to-sound rules used by a particular text-to-speech system were applied correctly. For most of our research, we simply ignored the suprasegmentals because at the time this work was initially carried out in the late 1970s, there were no behavioral techniques available to assess these attributes of synthetic speech.

Phoneme Intelligibility. The task that has been used most often in previous studies evaluating synthetic speech and the one we adopted to measure the segmental intelligibility was the modified rhyme test. In the MRT, subjects are required to identify a single English word by choosing one of six alternative responses that differ by a single phoneme in either initial or final position [HWHK65, NG74]. All the stimuli in the MRT are consonant-vowel-consonant (CVC) monosyllabic English words; on half the trials, the response alternatives share the vowel-consonant portion of the stimulus, and on the other half the response alternatives share the consonant-vowel portion. Thus, the MRT provides a measure of how well listeners can identify either the initial or the final phoneme from a set of spoken words. In recent years, new tests specifically for the assessment of synthetic speech have been developed using this approach [BP89, CG89, SAMW89]. Some examples of data obtained using in the MRT from [LGP89] for 10 text-to-speech systems are shown in figure 43.1. The intelligibility data shown here reveal a wide range of performance levels across different synthesis systems.

In addition to the standard forced-choice closed-response MRT, we have also explored the use of an open-response format. In this procedure, listeners are instructed simply to write down the word that they heard on each trial. The open-response test provides a measure of performance that minimizes the constraints on the response set; that is, all CVC words known to the listener are possible responses compared

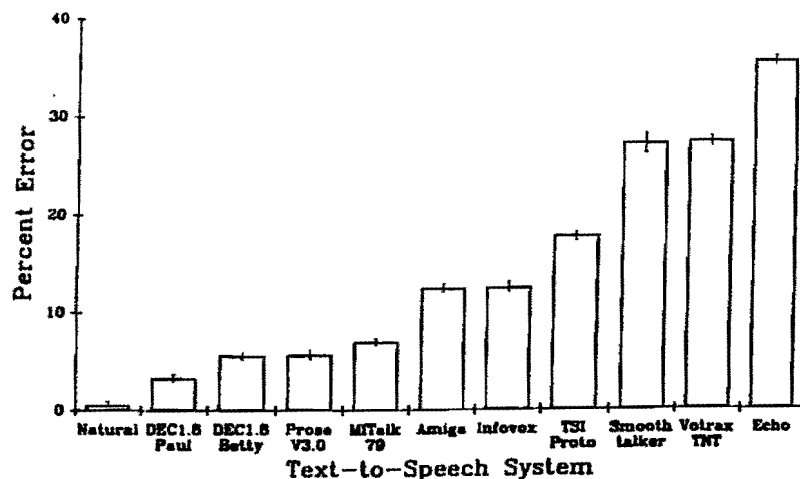


FIGURE 43.1. Overall error rates (in percent) for each of the 10 text-to-speech systems tested using the MRT. Natural speech is included here as a benchmark (from [LGP89]).

to the six alternative responses in the standard closed-response MRT. This open-set procedure also provides information about the intelligibility of vowels that is not available in the closed-response set version. Open-set tests require more cognitive effort and attention because the listener must first encode the auditory stimulus, search through his/her lexicon for one or more appropriate words, and then finally select one word as the match for the auditory stimulus [SKPO94]. By comparing the results obtained in the closed- and open-response versions of the MRT, we were able to obtain a great deal of useful information about the sources of error in a particular system [LGP89]. Indeed, detailed analyses of the stimulus-response confusions provided knowledge about the specific rules used to generate segmental contrasts in particular phonetic environments and how to modify them to improve intelligibility [Kla87].

Our results have shown large differences in segmental intelligibility between the open- and closed-response formats. Although the rank-ordering of intelligibility remains the same across the two forms of the MRT, it is clear that as speech becomes less intelligible, listeners rely more heavily on the response alternatives provided by the closed-set format to help their performance on this task [MHL51]. Comparisons between open- and closed-set performance are shown in figure 43.2 for 10 synthesis systems.

Nonnative Speakers of English. We have carried out several studies in which nonnative speakers of English listened to both natural and synthetic speech materials [Gre86]. Nonnative speakers reveal essentially the same pattern of results found for native speakers: Their performance is better when listening to natural speech than synthetic speech. Results were obtained for intelligibility of isolated words in the MRT task and for word recognition in sentences using a transcription

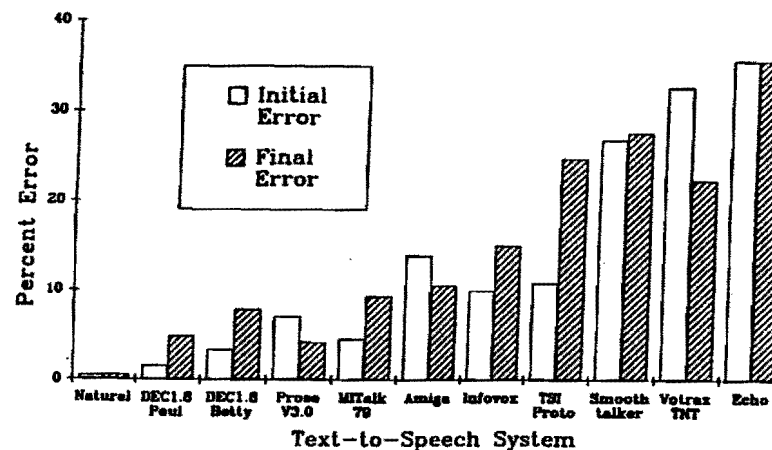


FIGURE 43.2. Error rates (in percent) for 10 systems tested using both the closed- and open-response format MRT. Open bars designate error rates for the closed-response format, and striped bars designate error rates for the open-response format (from [LGP89]).

task. However, the absolute levels of performance were substantially lower for these listeners than for native English speakers using the same materials.

Word Recognition in Sentences: Transcription Performance. To examine the contribution of sentence context and linguistic constraints on performance, we studied word recognition in two types of sentences: syntactically correct and meaningful sentences—"Add salt before you fry the egg"—and syntactically correct but semantically anomalous sentences—"The old farm cost the blood." Subjects listen to each sentence and simply write down what they hear on each trial.

In comparing word recognition performance for these two types of sentences, we found large and consistent effects of meaning and linguistic constraints on word recognition. For both natural and synthetic speech, word recognition was always better in meaningful sentences than in the semantically anomalous sentences. Not surprisingly, meaningful sentences narrow down response alternatives and help listeners understand words in context. Both top-down and bottom-up processes are required to carry out this transcription task. Furthermore, a comparison of correct word identification in these sentences revealed an interaction in performance, suggesting that semantic constraints are relied on much more by listeners as the speech becomes progressively less intelligible [Pis87, PH80]. Subjects also have great difficulty inhibiting the use of semantic constraints in word recognition even when it is not helpful to them. Some interesting examples of these response

TABLE 43.2. Consonant class labels used for American English.

Harvard and Haskins Sentence Targets and Examples of Responses

Harvard Sentences:

TARGET:

The juice of lemons makes fine punch.

RESPONSES:

The juice of lemons makes Hawaiian punch.

The goose lemon makes fine punch.

The juice of lemons makes a high punch.

Haskins Sentences:

TARGET:

The far man tried the wood.

RESPONSES:

The fireman dried the wood.

The fireman tried the wool.

The farm hand dried the wood.

strategies are shown in table 43.2. A recent modification of this task, called the semantically unpredictable sentences (SUS) task, uses five different syntactic structures for the anomalous sentences [Gri89, HG89]. More details of the methods and results of this test using other languages can be found in papers by Benoit [BvGHJ89, Ben90].

43.3 Comprehension of Synthetic Speech

In addition to our studies on segmental intelligibility and word recognition in sentences, we have had a long-standing interest in the process of comprehension. We carried out a series of experiments on the verification of isolated sentences as well as several studies on how listeners understand and answer questions about long passages of continuous synthetic speech produced by rule.

Sentence-Verification Studies. In the sentence-verification experiments, we used three- and six-word sentences that were either true or false: "Cotton is soft," "Snakes can sing." The sentences were pretested to determine whether the final word in each sentence was predictable and to insure that listeners could transcribe all the sentences correctly with no errors. In the sentence verification test, subjects were required to respond "true" or "false" after hearing each sentence. Results shown in figure 43.3 indicated that subjects were consistently faster in responding to natural speech than to synthetic speech. For both natural and synthetic speech, responses were faster for high-predictability sentences than for low-predictability sentences [PMD87]. The results showed that although the sentences were highly intelligible, even high-quality synthetic speech is not perceived in the same way

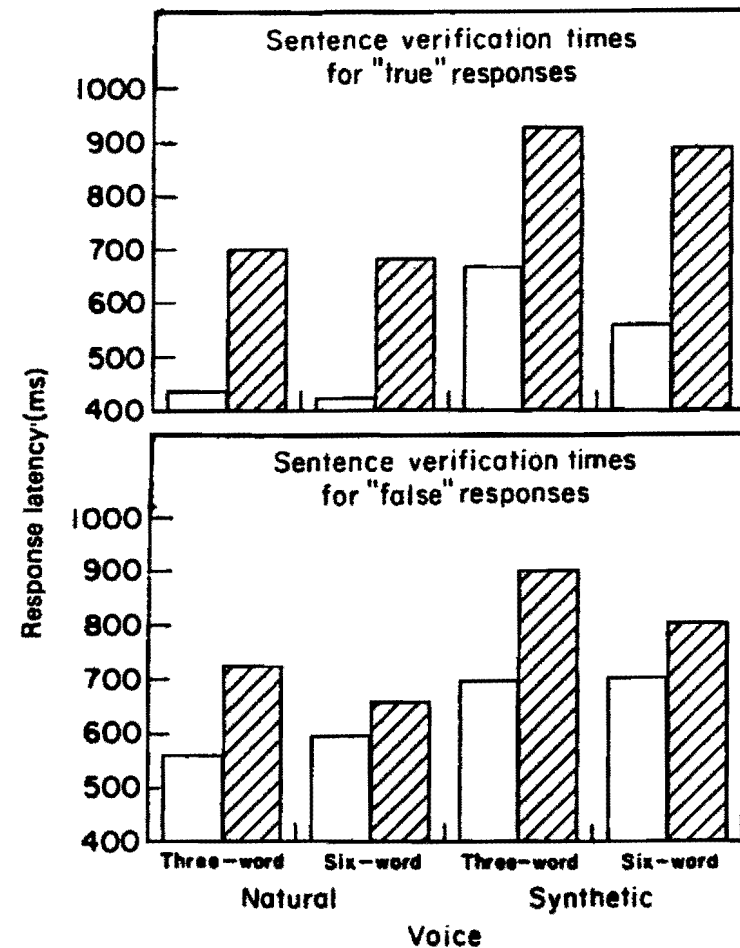


FIGURE 43.3. Mean sentence verification latencies (in ms) for the "true" responses (top panel) and "false" responses (bottom panel) for natural and synthetic speech for each of two sentence lengths. The high-predictability sentences are displayed with open bars; the low-predictability sentences are displayed with striped bars. The latencies displayed in this figure are based on only those trials in which subjects responded correctly and also transcribed the sentence correctly (from [Pis87]).

as natural speech. All of these sentences were easy to recognize and encode in memory as indexed by transcription scores, but there was additional cognitive effort required to understand the sentence and carry out the verification task, as shown by longer response times.

Comprehension of Connected Text. Spoken language understanding is a very complex cognitive process that involves the encoding of sensory information, retrieval of previously stored knowledge from long-term memory, and the subsequent interpretation and integration of various sources of knowledge available to a listener. Language comprehension therefore depends on a relatively large number of diverse and complex factors, many of which are still only poorly understood by cognitive psychologists. One of the factors that plays an important role in listening comprehension is the quality of the initial input signal—that is, the intelligibility of the speech itself, which is assumed to affect the earliest stage of processing, the encoding stage. But the acoustic-phonetic properties of the signal are only one source of information used by listeners in speech perception and spoken-language understanding. Additional consideration must also be given to the contribution of higher levels of linguistic knowledge to perception and comprehension.

In our first comprehension study, subjects either listened to synthetic or natural versions of narrative passages or they read the passages silently. All three groups then answered the same set of multiple-choice test questions immediately after each passage. Although there was a small advantage for the natural-speech group over the synthetic-speech group, the differences in performance appeared to be localized primarily in the first half of the test. The somewhat higher performance on natural speech was eliminated by the second half of the test. Performance by the subjects listening to synthetic speech improved substantially, whereas performance by the natural-speech group and the control group that simply read the texts and then answered questions remained about the same.

The finding of improved performance in the second half of the test for subjects listening to synthetic speech is consistent with our earlier results on word recognition in sentences. We found that recognition performance always improved for synthetic speech after only a short period of exposure and familiarization with the synthesis system [CGL76, NG74, PH80]. These results suggest that the overall differences in performance among the three comprehension groups are probably due to familiarity with the output of the synthesizer and not to any inherent differences in the basic strategies used in comprehending the linguistic content of these passages. One serious problem with this initial comprehension study was that we used multiple-choice questions presented immediately after listeners heard each passage. This test procedure therefore confounds the early stages of perceptual analysis and encoding with later stages of comprehension involving memory, inferencing and reconstruction, which are known to play an important role in studies of text processing, independent of which modality is used for input.

On-line Measures of Comprehension. Recently, we examined the comprehension process in greater detail using several on-line measurement techniques. In one study, Ralston and colleagues ([RPLGM91]) used a word-monitoring task to investigate comprehension of natural and synthetic speech. Subjects were required to monitor a spoken passage for a set of target words. Specifically, the listeners had to memorize a set of target stimuli, rehearse the items, and then press a response button whenever they heard one of the target words in a spoken passage. To make

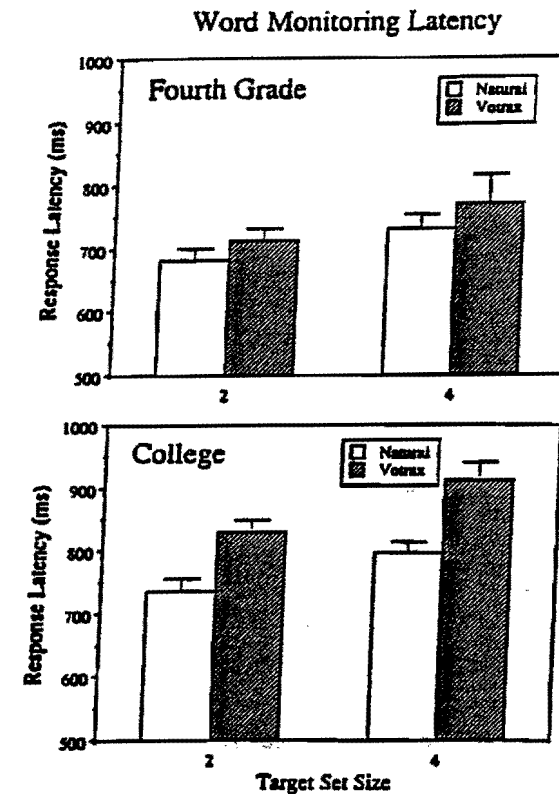


FIGURE 43.4. Word-monitoring latencies (in ms) as a function of target set size. The upper panel shows data for fourth-grade passages; the lower panel shows data for college-level passages. Open bars represent natural speech; striped bars represent latencies for passages of Votrax synthetic speech (from [RPLGM91]).

sure that subjects understood the content of the passage, we also had them answer a set of comprehension questions after each passage. As shown in figure 43.4, word-monitoring performance was better for subjects listening to natural speech compared to synthetic speech. Detection accuracy decreased and response latencies increased as the number of words in the target set became larger. Listeners were more accurate in answering questions following presentation of naturally produced passages than synthetic passages. Thus, both speech quality and memory load affected word-monitoring performance in this task.

In another comprehension study, we used a self-paced listening task to measure the amount of processing time subjects need to understand individual sentences in a passage of connected speech [RPLGM91]. As expected, we found that listeners

required more time to understand synthetic speech than natural speech. When the sentences in the passages were scrambled, listeners required even more processing time for *both* natural and synthetic speech [RLP90]. Moreover, the listening times were much larger for the passages of synthetic speech that required greater cognitive effort and processing resources than the natural passages that were less demanding.

43.4 Mechanisms of Perceptual Encoding

The results of the MRT and word-recognition studies revealed that synthetic speech is less intelligible than natural speech. In addition, these studies demonstrated that as synthetic speech becomes less intelligible, listeners rely increasingly on linguistic knowledge and response-set constraints to facilitate word identification. The findings from the comprehension studies are consistent with this conclusion as well [DP92]. However, the results of these studies are descriptive and do not provide an explanation for the differences in perception between natural and synthetic speech. Several different experiments were carried out over the years to pursue this problem.

Lexical Decision and Naming Latencies. In order to investigate differences in the perceptual processing of natural and synthetic speech, we carried out a series of experiments that measured the time needed to recognize and pronounce natural and synthetically produced words [Pis81b]. To measure the time course of the recognition process, we used a lexical decision task. As shown in figure 43.5, subjects responded significantly faster to *natural* words and nonwords than to *synthetic* words and nonwords. Because the differences in response latency were observed for both words *and* nonwords alike, and did not appear to depend on the lexical status of the test item, the extra processing effort appears to be related to the initial analysis and perceptual encoding of the acoustic-phonetic information in the signal and not to the process of accessing words from the lexicon. In short, the pattern of results suggested that the perceptual processes used to encode synthetic speech require more cognitive “effort” or resources than the processes used to encode natural speech. Thus, synthetic speech appears to be encoded less efficiently than natural speech, presumably because there is less redundancy in the acoustic signal.

Similar results were obtained in a naming task using natural and synthetic words and nonwords [SP82]. The naming results demonstrated that the extra processing time needed for synthetic speech does not depend on the type of response made by the listener; the pattern of latencies were comparable for both manual and vocal responses. Taken together, these two sets of findings suggest that early stages of perceptual encoding for synthetic speech are carried out more slowly and therefore require more processing time than natural speech.

Consonant-vowel (CV) Confusions. To account for the greater difficulty of encoding synthetic speech, some researchers have suggested that synthetic speech

AUDITORY LEXICAL DECISION TASK

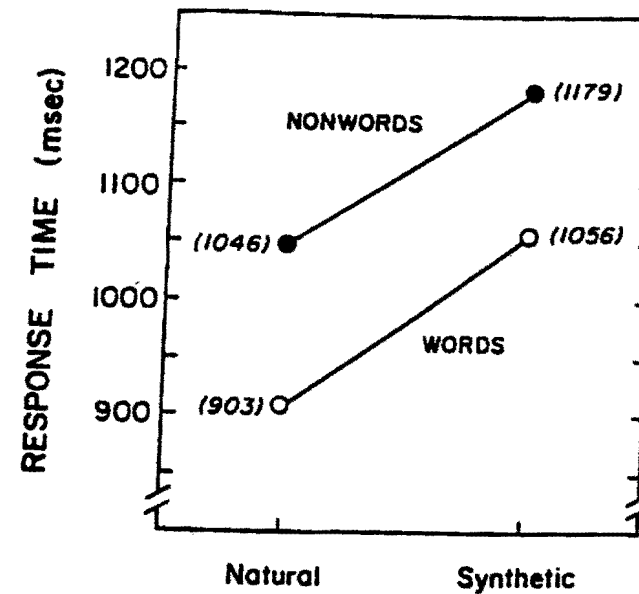


FIGURE 43.5. Response times (in ms) obtained in an auditory lexical decision task for words (open circles) and nonwords (filled circles) for natural and synthetic speech (from [Pis81b]).

should be viewed as natural speech that is degraded by noise. An alternative hypothesis proposes that synthetic speech is not like “noisy” or degraded natural speech at all, but instead may be thought of as a “perceptually impoverished” signal relative to natural speech because it lacks the additional redundancy and acoustic-phonetic variability found in natural speech. Thus, synthetic speech is fundamentally different from natural speech in both degree and kind because it contains only a minimal number of acoustic cues for each phonetic contrast. Recent findings have suggested that this redundancy is important for perceptual learning and retention of novel voices and novel words.

To test this proposal, we examined the perceptual confusions for a set of natural and synthetic consonant-vowel (CV) syllables [NDP84]. By comparing the confusion matrices for a particular text-to-speech system with the confusion matrices for natural speech masked by noise, we found that the predictions made by the “noise-degradation” hypothesis were incorrect. Some consonant identification er-

rors were based on the acoustic-phonetic similarity of the confused segments, but others followed a different pattern that could be explained only as “phonetic mis-cues.” These were confusions in which the acoustic cues used in synthesis simply specified the wrong segment in a particular phonetic environment.

Gating and Signal Duration. The results of the consonant-vowel confusion experiment support the conclusion that the differences in perception between natural and synthetic speech are largely the result of differences in the acoustic-phonetic properties of the signals and the initial process of encoding. In another study, we obtained further support for this proposal using the gating paradigm to investigate the perception of natural and synthetic words [Gro80]. This technique is used to manipulate the amount of stimulus information presented to a listener. We found that, on the average, natural words could be identified after only 67% of a word was heard, whereas for synthetic words it was necessary for listeners to hear more than 75% of a word for correct word identification [MP84]. These results demonstrate more directly that the acoustic-phonetic structure of synthetic speech conveys less information, per unit of time, than the acoustic-phonetic structure of natural speech, and thus the uptake of linguistic information from the signal appears to be less efficient [PNG85].

These results provide some converging evidence that encoding of the acoustic-phonetic structure of synthetic speech is more difficult and requires more cognitive effort and capacity than encoding of natural speech. Recognition of words and nonwords requires more processing time for synthetic speech compared to natural speech. The CV confusion study demonstrated that synthetic speech may be viewed as a phonetically impoverished signal. Finally, the gating results showed that synthetic speech requires *more* acoustic-phonetic information to correctly identify isolated monosyllabic words than natural speech.

Capacity Demands in Speech Perception. We also carried out a series of experiments to determine the effects of encoding synthetic speech on working memory and rehearsal processes [LFP83]. Subjects were given two different lists of items to remember: The first list consisted of a set of digits *visually* presented on a CRT screen; the second list consisted of a set of ten natural words or ten synthetic words. After the list of words was presented, the subjects were instructed to first write down all the visually presented digits in the order of presentation and then all the words they could remember from the list of items they heard. Recall for the natural words was significantly better than for the synthetic words. In addition, recall of both synthetic and natural words became worse as the size of the preload digit list increased. However, the most interesting finding was the presence of an interaction between the *type* of speech presented (synthetic versus natural) and the *number* of digits rehearsed (three versus six). As the size of the memory load increased, significantly fewer subjects were able to recall *all* the digits for the synthetic word lists compared to the digits from the natural word lists. Thus, processing of the synthetic speech impaired recall of the visually presented digits more than the processing of natural speech. These results demonstrate that synthetic speech requires more processing capacity and resources in short-term working memory than natu-

ral speech. The findings also suggest that synthetic speech should interfere much more with other concurrent cognitive processes because the perceptual encoding of synthetic speech imposes greater capacity demands on the human information processing system than the encoding of natural speech [Wic91].

Training and Experience with Synthetic Speech. We also carried out several experiments to study the effects of training on the perception of synthetic speech [SNP85, GNP88]. In the Schwab et al. study, three groups of subjects followed different procedures for eight days. When pre- and post-test scores were compared, the results showed that performance improved dramatically for only one group—the subjects who were specifically trained with the synthetic speech. Neither the first control group trained with natural speech nor the second control group, who received no training of any kind, showed any significant improvement in recognition of synthetic speech. These findings suggest that human listeners can easily modify their perceptual strategies and that substantial increases in performance can be realized in relatively short periods of time, even with poor-quality synthetic speech, if naive listeners become familiar with the synthesis system.

43.5 Some Cognitive Factors in Speech Perception

The literature in cognitive psychology over the last 40 years has identified several major factors that affect an observer's performance in behavioral tasks. These factors include: (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 consider here each of these in the context of our findings on the perception of synthetic speech.

Task Complexity. In some tasks, the response demands are relatively simple, such as deciding which of two words was presented. Other tasks are extremely complex, such as trying to recognize an unknown utterance from a virtually unlimited number of response alternatives, while simultaneously engaging in another activity that already requires attention. In carrying out any perceptual experiment, it is necessary to understand the requirements and demands of a particular task before drawing any strong inferences about an observer's performance. In our studies on the perception of synthetic speech, we have found that task complexity influences performance and affects the strategies that listeners use to complete their task. To take one example, large differences were observed in intelligibility of isolated words when the response set was changed from a closed-set to an open-set format.

Limitations on the Observer. Limitations exist on the human information-processing system's ability to perceive, encode, store, and retrieve information. The amount of information that can be processed in and out of short-term working memory is severely limited by the listener's attentional state, past experience, and the quality of the original sensory input that affects ease of encoding. These gen-

eral principles apply to the study of speech perception as well as other domains of human information processing.

Experience and Training. Human observers can quickly learn new strategies to improve performance in almost any psychological task. When given appropriate feedback and training, subjects can learn to classify novel stimuli, remember complex visual patterns, and respond to rapidly changing stimuli presented in different sensory modalities. It comes as no surprise then that listeners can benefit from short-term training and exposure to synthetic speech.

Message Set. The structure of the message set—that is, the constraints on the number of possible messages and the linguistic properties of the message set—play an important role in speech perception and language comprehension [MHL51]. The arrangement of speech sounds into words is constrained by the phonological rules of language; the choice and patterning of words in sentences is constrained by syntax; and finally, the meaning of individual words and the overall meaning of sentences in a text is constrained by the semantics and pragmatics of language. The contribution of these levels varies substantially in perceiving isolated words, sentences, and passages of continuous speech. In each case, listeners exploit constraints to recover the intended message.

Signal Characteristics. The segmental and prosodic structure of a synthetic utterance also constrains the choice of response. Synthetic speech is an impoverished signal that represents phonetic distinctions with only the minimum number of acoustic cues used to convey contrasts in natural speech. Under adverse conditions, synthetic speech may show serious degradation because of the lack in redundancy in the signal, which is the hallmark of natural speech.

43.6 Some New Directions for Research

Much of the research carried out on the perception of synthetic speech over the last 15 years has been concerned with the quality of the acoustic-phonetic output. Researchers have focused most of their attention on improving the segmental intelligibility of synthetic speech. At the present time, the available perceptual data suggest that segmental intelligibility is quite good for some commercially available systems. Synthetic speech is not at the same level of intelligibility as natural speech, and it may take a great deal of additional research effort to achieve relatively small gains in improvement in intelligibility. Thus, it seems entirely appropriate at this time to move the focus of research efforts to a number of other issues and to pursue several new directions in the future. In this section, I review several topics that have not been studied very much in the past. Research in these areas may yield important new knowledge that will help improve performance to levels approaching natural speech.

Naturalness. Improving the naturalness of synthetic speech has been a long-standing problem in synthesis research. Everyone working in the field today be-

lieves this is the next goal. Why do listeners want to listen to “natural”-sounding speech? Why do listeners “prefer” to listen to natural speech? We believe these are important research questions for the future because they suggest a close association between the linguistic properties of the signal and the so-called “indexical” attributes of a talker’s voice, which are carried in parallel in the speech signal. Recent studies have shown that knowledge of and familiarity with the talker’s voice affects intelligibility of speech and may contribute to more efficient encoding of the message in memory [NSP94]. In the past, researchers have treated these two sets of acoustic attributes as independent sources of information in the signal. Now we are beginning to realize the importance of naturalness and, in particular, the role of talker-familiarity in spoken language processing. Familiar voices are easier to process and understand because the listener can access talker-specific attributes from long-term memory, which facilitates encoding the signal into a linguistic representation [Pis93]. In the next few years, we will see increased research on synthesizing specific voices, dialects, and even different speaking styles in an effort to improve the naturalness of the synthesizer’s voice and therefore increase intelligibility.

Sources of Variability. In order to achieve naturalness, a great deal of new research will be needed on different sources of variability in speech and how to model and reproduce this information in the next generation of synthesis systems. Much of the early research on speech perception and acoustic-phonetics assumed the existence of abstract idealized representations for phonemes, words, and sentences. The research methodologies used over the years assumed that variability was a source of noise that had to be reduced or eliminated from the signal in order to recover the talker’s intended message. Several researchers have argued recently that variability in speech is not a source of noise but rather is informative and useful to the listener [EM86, Pis93]. If we are ever going to have synthesis systems that can model different speaking styles and adjust to the demands of the listener and the environment, it will be necessary to learn more about the different sources of variability that occur in natural speech and how these factors can be incorporated into synthesis routines [MA93].

Audio-Visual Integration. Along with improving naturalness, a number of researchers have suggested developing multimodal synthesis systems that are able to produce visual displays of a synthetic talking face along with synthesis of the speech signal [Des92, BLMA92]. The case for multimodal synthesis is convincing on several theoretical grounds, the most important of which is that listeners are able to use the additional information contained in the visual display of a talker’s face to improve intelligibility and recognition of the intended message [SP54]. Until recently, most of the research on speech synthesis has been concerned exclusively with the auditory modality despite the evidence that deaf and hearing-impaired listeners gain substantial information about speech from the optical display of a talker’s face [Sum87]. It appears very likely that synthetic faces will become an integral part of speech synthesis systems in the next few years.

New Assessment Methods. As we look back over the past 15 years, it is obvious from the research findings that much more behavioral research with human listeners will be needed to study the complex relations between traditional measures of segmental intelligibility, comprehension performance, and listener preferences. For example, whereas more basic research on prosody and speech timing will no doubt help to eventually improve synthesis in the long term, new behavioral tests will need to be developed to measure and assess these gains in performance. We believe that prosodic characteristics are not perceived directly by naive listeners; rather, they exert their influence indirectly on the processes used to recognize words and understand the meaning of sentences and discourse. Because of this, we believe new perceptual tests will have to be developed to study and measure the effects of prosody as they affect other aspects of speech perception and spoken language processing. These tests might include measures of processing load, attention, memory, or real-time comprehension.

A good example of the problems in measuring prosody can be seen in the recent experiments of van Santen [van94] on the development of duration rules in the Bell Laboratories synthesis-by-rule system. Using very sophisticated methodologies, van Santen showed that a group of naive listeners consistently preferred a "new" set of duration rules over an "old" set of rules and were able to make explicit quality judgments about sentences containing various kinds of problems in pronunciation, stress, voice quality, and timing. The question of interest about these new duration rules is whether they produce speech that is more intelligible than the old rules. Is the synthetic speech easier to process, that is, recognize or recall, and are these new rules less susceptible to degradation from noise, other competing voices, or tasks requiring higher cognitive load? If a naive listener prefers one durational rule system over another, is that system therefore "better" on a variety of behavioral performance measures or are the preferences and quality judgments simply domain-specific? Assuming that we could develop a sensitive on-line measure of comprehension, would there be any difference in comprehension performance between the "old" and "new" rule systems?

I believe these are the kinds of questions we will have to address in the future in designing the next generation of synthesis-by-rule systems. Because spoken language processing in human listeners is extremely robust under a wide variety of conditions, it has been and will continue to be very difficult to identify precisely which component or subcomponent in the system is responsible for a particular problem or what aspects of the system control a listener's preference in one direction or another. Humans are able to adjust and adapt their perceptual processing and response criteria rapidly on the fly to changing conditions in their listening and speaking environments. In the years to come, we will need to continue several broad-based programs of basic research on human speech perception and spoken language processing in parallel with research and development efforts on speech synthesis by rule.

These are a few of the questions and new research directions that should be pursued over the next few years. Research on naturalness, variability, audio-visual integration, prosody, and comprehension are not only topics of practical concern

with regard to the design and implementation of synthesis-by-rule systems, but these particular research issues are also at the center of current theoretical work in cognitive science and psycholinguistics. Answers to these questions will provide us with new insights into the reasons why synthetic speech is difficult to understand and why it requires more attention and effort for the listener to recover the intended meaning.

Acknowledgments: This research was supported, in part, by NIH Research Grant DC-00111 and NIH T32 Training Grant DC-00012 to Indiana University. I thank Beth Greene for her help over the years on the perceptual evaluation project and Steve Chin for his editorial comments.

REFERENCES

- [AHK87] J. Allen, S. Hunnicut, and D. H. Klatt. *From Text to Speech: The MITalk System*. Cambridge University Press, Cambridge, 1987.
- [BB92] G. Bailly and C. Benoît. *Talking Machines, Theories, Models, and Designs*. Elsevier, North-Holland, Amsterdam, 1992.
- [Ben90] C. Benoît. An intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity. *Speech Comm.* 9:293-304, 1990.
- [BLMA92] C. Benoît, T. Lallouache, T. Mohamadi, and C. Abry. A set of French visemes for visual speech synthesis. In *Talking Machines: Theories, Models and Designs*, G. Bailly, C. Benoît, and T. R. Sawallis, eds. Elsevier Science Publishers, North-Holland, 485-504, 1992.
- [BP89] R. V. Bezooijen and L. Pols. Evaluation of text-to-speech conversion for Dutch: From segment to text. In *Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, 1989.
- [BvGHJ89] C. Benoît, A. van Erp, M. Grice, V. Hazan, and U. Jekosch. Multilingual synthesizer assessment using unpredictable sentences. In *Proceedings of the First Eurospeech Conference*, Paris, France, 633-636, 1989.
- [CG89] R. Carlson and B. Granstrom. Evaluation and development of the KTH text-to-speech system at the segmental level. In *Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, 1989.
- [CGL76] R. Carlson, B. Granstrom, and K. Larssen. Evaluation of a text-to-speech system as a reading machine for the blind. *Quarterly Progress and Status Report, STL-QPSR 2-3*. Stockholm: Royal Institute of Technology, Department of Speech Communication, 1976.
- [Des92] R. Descout. Visual speech synthesis. In *Talking Machines: Theories, Models and Designs*, G. Bailly, C. Benoît, and T. R. Sawallis, eds. Elsevier Science Publishers, North-Holland, 475-477, 1992.
- [DP92] S. A. Duffy and D. B. Pisoni. Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech* 35:351-389, 1992.

- [EM86] J. L. Elman and J. L. McClelland. Exploiting lawful variability in the speech wave. In *Invariance and Variability in Speech Processes*, J. S. Perkell and D. J. Klatt, eds. Erlbaum, Hillsdale, NJ, 360–385, 1986.
- [FHBH89] A. Fourcin, G. Harland, W. Barry, and V. Hazan. *Speech Input and Output Assessment*. Ellis Horwood, Chichester, England, 1989.
- [GNP88] S. L. Greenspan, H. C. Nusbaum, and D. B. Pisoni. Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Human Learning, Memory and Cognition* 14:421–433, 1988.
- [Gre86] B. G. Greene. Perception of synthetic speech by nonnative speakers of English. In *Proceedings of the Human Factors Society*, Santa Monica, CA, 1340–1343, 1986.
- [Gri89] M. Grice. Syntactic structures and lexicon requirements for semantically unpredictable sentences in a number of languages. In *Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, 1989.
- [Gro80] F. Grosjean. Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics* 28:267–283, 1980.
- [HG89] V. Hazan and M. Grice. The assessment of synthetic speech intelligibility using semantically unpredictable sentences. In *Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, 1989.
- [HWHK65] A. S. House, C. E. Williams, M. H. L. Hecker, and K. Kryter. Articulation-testing methods: Consonantal differentiation with a closed-response set. *J. Acoust. Soc. Amer.* 37:158–166, 1965.
- [Kla87] D. H. Klatt. Review of text-to-speech conversion for English. *J. Acoust. Soc. Amer.* 82:737–793, 1987.
- [LFP83] P. A. Luce, T. C. Feustel, and D. B. Pisoni. Capacity demands in short-term memory for synthetic and natural word lists. *Human Factors* 25:17–32, 1983.
- [LGP89] J. S. Logan, B. G. Greene, and D. B. Pisoni. Segmental intelligibility of synthetic speech produced by rule. *J. Acoust. Soc. Amer.* 86:566–581, 1989.
- [MA93] I. R. Murray and J. L. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *J. Acoust. Soc. Amer.* 93:1097–1108, 1993.
- [MHL51] G. A. Miller, G. A. Heise, and W. Lichten. The intelligibility of speech as a function of the context of the test materials. *J. Experimental Psychology* 41:329–335, 1951.
- [Mil56] G. A. Miller. The perception of speech. In *For Roman Jakobson*, M. Halle, ed. Mouton, The Hague, 353–359, 1956.
- [MP84] L. M. Manous and D. B. Pisoni. Effects of signal duration on the perception of natural and synthetic speech. *Research on Speech Perception Progress Report No. 10*, Indiana University, Bloomington, IN, 1984.
- [NDP84] H. C. Nusbaum, M. J. Dedina, and D. B. Pisoni. Perceptual confusions of consonants in natural and synthetic CV syllables. *Speech Research Laboratory Technical Note 84-02*. Indiana University, Speech Research Laboratory, Bloomington, IN, 1984.
- [NG74] P. W. Nye and J. Gaitenby. The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. *Haskins Laboratories Status Report on Speech Research* 38:169–190, 1974.
- [NSP94] L. C. Nygaard, M. S. Sommers, and D. B. Pisoni. Speech perception as a talker-contingent process. *Psychological Science* 5:42–46, 1994.
- [PH80] D. B. Pisoni and S. Hunnicutt. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. In *1980 IEEE Conference Record on Acoustics, Speech and Signal Processing* IEEE Press, New York, 572–575, 1980.
- [Pis81a] D. B. Pisoni. Perceptual processing of synthetic speech: Implications for voice response systems in military applications. Paper presented at the *Conference on Voice-Interactive Avionics*, Naval Air Development Center, Warminster, PA, 1981.
- [Pis81b] D. B. Pisoni. Speeded classification of natural and synthetic speech in a lexical decision task. *J. Acoust. Soc. Amer.* 70:S98, 1981.
- [Pis87] D. B. Pisoni. Some measures of intelligibility and comprehension. In *From Text to Speech: The MITalk System*, J. Allen, S. Hunnicutt, and D. H. Klatt, Cambridge, Cambridge University Press, Cambridge, 1987.
- [Pis93] D. B. Pisoni. Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Comm.* 13:109–125, 1993.
- [PMD87] D. B. Pisoni, L. M. Manous, and M. J. Dedina. Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility. *Computer Speech and Language* 2:303–320, 1987.
- [PNG85] D. B. Pisoni, H. C. Nusbaum, and B. G. Greene. Perception of synthetic speech generated by rule. *Proceedings of the IEEE* 73(11):1665–1676, 1985.
- [Pol89a] L. C. W. Pols. Improving synthetic speech quality by systematic evaluation. In *Proceedings of the ESCA Tutorial Day on Speech Input/Output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, 1989.
- [Pol89b] L. C. W. Pols. Assessment of text-to-speech synthesis systems. In *Speech Input and Output Assessment*, A. J. Fourcin, G. Harland, W. Barry, and V. Hazan, eds. Ellis Horwood, Chichester, England, 55–81, 1989.
- [Pol92] L. C. W. Pols. Quality assessment of text-to-speech synthesis by rule. In *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, eds. Marcel Dekker, New York, 387–416, 1992.
- [RLP90] J. V. Ralston, S. E. Lively, and D. B. Pisoni. Comprehension of normal and scrambled passages using a sentence-by-sentence listening task. *Research on Speech Perception Progress Report No. # 16*. Indiana University, Speech Research Laboratory, Bloomington, IN, 1990.
- [RPLGM91] J. V. Ralston, D. B. Pisoni, S. E. Lively, B. G. Greene, and J. W. Mullennix. Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. *Human Factors* 33:471–191, 1991.
- [SAMW89] M. Spiegel, M. J. Altom, M. Macchi, and K. Wallace. A monosyllabic test corpus to evaluate the intelligibility of synthesized and natural speech. In *Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, 1989.
- [SKPO94] M. S. Sommers, K. I. Kirk, D. B. Pisoni, and M. J. Osberger. Some new direction in evaluating the speech perception abilities of cochlear implant patients: A preliminary report. Poster presented at *ARO*, St. Petersburg Beach, FL, 1994.
- [SNP85] E. C. Schwab, H. C. Nusbaum, and D. B. Pisoni. Effects of training on the perception of synthetic speech. *Human Factors* 27:395–408, 1985.
- [SP54] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Amer.* 26:212–215, 1954.

- [SP82] L. M. Slowiaczek and D. B. Pisoni. Effects of practice on speeded classification of natural and synthetic speech. *J. Acoust. Soc. Amer.* 71 Sup. 1, S95–S96, 1982.
- [Sum87] Q. Summerfield. Some preliminaries to a comprehensive account of audiovisual speech perception. In *Hearing by Eye: The Psychology of Lip-Reading*, B. Dodd and R. Campbell, eds. Erlbaum, Hillsdale, NJ, 3–51, 1987.
- [van94] J. P. H. van Santen. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language* 8:95–128, 1994.
- [Wic91] C. D. Wickens. Processing resources and attention. In *Multiple Task Performance*, D. Damos, ed. Taylor & Francis, Washington, DC, 334, 1991.

Section VIII

Systems and Applications