Behavior Research Methods, Instruments, & Computers
1986, 18 (2), 100-107

Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems

BETH G. GREENE, JOHN S. LOGAN, and DAVID B. PISONI
Indiana University, Bloomington, Indiana

We present the results of studies designed to measure the segmental intelligibility of eight text-to-speech systems and a natural speech control, using the Modified Rhyme Test (MRT). Results indicated that the voices tested could be grouped into four categories: natural speech, high-quality synthetic speech, moderate-quality synthetic speech, and low-quality synthetic speech. The overall performance of the best synthesis system, DECtalk-Paul, was equivalent to natural speech only in terms of performance on initial consonants. The findings are discussed in terms of recent work investigating the perception of synthetic speech under more severe conditions. Suggestions for future research on improving the quality of synthetic speech are also considered.

There has always been a practical need for devices that can produce and understand spoken language automatically without human intervention. At the present time, the development and use of such automated voice-response systems is no longer a matter of basic research in linguistics, engineering, or product development—the technology is now available in the form of specialized microprocessor-based speech-processing devices that can be easily integrated into numerous computer-based systems to support user-machine communication via spoken language.

Speech is, without question, the most natural means of communication (Lindgren, 1967). It is automatic, requires little conscious effort or attention, and creates few, if any, demands while other tasks are carried out concurrently, especially tasks which require active use of the hands or eyes in demanding conditions. One potential use of speech is as an interface to computers. At the present time, most users interact with computers using traditional screens and keyboards. However, these systems can and will eventually be replaced by speech input/output (I/O). Speech is not only more natural for humans to use, but is also faster and less prone to errors. Although speech interfaces to computers are not yet widely available, extensive research efforts have been carried out over the last few years to develop speech recognition and synthesis technology. In this paper, we examine the use of one aspect of this technology—speech synthesis by rule, using automatic text-to-speech conversion. With a text-to-speech system, any computer can generate spoken output from a string of characters, and therefore can provide the user with a novel speech display instead of the more traditional screen. In some applications, this display may significantly reduce the user's workload and increase operator efficiency in getting information from a computer. In other applications, it may provide entirely new methods for retrieving data and other kinds of information from the computer using standard telephone voice and data channels. At the present time, speech output from computers using some form of text-to-speech conversion is still in its infancy. However, as the technology becomes more widely known and the costs decrease, much wider usage can be anticipated, because the benefits of this output channel are often quite substantial in many applications.

The research reported in this paper was supported in part by NIH research Grant No. NS 12179; in part by contract No. AF-F 33615-83-K-0501 with the Air Force Systems Command, AFOSR, through the Aerospace Medicine Research Laboratory, Wright-Patterson AFB, OH, to Indiana University in Bloomington; and in part by a James McKeen Cattell award to the third author. B. G. Greene's address is: Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405.

WHAT IS TEXT-TO-SPEECH?

A text-to-speech system is a device that automatically converts printed orthographic text into spoken output without human intervention of any kind. This process usually takes place immediately in real time and will accept any text that can be typed at a computer terminal and converted into ASCII code. Several currently available text-to-speech systems convert unrestricted English text to intelligible speech in real time. There are systems for other languages as well (Carlson, Granstrom, & Hunnicutt, 1982). The speech output generated by a text-to-speech system is synthesized or created anew in real time by the device in response to a phonetic representation of the specific typed input (see Allen, 1973a, 1973b, 1981, in press; Studdert-Kennedy & Cooper, 1966). Most text-to-speech systems are designed to allow the user to customize certain features. For example, there is a phonetic mode that allows the user to specify the correct pronunci-

ation for proper names or to enter a specialized vocabulary that may have unusual pronunciations.

Stored Speech

Most readers are already familiar with a form of synthetic speech known as stored speech. Natural speech is recorded on audio tape using a microphone and tape recorder. This speech is then digitized with a computer using an analog-to-digital converter. The actual process involves sampling the speech waveform at a rapid rate and storing the samples in digital form. Typically, from 8,000 to 10,000 samples are taken for every 1 sec of speech. These digital samples are then stored in the computer memory as a series of numerical parameters. Thus, a 5-sec sentence will have at least 40,000 samples associated with it; each of these samples will be stored digitally in the computer. Unfortunately, for long passages of speech, the storage requirements are enormous. However, there is a good reason to use stored speech. All the digital samples can be retrieved from the computer memory and then reconverted to analog form using a digital-to-analog converter. This process reproduces the speech that was orginally recorded with little or no degradation or effects on intelligibility. Although there may be some loss in speech quality due to the sampling rate and the number of bits used to code the speech waveform, the resulting speech quality is acceptable and often sounds better than speech transmitted over the telephone. When listeners hear stored speech, they typically have little difficulty perceiving or understanding it.

However, every message that is to be stored has to be recorded, digitized, and stored in computer memory, and then retrieved and played out. If a message needs to be changed or updated, the entire process must be repeated. Thus, stored speech is useful for very limited message sets, such as the letters of the alphabet, the digits 0 through 9, or a very small vocabulary of key words or instructions. When the vocabulary becomes very large and the potential set of messages is theoretically unrestricted, a voice output system using stored speech becomes impractical and extremely expensive (see Cooper, 1963; Cooper, Gaitenby, Mattingly, & Umeda, 1969; Studdert-Kennedy & Cooper, 1966). Furthermore, when individual stored items are combined into word strings without additional processing and smoothing, the resulting speech lacks normal pitch and intonation; listeners often describe this type of speech as unnatural and mechanical sounding. The intelligibility of this kind of connected speech is often quite poor, even though the intelligibility of individual words is typically quite high.

Synthesis by Rule

Voice output using stored speech may be contrasted with voice output using various synthesis-by-rule techniques. In this case, the speech is generated by a series of rules which are used to create utterances on demand (Allen, 1973a, 1973b, in press; Cooper, 1963). These voice output systems are very sophisticated and consist of a number of modular subsystems, each of which has a special set of rules. The initial typed input is first converted into ASCII code. In most current systems, the ASCII code is then processed through several modules which serve to produce a detailed phonetic description (see Allen, 1981).

In one system, MITalk-79, this analytic process involves the determination of the underlying phonemic, syllabic, morphemic, and syntactic form of the input message, as well as adjustment of the input when numerals, abbreviations, and special symbols are present. After the basic modules have operated on the input message, any word that has not been analyzed is processed through a set of letter-to-phoneme rules. Once the text has been converted into a phonetic transcription, other modules containing detailed phonological, pitch, stress, and timing adjustments operate on this representation. Additional rules are included to make the speech sound less mechanical. Some rules "smooth" the speech and lead to more natural-sounding output. Other rules serve to disambiguate words such as "read," which can be pronounced like "red" or like "reed."

After the input text has been analyzed, it is converted into spoken output. The output process is also modular in nature. Several modules are used to specify the way each speech sound is to be pronounced, how certain speech sounds are modified by specific contexts, and where stress is to be placed. The more detailed the rule system, the more nearly the synthesized speech approximates natural speech. All the parametric information that has been accumulated in the various modules is then input to a digital speech synthesizer and a speech waveform is generated. Finally, the speech samples are converted to analog form via a digital-to-analog converter and are low pass filtered. The text-to-speech systems that are available at this time all work in real time, performing the analysis and synthesis immediately after the text is input to the device (for further details see Allen, 1981; Bruckert, 1984; Groner, Bernstein, Ingber, Pearlman, & Toal, 1982).

For the last 6 years, we have been engaged in a program of research designed to study the perception of synthetic speech produced by rule. As part of this work we have had the opportunity to collect behavioral data from eight text-to-speech systems that produce speech automatically by rule. In the remainder of this paper, we describe the systems we have tested, the procedures used to measure segmental intelligibility, and the results obtained. Finally, we consider some limitations of the present approach and suggest several directions for future studies. Our perceptual findings have led to consistent improvements in the performance of several commercial systems and have suggested areas of further work to improve the quality of speech synthesis by rule.

THE TEXT-TO-SPEECH SYSTEMS TESTED

The text-to-speech systems used in our evaluations ranged from a research system running on a large main-

Table 1
Text-to-Speech Systems Tested in SRL with MRT

System	Source	Date Tested
MITalk-79	Natural Language Processing Group, MIT	4/79
TSI Prototype-1	Telesensory Systems, Inc.	11/79
DECtalk V1.8	Digital Equipment Corporation	4/84, 11/84, 2/85
Berkeley (Prototype)	Berkeley Systems Works	11/84
Infovox SA 101	Infovox AB, Sweden	3/85
Prose 2000 V3.0	Speech Plus, Inc.	4/85
Votrax Type'n'Talk	Votrax, division of Federal Screw Works, Inc.	7/85
Echo	Street Electronics, Inc.	7/85

frame computer that did not run in real time to several relatively inexpensive units designed primarily for hobbyists. Table 1 provides a descriptive summary of the text-to-speech systems we have tested. A brief description of each system is given below.

MITalk-79

The MITalk-79 system was designed as a research tool. It was implemented on a DECSYSTEM-20 computer at the Massachusetts Institute of Technology (MIT), and was the product of a 10-year effort to convert unrestricted English text input into high-quality speech output (Allen, 1976, 1981). MITalk consisted of a number of program modules which first analyzed the text input in terms of morphological composition and performed a lexical look-up operation to determine whether or not each morpheme was present in a 12,000-item dictionary. If the morphemes composing the words were not found in the dictionary, another module containing approximately 400 letter-to-sound rules was used to arrive at a pronunciation of the text (Hunnicut, 1976). In addition, sentence-level syntactic analysis was also carried out in order to determine prosodic information such as timing, duration, and stress. The parameters resulting from these analyses of the text were then used to control a formant synthesizer designed by Klatt (1980). The MITalk system ran in about 10 times real time due to the time required for input/output operations.

Telesensory Systems, Inc., TSI Prototype-1 of the Prose 2000

The TSI system was an early prototype of the current Prose 2000 text-to-speech system developed by Telesensory Systems, Inc. The Prose 2000 and other Prose products are now produced by Speech Plus, Inc. The TSI Prototype-1 was based in part on the MITalk-77 system. However, it used only a 1,100-unit dictionary for lexical look-up, omitted the parsing system, and replaced the MITalk fundamental frequency module with a "hat and declination" routine. In addition, the TSI system was implemented using IC technology and ran in real time (see Bernstein & Pisoni, 1980, for further details).

Digital Equipment Corporation DECtalk V1.8

DECtalk V1.8 is a stand-alone text-to-speech system produced commercially by Digital Equipment Corporation (DEC). A prototype of the system, klattalk, was developed by Dennis Klatt at MIT and was licensed to DEC for commercial use (Klatt, 1982). DECtalk V1.8 has seven voices, two of which were tested in the evaluation reported here. This device was designed to produce high-quality synthetic speech. It also has a wide range of useful features, such as the diversity of voices available, the flexibility of a user-defined dictionary, and standard telephone interfaces.

Votrax Type'n'Talk

The Votrax Type'n'Talk is a relatively inexpensive text-to-speech system manufactured by the Votrax division of Federal Screw Works, Inc. (now Votrax, Inc.). Text is converted to phoneme control codes by a text-to-speech translator module. These codes serve as input to the Votrax SC01 phoneme synthesizer chip (Votrax, 1981), which utilizes formant synthesis techniques to produce speech. All speech is generated by rule.

Street Electronics Echo

The Echo text-to-speech system is an inexpensive system manufactured by Street Electronics and designed primarily for the computer hobbyist market. Using an algorithm developed at the Naval Research Laboratory (Elovitz and Johnson, 1976, cited in Morris, 1979), text is converted into allophonic control codes which are then converted to speech using linear predictive coding (LPC) synthesis by a Texas Instruments TMS-5200 chip (Echo, 1982).

Speech Plus Prose 2000 V3.0

The Prose 2000 V3.0 is a further development of the TSI system tested in 1979. It is manufactured by Speech Plus, Inc., and is a high-quality modular text-to-speech system designed to provide voice output from unrestricted English text. Like DECtalk and the earlier TSI prototype, this sytem has both precompiled and user-defined dictionaries.

Berkeley Systems Works

The Berkeley system is a prototype device that used the General Instruments SP1000 chip to carry out LPC synthesis of allophonic segments generated by a set of proprietary rules.

Infovox SA 101

The Infovox SA 101 text-to-speech system is another

stand-alone unit based on synthesis rules developed for Swedish and English by Carlson and Granstrom. It was developed in Sweden at the Royal Institute of Technology (Carlson et al. 1982) and was commercially implemented by Infovox AB (Magnusson, Blomberg, Carlson, Elenius, & Granstrom, 1984). The most distinctive feature of this system is its multilingual capability; at this time, Infovox can process text using spelling-to-sound and phoneme-to-speech rules for English, French, Spanish, German, Italian, and of course, Swedish. Only the English version of this system was tested.

PERCEPTUAL STUDIES USING THE MODIFIED RHYME TEST

In 1958, Fairbanks developed the Rhyme Test, an easily scored speech intelligibility test that could be administered to groups of untrained subjects in a short period of time (Fairbanks, 1958). Each stimulus item was given in stem form (e.g., _ot, _ay) on the answer sheet and the subject was required to supply the missing letter based on his/her perception of the stimulus item. Only initial-consonant phonemes that could be spelled with one letter were tested, which excluded the phonemes /ɵ/, /ð/, /ʧ/, and /ʤ/. As stated in his objectives, Fairbanks intentionally did not include final consonants in the Rhyme Test, clearly a relevant shortcoming.

In response to some of the deficiencies that existed in the Rhyme Test and in other earlier tests, House, Williams, Hecker, and Kryter (1965) developed the Modified Rhyme Test (MRT). The MRT was designed to be easily administered and scored. Unlike the Rhyme Test, however, the MRT provided information on consonants in both initial and final positions. The test also included the phonemes omitted from the original Rhyme Test. House and his colleagues designed the MRT as a closed-response test, with six response alternatives available to subjects for each stimulus presentation.

Until recently, no tests of segmental intelligibility using isolated words have been designed specifically to assess the performance of systems that generate synthetic speech, particularly systems that generate synthetic speech by rule, nor has a comprehensive evaluation of the suitability of any of the existing tests of intelligibility for synthetic speech been carried out. Nye and Gaitenby (1973) were the first to use the MRT to examine the segmental intelligibility of synthetic speech. They used the MRT to evaluate the speech produced by the Haskins Laboratories speech-synthesis system and to compare its performance to the intelligibility of natural speech. They found that segmental intelligibility of synthetic speech showed a higher error rate than natural speech (7.6% and 2.7%, respectively).

During the period between the testing of the MITalk system in June of 1979 and the present report, a number of text-to-speech systems have been tested in our laboratory at Indiana University. All of these tests have followed the same basic procedures employed in the initial evaluation of the MITalk-79 system. In carrying out these tests, we have also developed several variations of the standard MRT that were useful for providing additional information as well as alleviating some of the problems cited above.

In order to learn more about the perceptual confusions among segments and the pattern of errors, we have modified the MRT for use with an open-response format (Pisoni, in press). This change addresses a criticism first noted by Nye and Gaitenby (1973, 1974) concerning the limitations of the closed-response set in the standard MRT. Both the closed- and open-response formats are described in greater detail below.

Procedures

The subjects were obtained from two sources: (1) undergraduate students at Indiana University who received course credit for their participation as part of a requirement in an introductory psychology class, and (2) paid subjects obtained from a subject pool maintained by the Speech Research Laboratory. All subjects were native speakers of English, with no history of a speech or hearing disorder and no previous experience listening to synthesized speech. A total of 72 subjects participated in each evaluation.

The stimulus items were generated by each text-to-speech system and recorded on audiotape for later playback to subjects. Most of the tapes were made in our laboratory; however, MITalk-79, TSI Prototype-1, and Berkeley systems tapes were provided to us by the manufacturer according to our specifications. The tapes containing natural speech were recorded by Klatt at MIT in 1979.

Subjects were tested in groups of 6 in a quiet room containing individual cubicles, each equipped with a desk and a set of high-quality headphones. Subjects were informed that they would hear a single isolated English word on each trial of the test, and that their task was to indicate the word they heard on the answer sheet provided. Subjects were told to respond on every trial and to guess if they were uncertain.

As originally designed by House et al. (1965), the closed format MRT consisted of 300 words arranged into six lists of 50 words each. On each test trial, subjects were provided with six alternative responses on specially prepared answer sheets. Each of the alternatives was the correct response on one of the stimulus lists.

The data obtained from the closed-format MRT provide information about phonemes appearing only in initial and final positions. No information concerning medial vowels is available, because the response alternatives available to subjects on each trial always contained the same vowel. In order to obtain information on vowels and to reduce the response constraints, an open-response version of the MRT was used. The open MRT can be very useful as a diagnostic aid in identifying poorly synthesized phonemes by providing an unbiased estimate of the most common types of perceptual confusions possible with each

phoneme. For the open-format test, subjects were told to write down the English word they heard on each trial. The response sheet for the open test simply contained 50 blank lines.

Results

The data from each system were tabulated in terms of the percentage error for both initial and final consonants, as well as overall error rates. These data were calculated for each subject tested with each test form. In addition, the overall pattern of errors was examined with respect to initial versus final error rates. Figure 1 shows the mean overall error rates and confidence intervals for each of the text-to-speech systems tested. Comparable data from the natural speech control condition are also shown here. Table 2 shows the mean error rates for consonants in initial and final position as well as the numerical values for the overall error rates shown in Figure 1.

Examination of the data in Table 2 reveals a fairly wide range of performance for the different systems. The best performance for synthetic speech was obtained with DECtalk Paul in initial position (1.6%) and Prose V3.0 in final position (4.3%); the worst performance was obtained with Echo for both initial and final consonant positions (35.6% error rate for each position). The scores shown in Table 2 were calculated by dividing the number of errors by the total number of responses available for each category. (For detailed descriptions of each of these systems see Bernstein & Pisoni, 1980; Greene, Logan, & Pisoni, 1984, 1985; Greene, Manous, & Pisoni, 1984; Logan, Greene, & Pisoni, 1985; Logan, Pisoni, & Greene, 1985; Pisoni, 1982, in press; Pisoni & Hunnicutt, 1980.)

Specific Phonemic Errors

More detailed information about each system is provided by an analysis of the individual phoneme errors.

Table 2
MRT Error Rates Overall and Error Rates for Consonants in Initial and Final Position

	Error Rate (in Percent)		
Voice	Initial	Final	Overall
Natural Speech	0.50	0.56	0.53
DECtalk 1.8, Paul	1.56	4.94	3.25
DECtalk 1.8, Betty	3.39	7.89	5.72
MITalk-79	4.61	9.39	7.00
Prose 2000 V3.0	7.11	4.33	5.72
Infovox SA 101	10.00	15.00	12.50
Berkeley	9.78	18.50	14.14
TSI-Prototype 1	10.78	24.72	17.75
Votrax Type'n'Talk	32.56	22.33	27.44
Echo	35.56	35.56	35.56

Examination of the patterns of errors obtained in these tests shows a remarkable degree of regularity in the common error types for initial position: the stops /k/, /g/, /b/, and /p/, the approximants /h/ and /w/, and the fricative /f/ account for most of the errors observed across the different voices. In final position, a slightly different error pattern was observed across the different voices. The stops no longer dominate the errors; along with the stops /k/, /p/, /t/, and /d/, there is also a wider variety of fricative errors than occurred in initial position, including the phonemes /ɵ/, /f/, and /v/. In addition, the nasals /n/, /m/, and /ŋ/ also contribute a large proportion of the total error for each voice.

Closed versus Open MRT Results

We have tested a subset of the synthetic voices used in the closed version of the MRT using the open-response format MRT. These eight voices were DECtalk 1.8 Paul, DECtalk 1.8 Betty, Prose 2000 V3.0, MITalk-79, Infovox, Votrax, and Echo. The natural speech condition was included as well. Figure 2 shows the mean error rates for

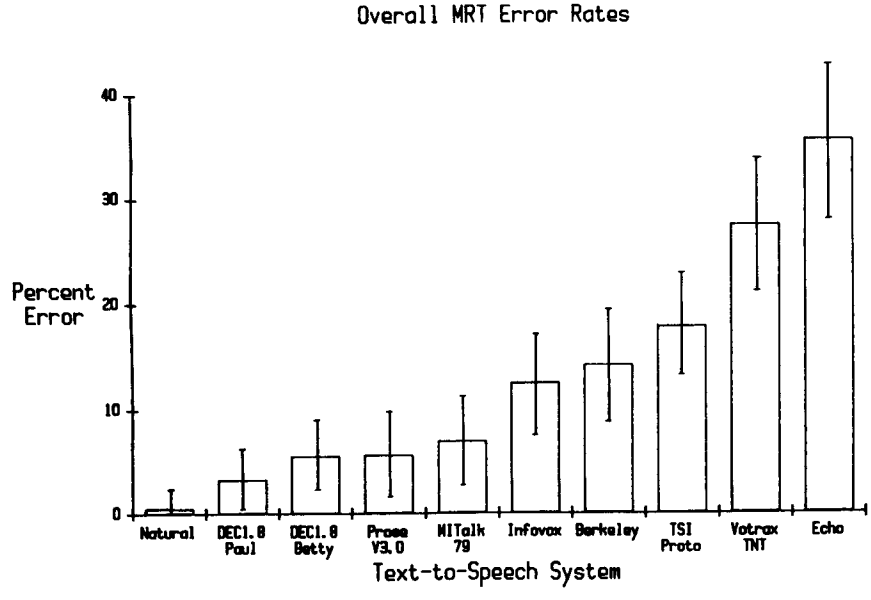

Figure 1. Overall error rates (in percent) for each of the 10 voices tested in the MRT.

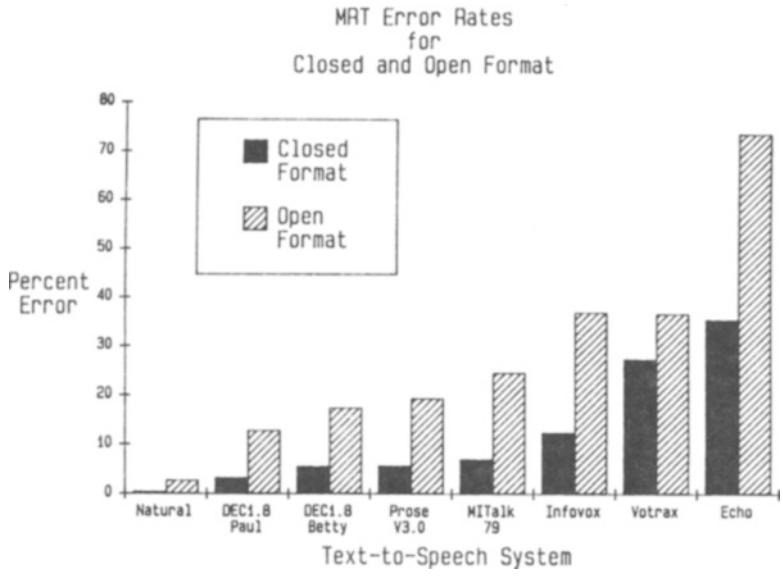

Figure 2. Error rates (in percent) for 10 voices in both the closed- and open-response format MRT. Open bars designate error rates for the closed-response format and striped bars designate error rates for the open-response format.

the eight voices in both the closed- and open-response versions of the MRT. Examination of Figure 2 shows that as the error rate increases from left to right in the figure, the increase in the error rate for the open version of the MRT is much greater than the rate of increase for the closed version of the MRT. When response constraints are removed in the open test, the increase in error rate is as much as 40% greater than the error rate obtained in the closed version.

Summary of Overall Results

An examination of the overall error rates for the 10 voices tested in our laboratory suggests the presence of four distinct groupings: (1) natural speech, (2) high-quality synthetic speech (DECtalk 1.8 Paul and Betty, Prose 3.0 and MITalk-79), (3) moderate-quality synthetic speech (Infovox SA101, Berkeley, and TSI proto-1), and (4) low-quality synthetic speech (Votrax Type'n'Talk and Echo). In comparing the error rates for consonants in initial and final positions, the same general pattern was observed as for overall performance. These four groupings reflect the adequacy of the phonetic implementation rules used in the individual text-to-speech systems, which in turn is directly related to the amount of speech knowledge incorporated into each system (Nusbaum & Pisoni, 1985; Pisoni, Nusbaum, & Greene, 1985).

The largest proportion of errors occurring in initial position were due to only a few stops and fricatives. Similarly, the largest proportion of errors in final position were due to only a few stops, fricatives and nasals. These common error patterns, across a wide range of synthetic voices, suggest that some phonemes may be inherently difficult to perceive, especially since the phonemes typically misperceived in natural speech also tend to be those misperceived in synthetic speech. However, the error rates for synthetic speech are still substantially higher than those observed for natural speech. More importantly, those phonemes contributing the greatest proportion of the errors are not all phonemes with a low frequency of occurrence, suggesting that an imperfect knowledge concerning the synthesis of these sounds still exists. The phonemes with the highest error rates are typically those with complex spectra or those showing the greatest amount of coarticulation in speech. Although the human perceptual mechanism is capable of compensating for some of the problems in synthesizing these phonemes, a major share of the difficulty lies in the lack of knowledge in synthesizing certain phonemes in specific phonetic environments.

GENERAL CONCLUSIONS

The results of our tests of segmental intelligibility of synthetic speech produced by several text-to-speech systems reveal a strong relationship between the amount of speech knowledge incorporated into the system and the perceptual performance as measured by human observers. Basically, "you get what you pay for!" Text-to-speech systems that come closest to the benchmark results obtained with natural speech are the most expensive. Of course, this is not a surprising finding. High-end systems have had the greatest amount of research and development and have been tested and evaluated more systematically prior to being offered to consumers. Moreover, and perhaps more importantly, these systems include more formal knowledge about the acoustic-phonetic properties of speech in the rule systems used to generate the synthetic speech.

In the present studies, segmental intelligibility of syn-

thetic speech produced by several text-to-speech systems was measured under benign laboratory conditions. Similar perceptual studies under more demanding conditions, such as in noisy environments or under increased cognitive load, would, in all likelihood, yield different results. Indeed, several studies of the perception of synthetic speech in noise carried out in our laboratory have shown that listeners have greater difficulty identifying synthetic targets than natural targets. Moreover, under these adverse conditions, subjects show a higher error rate with poor-quality synthetic speech, such as Votrax, than with higher quality synthetic speech, such as the Prose 2000 or DECtalk (see, e.g., Pisoni & Koen, 1981; Pisoni, Nusbaum, & Greene, 1985; Yuchtman, Nusbaum, & Pisoni, 1985).

Noise is only one of several factors that may produce decrements in performance. Stress, fatigue, and acceleration are likely to affect the perceptual process. In our laboratory, we have also been interested in studying the perception of synthetic speech under conditions of high cognitive load. Both the attentional load required by the task and the intelligibility of the speech appear to interact to affect performance (Greenspan, Nusbaum, & Pisoni, 1985).

In other studies, we have also presented listeners with sentences produced by various text-to-speech systems. In these experiments, listeners use the sentence context to help them understand ambiguous or poorly synthesized segments and, therefore, overcome one of the problems encountered in listening to poorly synthesized isolated words. The results obtained in these sentence perception studies show the same general pattern of results that we found in the MRTs—the rank ordering of the different text-to-speech systems falls into the same four categories observed earlier (Manous, Pisoni, Dedina, & Nusbaum, 1985; Pisoni, Nusbaum, & Greene, 1985).

A number of important questions still remain to be examined in future research on the perception of synthetic speech. How will listeners respond when listening to long passages of synthetic speech? Will they find it more tiring or boring than listening to natural speech? Will there be a loss of attention and concentration? Are the comprehension processes affected by the quality of the acoustic-phonetic input? These are just a few of the questions we are interested in examining in the future.

Generalizations of the MRT results reported here need to be made cautiously since the present tests were carried out under ideal circumstances in a laboratory environment. Substantial differences in performance can be anticipated if synthetic speech is presented in noise, under conditions of high cognitive load, or in applications that require differential attentional demands to several input signals at the same time. Obviously, further work is needed on these topics. To our knowledge, there has been little if any basic research directed at these problems in the past. Hopefully, more research will be carried out on problems such as these in order to learn more about the differences in perception between natural speech and various kinds of synthesized speech generated by rule.

NEW DIRECTIONS AND IMPROVEMENTS

With error rates for segmental intelligibility of isolated monosyllabic words in the range of 3% to 4% for the best text-to-speech system we have tested to date, performance is rapidly approaching asymptote. A great deal of further refinement and research probably will be necessary to improve segmental intelligibility much above these levels of performance. At this time, it is probably more productive to look for ways to improve prosody—the amplitude, timing, and durations of individual sounds and words in sentences and the perceived naturalness of synthetic speech. There is a belief among speech researchers that the mechanical sounding quality of synthetic speech is primarily related to the poor knowledge of prosody and the relatively simple algorithms that are currently used to compute pitch and duration in sentences. There is also a need to improve the naturalness of synthetic speech and to further investigate the factors that control a listener's preference for one synthetic voice over another. It is also very likely that the specific application will play an important role in influencing judgments of naturalness and preference among synthetic voices. For the present, however, our studies demonstrate that very high-quality synthetic speech is commercially available and can be incorporated into a wide variety of applications requiring voice output of unrestricted English text.

REFERENCES

ALLEN, J. (1973a). Reading machines for the blind: The technical problems and the methods adopted for their solution. *IEEE Transactions on Audio and Electroacoustics, AU-21*, 3, 259-264.

ALLEN, J. (1973b). Speech synthesis from unrestricted text. In J. L. Flanagan & L. R. Rabiner (Eds.), *Speech synthesis* (pp. 416-428). Stroudsburg, PA: Dowden, Hutchinson, & Ross.

ALLEN, J. (1976). Synthesis of speech from unrestricted text. *Proceedings of the IEEE*, **64**, 433-442.

ALLEN, J. (1981). Linguistic-based algorithms offer practical text-to-speech systems. *Speech Technology*, **1**, 12-16.

ALLEN, J. (in press). *From text to speech*. Cambridge, MA: Cambridge University Press.

BERNSTEIN, J., & PISONI, D. B. (1980). Unlimited text-to-speech device: Description and evaluation of a microprocessor-based system. In *1980 IEEE International Conference Record on Acoustics, Speech, and Signal Processing* (pp. 576-579). New York: IEEE Press.

BRUCKERT, E. (1984). A new text-to-speech product produces dynamic human-quality voice. *Speech Technology*, **2**, 114-119.

CARLSON, R., GRANSTROM, B., & HUNNICUTT, S. (1982). A multilanguage text-to-speech module. *Proceedings of the International Conference on Acoustics, Speech, & Signal Processing*, **3**, 1604-1607.

COOPER, F. S. (1963). Speech from stored data. *IEEE International Convention Record*, **7**, 137-149.

COOPER, F. S., GAITENBY, J. H., MATTINGLY, I. G., & UMEDA, N. (1969). Reading aids for the blind: A special case of machine-to-man communication. *IEEE Transactions on Audio and Electroacoustics, AU-17*, 266-270.

ECHO. (1982). *User's Manual*, Echo Synthesizer. Carpinteria, CA: Street Electronics Corp.

FAIRBANKS, G. (1958). Test of phonemic differentiation: The rhyme test. *Journal of the Acoustical Society of America*, **30**, 596-600.

GREENE, B. G., LOGAN, J. S., & PISONI, D. B. (1984). *Perceptual evaluation of the Berkeley Systems Works text-to-speech system.* (SRL Tech. Note No. 84-05). Bloomington: Indiana University, Speech Research Laboratory, Department of Psychology.

GREENE, B., G., LOGAN, J. S., & PISONI, D. B. (1985). *Perceptual evaluation of the Infovox text-to-speech system.* (SRL Tech. Note No. 85-03). Bloomington: Indiana University, Speech Research Laboratory, Department of Psychology.

GREENE, B. G., MANOUS, L. M., & PISONI, D. B. (1984). Perceptual evaluation of DECtalk: Final report on version 1.8. *Research on Speech Perception Progress Report No. 10.* Bloomington: Indiana University, Speech Research Laboratory, Department of Psychology.

GREENSPAN, S., NUSBAUM, H. C., & PISONI, D. B. (1985). Perception of synthetic speech: Some effects of training and attentional limitations. *Research on Speech Perception Progress Report No. 11.* Bloomington: Indiana University, Speech Research Laboratory, Department of Psychology.

GRONER, G. F., BERNSTEIN, J., INGBER, E., PEARLMAN, J., & TOAL, T. (1982). A real-time text-to-speech converter. *Speech Technology*, **1**, 73-76.

HOUSE, A. S., WILLIAMS, C. E., HECKER, M. H., & KRYTER, K. D. (1965). Articulation testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, **37**, 158-166.

HUNNICUTT, S. (1976). Phonological rules for a text-to-speech system. *American Journal of Computational Linguistics*, Microfiche **57**.

KLATT, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, **67**, 971-995.

KLATT, D. H. (1982). The Klattalk text-to-speech system. *Proceedings of the International Conference on Acoustics, Speech, & Signal Processing*, **3**, 1589-1592.

LINDGREN, N. (1967). Speech: Man's natural communication. *IEEE Spectrum*, **4**, 75-86.

LOGAN, J. S., GREENE, B. G., & PISONI, D. B. (1985). Perceptual evaluation of the Prose 3.0 text-to-speech system. (SRL Tech. Note No. 85-05). Bloomington: Indiana University, Speech Research Laboratory, Department of Psychology.

LOGAN, J. S., PISONI, D. B., & GREENE, B. G. (1985). Measuring the segmental intelligibility of synthetic speech: MRT results for eight text-to-speech systems (Research on Speech Perception Progress Report No. 11). Bloomington: Indiana University, Speech Research Laboratory, Department of Psychology.

MAGNUSSON, L., BLOMBERG, M., CARLSON, R., ELENIUS, K., & GRANSTROM, B. (1984). Swedish speech researchers team-up with electronic venture capitalists. *Speech Technology*, **2**, 15-24.

MANOUS, L. M., PISONI, D. B., DEDINA, M. J., & NUSBAUM, H. C. (1985). Comprehension of natural and synthetic speech using a sentence verification task. *Research on Speech Perception Progress Report No. 11.* Bloomington: Indiana University, Speech Research Laboratory, Department of Psychology.

MORRIS, L. R. (1979). A fast FORTRAN implementation of the U.S. Naval Research Laboratory algorithm for automatic translation of English text to Votrax parameters. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 907-913). Washington, DC: IEEE Press.

NUSBAUM, H. C., & PISONI, D. B. (1985). Constraints on the perception of synthetic speech generated by rule. *Behavior Research Methods, Instruments, & Computers*, **17**, 235-242.

NYE, P. W., & GAITENBY, J. (1973). Consonant intelligibility in synthetic speech and in a natural speech control (Modified Rhyme Test results). *Haskins Laboratories Status Report on Speech Research* (SR-33, pp. 77-91). New Haven, CT: Haskins Laboratories.

NYE, P. W., & GAITENBY, J. H. (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. *Haskins Laboratories Status Report on Speech Research* (SR-37/38, pp. 169-190). New Haven, CT: Haskins Laboratories.

PISONI, D. B. (1982). Perception of speech: The human listener as a cognitive interface. *Speech Technology*, **1**, 10-23.

PISONI, D. B. (in press). Some measures of intelligibility and comprehension. In J. Allen (Ed.), *From text to speech.* Cambridge, MA: Cambridge University Press.

PISONI, D. B., & HUNNICUTT, S. (1980). Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 572-575). New York: IEEE Press.

PISONI, D. B., & KOEN, E. (1981). Some comparisons of intelligibility of synthetic and natural speech at different speech-to-noise ratios. *Research on Speech Perception Progress Report No. 7.* Bloomington: Indiana University, Speech Research Laboratory, Department of Psychology.

PISONI, D. B., NUSBAUM, H. C., & GREENE, B. G. (1985). Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, **73**, 1665-1676.

STUDDERT-KENNEDY, M., & COOPER, F. S. (1966). High-performance reading machines for the blind. In R. Dufton (Ed.), *Proceedings of the International Conference on Sensory Devices for the Blind* (pp. 317-342). London: St. Dunstans.

VOTRAX, INC. (1981). *User's Manual.* Troy, MI: Author.

YUCHTMAN, M. NUSBAUM, H. C., & PISONI, D. B. (1985, November). *Consonant confusions and perceptual spaces for natural and synthetic speech.* Paper presented at 110th meeting of Acoustical Society of America, Nashville, TN.