Published in final edited form as:
J Acoust Soc Am. 1989 August ; 86(2): 566–581.

Segmental intelligibility of synthetic speech produced by rule

John S. Logan, **Beth G. Greene**, and **David B. Pisoni**
Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405

Abstract

This paper reports the results of an investigation that employed the modified rhyme test (MRT) to measure the segmental intelligibility of synthetic speech generated automatically by rule. Synthetic speech produced by ten text-to-speech systems was studied and compared to natural speech. A variation of the standard MRT was also used to study the effects of response set size on perceptual confusions. Results indicated that the segmental intelligibility scores formed a continuum. Several systems displayed very high levels of performance that were close to or equal to scores obtained with natural speech; other systems displayed substantially worse performance compared to natural speech. The overall performance of the best system, DECtalk—Paul, was equivalent to the data obtained with natural speech for consonants in syllable-initial position. The findings from this study are discussed in terms of the use of a set of standardized procedures for measuring intelligibility of synthetic speech under controlled laboratory conditions. Recent work investigating the perception of synthetic speech under more severe conditions in which greater demands are made on the listener's processing resources is also considered. The wide range of intelligibility scores obtained in the present study demonstrates important differences in perception and suggests that not all synthetic speech is perceptually equivalent to the listener.

INTRODUCTION

Standardized tests of segmental intelligibility for speech communication systems have been in existence since the advent of widespread telephone usage during the early years of the present century. Faced with evaluating the output of various communication devices, engineers needed formal tests that could be used to assess listeners' perception of the fidelity of signal transmission (Fletcher and Steinberg, 1929). During the Second World War, Egan and his associates at the Psycho-Acoustics Laboratory at Harvard University developed the phonetically balanced (PB) word lists to measure speech intelligibility (Egan, 1948). Each of the PB lists contained words that reflected the frequency distribution of phonemes used in spoken English in order to validate the test materials and to generalize findings to the language as a whole. These stimulus items were all monosyllabic words with a limited range of consonant–vowel combinations. In addition, the relative difficulty of the stimulus items was constrained so that items that were always missed or were always correct were removed, leaving only those items that provided useful information. These test words were then administered to trained groups of listeners under controlled conditions to facilitate comparison of performance among various devices.

The factors that Egan and his associates took into account when they designed the PB word lists reflected many of the basic attributes that are important in constructing intelligibility tests. Some further pragmatic requirements for intelligibility tests were described by Fairbanks (1958). These included ease and speed of scoring, use of untrained listeners, and

the length of time needed for administration. To meet these requirements, Fairbanks (1958) developed the rhyme test, which consisted of five lists of 50 words each. 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 and those that could be spelled with one letter were tested, excluding the phonemes /ŋ/, /ʒ/, /θ/, /ð/, /š/, and /č/. The rhyme test was therefore limited in its coverage of English phonology.

In response to some of the deficiencies that existed in the PB word lists and the rhyme test, House and his co-workers (House *et al.*, 1965) developed the modified rhyme test (MRT). Whereas the PB words required relatively complex scoring procedures, 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 forced-choice closed-response test with six response alternatives available to subjects for each stimulus presentation.[1,2]

Until the present time, only a small number of tests have been specifically designed to assess the segmental intelligibility of systems that generate synthetic speech. In their initial evaluation of the Haskins speech synthesis system, Nye and Gaitenby (1973) used the MRT to examine the segmental intelligibility of the Haskins system and to compare its performance to natural speech. They reported that the MRT had several deficiencies that, in their view, made it less than optimal as a diagnostic tool for isolating poorly synthesized phonemes. Although some poorly synthesized phonemes could be identified with the MRT, the closed response set limited the number of possible perceptual confusions that subjects could generate. Consequently, Nye and Gaitenby developed a set of special test sentences, often referred to in the literature as the Haskins semantically anomalous sentences, that were used to evaluate the intelligibility of the Haskins synthesis system (Nye and Gaitenby, 1974). These materials were the first attempt to develop a test that was designed specifically for use in evaluating synthetic speech produced by rule.

Tests of speech intelligibility are often compromises between a number of competing criteria. When considering intelligibility data obtained using the MRT, several factors, both positive and negative, should be taken into account. Two factors can be distinguished: those intrinsic to the MRT itself, such as how effectively it detects poorly synthesized phonemes, and those that are extrinsic to the MRT, such as how well it facilitates comparisons of intelligibility among different systems. Several issues related to these factors are discussed below.

First, it should be emphasized that the MRT only provides information on segmental intelligibility of isolated monosyllabic words, limiting inferences regarding intelligibility of more complex words and words in sentences. Also, in the closed format, the most widely

[1]Kryter and Whitman (1965) compared the PB lists and the MRT and found that, if the complete set of 1000 PB words was used, performance using the MRT was better, especially at low signal-to-noise ratios. If the number of the PB words was reduced (Miller *et al.*, 1951), the smaller response set resulted in performance approximately equal to that obtained with the MRT.

[2]There exist a number of additional tests of intelligibility, each designed to satisfy the requirements of efficiency, ease of administration and scoring, generalizability, and comprehensiveness that are seen as necessary for an adequate test of speech intelligibility. Most of these tests have been developed for use with a specific purpose in mind. For example, the CID W-22 list (Hirsh *et al.*, 1952) was developed to assess the performance of hearing-impaired listeners. A number of other tests were also designed to evaluate speech processing devices, including vocoders. These tests include the diagnostic rhyme test (Voiers *et al.*, 1965; Voiers, 1983), the phoneme-specific intelligibility test (Stevens, 1962), and the phoneme-specific sentences (Huggins and Nickerson, 1985), CNClists (Lehiste and Peterson, 1959), the Mitchell lists (Mitchell, 1974), the consonant recognition test [Voiers *et al.*, 1976 (cited in Kalikow *et al.*, 1976)], the four alternative auditory feature test (Foster and Haggard, 1984), and Clark's CV syllables (Clark *et al.*, 1985). For a more exhaustive examination of tests of intelligibility, see Kalikow *et al.* (1976).

used version of the test, information is available only for consonants in initial and final position (House *et al.*, 1965; Nye and Gaitenby, 1974). Furthermore, the phonemes in these positions are nearly all singleton consonants; no information on consonant clusters or other syllable structures is provided. The absence of this information may be a serious limitation on the extent to which the results of the MRT can be generalized to a wider range of linguistic materials, particularly because knowledge of the contextual effects of coarticulation in many of the possible phonetic environments is absent from the monosyllabic vocabulary used in the MRT. The application of many phonological rules is not possible in the highly constrained CVC environment of the MRT, a limitation that exists not only for closed versions of the MRT but also for versions that do not have a closed response set (i.e., an open-response test in which the possible response set is limited only by the number of entries in the listener's own lexicon). In addition, the stimulus items used in the MRT are not entirely representative of the distribution of phonemes found in English (Nye and Gaitenby, 1974). Finally, questions related to listener preference, naturalness, and comprehension of fluent passages of synthetic speech cannot be addressed by examination of intelligibility scores obtained using the MRT (Nusbaum *et al.*, 1984; Logan and Pisoni, 1986). Most of the limitations of the MRT described above are factors intrinsic to the MRT itself and hence are difficult to address without designing an entirely new test.

Setting these criticisms aside, there are many factors that make the MRT a useful instrument for assessing the intelligibility of synthetic speech. As a first approximation, MRT scores give a good estimate of the overall system performance for a well-defined vocabulary. In addition, the results described below demonstrate that the MRT is a reliable test by which the segmental intelligibility of a given system can be measured with some confidence. Another useful characteristic of the MRT is that it is easily administered to untrained listeners. No special training in phonetic transcription is required in order to take the test. Moreover, conclusions obtained with a group of untrained subjects using the MRT may be more readily generalizable to the population of potential users than tests requiring specially trained listeners.

When compared to other tests of segmental intelligibility, such as the diagnostic rhyme test (DRT) (Voiers, 1966) and the Haskins anomalous sentences, the MRT has additional characteristics that place it on equal footing with these other tests and, in some cases, may make it even more suitable for measuring segmental intelligibility. For example, assessing segmental intelligibility in the context of sentence length material makes interpretation of phonetic errors more difficult than in tests like the MRT because of factors such as timing, prosody, and the allophonic variation present in sentences. These additional factors make it more difficult to determine precisely what factor was actually responsible for the particular errors observed. Thus, for purposes of assessing segmental intelligibility, test materials such as the Haskins sentences may be best viewed as complementary to tests using isolated words such as the MRT. In relation to other tests of intelligibility that use isolated words, such as the DRT, the MRT provides comparable information (see Pratt, 1987). The DRT's strength is that it provides information on perceptual confusions based on one-feature differences between the alternative choices in its response set. However, it is not clear that the two-alternative forced-choice procedure employed in the DRT has an obvious advantage over the procedure used in the MRT. If detailed information concerning perceptual confusions is required, an open-response MRT may be administered to supplement the information obtained using the closed-format MRT. In general, the rank ordering of MRT scores for several of the systems reported here correlates well with performance on other tasks, including perception of Harvard and Haskins sentences (Greene *et al.*, 1984), speeded verification tests (Pisoni *et al.*, 1987), subjective evaluations (Logan and Pisoni, 1986), and the DRT (Pratt, 1987). Finally, the results of the present study have generated a modest body

of data, making comparisons among a relatively large number of systems straightforward and standardized.

The studies reported in the present paper began in 1979 when one of the authors carried out a perceptual evaluation of the MITalk-79 text-to-speech system using the MRT along with several other measures of word recognition and listening comprehension (see Pisoni and Hunnicutt, 1980). The importance of the MITalk text-to-speech system was that, although originally designed for research purposes, its general architecture later served as the basis for a number of commercial text-to-speech systems (see Klatt, 1987). Results of this initial evaluation were useful because they provided a baseline for comparing the performance of later systems (see Pisoni, 1987). Since the time of this first investigation, a number of other text-to-speech systems have been developed and tests were carried out in our laboratory to assess their intelligibility and to compare their performance with natural speech.

In the sections below, we present data from the MRT obtained with ten different text-to-speech systems[3]: MITalk-79, TSI (prototype Prose 2000), DECtalk 1.8 Perfect Paul, DECtalk 1.8 Beautiful Betty, Prose 2000 V3.0, Infovox SA 101, Votrax Type'n'Talk, Echo, Amiga SoftVoice, and Smoothtalker. We also collected data using natural speech to serve as a benchmark condition for comparison purposes. The primary questions of interest were whether the synthetic speech produced by these various text-to-speech systems differed significantly from each other in terms of segmental intelligibility and also whether the speech of the most intelligible systems differed from natural speech. We were also interested in the common and distinctive error patterns produced by these systems. Several questions regarding the use of the MRT were also examined, such as its reliability as a test instrument and the kinds of differences that exist between the open- and closed-response formats. The results of this investigation should therefore be of interest to a variety of researchers who are interested in comparing the intelligibility of different systems studied under the same laboratory conditions. The results should also be useful to researchers developing new tests of intelligibility.

I. METHOD

The same general experimental methodology has been followed in all the tests carried out in our laboratory.

A. Subjects

Subjects for all of the perceptual tests were obtained from two primary sources: undergraduate students at Indiana University who received course credit for their participation as part of a requirement for an introductory psychology class, or paid subjects obtained from a voluntary subject pool of students maintained by the Speech Research Laboratory. Both sources were drawn from the same undergraduate population in Bloomington, Indiana. All subjects were native speakers of English who reported no prior history of a speech or hearing disorder or any previous experience listening to synthesized speech. A total of 72 subjects participated in each evaluation (where "evaluation" refers to a test of an individual system using a specific test format). All comparisons reported below among systems, test formats, etc., were between-subjects comparisons.

[3]For purposes of clarity, a "system" refers not only to a specific text-to-speech system but also to the individual voices produced by a text-to-speech system. Thus DECtalk Paul and DECtalk Betty are considered different systems (as well as different voices) in the context of the present paper.

B. Materials

The stimulus materials were obtained from the MRT lists developed by House *et al.* (1965) and consisted of 300 English words arranged into six lists of 50 words each. In the closed-format MRT, subjects were provided with six alternative responses for each test trial on specially prepared answer sheets. Each of the alternatives was the correct response on one of the stimulus lists. In the open-response version of the test, subjects were provided with an answer booklet containing a numbered set of blank lines on which to write down each word that they heard. The words in each list were arranged in one of two possible ways. In the blocked form of the test, the first half of each list contained words varying in final consonant and the second half of each list contained words varying in initial consonant. We also used a mixed version of the test in which items were randomly mixed from trial to trial. Each list was randomized twice, resulting in a total of 12 lists. The blocked version of the MRT was used in the evaluation of natural speech, MITalk-79, TSI, and DECtalk 1.8 Perfect Paul. The mixed version of the MRT was used in the evaluation of DECtalk 1.8 Perfect Paul, DECtalk 1.8 Beautiful Betty, Prose 2000 V3.0, Infovox SA 101, Votrax Type'n'Talk, Echo, Amiga Softvoice, and Smoothtalker.

The stimulus items were generated by each text-to-speech system and recorded on audio tape for later playback to subjects. Each test tape began with a 20-s sustained vowel /a/ for use in calibrating playback levels. This was followed by a short description of the experimental task in order to familiarize subjects with the speech quality of the system to be tested. The test items were then presented with an interstimulus interval of 4 s. All audio tapes were made in our laboratory on a Crown 800 series tape recorder except for the MITalk-79 and TSI Prototype-1 tapes, which were prepared according to our specifications. The tapes for the Amiga and Smoothtalker systems were made from digitized files using a PDP 11/34 computer with 12-bit A/D and D/A converters. The tapes containing natural speech were originally recorded on audio tape by Dennis Klatt at MIT in 1979. The experimental conditions used to assess performance with natural speech were identical to those used with the synthetic speech.

C. Equipment

The text-to-speech systems used in the present investigation ranged from an experimental research system running on a large mainframe computer that did not run in real time to relatively inexpensive units designed primarily for hobbyists. Table I provides some of the technical details associated with the text-to-speech systems we tested. A brief description of each system is given below. The order of presentation is determined roughly by the date of testing. Descriptions of some systems are necessarily more brief than others owing to the paucity of information in the public domain available for those particular systems or because of redundancy with earlier descriptions. Further general information about the design of many of the systems and their history can be found in a recent comprehensive review article by Klatt (1987).

1. MITalk-79—The MITalk-79 system was developed as a research tool at MIT (see Allen *et al.*, 1987) and was implemented on a DECSYSTEM-20 computer. The system was the product of a 10-year effort at MIT to develop a system that would automatically convert unrestricted English text into high-quality synthetic speech (Allen, 1976, 1981; Allen *et al.*, 1987). MITalk consisted of a number of programs that first analyzed the text input in terms of morphological composition and performed a lexical lookup operation to determine whether each morpheme was present in a large, precompiled 12 000-item dictionary. If the morphemes comprising the words were not found in the dictionary, another module containing approximately 400 letter-to-sound rules was used to determine a pronunciation (a phonetic representation) of the morphemes (Hunnicutt, 1976). 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). Due primarily to the time required for input/output operations, the MITalk system ran in about ten times real time (see Allen *et al.*, 1987, for a history and complete description).

2. Telesensory Systems, TSI Prototype-1 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 on the MITalk-77 system but used only an 1100-word dictionary for lexical lookup. The MITalk parsing system was omitted and the fundamental frequency module was replaced with a "hat and declination" routine. In addition, the TSI system was implemented using IC technology and ran in real time (see Bernstein and Pisoni, 1980; Groner *et al.*, 1982, for further details).

3. Digital Equipment DECtalk V1.8—DECtalk 1.8 was a text-to-speech system produced commercially by Digital Equipment Corporation (DEC). It was originally based on the MITalk-79 system but new letter-to-phoneme rules developed by Hunnicutt (1980) were added by Klatt to produce a system known as Klattalk (Klatt, 1982). In 1982, Klattalk was licensed to DEC for commercial use. DECtalk 1.8 had seven different voices, two of which were studied in the present investigation. For further details, see Bruckert *et al.* (1983) and Klatt (1987).

4. Infovox SA 101—The Infovox SA 101 text-to-speech system was developed in Sweden at the Royal Institute of Technology by Carlson *et al.* (1982) and commercially implemented by Infovox, a company owned by the Swedish National Development Company and the Ventronic Company (Magnusson *et al.*, 1984). A unique feature of this system was its multilingual capabilities; text could be processed using spelling-to-sound rules for English, French, Spanish, German, Italian, and Swedish. Only the English version of this system was tested. See Klatt (1987) for additional information.

5. Speech Plus Prose 2000 V3.0—The Prose 2000 V3.0 was a complete redesign of the prototype TSI system tested in 1979, sharing only the Klatt synthesis routines. See Groner *et al.* (1982) and Klatt (1987) for further details.

6. Votrax Type'n'Talk—The Votrax Type'n'Talk was 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 uses formant synthesis techniques to produce speech. See Klatt (1987) for further details.

7. Street Electronics Echo—The Echo text-to-speech system was 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 *et al.*, 1976), text was converted into allophonic control codes, which were then converted to speech using linear predictive coding (LPC) synthesis by a Texas Instruments TMS-5220 chip (Echo, 1982). See Klatt (1987) for further details.

8. Amiga—The Amiga SoftVoice system was a feature of the Amiga series of personal computers produced by Commodore Business Machines. It consists of software that translates text into phoneme codes, which are then used to generate time-varying parameters

for a three-formant synthesizer. For exception words, a dictionary is used to provide the phoneme codes. The resultant waveform is then sent to the 8-bit internal D/A system of the Amiga (Commodore Business Machines, 1987).

9. Smoothtalker—The Smoothtalker system was a software package for personal computers produced by First Byte, Inc. Because it uses the hardware contained in the host computer to generate synthesized speech, the actual implementation of the system varies from computer to computer. The version we tested was designed for the Apple Macintosh personal computer and used the internal D/A system of the Macintosh. Text is parsed using proprietary letter-to-sound rules which serve to generate control codes for prestored allophonic segments. The segments are then concatenated together to produce a speech waveform.

D. Procedure

Subjects were tested in groups of six or less in a quiet room containing six individual cubicles, each equipped with a desk and a set of high-quality headphones. Subjects were told that the experiment in which they were about to participate was concerned with the perception of synthetic speech produced by a computer. They were informed that they would hear a single isolated English word on each trial of the test and that their task was to indicate on the answer sheet the word they heard on each trial. Subjects were told to respond on every trial and to guess if they were uncertain about a response. Both a closed-format and an open-format version of the MRT were used in the perceptual evaluations described in the present report. For the closed-format tests, subjects were provided with a response form containing two sheets, each with 25 trials. For each trial, the form contained six response alternatives, one of which was the correct response. All ten types of speech described above were tested using the closed-format MRT. For the open-response test, subjects were told to write down the English word they heard on each trial. The response sheet for the open test contained 50 blank lines. A subset of the systems that were tested using the closed-format MRT was also tested using the open-response test.

The stimulus tapes were reproduced using an Ampex AG-500 tape recorder and presented binaurally over matched and calibrated Telephonics TDH-39 headphones. The tapes were played back at approximately 80 dB SPL measured for the calibration vowel by a Hewlett–Packard 400H VTVM. Broadband white noise generated by a Grason-Stadler 1724 noise generator was presented at 55 dB SPL and was mixed with the speech to mask tape noise, extraneous electronic noises produced by the synthesizers, and any low-level environmental noise in the testing area.

II. RESULTS

A. Closed-format MRT performance

1. Overall error rates—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 performance on initial versus final consonants. Figure 1 shows the mean overall error rate and standard error for each of the text-to-speech systems tested. Comparable data from the natural speech control condition are also shown here. Figure 2 shows the mean error rate and corresponding standard error separately for consonants in initial and final position. Examination of the data in both figures reveals a fairly wide range of performance for the different systems. The best performance for synthetic speech was obtained with DECtalk Paul for consonants in initial position and Prose 2000 V3.0 for consonants in final position. The worst performance was obtained with

Echo in both initial and final position. Table II provides a numerical summary of the error rates shown in Figs. 1 and 2. Each value was calculated by dividing the number of errors by the total number of responses available for each category.

One of the questions of major interest in this study was whether any statistically reliable differences would be observed in MRT scores for the different text-to-speech systems. In addition, we were interested in determining whether or not there were any differences in performance between the best text-to-speech systems and natural speech. Finally, the data were also used to assess the reliability of the MRT with synthetic speech. This was done for one voice generated by one system using a test–retest procedure.

To assess differences in performance between systems, the error rates for initial and final consonants were analyzed in an 11 × 2 mixed ANOVA in which type of speech was treated as a between-subjects factor (ten text-to-speech systems and natural speech) and syllable position (initial versus final consonants) was the within-subjects factor. This analysis revealed highly significant main effects for type of speech [$F(10, 781) = 386.54$, $p < 0.0001$] and syllable position [$F(1, 781) = 22.73$, $p < 0.0001$]. In addition, a significant interaction between type of speech and syllable position [$F(10, 781) = 35.76$, $p < 0.0001$] was also obtained. *Post hoc* Newman–Keuls tests comparing the overall error rates revealed significant differences in performance ($p < 0.05$ or greater) between many of the systems. Table III provides a summary of the results of the *post hoc* tests comparing error rates across systems.

Another set of *post hoc* tests was also carried out to assess differences in error rates as a function of syllable position for the different systems. Comparisons of the error rates for consonants in initial position indicated no significant differences between natural speech and DECtalk 1.8 Paul, DECtalk 1.8 Paul and Betty, DECtalk 1.8 Betty and MITalk-79, Amiga and TSI, Infovox and TSI, and Votrax and Echo. All other comparisons yielded significant differences ($p < 0.05$ or greater) among the types of speech in terms of error rates for consonants in initial position. Additional *post hoc* tests were carried out to examine errors for consonants in final position. These tests showed no significant differences between DECtalk 1.8 Paul and Prose 3.0, as well as DECtalk 1.8 Betty and MITalk-79. All other comparisons revealed significant differences ($p < 0.05$ or greater) among systems for consonants in final position. Table IV displays in tabular form the results of the *post hoc* tests comparing the error rates among systems as a function of consonant position.

2. Specific segmental errors—More detailed information about the performance of each system was provided by an examination and analysis of the individual phoneme errors and the pattern of errors observed in the MRT. Tables V and VI show the phonemes that accounted for the greatest proportion of the total error for each of the systems for initial and final position, respectively. The overall error rates for each system should be kept in mind as a baseline when examining these tables. For example, /k/ accounts for 33.33% of the errors in initial position for natural speech. However, the overall error rate for natural speech was only 0.5%. Since there were only nine errors out a total of 1800 possible responses in initial position for the natural speech control condition, the error rate for this phoneme was based on only three responses.

Examination of the patterns of errors in Table V shows a remarkable degree of regularity in the common error types observed for initial position. All of the errors observed in the natural speech condition were present in the synthesized speech as well. The stops /k/, /g/, /b/, and /p/, the approximants /h/ and /w/, and the fricative /f/ account for most of the errors across the different systems. In final position, shown in Table VI, a somewhat different error pattern was obtained. 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 /θ/, /f/, and /v/. In addition, the nasals /n/, /m/, and /ŋ/ also contribute a large proportion of the total error for each system in final position.

3. Blocked versus mixed formats—The analyses discussed above were primarily designed to examine differences among the text-to-speech systems tested. In addition, several analyses were also carried out to study properties of the MRT when used with synthetic speech. First, a comparison of the blocked versus mixed forms of the MRT was undertaken. This analysis compared the error rates in initial and final position for DECtalk 1.8 Paul in both blocked and mixed formats. DECtalk 1.8 Paul was chosen for the comparison of blocked and mixed formats because we were evaluating this system at the same time we were also considering adoption of the mixed-format MRT to replace the blocked-format MRT. The analysis revealed no significant effect of test format [$F(1, 142) = 0.94$, $p < 0.336$] but a significant main effect of consonant position [$F(1, 142) = 34.52$, $p < 0.0001$] due to the superior intelligibility of consonants in initial position. No significant interaction was obtained between test format and consonant position. Another factor related to the mixed versus blocked variable was the effect of randomization of the stimulus lists in each form of the MRT. In the mixed-format MRT, each test form had two randomizations (randomization A versus randomization B), making a total of 12 different test forms. A one-way ANOVA comparing the error rate across seven sets of data using the mixed MRT was carried out to determine if any difference existed between the A and B randomizations. This analysis also revealed no significant effect [$F(1, 502) = 0.0024$, $p < 0.9612$].

4. Test–retest results—An assessment of the reliability of the MRT as a test instrument for evaluating synthetic speech was carried out using the DECtalk 1.8 Paul voice. Performance on two separate administrations of the mixed-format MRT using 72 subjects each was compared. These data were analyzed in a 2 × 2 ANOVA in which test (test versus retest) was the between-subjects factor and consonant position (initial versus final segment) was the within-subject factor. This analysis revealed no significant effect for the test versus retest variable [$F(1, 142) = 0.0$, $p < 1.0$], but a significant main effect of consonant position [$F(1, 142) = 34.51$, $p < 0.0001$] due to the greater intelligibility of initial consonants. No significant interaction between test and consonant position was obtained.

B. Open-format performance

1. Overall error rates—Several of the systems tested using the closed-response version of the MRT were also studied using an open-response format with the same vocabulary. The open-format test was used with natural speech, MITalk-79, DECtalk 1.8 Perfect Paul, DECtalk 1.8 Beautiful Betty, Prose 2000 V3.0, Infovox SA 101, Votrax Type'n'Talk, Echo, Amiga SoftVoice, and Smoothtalker. With the open-response format test, a much wider range of response confusions can be obtained because subjects are no longer restricted to a set of fixed response alternatives. Figure 3 shows the mean error rates for the nine systems in both the closed and open versions of the MRT. Examination of Fig. 3 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 test is generally much greater than the increase in the error rate for the closed version. An ANOVA comparing the performance of these nine systems in both the closed- and open-response formats revealed highly significant main effects of system [$F(9, 1420) = 1203.51$, $p < 0.0001$] and test format [$F(1,420) = 4400.28$, $p < 0.0001$], and a significant interaction between system and test format [$F(9, 1420) = 152.87$, $p < 0.0001$], confirming the trends displayed in Fig. 3. When response constraints are removed by using the open format, the increase in error rate can be as much as 41 % greater than the error rate obtained in the closed version of the test. Table VII shows the error rates for initial and final consonants for the nine systems tested using the open-response test. These values are

provided so that direct comparisons can be made with the initial and final consonant errors obtained in the closed-format MRT.

2. Specific segmental errors—As was the case with the closed-format MRT, more detailed information about the performance of each system was provided by an examination and analysis of the individual phoneme errors and the pattern of errors observed in the open-response test. Table VIII shows the phonemes that accounted for the largest proportion of errors in initial position; Table IX shows the most common errors in medial position; and Table X shows the most common errors in final position.

An examination of the pattern of errors observed with natural speech is useful as a starting point for comparing the errors associated with each text-to-speech system (although the small number of overall errors obtained with natural speech should be kept in mind at the same time). In initial position, natural speech errors included stops, nasals, and the fricative /f/. For the synthetic speech, on the other hand, three general patterns emerged: First, the bulk of the errors found across the different text-to-speech systems occurred with stops; second, relatively few errors were made on nasals; and third, errors for the phoneme /h/ were found in almost all of the text-to-speech systems, yet they were absent from the set of errors observed with natural speech. Taken together, these results not only demonstrate important differences in perception between natural and synthetic speech but they also show that some common error patterns are also found in both natural and synthetic speech.

In final position, errors in the natural speech condition occurred primarily for nasals, followed by the phonemes /p/ and /ð/. Nasals also accounted for a large proportion of the errors in the synthesized speech. The remaining natural speech errors were also found among the synthesized speech, but to a much lesser degree. A wide variety of errors, including /k/, A/, /v/, /d/, and /g/, were observed in the synthesized speech, although they were absent or rare in natural speech.

The open-response test also provided information about the intelligibility of medial vowels. More than half of the total number of vocalic errors were common to both natural and synthetic speech. The remainder of the vowel errors were unique to synthetic speech. More than three-quarters of these synthetic vowel errors were due to just two phonemes, /e'/ and /æ/.

The overall pattern of errors for phonemes in initial position using the open-response test was comparable to the pattern of errors obtained using the closed-format MRT, with the stops /k/, /b/, and /p/, the approximants /h/ and /w/, and the fricative /f/ accounting for most of the errors. In final position, the pattern of errors was very similar in the open- and closed-format tests, except that a wider range of errors was present in the open-response test.

Despite the greater number of errors in the open-response test, the proportion of phonemes responsible for the largest number of errors was roughly comparable in the two sets of data. Exceptions to this rule, however, were also present. For example, consider nasal errors in initial position. Using the closed-format test, virtually no nasal errors were obtained in the natural speech condition. In contrast, initial errors from two nasals were obtained when the open-response test was used. For synthetic speech, the number of nasal errors in initial position was relatively small for both the closed- and open-response tests. Thus, if only results from the closed test were considered, natural speech and synthetic speech would appear to be perceptually very similar. However, if the results of the open-response test are considered, natural and synthetic speech do differ in the pattern of nasal errors in initial position. Undoubtedly, some of the confusions observed in the closed-format MRT results were due to the limited set of alternatives provided as possible responses. As a consequence,

the constraints imposed by the closed-format MRT yielded a more homogeneous set of errors than those obtained using the open-reponse test. In any event, the overall similarities between the patterns of perceptual confusions for the closed- and open-response tests appear to be greater than the differences between the two tests.

III. DISCUSSION

An examination of the overall error rates for the ten text-to-speech systems tested using the closed-format MRT revealed a wide range of performance. Several text-to-speech systems had error rates close to that of natural speech, whereas others had large error rates, up to 35% worse than natural speech. When the open-format MRT results are considered, the differences in performance between synthetic speech and natural speech were even larger, with the error rates for some text-to-speech systems as much as 70% greater than the error rate observed for natural speech. This wide range of performance reflects a number of factors, including the adequacy and the sophistication of the phonetic implementation rules used in the individual text-to-speech systems and the methods used for synthesis (see Klatt, 1987; Nusbaum and Pisoni, 1985, for further discussion). Attributing differences in intelligibility to specific factors is not necessarily straightforward, however. For example, consider the relationship between synthesis technology and intelligibility performance. At first glance, the results of the present investigation seem to show that systems using formant synthesis are, in general, more intelligible than systems using other types of synthesis technologies, such as LPC synthesis. However, if the cost of the different text-to-speech systems is taken into account, it is clear that those systems demonstrating good to excellent performance were consistently much more expensive than those systems that exhibited poorer performance. Perhaps a more appropriate way to characterize the differences in performance between synthesis systems is to focus on the amount of detailed acoustic–phonetic information that the system uses. For example, synthesis methods that take into account the context-sensitive nature of phonetic segments and coarticulation phenomena in speech are just as important to intelligibility as detailed formal rules that select and modify the correct target values for the individual phonetic segments themselves. Thus systems that simply concatenate prestored segments such as allophones, diphones, or demisyllables can display performance ranging from relatively poor quality (such as obtained with the Echo system in the present study) to very high quality [such as observed with the AT&T Bell Laboratories text-to-speech system (Olive, 1977; Olive and Liberman, 1985)]. In the case of formant synthesis used in high-quality systems such as DECtalk and Prose, a great deal of very detailed acoustic–phonetic knowledge has been formalized in the form of a set of phonetic implementation rules that control the parametric input to the synthesis routines. Substantially less acoustic–phonetic knowledge is present in the rules used in formant synthesis systems like Votrax, which not only sounds less natural but also shows substantially lower intelligibility scores. In short, the most important factor affecting synthesis performance appears to be related to the degree to which the synthesis system is capable of modeling both the temporal and spectral changes found in natural speech.

A. Specific phoneme errors

The overall pattern of phoneme errors that was observed in both the closed- and open-format was similar for both natural and synthetic speech. In general, the same phoneme errors found in natural speech also tended to occur in synthetic speech. The fact that natural and synthetic speech share many common phoneme errors suggests that some phonemes may be inherently more confusable than other phonemes, a result that is consistent with the well-known findings of Miller and Nicely (1955) and others. Other phoneme errors, however, were unique to synthetic speech. Some phonemes, such as /h/, were commonly misperceived across all ten systems. Other phonemes tended to be misperceived on a system-by-system

basis, a finding that was more likely due to the idiosyncracies of individual devices than a general design property common to all synthesis systems.

The open-response MRT proved to be useful in identifying some of the differences between natural and synthetic speech that were not revealed when the closed-format MRT was used. In using an open-response test, listeners were forced to rely entirely on the information in the speech signal to recognize each word. The absence of response constraints in the open-response test provided more information about the perception of phonemes in initial and final position than the closed-format MRT. Moreover, the open-response test also provided useful information about the perception of vowels. In addition to differentiating natural and synthetic speech, the open-response test also proved to be useful in identifying patterns of errors that were unique to specific text-to-speech systems. For example, the only system in which vowels accounted for the largest proportion of errors in final position was Infovox. No other system showed a similar pattern of errors in final position, suggesting the need for revisions and modifications of the rules used to generate vowels in syllable-final position with this system.

B. Functions of the MRT

General statements regarding the present results using the MRT should be viewed in light of some of the limitations described earlier. Although the MRT provides useful information about the segmental intelligibility of isolated monosyllabic words, other information is unavailable from these scores. For example, the perception of phonemes in more complex phonetic environments, such as those found in sentences or in words containing consonant clusters and other syllable structures, is not assessed by the MRT. Information related to user preferences and listener comprehension is also not provided by the MRT. The results of a recent study by Logan and Pisoni (1986) suggest that the segmental intelligibility of a text-to-speech system is positively correlated with the degree of preference for the speech generated by that system. To the extent that these are important factors for both users and developers of text-to-speech systems, the results of the present investigation should be useful as a starting point, a basis for using other types of perceptual tests that would add to the basic information about segmental intelligibility provided by the MRT.

C. Intelligibility under adverse conditions

It is important to emphasize here that the present results were all obtained under benign laboratory conditions, optimizing the level of intelligibility of each system. Signal-to-noise ratios were very high, permitting unimpaired reception of the acoustic–phonetic properties of the signal generated by each system. Performance of the most intelligible text-to-speech systems was close to that found for natural speech. However, under less favorable conditions corresponding to what might be encountered in real-world listening situations, the level of performance obtained for the synthetic speech would, in all cases, be significantly worse. Previous studies have shown that decreasing the signal-to-noise ratio has more deleterious effects on the perception of synthetic speech than on the perception of natural speech (Yuchtman *et al.*, 1985). Furthermore, factors related to increases in attentional load have been shown to have differential effects on the perception of synthetic and natural speech (Luce *et al.*, 1983). These factors were small or nonexistent in the data reported here.

Effects of attentional load have been shown to be greatest in the transfer of the acoustic representation of the speech signal from short-term memory to more permanent storage in long-term memory (Luce *et al.*, 1983). The limited attentional capacity of the human listener appears to be taxed more heavily when decoding the acoustic-phonetic representation of synthetic speech than natural speech (Greenspan *et al.*, 1985). The differential effects of both noise and attentional load on the perception of synthetic speech appear to be due, in

part, to the small number of acoustic cues used to specify phonetic segments in synthetic speech (Nusbaum and Pisoni, 1985). Whereas natural speech contains many redundant cues that help to specify a particular phoneme, synthetic speech uses only a minimal subset of these cues, therefore increasing the likelihood that the perception of synthetic speech will be impaired more than natural speech under adverse conditions where the cues may be differentially masked or obliterated. The lack of redundancy of cues in synthetic speech may be one major reason why it becomes degraded so easily in noise and therefore requires greater processing resources by the listener (Pisoni *et al.*, 1985).

Previous research has shown that the perception of synthetic speech generated by a given text-to-speech system depends on several factors including the task, properties of the signal, and the experience of the observer (Pisoni, 1982; Pisoni *et al.*, 1985). Some preliminary data on the effects of noise on the performance of subjects tested in the MRT using synthetic speech have been reported by Pisoni and Koen (1981). They found that the intelligibility of synthetic speech generated by the MlTalk-79 system was impaired more than the intelligibility of natural speech under several different signal-to-noise ratios. More recently, using synthetic speech, Clark (1983) found that the segments most affected by masking noise were stops and fricatives, speech sounds with relatively low-amplitude spectra. Pratt (1987) also carried out an investigation of the effect of noise on the perception of synthetic speech. Using the DRT, he found that the intelligibility of synthetic speech was more impaired than natural speech when presented in noise. Details of Pratt's results are described below.

D. Perceptual confusions in natural and synthetic speech

An additional factor that needs to be explored in comparisons of natural and synthetic speech is the observed pattern of perceptual confusions. Nusbaum *et al.* (1984) tested the hypothesis that listening to synthetic speech is similar to listening to natural speech embedded in noise. According to this view, the pattern of confusions for natural and synthetic speech should be similar when natural speech is presented in noise. To test this prediction, Nusbaum *et al.* presented CV nonsense syllables to subjects under two conditions: natural speech at several signal-to-noise ratios and synthetic speech in the quiet. The results showed that the pattern of perceptual confusions differed in the two conditions. Analyses of the confusion matrices generated in their study revealed that synthetic speech was perceptually distinct from natural speech. Some errors were common to both types of speech, indicating that these phonetic segments were confused on the basis of acoustic–phonetic similarity. However, the remainder of the errors were unique to the synthetic speech and appeared to be the result of phonetic miscues, that is, errors arising from the application of ambiguous or incorrect phonetic implementation rules during synthesis. Even in the present study, in which the testing conditions were relatively benign, consistent differences were observed in the pattern of phonetic confusions for natural and synthetic speech, suggesting important differences in the acoustic–phonetic properties of these stimuli.

E. Other studies on the perception of synthetic speech

It is of some interest to compare the present results using the MRT with the results of a study reported recently by Pratt (1987) who used the DRT. As described in the Introduction, the DRT is similar to the MRT. Both are forced-choice rhyming tasks in which the listener chooses the correct response alternative on each trial. In the case of the DRT, only two choices are provided, each alternative differing from the other by only one segmental feature. Pratt (1987) tested 12 types of synthetic speech and a natural speech control condition using the DRT under two signal-to-noise (S/N) conditions, namely, no noise and 0-dB S/N. Five of the ten systems that we tested were also included in the tests carried out

by Pratt. It is important to note that, although Pratt tested five systems using the DRT that overlapped with the present investigation, he used different versions of four of these five systems.[4] In general, the overall rank ordering of the intelligibility of the systems tested by Pratt was similar to the present findings. The scores for DECtalk Betty and Paul were reversed, as were the scores for Prose and Infovox, indicating that agreement between the two sets of tests was not perfect. However, the same systems in which these reversals were noted were those that differed in version number from those we tested. Despite these inconsistencies, the differences that separated the systems were extremely small, suggesting that the MRT and the DRT are providing essentially the same information about the perception of these segmental contrasts. Not surprisingly, performance decreased for all systems when noise was added. One particularly interesting result observed in Pratt's study was an interaction between the intelligibility of natural and synthetic speech as a function of S/N ratio. The addition of noise produced significant differences in intelligibility between natural speech and several types of synthetic speech. These differences in performance were not present when the speech was presented in the quiet.

Several additional studies have also examined intelligibilty of low-priced text-to-speech systems. Hoover *et al.* (1987) compared the intelligibility of Votrax and Echo using single words and sentences. They found that natural speech was always more intelligible than either of the two text-to-speech systems, and that, although Votrax and Echo had comparable error rates for single words, Votrax was slightly more intelligible than Echo when sentence materials were used. These findings are consistent with the results reported here.

In another study, Mirenda and Beukelman (1987) compared the intelligibility of Votrax and a newer version of Echo (Echo II +), plus three DECtalk voices (Paul, Betty, and a child's voice, Kit the Kid) using isolated word and sentence materials. Three age groups were tested: adults, second and third graders, and fifth and sixth graders. The authors found that, for isolated words, DECtalk was the most intelligible, followed by Votrax and then Echo. This was true for all age groups. However, for the sentence-length materials, they found that, in addition to a main effect for the synthesis system, there was also an effect of age, and also an age by system interaction. These results indicate that the rank ordering of systems based on tests of intelligibility administered to adults does not necessarily reflect the performance of children with synthetic speech (see Greene and Pisoni, 1988, for extensive data and discussion of children's perception of synthetic speech produced by rule). The tests carried out by Hoover *et al.* and Mirenda and Beukleman show that the overall rank ordering of the synthetic speech obtained using the MRT compares favorably with the results of other tests of intelligibility, at least with respect to adult performance. Studies with populations other than native adult speakers of English have reported consistently lower scores reflecting the degree of linguistic knowledge and experience that listeners bring to the task (see also Greene, 1986; Ozawa and Logan, 1989).

IV. CONCLUSIONS

The segmental intelligibility of synthetic speech generated by ten text-to-speech systems was examined and compared with the intelligibility of natural speech using the MRT. Overall, the error rates for synthetic speech were higher than those obtained with natural speech. Only in comparing the error rates for initial consonants did any system display performance that was comparable to that obtained with natural speech. Significant

[4]The specific models of the text-to-speech systems used by Pratt (1987) differed from the versions we used in our assessment. In Pratt's study, the following devices were used: DECtalk 2.0 Paul, DECtalk 2.0 Betty, Infovox SA 201, Prose 2000 v1.2, and Votrax Type'n'Talk. A natural speech control condition was also included in Pratt's study.

differences in overall error rate were found among many of the systems. Common patterns of phoneme errors were observed across all the different systems, indicating an imperfect knowledge of how to effectively generate the acoustic cues responsible for the perception of certain phonemes. However, distinctive patterns of phoneme errors were also observed, indicating that not all text-to-speech systems are the same and that references to *some generic* form of synthetic speech are incorrect and potentially misleading. In short, these findings demonstrate a very wide range of performance levels for different kinds of synthetic speech produced by rule.

The results of the present investigation indicate that the MRT can be used as a reliable measure of the segmental intelligibility of synthetic speech produced by a variety of text-to-speech systems. Several variations of the closed-format MRT were found to be functionally equivalent, therefore facilitating comparisons across different versions of the test. Results obtained using the open-format MRT provided additional information supplementing data obtained using the more traditional closed-format MRT. Although the MRT has several limitations as a diagnostic tool, it does provide useful information regarding segmental intelligibility for a constrained vocabulary and therefore permits potential users to make direct comparative judgments about the relative performance of different systems tested under the same laboratory conditions. Using the open and closed versions of the MRT, a general picture of the segmental intelligibility of many phonemes in English can easily be obtained with untrained listeners. Generalizations of the MRT results reported here need to be made cautiously, however, since the present tests were carried out in a benign laboratory environment using adult listeners who were native speakers of English. Substantial differences in performance can be anticipated with other populations of listeners or when synthetic speech, even very high-quality synthetic speech, is presented in noise, under conditions of high cognitive load, or in real-world applications that require differential attentional demands to several competing signals in an observer's environment.

Acknowledgments

We wish to acknowledge Digital Equipment Corporation, Speech Plus, Street Electronics, and Swedish Telecom for their help and cooperation in this research. Without their support this study could not have been carried out. We acknowledge the contribution of Jonathan Allen, Jared Bernstein, Rolf Carlson, Bjorn Granstrom, Sheri Hunnicutt, Dennis Klatt, and many other people at Bell Labs, Haskins, MIT, Royal Institute of Technology, and elsewhere, who spent years developing the hardware and software that led to successful text-to-speech devices. We gratefully acknowledge the important contribution of Amanda Walley, Thomas Carrell, Laura Manous, Kimberly Baals, Jennifer Hearst, and Amy Lawlor in testing subjects and scoring data. We also thank Michael Dedina and Howard Nusbaum for supplying us with speech from the Amiga and Smoothtalker systems, respectively. The helpful comments of Paul Luce on an earlier version of the manuscript are also acknowledged. The research reported in this paper was supported, in part, by NIH Research Grant NS 12179, in part, by Air Force Contract No. AF-F-33615-83-K-0501 and, in part, by NSF Research Grant IRI-86-17847 to Indiana University in Bloomington.

References

Allen J. Synthesis of speech from unrestricted text. Proc. IEEE. 1976; 64:433–442.

Allen J. Linguistic-based algorithms offer practical text-to-speech systems. Speech Tech. 1981; 1:12–16.

Allen, J.; Hunnicutt, S.; Klatt, DH. From Text to Speech: The MITalk System. Cambridge, United Kingdom: Cambridge U. P.; 1987.

Bernstein J, Pisoni DB. Unlimited text-to-speech system: Description and evaluation of a microprocessor based device. Proc. Int. Conf. Acoust. Speech Signal Process. 1980; ICASSP-80:576–579.

Bruckert E, Minow M, Tetschner W. Three-tiered software and VLSI aid developmental system to read text aloud. Electronics. 1983; 56:133–138.

Carlson R, Granstrom B, Hunnicutt S. A multi-language text-to-speech module. Proc. Int. Conf. Acoust. Speech Signal Process. 1982; ICASSP-82:1604–1607.

Clark JE, Dermody P, Palethorpe S. Cue enhancement by stimulus repetition: Natural and synthetic speech comparisons. J. Acoust. Soc. Am. 1985; 78:458–462.

Clark JE. Intelligibility for two synthetic and one natural speech source. J. Phon. 1983; 11:37–49.

Commodore Business Machines, Inc.. Amiga ROM Kernal Reference Manual: Libraries and Devices. Reading, MA: Addison-Wesley; 1987.

Echo. User's Manual. Echo Synthesizer, Street Electronics Corp.; 1982.

Egan JP. Articulation testing methods. Laryngoscope. 1948; 58:955–991. [PubMed: 18887435]

Elovitz H, Johnson R, McHugh A, Shore J. Letter-to-sound rules for automatic translation of English text lo phonetics. IEEE Trans. Acoust. Speech Signal Process. 1976; ASSP-24:446–459.

Fairbanks G. Test of phonemic differentiation: The rhyme test. J. Acoust. Soc. Am. 1958; 30:596–600.

Fletcher H, Steinberg JC. Articulation testing methods. Bell Syst. Tech. J. 1929; 8:806–854.

Foster JR, Haggard MP. Introduction and test manual for FAAF II: The four alternative auditory feature lest (Mark II). MRC Instit. Hear. Res. IHR Internal Rep., Ser. B, No. 11. 1984

Greene BG. Perception of synthetic speech by nonnative speakers of English. Proc. Human Factors Soc. 1986; 2:1340–1343.

Greene, BG.; Manous, LM.; Pisoni, DB. Res. Speech Percept. Progr. Rep. No. 10. Bloomington: Speech Research Laboratory, Indiana University; 1984. Perceptual evaluation of DECtalk: A final report on Version 1.8.

Greene, BG.; Pisoni, DB. Perception of synthetic speech by adults and children: Research on processing voice output from text-to-speech systems. In: Bernstein, LE., editor. The Vocally Impaired: Clinical Practice and Research. Philadelphia: Grune & Stratton; 1988. p. 206-248.

Greenspan, S.; Nusbaum, HC.; Pisoni, DB. Res. Speech Percept. Progr. Rep. No. 11. Bloomington: Speech Research Laboratory, Indiana University; 1985. Perception of synthetic speech generated by rule: Effects of training and attentional limitations.

Groner GF, Bernstein J, Ingber E, Pearlman J, Toal T. A real-time text-to-speech converter. Speech Tech. 1982; 1:73–76.

Hirsh IJ, Davis H, Silverman SR, Reynolds EG, Eldert E, Benson RW. Development of materials for speech audiometry. J. Speech Hear. Disord. 1952; 17:321–337. [PubMed: 13053556]

Hoover J, Reichle J, van Tasell D, Cole D. The intelligibility of synthesized speech: Echo II versus Votrax. J. Speech Hear. Res. 1987; 30:425–431. [PubMed: 2959817]

House AS, Williams CE, Hecker MH, Kryter KD. Articulation-testing methods: Consonantal differentiation with a closed-response set. J. Acoust. Soc. Am. 1965; 37:158–166. [PubMed: 14265103]

Huggins AW, Nickerson RS. Speech quality evaluation using "phoneme-specific" sentences. J. Acoust. Soc. Am. 1985; 77:1896–1906. [PubMed: 3998299]

Hunnicutt S. Phonological rules for a text-to-speech system. Am. J. Computational Ling. Microfiche. 1976; 57:1–72.

Hunnicutt, S. Grapheme-to-phoneme rules, a review. Vol. QPSR 2–3. Stockholm, Sweden: Speech Transmiss. Lab., R. Institute Techn.; 1980. p. 38-60.

Kalikow DN, Huggins AW, Blackman E, Vishu R, Sullivan F. Speech intelligibility and quality measurement. Speech Compression Techniques for Secure Communication, BBN Rep. No. 3226. 1976

Klatt DH. Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am. 1980; 67:971–995.

Klatt DH. The Klattalk text-to-speech system. Proc. Int. Conf. Acoust. Speech Signal Process. 1982; ICASSP-82:1589–1592.

Klatt DH. Review of text-to-speech conversion for English. J. Acoust. Soc. Am. 1987; 82:737–793. [PubMed: 2958525]

Kryter KD, Whitman EC. Some comparisons between rhyme and PB-word tests. J. Acoust. Soc. Am. 1965; 37:1146. [PubMed: 14339726]

Lehiste I, Peterson GE. Linguistic considerations in the study of speech intelligibility. J. Acoust. Soc. Am. 1959; 31:280–287.

Logan, JS.; Pisoni, DB. Res. Speech Percep., Progr. Rep. No. 12. Bloomington: Speech Research Laboratory, Indiana University; 1986. Preference judgements comparing different synthetic voices.

Luce PA, Feustel TC, Pisoni DB. Capacity demands in short-term memory for synthetic and natural word lists. Hum. Factors. 1983; 25:17–32. [PubMed: 6840769]

Magnusson L, Blomberg M, Carlson R, Elenius K, Granstrom B. Swedish speech researchers team-up with electronic venture capitalists. Speech Tech. 1984; 2:15–24.

Miller GA, Heise GA, Lichten W. The intelligibility of speech as a function of the context of the test materials. J. Exp. Psychol. 1951; 41:329–335. [PubMed: 14861384]

Miller GA, Nicely PE. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. 1955; 27:338–352.

Mirenda P, Beukelman DR. A comparison of speech synthesis intelligibility with listeners from three age groups. Augmentive Alternative Commun. 1987; 3:120–128.

Mitchell PD. Test of differentiation of phonetic feature contrasts. J. Acoust. Soc. Am. Suppl. 1. 1974; 55:S55.

Nusbaum, HC.; Dedina, MJ.; Pisoni, DB. Speech Res. Lab. Tech. Note. 84-02. Bloomington: Speech Research Laboratory, Indiana University; 1984. Perceptual confusions of consonants in natural and synthetic CV syllables.

Nusbaum HC, Pisoni DB. Constraints on the perception of synthetic speech generated by rule. Behav. Res. Meth. Instr. Comp. 1985; 17:235–242.

Nusbaum, HC.; Schwab, EC.; Pisoni, DB. Res. Speech Percept. Progr. Rep. No. 10. Bloomington: Speech Research Laboratory, Indiana University; 1984. Subjective evaluation of synthetic speech: Measuring preference, naturalness, and acceptability.

Nye, PW.; Gaitenby, JH. Stat. Rep. Speech Res. Vol. SR-33. New Haven, CT: Haskins Laboratories; 1973. Consonant intelligibility in synthetic speech and in a natural speech control (Modified Rhyme Test results); p. 77-91.

Nye, PW.; Gaitenby, JH. Stat. Rep. Speech Res. Vol. SR-37/38. New Haven, CT: Haskins Laboratories; 1974. The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences; p. 169-190.

Olive JP. Rule synthesis of speech from diadic units. Proc. Int. Conf. Acoust. Speech Signal Process. 1977; ICASSP-77:568–570.

Olive JP, Liberman MY. Text-to-speech—An overview. J. Acoust. Soc. Am. Suppl. 1. 1985; 78:S6.

Ozawa K, Logan JS. Perceptual evaluation of two speech coding methods by native and non-native speakers of English. J. Comput. Speech Lang. 1989; 3:53–59.

Pisoni DB. Perception of speech: The human listener as a cognitive interface. Speech Tech. 1982; 1:10–23.

Pisoni, DB. Some measures of intelligibility and comprehension. In: Allen, J.; Hunnicutt, S.; Klatt, DH., editors. From Text to Speech: The MITalk System. Cambridge, United Kingdom: Cambridge U. P.; 1987. p. 151-171.

Pisoni DB, Hunnicutt S. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. Proc. Int. Conf. Acoust. Speech Signal Process. 1980; ICASSP-80:572–575.

Pisoni, DB.; Koen, E. Res. Speech Percept. Progr. Rep. No. 7. Bloomington: Speech Research Laboratory, Indiana University; 1981. Some comparisons of intelligibility of synthetic and natural speech at different speech-to-noise ratios.

Pisoni DB, Manous LM, Dedina MJ. Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility. Comp. Speech Lang. 1987; 2:303–320.

Pisoni DB, Nusbaum H, Greene BG. Perception of synthetic speech generated by rule. Proc. IEEE. 1985; 73:1665–1676.

Pratt RL. Quantifying the performance of text-to-speech synthesizers. Speech Tech. 1987; 5:54–63.

Stevens KN. Simplified nonsense-syllable tests for analytic evaluation of speech transmission systems. J. Acoust. Soc. Am. 1962; 34:729.

Voiers WD. Evaluating processed speech using the Diagnostic Rhyme Test. Speech Tech. 1983; 1:30–39.

Voiers, WD.; Cohen, MF.; Mickunas, J. Final Rep., Contract No. AF. Vol. 19. OAS; 1965. Evaluation of speech processing devices, I. Intelligibility, quality, speaker recognizability; p. 4195

Votrax. Use's Manual. Votrax Type 'n' Talk, Votrax, Div. of Federal Screw Works; 1981.

Yuchtman M, Nusbaum HC, Pisoni DB. Consonant confusions and perceptual spaces for natural and synthetic speech. J. Acoust. Soc. Am. Suppl. 1. 1985; 78:S83.

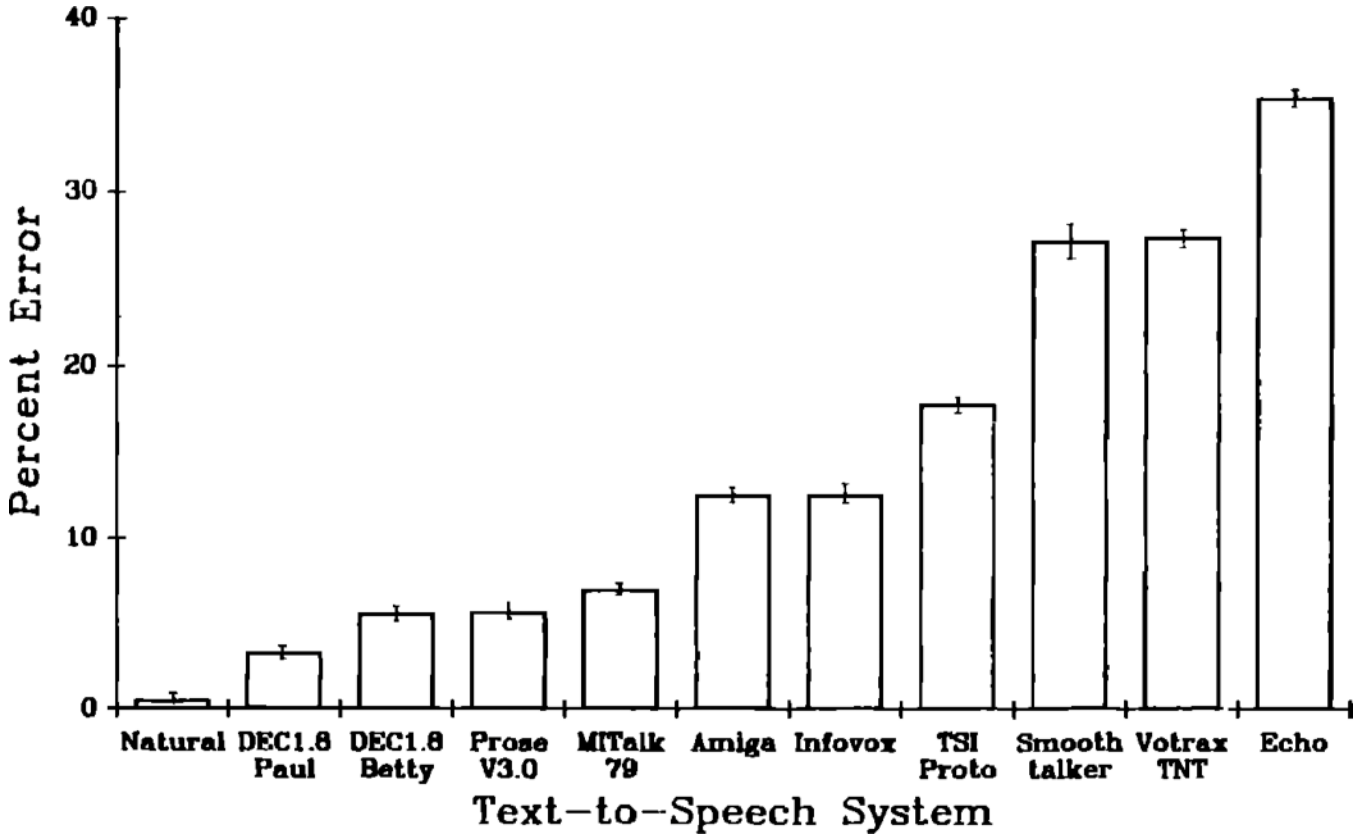

FIG. 1.
Overall error rates (in percent) for each of the ten text-to-speech systems tested using the MRT. Natural speech is included here as a control condition and benchmark.

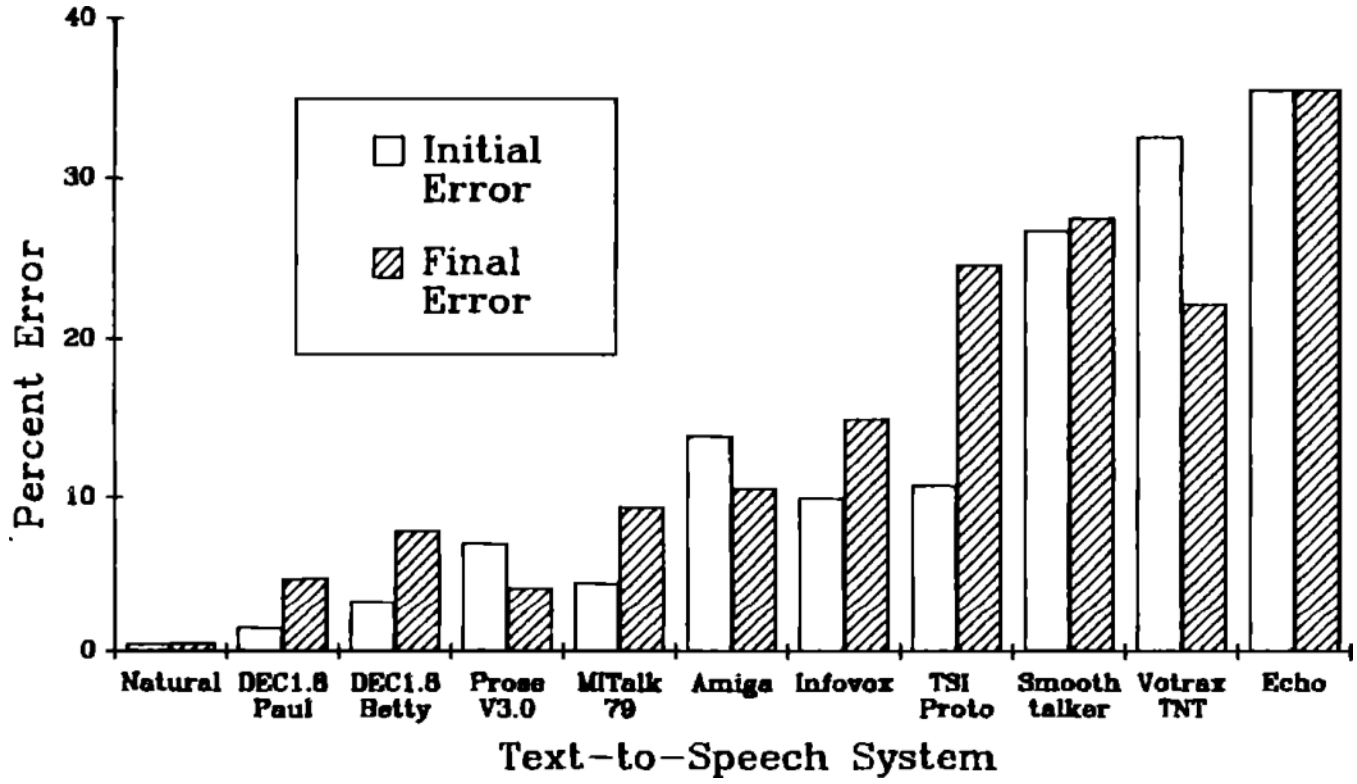

FIG. 2.
Error rates (in percent) for consonants in initial and final position for each of the systems tested in the MRT. Open bars designate initial consonant error rates and striped bars designate final consonant error rates.

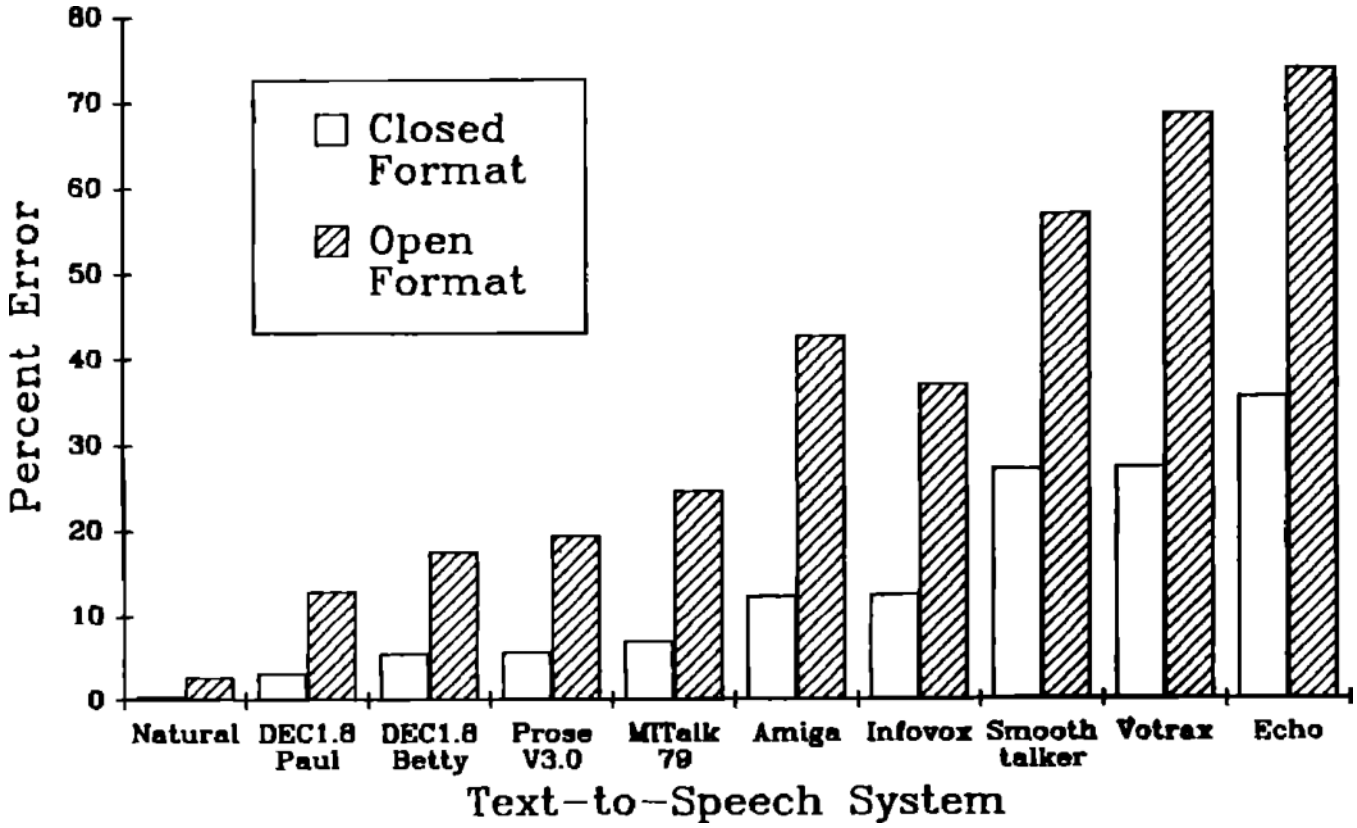

FIG. 3.

Error rates (in percent) for all systems tested 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.

TABLE I

Text-to-speech systems tested in SRL with modified rhyme test.

System	Type of synthesizer	No. of parameters[a]	D/A and rate
MITalk-79 Natural Language Processing Group, MIT 1979—4/79[b]	formant	20	12 bit 5-ms frame
TSI Prototype-1 Telesensory Systems, Inc. 1979—11/79	formant	18	10 bit 10-ms frame
DECtalk 1.8[c] Digital Equipment Corp. 1984—4/84, 11/84, 2/85	formant	18	12 bit 6.4-ms frame
Infovox SA 101 Infovox, div. of Swedish National Development Corp. 1985—3/85	formant	NA	NA 10 ms-frame
Prose 3.0 Speech Plus, Inc. 1985—4/85	formant	18	12 bit 10-ms frame
Votrax Type'n'Talk Votrax, div. of Federal Screw Works, Inc. 1981—7/85	formant	NA	NA
Echo Street Electronics, Inc. 1982—7/85	LPC	12	8 bit 21-ms frame
Amiga 500 Commodore Business Machines, Inc. 1986—10/87	formant	NA	8 bit 8-ms frame
Smoothtalker First Byte, Inc. 1984—11/87	allophonic segment concatenation	NA	8 bit[d] NA

[a]The number of parameters indicates the minimum number of variables needed to describe the speech signal. Some systems, such as DECtalk 1.8, had additional parameters that were user changeable and could be used to change the default voices.

[b]The first date is when the system was released; the second date is when the system was tested in our laboratory.

[c]DECtalk 1.8 was tested using two voices: Perfect Paul, the default voice; and Beautiful Betty, a female voice.

[d]A Macintosh Plus was used as the host computer for Smoothtalker.

TABLE II

MRT overall error rates and error rates for consonants in initial and final position.

	Error rate (in percent)		
System	**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 3.0	7.11	4.33	5.72
Amiga	13.89	10.61	12.25
Infovox SA 101	10.00	15.00	12.50
TSI-Proto 1	10.78	24.72	17.75
Smoothtalker	26.83	27.61	27.22
Votrax Type'n'Talk	32.56	22.33	27.44
Echo	35.56	35.56	35.56

TABLE III

Differences between systems obtained from *post hoc* comparisons on overall error rates.

	System									
	DECPaul	**DECBetty**	**Prose**	**MITalk**	**Amiga**	**Infovox**	**TSI**	**SmoothT**	**Votrax**	**Echo**
Natural	*a*	*a*	*a*	*a*	*a*	*a*	*a*	*a*	*a*	*a*
DEC Paul		*b*	*b*	*a*	*a*	*a*	*a*	*a*	*a*	*a*
DEC Betty			*c*	*c*	*a*	*a*	*a*	*a*	*a*	*a*
Prose				*c*	*a*	*a*	*a*	*a*	*a*	*a*
MITalk					*a*	*a*	*a*	*a*	*a*	*a*
Amiga						*c*	*a*	*a*	*a*	*a*
Infovox							*a*	*a*	*a*	*a*
TSI								*a*	*a*	*a*
SmoothT									*c*	*a*
Votrax										*a*

[a] $p < 0.01$

[b] $p < 0.05$

[c] Not significantly different.

TABLE IV

Differences between systems for initial and final position error obtained from *post hoc* comparisons. I and F denote initial position, respectively.

	Systems									
	DECPaul	**DECBetty**	**Prose**	**MITalk**	**Amiga**	**Infovox**	**TSI**	**SmoothT**	**Votrax**	**Echo**
Natural	$_{I}a_{F}b$	$_{I}b_{F}c$	*c*	*c*	*c*	*c*	*c*	*c*	*c*	*c*
DECPaul		$_{I}a_{F}b$	$_{I}c_{F}a$	*c*	*c*	*c*	*c*	*c*	*c*	*c*
DECBetty			*c*	*a*	*c*	*c*	*c*	*c*	*c*	*c*
Prose				$_{I}b_{F}c$	*c*	*c*	*c*	*c*	*c*	*c*
MITalk					*c*	*c*	*c*	*c*	*c*	*c*
Amiga						*c*	$_{I}a_{F}c$	*c*	*c*	*c*
Infovox							$_{I}a_{F}b$	$_{I}a_{F}b$	*c*	*c*
TSI								$_{I}c_{F}b$	*c*	*c*
SmoothT									*c*	*c*
Votrax										$_{I}a_{F}c$

[a] $p < 0.01$

[b] $p < 0.05$

[c] Not significant difference.

TABLE V

Phonemes accounting for the highest percentage of error in initial position.

System	Phonemes (Percentage of error accounted for in each evaluation)				
Natural speech	k (33.33)	g (11.11)	t (11.11)	p (11.11)	f (11.11)
DECtalk Paul	b (20.69)	k (17.24)	h (13.79)	p (10.34)	r (10.34)
DECtalk Betty	h (47.37)	b (24.56)	v (14.04)	p (07.02)	k (05.26)
Prose 2000 V3.0	h (22.22)	b (19.05)	k (13.49)	g (11.11)	w (10.32)
MITalk-79	h (22.89)	b (22.89)	k (19.28)	t (07.22)	θ (06.02)
Amiga	p (34.80)	h (20.40)	k (10.00)	J (07.20)	b (04.80)
Infovox	w (28.26)	f (18.48)	k (10.33)	h (09.78)	b (07.61)
TSI-Proto 1	h (18.04)	b (17.53)	n (10.82)	k (08.76)	p (07.73)
Smoothtalker	b (19.25)	f (13.87)	g (11.39)	h (09.11)	k (08.90)
Votrax Type'nTalk	h (21.00)	k (15.66)	g (07.57)	p (07.23)	f (05.34)
Echo	b (17.78)	h (13.73)	g (13.42)	p (12.48)	k (11.86)

TABLE VI

Phonemes accounting for the highest percentage of error in final position.

System	Phonemes (Percentage of error accounted for in each evaluation)				
Natural	t (18.18)	d (18.18)	θ (18.18)	m (18.18)	b (09.09)
Prose 2000 V3.0	k (38.46)	θ (26.92)	m (07.69)	z (06.41)	v (06.41)
DECtalk Paul	k (20.22)	n (19.10)	ŋ (15.73)	m (10.11)	p (08.99)
DECtalk Betty	n (16.11)	ŋ (15.43)	θ (14.09)	p (12.75)	m (08.72)
MITalk-79	n (27.54)	ŋ (19.76)	θ (11.98)	m (11.38)	f (07.19)
Amiga 500	v (16.75)	k (16.23)	t (14.14)	p (13.09)	ŋ (09.42)
Infovox	v (13.38)	f (12.64)	m (09.67)	d (08.92)	n (08.18)
Smoothtalker	d (21.37)	p (12.88)	g (11.47)	b (10.26)	k (08.65)
Votrax	p (18.32)	m (15.59)	k (13.12)	θ (09.90)	n (09.16)
TSI-Proto 1	n (34.38)	t (26.07)	m (12.13)	ŋ (06.29)	p (04.49)
Echo	k (15.20)	p (13.64)	d (12.23)	n (09.72)	b (08.62)

TABLE VII

MRT overall open error rates and error rates for consonants in initial and final position

	Error rate (in percent)		
System	**Initial**	**Final**	**Overall**
Natural Speech	0.5	0.8	2.78[a]
DECtalk 1.8, Paul	5.1	5.7	12.92
DECtalk 1.8, Betty	3.8	11.4	17.50
MITalk-79	9.97	12.5	24.56
Prose 2000 3.0	10.1	6.4	19.42
Amiga 500	23.94	21.17	42.89
Infovox SA 101	15.00	26.25	37.14
Smoothtalker	34.14	31.64	56.89
Votrax Type'n'Talk	54.5	33.2	68.47
Echo	51.06	43.39	73.97

[a]Note: Overall error was derived from exact word match errors. Scoring according to this criterion requited that listeners respond with the correct word or a homophone.

TABLE VIII

MRT open-response phonemes accounting for the highest percentage of error in initial position.

System	Phonemes (Percentage of error accounted for in each evaluation)				
Natural	f (29.41)	k (23.50)	m (11.76)	b (11.76)	n (05.88)
Prose 2000 V3.0	b (24.79)	k (22.59)	h (12.40)	g (08.82)	w (06.06)
DECtalk Paul	b (42.90)	k (16.30)	r (16.30)	h (06.00)	l (06.00)
DECtalk Betty	h (23.70)	p (22.20)	b (16.30)	t (07.40)	ð (05.20)
MITalk-79	k (19.50)	b (17.30)	h (13.10)	t (11.70)	f (06.70)
Amiga 500	p (44.19)	h (11.25)	k (09.28)	b (09.16)	t (04.87)
Infovox	f (17.60)	w (14.60)	b (11.90)	s (10.40)	h (08.90)
Smoothtalker	p (18.88)	b (17.58)	k (12.29)	f (06.99)	h (06.51)
Votrax	k (15.39)	h (10.75)	f (08.41)	t (06.32)	d (05.30)
Echo	p (22.91)	b (18.88)	k (14.04)	h (09.03)	d (06.96)

TABLE IX

MRT open-response phonemes accounting for the highest percentage of error in medial position (vowels).

System	Phonemes (Percentage of error accounted for in each evaluation)				
Natural	ɪ (29.41)	ɛ (17.65)	ˆ (17.65)	u (11.76)	i (11.76)
Prose 2000 V3.0	ɪ (25.10)	e' (22.78)	a (13.13)	æ (13.13)	ˆ (12.74)
DECtalk Paul	æ (32.30)	e' (16.20)	ˆ (12.60)	i (12.10)	ɪ (10.10)
DECtalk Betty	æ (30.00)	ɪ (23.60)	ˆ (13.60)	e' (12.30)	i (08.60)
MITalk-79	e' (49.50)	ˆ (12.90)	i (10.00)	o (09.00)	æ (06.80)
Amiga 500	ɪ (43.16)	i (18.09)	e' (10.67)	ˆ (08.35)	æ (07.42)
Infovox	æ (36.00)	e' (24.30)	a (10.00)	ɪ (09.40)	ɔ (05.10)
Smoothtalker	ɪ (21.53)	æ (21.09)	ɛ (15.61)	i (26.99)	ˆ (11.13)
Votrax	æ (28.21)	ɪ (24.59)	e' (10.49)	ˆ (10.11)	u (06.37)
Echo	ɪ (14.58)	ˆ (16.76)	ɛ (16.34)	e' (14.79)	a (09.44)

TABLE X

MRT open-response phonemes accounting for the highest percentage of error in final position.

System	Phonemes (Percentage of error accounted for in each evaluation)				
Natural	ŋ (26.67)	m (13.33)	p (13.13)	n (13.13)	ð (06.67)
Prose 2000 V3.0	k (28.02)	θ (11.64)	s (07.76)	p (06.47)	t (05.60)
DECtalk Paul	n (17.20)	k (13.70)	ŋ (11.80)	s (10.30)	m (05.90)
DECtalk Betty	p (15.50)	ŋ (12.90)	s (12.40)	n (10.40)	t (08.70)
MITalk-79	n (23.60)	ŋ (20.00)	vowel[a] (10.20)	m (08.90)	f (06.20)
Amiga 500	k (18.64)	t (14.57)	ŋ (08.53)	ld (08.01)	v (06.43)
Infovox	vowel (09.30)	ŋ (06.80)	m (05.70)	n (05.60)	v (05.20)
Smoothtalker	d (13.43)	g (10.45)	t (09.57)	p (09.13)	n (08.25)
Votrax	p (14.05)	k (13.04)	n (10.03)	ŋ (09.03)	m (08.11)
Echo	k (15.81)	n (12.99)	p (12.09)	d (10.44)	g (08.89)

[a]The MRT contains six words (way, may, say, gay, day, and pay) with the vowel /e'/ in final position. Errors for the vowels in these words are also included in the calculation of the error rates for the vowel /e'/ in Table IX.