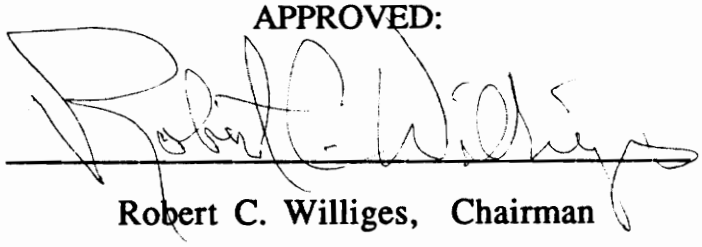
The Use of the Auditory Lexical Decision Task as a Method for Assessing the Relative Quality of Synthetic Speech

by

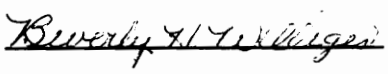
Reni L. Jenkins

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
Master of Science
in
Industrial and Systems Engineering

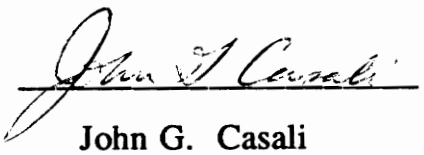
APPROVED:



Robert C. Williges, Chairman



Beverly H. Williges



John G. Casali

The Use of the Auditory Lexical Decision Task as a Method for Assessing the Relative Quality of Synthetic Speech

by
Reni L. Jenkins
Committee Chairman: Robert C. Williges
Industrial and Systems Engineering
(ABSTRACT)

This study evaluates a method for determining the quality of synthetic speech systems. The method involves the use of an auditory lexical decision task to assess the quality of synthetic speech generators relative to each other and to natural speech by using reaction time differences and error rates. Seven voices were evaluated; four synthesizers provided six voices (DECtalk 1.8 Perfect Paul, DECtalk 1.8 Beautiful Betty, DECtalk 2.0 Perfect Paul, DECtalk 2.0 Beautiful Betty, Votrax Personal Speech, Votrax Type'n'Talk) and natural speech provided the seventh voice. Both reaction times and error rates were higher for the low quality synthetic speech systems. The results document that the DECtalk can currently be considered a high quality synthesizer and that the Personal Speech and the Type'n'Talk can be considered low quality synthesizers. The results obtained by using this method can be explained by use of the Activation-Verification model (Paap, McDonald, Schvaneveldt, and Noel, 1986). Within the framework of this model, the results of this study suggest that the verification phase is the bottle-neck in processing words produced by synthetic speech generators. This interpretation suggests that by emphasizing the differences between different phonemes, to make them more uniquely identifiable, rather than concentrating on making them more "natural" might lead to improved results with synthesized speech.

ACKNOWLEDGEMENTS

I would like to extend my gratitude to my committee members, R. C. Williges, B. H. Williges, and J. G. Casali, for providing the helpful comments, equipment support and technical expertise that was vital in completion of this research. I would also like to thank the Ford Foundation for giving me the opportunity to pursue my graduate study. A special thanks goes to Dan Mauney and Gary Robinson who took out time from their studies to help perform sound measurements.

In addition, I would like to thank my parents for their constant love and encouragement. Finally, I would like to thank my husband, Joe Jenkins, for his love, emotional support, and ability to find the humor in all situations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Background

The capability to communicate through language is one of the most complex and well performed human information processing capabilities. Since communication through spoken language is considered to be more "natural" than communication through written language, many people consider verbal communication with computers to be a "more natural" and possibly ideal way for humans to interact with computers. Future natural language interfaces have been conceived of as including both speech recognition by computers in place of keyboard or other manual input methods, and speech generation by computers in place of the usual visual output methods.

While speech recognition technology has made much progress, it has not reached the sophistication level of speech synthesis technology (O'Shaughnessy, 1987). Speech output interfaces have been introduced in a variety of applications such as, telephone menu systems, teaching tools, and computer interaction aids for the blind.

Speech output, given the right application, has several advantages over traditional visual types of interaction with computers. Speech communication is the most natural means of communication between humans and, given non-extreme conditions, it is performed virtually error free. Additionally, the use of speech

output can be helpful in allowing the user more freedom of movement, since auditory signals can be designed so that they can be heard several feet away from their source. Finally, including use of the auditory mode may lessen the extent to which users must "tie-up" their hands and eyes while performing a given task. Thus the visual and manual loads necessary to interact with a computer system may be decreased.

The use of computer-generated synthetic speech also offers advantages to companies who choose to implement such systems. For example, companies can implement telephone information systems to direct incoming calls to the appropriate departments in a large company. These kinds of systems can shorten customers' time on hold and possibly decrease the number of main operators the company will need to employ.

Quality Assessment

Quality of synthetic speech has traditionally been measured by testing intelligibility or by obtaining subjective preference measures (Streeter, 1988). Intelligibility is usually defined in terms of a listener's ability to identify units of speech correctly. Many research studies have focused on optimizing specific parameters in order to obtain better intelligibility of text-to-speech systems (i.e. Herlong and Williges, 1988; Merva, 1987; Simpson and Marchionda-Frost, 1984; Slowiaczek and Nusbaum, 1985). Some physical characteristics that affect the intelligibility of synthetic speech are

fundamental frequency, speech rate, prosodics, intonation, voice type, phonetic accuracy, and context effects (Simpson, McCauley, Roland, Ruth, and Williges, 1985). The information gained through such studies is then used to determine how synthesized speech intelligibility can be optimized by manipulating the relevant parameters in specific situations.

Intelligibility and naturalness of synthesized speech have been shown to be somewhat independent (see, Nusbaum, Schwab, and Pisoni, 1984). Therefore, synthesized speech can be intelligible but far from natural (O'Shaughnessy, 1987). There has been an overwhelming amount of data collected evaluating the intelligibility of synthetic speech systems, but there have been relatively few studies addressing naturalness and acceptability of synthetic speech (Nusbaum *et al.*, 1984). One study addressing this issue (See Nusbaum *et al.*, 1984) used objective intelligibility measures along with subjective preference measures to determine the relative naturalness and acceptability of different text-to-speech systems.

In addition to the intelligibility and naturalness research, there have been numerous studies in the psycholinguistics domain investigating the capability of humans to generate and understand spoken language (i.e. Connine, Mullenix, Shernoff, and Yelen, 1990; Luce, 1986; Paap, Newsome, McDonald, and Schvaneveldt, 1982; Paap, McDonald, Schvaneveldt, and Noel, 1986). Although psycholinguistic models are not yet complete, they can be used in the context of assessing the quality of text-to-speech systems, because

the basic mechanisms of lexical processing are well modeled (see Paap *et al.*, 1982; and Paap *et al.*, 1986).

If the goal of improving the quality of text-to-speech systems is to make their output indistinguishable from natural speech, it follows that evaluation of quality should include comparisons between the way in which processing of synthetic speech differs from processing of natural speech. Examining synthetic speech intelligibility within the context of a psycholinguistic model, such as the Activation-Verification model (Paap *et al.*, 1986) will enable these comparisons.

Word Frequency Effects

There are a number of phenomena that occur in spoken language. One of the most studied is the word frequency effect. The role of word frequency in visual word recognition has been studied by many investigators. The results of studies using both lexical decision tasks and naming tasks indicate that frequency influences response time, with high-frequency words having shorter times (Rubenstein, Garfield, and Milliken, 1970). Recent studies of spoken word recognition have also found frequency effects with auditory presentation of stimuli (e.g., Connine *et al.*, 1990; and Luce, 1986).

Several different descriptive models have accounted for the frequency effect (e.g., Paap *et al.*, 1986; and Seidenberg, Waters, Barnes, and Tanenhaus, 1984). It thus seems that there is agreement between researchers that words that are used more frequently are,

in some sense, more accessible than words used less frequently (Carroll, 1986). However, one of the major problems facing researchers in the area of assessing synthetic speech quality, is the abundance of modality-specific theories, and the lack of general theories of word recognition. However, the idea that intlligibility and speech perception must be studied within the context of a lexical processing system, rather than just within a phonetic domain has been spreading among speech researchers (Frauenfelder and Tyler, 1987).

Lexical Processing Phases

Because there are different theories of lexical processing, it follows logically that each theory may account for word frequency effects differently. The stages of processing involved according to each model also differ slightly. However, in an overview paper, Frauenfelder and Tyler (1987) argue that there are only five phases involved in lexical processing, and these five are found in most of the theories cited in their review, though they may be labeled differently;

1. Initial lexical contact

Initial lexical contact is described as the process of generating a representation, which contacts the internally stored representations associated with lexical entries, from the sensory input.

2. Activation

When lexical entries match the contact representation to some critical degree, they change in state. This change is referred to as activation.

3. Selection

Selection occurs when accumulating input narrows the subset of lexical items activated to one lexical entry.

4. Word recognition

Word recognition refers to the outcome of the selection process when a listener has determined which lexical entry was heard.

5. Lexical access

Lexical access is the point at which the phonological, syntactic, semantic, and pragmatic properties of a word become available to the listener. Lexical access is the goal of lexical processing.

The Activation-Verification Model

The activation-verification model (AV) for word recognition (Paap *et al.*, 1986; see also Paap *et al.*, 1982) is one of the most complete and refined models for visual as well as auditory word recognition. Unlike many of the models proposed in the literature, it is possible to make solid predictions based on this model. Since terminology is one of the problems plaguing the word recognition

literature, an introduction to the terminology used by Paap *et al.* (1986) is in order.

Paap *et al.* (1986) state that there are three operations involved in their model; encoding, verification, and decision, see Figure 1. The encoding process leads to unconscious activation of learned units in memory. For spoken stimuli, this operation involves activation of phoneme units by phonemic feature detectors. Activation levels are then determined by the number of matching and mismatching phonemic features. Finally, the most highly activated lexical entries are placed in a verification list.

Verification immediately follows encoding and involves independent, top-down analysis of the items in the verification list (also called candidate words). The verification process is a serial comparison between the items in the verification list and a stored representation of the stimulus. The items are verified in descending order of frequency.

The verification process can have two outcomes, a mismatch if the degree of fit between the candidate word and a stored representation of the word does not exceed some criterion, or a match if the degree of fit does exceed the criterion. A mismatch results in verification of the next candidate from the verification list, and a match results in instantaneous recognition of the candidate word. The AV model assumes that once word recognition is achieved, lexical access is automatic and instantaneous. Therefore,

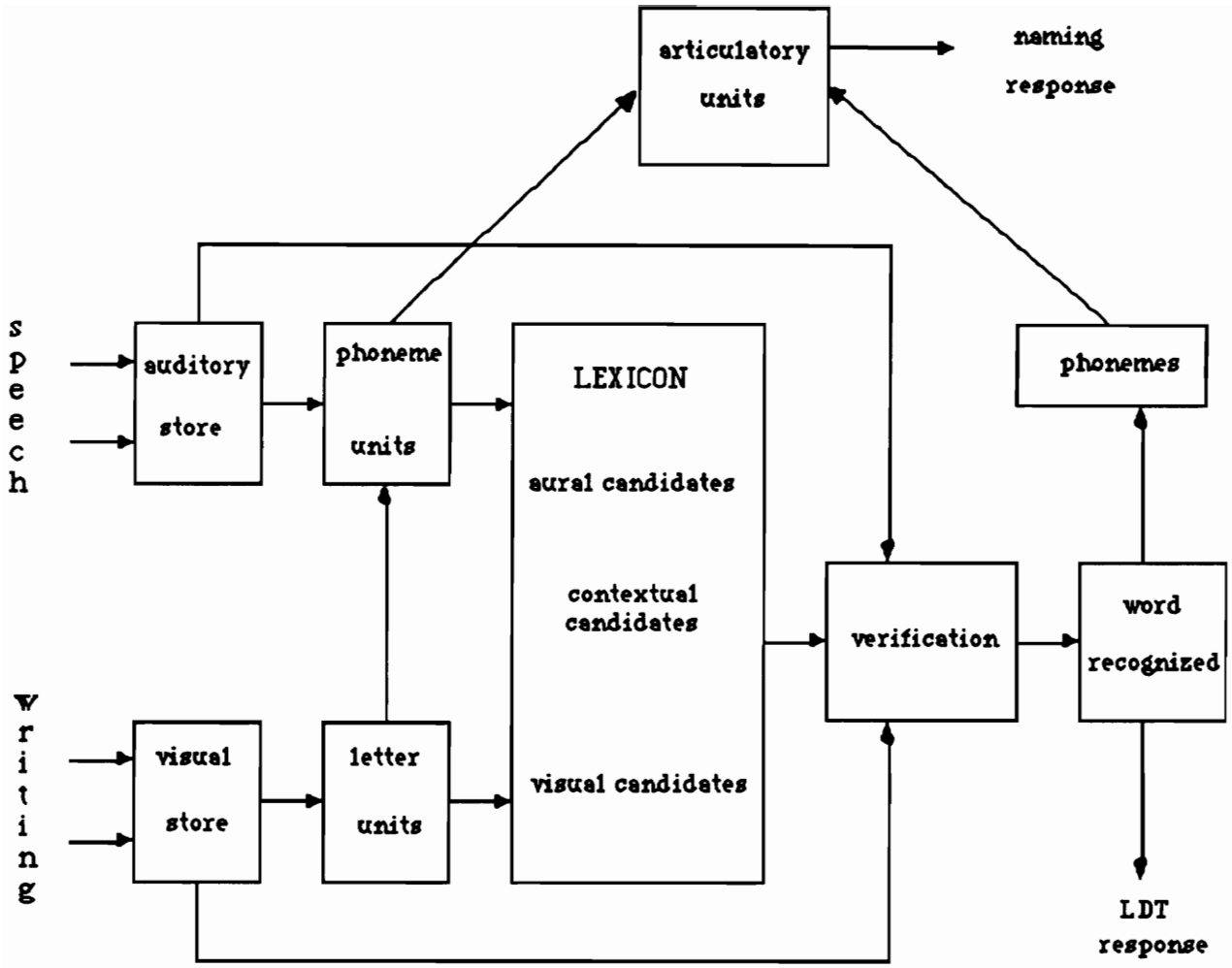


Figure 1. Paap *et al.*, (1986) model (Jenkins, 1989).

there is no distinction between word recognition and lexical access in the AV model.

The decision processes in lexical access involve many processes, including all decisions that are made about comparisons at various stages of word recognition and criterion setting. The lexical decision task requires that successful encoding, verification, and word recognition be completed before a response can be made.

In terms of Frauenfelder and Tyler (1987), the encoding operation of the AV model can be thought of as encompassing both the initial lexical contact stage and the activation stage. The verification operation of the AV model is the process by which activated items are selected, recognized, and accessed from the lexicon and thus can be thought of as including the selection, word recognition, and the lexical access stage of Frauenfelder and Tyler (1987).

Current Intelligibility Testing Methods

Although many testing procedures and methods exist for assessing speech intelligibility (e.g. the Articulation Index and Speech Interference Level), most of the synthetic speech intelligibility research involves use of one of six testing methods. These methods are the Rhyme Test, the Modified Rhyme Test, the Diagnostic Rhyme Test, sentence verification testing, Phonetically Balanced (PB) word lists, and subjective testing measures. Since these tests have

traditionally been used to assess synthetic speech intelligibility and quality, a brief description of them is in order.

The Rhyme Test

The Rhyme Test was developed by Fairbanks (1958). The test was motivated by a need for experimental materials which:

1) the spoken word would be the stimulus unit,
2) recognition of the word would be the response,
3) the response would depend on the initial consonant and consonant-vowel (CV) transition, and
4) the subject's task would bear valid relation to the discrimination demands of real speech.

This test involves presenting subjects with a three-letter, mono-syllabic word stem and a spoken word which uses that stem. The subject's task is to identify the stimulus word based on its initial consonant. For example, the word stem might be "-op" and the spoken word, "mop".

The rhyme test lists contain 18 consonant phonemes, that account for 90% of all consonant occurrences (Fairbanks, 1958). However, seven consonants are not included in the lists. The first two, /ŋ/ (ri*ng*) and /ž/ (azure), never occur in the initial consonant position and the other seven, /θ/ (*th*in), /ð/ (*th*en), /š/ (*sh*op), /č/ (*ch*op), and /hw/ (*wh*en), require a two letter spelling.

In an attempt to control for frequency effects, the stimuli contained in the lists are among the 1000 - 9000 most frequently

used words in written English. The frequency data were obtained by using the "The Teacher's Book of 30,000 Words" by Thorndike and Lorge (1944).

There are a few disadvantages associated with using this test for evaluating segmental intelligibility of synthetic speech systems. The first is that the auditory-phonemic factors are weighted more heavily than the higher order linguistic factors (Fairbanks, 1958). Since word recognition is the desired response, higher order linguistic processes must be given proper consideration. Auditory-phonemic processing begins at the onset of stimuli presentation, while word recognition is the last step involved in auditory processing of word stimuli.

Another disadvantage of this test, is that because it focuses on the initial consonant. Several phonemes are not tested and those may be the ones that define poor quality or good quality synthetic speech. For example, /θ/ and /ð/ are among the most highly confusable phonemes (Miller and Nicely, 1955), but they are not tested using this method because they require a two letter spelling.

The Modified Rhyme Test

The Rhyme Test was modified by House, Williams, Hecker, and Kryter (1965). The Modified Rhyme Test (MRT) was intended to be short and reliable, as well as easier to administer and score than the earlier speech intelligibility and articulation testing methods.

Although the procedure for administering the test is the same, the MRT differs from the original Rhyme Test in several ways. First, although the words are mono-syllabic with a CVC form as in the original Rhyme Test, the one-letter consonant representation was not employed in the MRT, therefore, phonemes that require two-letter spellings were not eliminated from the lists. Second, the MRT includes subject responses to both the initial and final consonants of the word. Third, the MRT has a close-ended response format, as opposed to Fairbank's original open-ended format. This eliminates the learning time involved in testing where subjects must become familiar with the message set, because listeners already have access to it (House *et al.,* 1965). Fourth, the lists constructed for the MRT are not balanced according to their frequency of occurrence in written English.

The MRT has several disadvantages. One disadvantage is that the MRT fails to take into consideration word frequency. Word frequency effects have been well documented and researched, and the frequency of occurrence of a word has been shown to play a significant role in word recognition. Secondly, because the listener has available to him/her the complete message set, the results may not be generalizable to real world situations where the message set is not known. Third, the MRT lists are arranged so that the first 25 items vary on the initial consonant and the last 25 vary on the final consonant. Because of this regularity in the test items, it is possible for the subjects to base their responses on factors other than

perceptual processes (Logan, Pisoni, and Greene, 1985). In order to minimize this response bias, Logan *et al.* (1985) developed a version they call the Mixed Format MRT, that randomly orders the variation of the initial and final consonants.

Diagnostic Rhyme Test

The Diagnostic Rhyme test (DRT) was first described by Voiers, Cohen, and Mickunas in 1965 (Voiers, 1977). Its development was motivated by the interest in controlling all confounding attributes contributing to word intelligibility. Of particular interest was the control of context effects; factors that affect a listener's *a priori* uncertainty of the contents of a message (Voiers, 1977).

A phonemic taxonomy of distinctive features was compiled and used to create the test materials. There are seven features included in the taxonomy: voicing, nasality, sustention, sibilation, graveness, compactness, and vowel-like (see Voiers, 1977 for definition of these terms). The corpus of test materials was created such that each item consists of a set of two rhyming words, initial phonemes of which vary only by a single distinctive feature. For example, one set of words might be VEAL-FEEL, which vary only in the feature of voicing.

The procedure for administering the test is similar to the previously mentioned tests. A word is presented to a subject and the subject indicates which word of the test item he/she heard. Thus, for each item, the intelligibility of a specific phonemic feature is being

tested. This test has the advantage of being able to pinpoint exact properties of phonemes that are responsible for poor intelligibility.

The DRT has some of the same disadvantages as the MRT. It fails to take into consideration word frequency, and the message set is known and limited.

Phonetically balanced word lists

Phonetically balanced word lists are also used to test speech intelligibility. The lists were constructed by the Psycho-Acoustic Laboratory at Harvard University (Egan, 1948) in an effort to increase the reliability and validity of natural speech intelligibility testing.

The 20 lists each contain 50 monosyllabic words. Each list contains a representative sample of the sounds found in conversational speech. Both the types of sounds and the frequency of occurrence of those sounds in "average speech" are taken into account. The phonetic composition of the initial part of a word determines its placement in a list. Lists are designed to be equal on average difficulty and have an equal range of difficulty. In addition, only "common" words are included in the lists.

The task of the listener is to indicate by transcription, the word he/she heard. The intelligibility score is then a measure of the number of correct responses. Tests can be scored by number of successful word transcriptions or number of successful phoneme transcriptions depending on the needs of the experiment.

This test has some disadvantages for testing synthetic speech intelligibility. One disadvantage of this test is that lists are constructed by taking only the initial part of the word into consideration. Another disadvantage is that training, testing, and scoring procedures are extremely time intensive and thus may make this test less desirable, especially in situations where several synthesizers are to be tested.

Sentence verification testing

Sentence testing is another method used to evaluate the intelligibility of synthetic speech. The procedure involves presentation of a spoken sentence and transcription of that sentence by the subject. The two most commonly used sets of sentences for these tests are the Harvard Psycho-Acoustic Sentences and the Haskins Laboratory Syntactically Normal Sentences (Merva, 1987).

The Harvard sentences are meaningful sentences that contain a wide variety of syntactic structures (Pisoni, 1979). They are balanced on difficulty, and segmental phonemes are represented in accordance with their frequency of occurrence in English. Two factors that help determine the difficulty of a word are length of the word and phonetic structure of the word. The Harvard sentences are useful for obtaining an estimate of how well word recognition is expected to proceed in sentences where both syntactic and semantic information is present (Pisoni, 1979).

The Haskins sentences are syntactically normal anomalous sentences. In this test, listeners have only acoustic-phonetic information available to them. Therefore, this test could provide an upper bound estimate of the contribution of phonetic information to word recognition in sentences (Pisoni, 1979).

Like the Phonetically Balanced word lists method, there are two ways of obtaining an intelligibility rating using sentence testing. The first involves counting the number of incorrect whole word transcriptions. The second is a more detailed analysis and involves recording and summarizing the number of specific phonetic errors.

Subjective tests

Another method of estimating the quality of synthesized speech is through listeners' ratings of naturalness and difficulty. Naturalness of synthetic speech refers to a listener's judgement, on some scale, that indicates the degree to which it appears to sound as if it was spoken by a human (Herlong and Williges, 1988). Data on naturalness may accompany tests of intelligibility. Information is obtained by asking subjects to complete a questionnaire addressing the naturalness of the synthetic speech system either during or after intelligibility testing.

Preference testing methods can be separated into two groups, relative methods and absolute methods. Relative methods require the listener to express his/her preference for one signal of a pair, the test signal and the reference signal (Rothauser, Urbanek, and Pachl,

1971). The absolute methods require the listener to make preference judgements about some signal in terms of some absolute value system (Rothauser *et al.*, 1971).

Preference measures, however, do not allow inferences about the synthesizers, only descriptions. In addition, when a subject responds that he/she prefers one synthesizer to another, the reasons for this preference may not be known. This may be due to the researcher's non-interest in the reasons, but more likely it is because the reasons are inaccessible to the subject. Therefore, preference information is not particularly useful in and of itself in determining how to improve the synthesis system.

The Need For New Methods

The methods described above have been invaluable in past intelligibility research, but what is really needed is to incorporate a cognitive model into the intelligibility testing. Intelligibility is not independent of speech processing, and it should be examined within the framework of a speech processing model.

When the processing of synthetic speech is sufficiently similar to the processing of natural speech, natural language phenomena occurring for spoken language, such as the word frequency effect, should also occur for synthetic speech. This thesis proposes that it is possible to use traditional psycholinguistic techniques and theories to determine whether phenomena such as word frequency effects occur to the same degree for synthetic speech as they do for natural

speech. The measurement of speech quality would be therefore, based upon the extent to which the processing of synthetic speech is the same as the processing of natural speech.

This method involves determination of the extent to which naturally occurring frequency effects found in human speech were found to exist in synthetic speech. It is more quantifiable than some of the previous methods proposed for determining synthetic speech quality for two reasons. First, two numerical values are obtained through testing, a reaction time score for direct comparison to the natural language condition as well as to other synthesizers, and an error rate score, also for comparison to natural language and other synthesizers. Secondly, the criteria upon which the selection of stimuli are based, frequency of occurrence in written English, is a numerical quantity.

Using the frequency effect information, it is possible to evaluate the quality of any particular synthetic speech generator relative to natural speech as well as to other synthetic speech generators. Quality can then be thought of as a measure of the difference between the processing of natural speech and the processing of synthetic speech. Theoretically, the more similar the processing between synthesized speech and natural speech, the less the cognitive burden on the listener.

One advantage of the proposed method is that it indicates important differences between the processing of natural and synthetic speech. In addition it provides, in one testing session,

information about the relative differences among samples of speech produced by different text-to-speech systems. Another advantage is that the auditory lexical decision task results in two quantitative measures, error rate and reaction time, upon which to base the results of the test. Finally, it includes only word recognition processes thus problems in quality can be attributed to a specific point or operation in processing.

Purpose

This thesis proposes an alternate method of assessing the quality of synthetic speech. The study attempted to replicate the relative intelligibility order of four of the ten voices used in a study by Greene, Logan, and Pisoni (1986). This was accomplished by determining the relative quality order through examination of the differences in reaction time, error rate, and word frequency effects between each synthesizer and the natural control condition.

In their study of segmental intelligibility, Greene *et al.* (1986), used a Mixed Format MRT to determined the relative intelligibility order of different default voices of eight text-to-speech systems. The text-to-speech converters used in the study included, the MITalk-79, the Digital Equipment Corporation DECtalk version 1.8, the Telesensory Systems, Inc. TSI Prototype-1 of the Prose 2000, the Votrax Type'n'Talk, the Street Electronics Echo, the Speech Plus Prose 2000 V3.0, the Berkely Systems Works, and the Infovox SA 101. The results of the study indicate that the lowest overall error rates of

the eight systems evaluated in the study were the DECtalk 1.8 Perfect Paul, the DECtalk 1.8 Beautiful Betty, and the Prose 2000 V3.0 default setting.

The goal of this thesis study was to use an auditory lexical decision task to determine the differences in processing time involved between the text-to-speech systems and natural speech. In particular, this study focused on the extent to which frequency effects are observed when listeners process synthetic speech from four text-to-speech systems and natural speech.

The present study included four text-to-speech converters, the Digital Equipment Corporation DECtalk 1.8 and DECtalk 2.0, the Votrax Type'n'Talk, and the Votrax Personal Speech System. Seven voices were tested: DECtalk 1.8 Perfect Paul, DECtalk 2.0 Perfect Paul, DECtalk 1.8 Beautiful Betty, DECtalk 2.0 Beautiful Betty, the Votrax Type'n'Talk, the Votrax Personal Speech System, and natural speech.

It was hypothesized that if frequency effects (reaction time differences between high frequency words and low frequency words) were observed for the synthetic speech, they would be considerably less than those observed in natural speech. That is, frequency effects observed for stimuli produced by a speech synthesizer, would differ from the frequency effects observed for stimuli produced by a human speaker. Those differences were used as a relative quality metric for each particular synthetic speech system. Error rate data were also used to determine the relative intelligibility of the six voices and the natural voice condition.

METHOD

This study required subjects to make lexical decisions about stimuli produced by speech synthesizers. The auditory stimuli were presented through headphones, the subject's task was to indicate, by pressing a computer key, whether the stimulus was a word or a non-word. Reaction times and error rates were recorded by computer.

Experimental Design

The experimental design was a 3x7 within-subjects factorial design.

Independent variables

The two independent variables used in this study were Word Frequency and Voice Type. Word frequency had three levels corresponding to high frequency, medium frequency, and low frequency, as determined by Francis and Kucera (1982). Voice type had seven levels, corresponding to DECtalk 2.0 Perfect Paul (DEC P2.0), DECtalk 2.0 Beautiful Betty (DEC B2.0), DECtalk 1.8 Perfect Paul (DEC P1.8), DECtalk 1.8 Beautiful Betty (DEC B1.8), the Votrax Personal Speech System (Pers. Speech), the Votrax Type N' Talk (TNT), and natural speech (Nat.). The three levels of word frequency were fully crossed with the seven levels of voice type, see Figure 2.

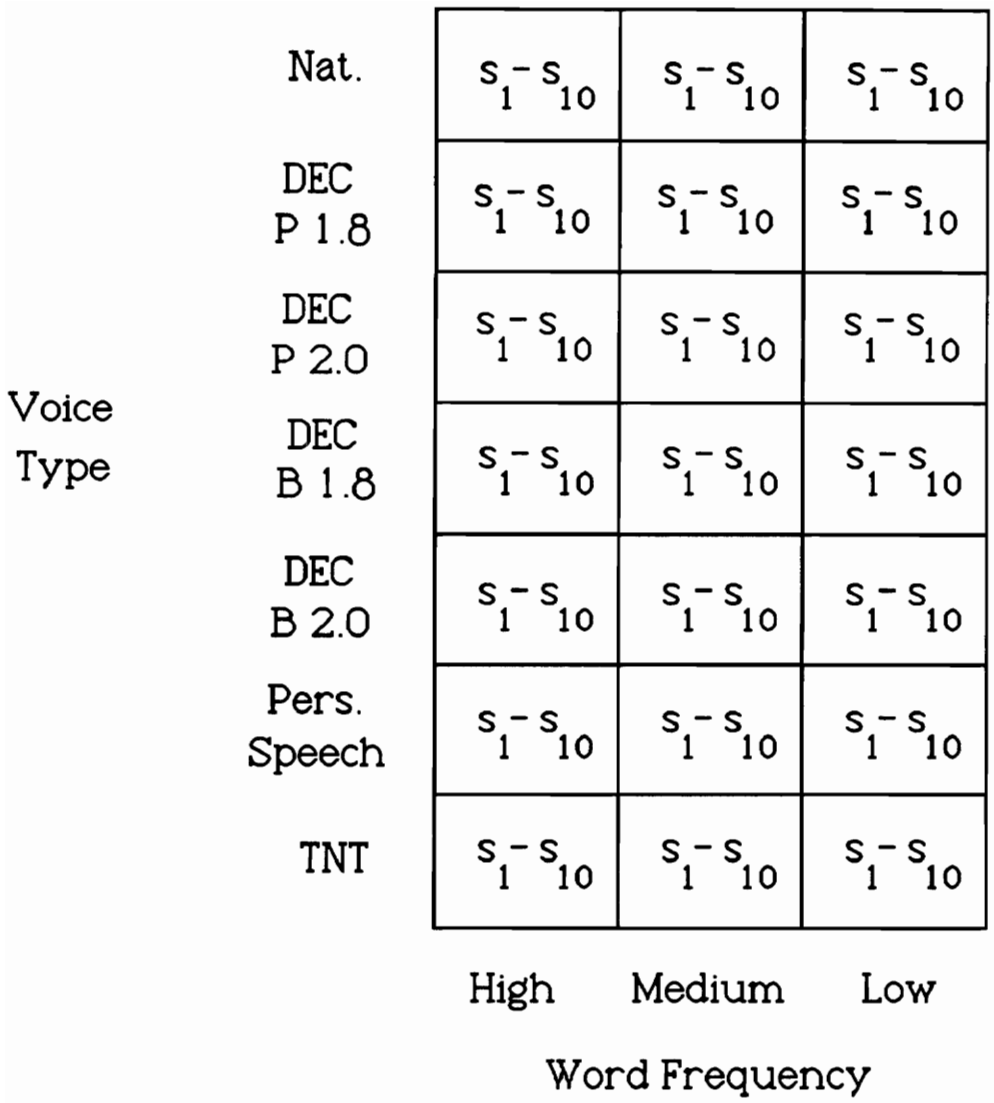


Figure 2. Experimental Design.

Dependent Variables

For each trial, the subject's reaction time and number of errors were recorded by software written for the experiment. Reaction time was calculated as the time between the end of the stimulus presentation and the subject's depression of a response key.

Stimuli

The word stimuli consisted of 60 four to six letter words based on their frequency of occurrence in written English (20 high frequency, 20 medium frequency and 20 low frequency), as determined by Francis and Kucera (1982). The high frequency word stimuli had a mean frequency of 804.85 occurrences per million words and a range of 451-1661 occurrences per million words, medium frequency word stimuli had a mean frequency of 137.6 occurrences per million words and a range of 100-212 occurrences per million words, and low frequency word stimuli had a mean frequency of 10.45 occurrences per million words and a range of 2-15 occurrences per million words. A visual lexical decision task was performed as a pre-study to ensure that frequency effects were obtained from the stimuli.

In addition, 60 four to six letter non-word strings were created to serve as noise in the experiment. The non-word stimuli were constructed by substituting one phoneme for another in a four to six letter word to form a pronounceable non-word letter string. For

example, the non-word "souncil" was formed by substituting the phoneme /s/ for the phoneme /k/ in "council".

Considerable thought was given to using frequency of occurrence in spoken English as opposed to written English; however, according to Connine *et al.* (1990) there is only one source for this information and it has such a limited scope, only 416 words, that it was not a reasonable alternative to the Francis and Kucera (1982) count. Previous studies (Connine *et al.*, 1990; Luce, 1986) have also set precedence a for using written frequency counts and obtaining frequency effects with auditory lexical decision tasks.

Each word and non-word in this study was generated by each voice type, resulting in a total of 840 stimuli, see Figure 3. The stimuli were digitally recorded and stored on disk for later experimental presentation. A list of the stimuli is in Appendix A.

An auditory confusion matrix reported by Miller and Nicely (1955) was used to ensure that the most confused phonemes would be adequately represented so as not to bias any particular synthesizer. In particular, the confusion matrix corresponding to a speech-peak to noise-peak (i.e., signal-to-noise) ratio of +12 dB and bandwidth of 200-6500 Hz, was used to determine phoneme pairs that were highly confusable. See Table 1 for the confused phonemes and their respective confusion percentages. A best attempt was made to include each phoneme pair, at each level of frequency.

Voice Type	High	Medium	Low	Non-words
Nat.	20	20	20	60
DEC P 1.8	20	20	20	60
DEC P 2.0	20	20	20	60
DEC B 1.8	20	20	20	60
DEC B 2.0	20	20	20	60
Pers. Speech	20	20	20	60
TNT	20	20	20	60

Stimulus Category

Figure 3. Experimental Stimuli.

TABLE 1

Phoneme presented and incorrect response† and the percent response error rate for the confusion (Adapted from Miller and Nicely, 1955).

Confusion Pair	Error Rate
p / k	14.4%
k / p	7.5%
f/θ	9.5%
θ/f	25.8%
b / v	8.6%
d / g	9.3%
g / d	13.8%
v/ð	14.4%
ð / v	9.2%

† first phoneme is presentation phoneme, second phoneme is response made by the subject.

Experimental Procedure

This study involved the use of an auditory lexical decision task to assess the presence of word frequency effects. The 840 digitally recorded stimuli were randomized and presented binaurally through a set of headphones.

The subject's task was to indicate whether the item presented was a word or a non-word by depressing the appropriate key on the keyboard. The keyboard was modified by placing a label over the "?" key reading "WORD" and another label over the "Z" key reading "NONWORD".

A trial was initiated by a 500 ms warning light on the computer display. This warning signal preceded the stimulus presentation by 500 ms. The stimulus was then played by a Macintosh II computer and presented through the headphones. Computer software initiated timing of the reaction time. The clock was started at end of the auditory presentation and was stopped upon depression of the "WORD" key or the "NONWORD" key by the subject. The next trial was begun after a two second delay. The stimuli were presented continuously for five minutes at which time the subject was given a one minute break.

The computer software calculated the reaction time as the time between the end of presentation of the stimulus and the depression of a response key. Although reaction times for the non-word stimuli were collected, they were not included in any of the analyses. In addition to the calculation and recording of reaction times, error

rates were also recorded and all information was written to a text file for later analysis.

Equipment

A HyperCard 2.0 stack running on an Apple Macintosh II computer controlled the experiment. The software was written using HyperTalk with XCMD's written in Lightspeed Pascal. A Digital Equipment Corporation DECtalk version 1.8 speech synthesizer provided voices "Perfect Paul 1.8" and "Beautiful Betty 1.8"; a Digital Equipment Corporation DECtalk version 2.0 speech synthesizer provided voices "Perfect Paul 2.0" and "Beautiful Betty 2.0"; a Votrax Personal Speech System provided the fifth voice; a Votrax Type'n'Talk speech synthesizer provided the sixth voice; and a male speaker provided the seventh voice. A male voice was chosen in order to be more consistent with past literature (e.g. Connine *et al.*, 1990; Greene *et al.*, 1986), in particular, the Greene *et al.* (1986) study, which used a male voice as the control.

The stimuli were entered into each of the synthesis systems and the resulting speech generation was digitally recorded at 22 KHz using a Macintosh II with Farallon's MacRecorder software and a Realistic cardioid microphone. The microphone was placed directly in front of the system's speaker, or so that the angle of incidence of sound on the microphone was not more than 45° off of 0°. The human voice was digitally recorded at the same sample rate with the same apparatus and procedures.

A set of Realistic NOVA 40 circumaural headphones were used for presentation of the stimuli to the subjects. The headphones had a ±3 dB frequency response for the range of 250-10 000 Hz.

Ambient Environment

The ambient noise level of the room was measured to be less than 60 dB(A) by a Realistic sound level meter.

Subjects

Ten native English speaking students (i.e., five male and five female) at Virginia Polytechnic Institute and State University served as subjects for this study. Participants had no previous experience with synthesized speech and were paid ten dollars for their time.

Hearing Tests

Each subject received a hearing test prior to selection for the experiment. The modified Houghson-Westlake procedure (Morrill, 1986) was used to conduct the hearing tests on a Beltone 100 Series Audiometer. In order to qualify for the experiment, the subject had to have normal hearing in both ears. Normal hearing was defined as hearing threshold levels (HTLs) in either ear no higher than 26 dB for pure tones at 500 Hz, 1000 Hz, and 2000 Hz (Davis and Kranz, 1964). Each subject was tested at the reference frequencies as well as at 250 Hz, 1500 Hz, 4000 Hz, and 6000 Hz, to adequately test the range of critical frequencies of speech signals.

Each subject was informed that the hearing test was not designed to assess or diagnose any physiological or anatomical hearing disorders and that it would only be used to determine their qualification to participate in the study. All subjects passed the hearing test.

Instructions and Informed Consent

Subjects were instructed to indicate whether each item was a word or a non-word by pressing the appropriate key on the keyboard. Additionally, they were instructed to make their responses as quickly and accurately as possible (see Appendix B for Instructions). Subjects were also presented with an informed consent document to read and sign. This document clearly informed them of their rights as a participant in experimental studies and briefly described the hearing test to be used for screening (see Appendix C for Informed Consent Document).

Training

Subjects were presented with 70 practice trials. The stimuli were 70 preselected and digitized words (using all seven voice types) of four to six letters long. These practice trials were followed by the 840 experimental trials.

RESULTS

The raw data were first trimmed to reduce noise. Specifically, for each subject, all reaction times greater than 2.5 standard deviations from the mean for that particular subject were dropped from the analysis. This data trimming technique, employed frequently in the previous literature, (see Jenkins, 1989; Schvaneveldt and McDonald, 1981) was used to remove the effect of words unknown by the subject. Since the auditory lexical decision task involves word recognition, the stimulus word must be present in the subject's lexicon for him/her to access it. Therefore, unknown words are considered noise. The cutoff value, 2.5 standard deviations, is set at approximately the 95th percentile and has been used as the value for trimming lexical decision task data by Paap and associates for about fifteen years. Table 2 shows the ratio of eliminated data points to the total number of data points in each voice type/word frequency category for each analysis in this study.

Subsequent to data trimming, mean reaction times were computed for each of the 21 conditions for each subject. Mean reaction times were used instead of individual reaction times in the analysis due to the amount of between subject variability, namely the differences in exposure to written literature. Four analyses of variance (ANOVA) were performed, an analysis of mean reaction

TABLE 2

Number of data points dropped from each analysis per total data points for each voice type/word frequency category.

Voice Type	Word Freq.	Elimination Ratio for Mean Reaction Time Analysis	Elimination Ratio for Error Rate Analysis
DEC P 1.8	H	7/156	0/ 44
DEC P 1.8	M	5/151	1/ 49
DEC P 1.8	L	2/145	2/ 55
DEC P 2.0	H	1/138	2/ 62
DEC P 2.0	M	1/146	2/ 54
DEC P 2.0	L	2/128	2/ 72
DEC B 1.8	H	2/140	4/ 60
DEC B 1.8	M	2/134	0/ 66
DEC B 1.8	L	5/131	2/ 69
DEC B 2.0	H	4/137	0/ 63
DEC B 2.0	M	4/133	0/ 67
DEC B 2.0	L	0/130	0/ 70
Pers. Speech	H	4/ 86	6/114
Pers. Speech	M	2/ 79	4/121
Pers. Speech	L	3/ 84	4/116
TNT	H	2/ 90	3/110
TNT	M	1/ 77	4/123
TNT	L	2/ 75	4/125
Nat.	H	1/199	0/ 1
Nat.	M	2/198	0/ 2
Nat.	L	3/180	0/ 20

time, an analysis of errors, an analysis of mean reaction time excluding the medium frequency condition, and an analysis of errors excluding the medium frequency condition. Since it is difficult to select an appropriate range to label medium frequency, the medium frequency condition was excluded after the initial analysis to determine whether differences between the high frequency and low frequency conditions could be highlighted.

Since many comparisons were to be made, the post-hoc analyses were performed using the Newman-Keuls paired comparison test. The main advantage associated with the Newman-Keuls is that it corrects for inflated α error across a set of comparisons.

Because of the possibility of violation of the assumption of homogeneity of covariance when using a within-subjects design, the Greenhouse-Geisser correction was used to correct p-values for the ANOVAs. All summary tables report Greenhouse-Geisser corrected p-values and epsilon correction factors. The degrees of freedom (df) for the numerator and denominator associated with the F value are multiplied by the epsilon correction factor to obtain the "effective df". These new values are used in consulting the F table. See Greenhouse and Geisser (1959) for details of this correction.

For all statistical tests, the level of significance was set at $p < 0.05$. Analyses were performed using CLR Anova and SuperAnova for the Macintosh.

Reaction Time

The summary table for the mean reaction time analysis including the medium frequency condition, is reported in Table 3. The main effects of word frequency and voice type were significant. The means for each word frequency category and each voice type are shown in Figure 4 and Figure 5, respectively. All reaction times are in seconds. The word frequency by voice type interaction was not significant.

Post hoc Newman-Keuls analysis showed that there was a significant difference between mean reaction times for high frequency words and low frequency words and between medium frequency words and low frequency words. The difference between mean reaction time for high frequency words and medium frequency words was not significant.

Another Newman-Keuls analysis showed that the natural voice condition was significantly better than all of the other conditions in terms of reaction time. In addition, the DECtalk voices (Perfect Paul 1.8, Perfect Paul 2.0, Beautiful Betty 1.8, and Beautiful Betty 2.0) were not found to differ significantly from each other in terms of reaction time. These voices did, however, differ significantly from the Personal Speech and the Type'n'Talk voices. Similarly, the Personal Speech and Type'n'Talk voices were not significantly different from each other in terms of reaction time.

The summary table for the mean reaction time analysis excluding the medium frequency condition, is reported in Table 4.

TABLE 3

ANOVA Summary Table for Mean Reaction Time (Including Medium Frequency)

Source	*df*	*SS*	*F*	*p*	*Epsilon*
Between Subjects					
Subjects (S)	9	8.333			
Within Subjects					
Word Frequency(WF)	2	0.161	14.815	0.0005	0.83
WF x S	18	0.098			
Voice Type (VT)	6	1.522	40.633	0.0001	0.67
VT x S	54	0.337			
WF x VT	12	0.170	1.869	0.1341	0.37
WF x VT x S	108	0.817			
Total	209	11.438			

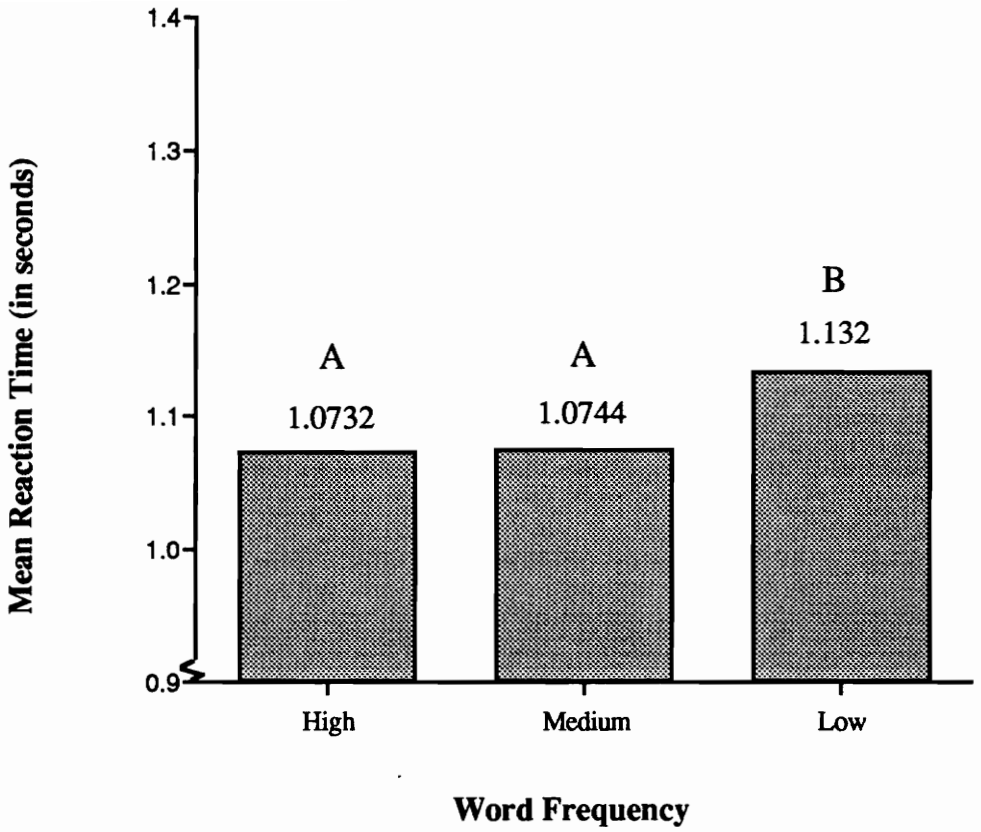


Figure 4. Reaction time means for word frequency (including medium frequency). Reaction times for word frequencies with different letters were found to be significant at $p < 0.05$.

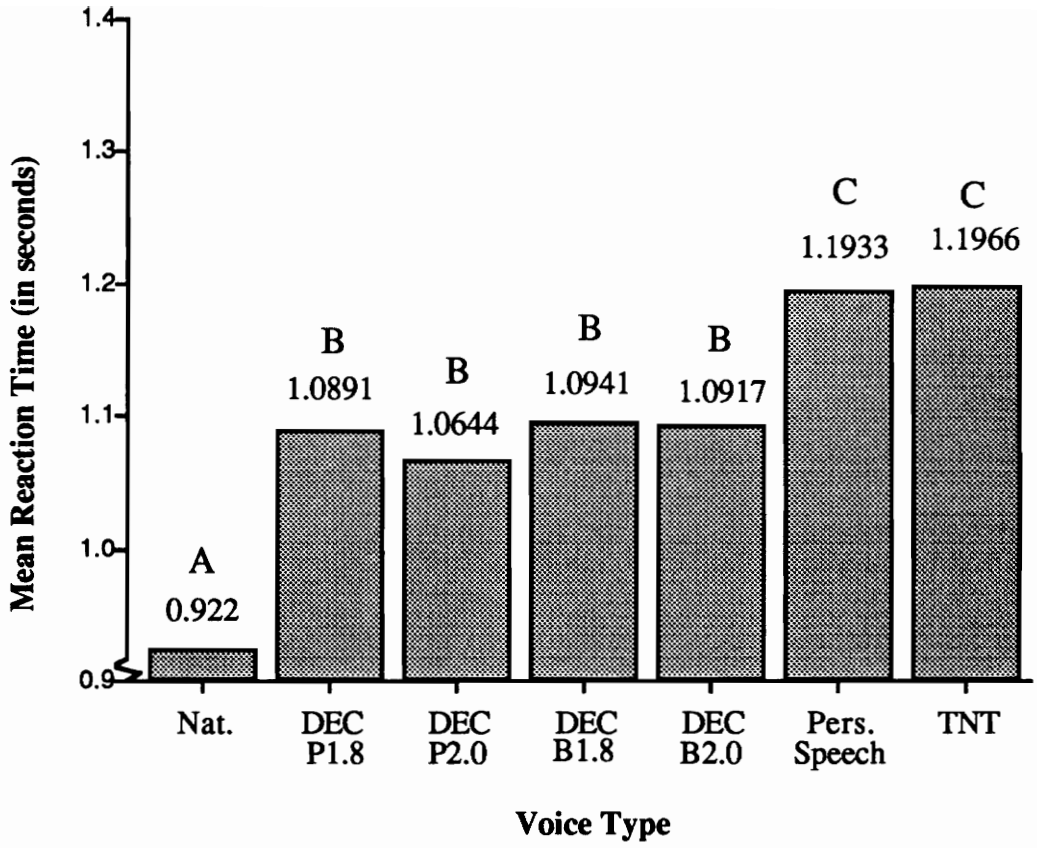


Figure 5. Reaction time means for voice type (including medium frequency). Reaction times for voice types with different letters were found to be significant at $p < 0.05$.

TABLE 4

ANOVA Summary Table for Mean Reaction Time (Excluding Medium Frequency).

Source	*df*	*SS*	*F*	*p*	*Epsilon*
Between Subjects					
Subjects (S)	9	5.701			
Within Subjects					
Word Frequency(WF)	1	0.123	17.545	0.0023	1.0
WF x S	9	0.063			
Voice Type (VT)	6	1.125	28.041	0.0001	0.52
VT x S	54	0.361			
WF x VT	6	0.148	2.918	0.0477	0.54
WF x VT x S	54	0.458			
Total	139	7.979			

The main effects of word frequency and voice type were significant. The Newman-Keuls analysis results for voice type are shown in Figure 6.

The word frequency by voice type interaction was also significant. Newman-Keuls post hoc analysis was used to evaluate the interaction, however, only 7 unconfounded comparisons are used to determine frequency effects (i.e. a significant difference between the observed mean reaction time for the low frequency words and the observed mean reaction time for the high frequency words, for a particular voice type). The comparisons that indicate the presence of frequency effects are the comparisons between the mean reaction times to high frequency words for a voice type and the mean reaction times to low frequency words for the same voice type. The results showed that there were no frequency effects observed for the four DECtalk voices or natural speech. Frequency effects were found, however, for the Personal Speech and the Type'n'Talk. Figure 7 shows the unconfounded comparisons used in evaluating the word frequency by voice type interaction and frequency effects.

Error Rate

The summary table for the error analysis including medium frequency is reported in Table 5. The main effects of word frequency and voice type were significant. Mean error rates for each frequency and each voice type are shown in Figure 8 and Figure 9,

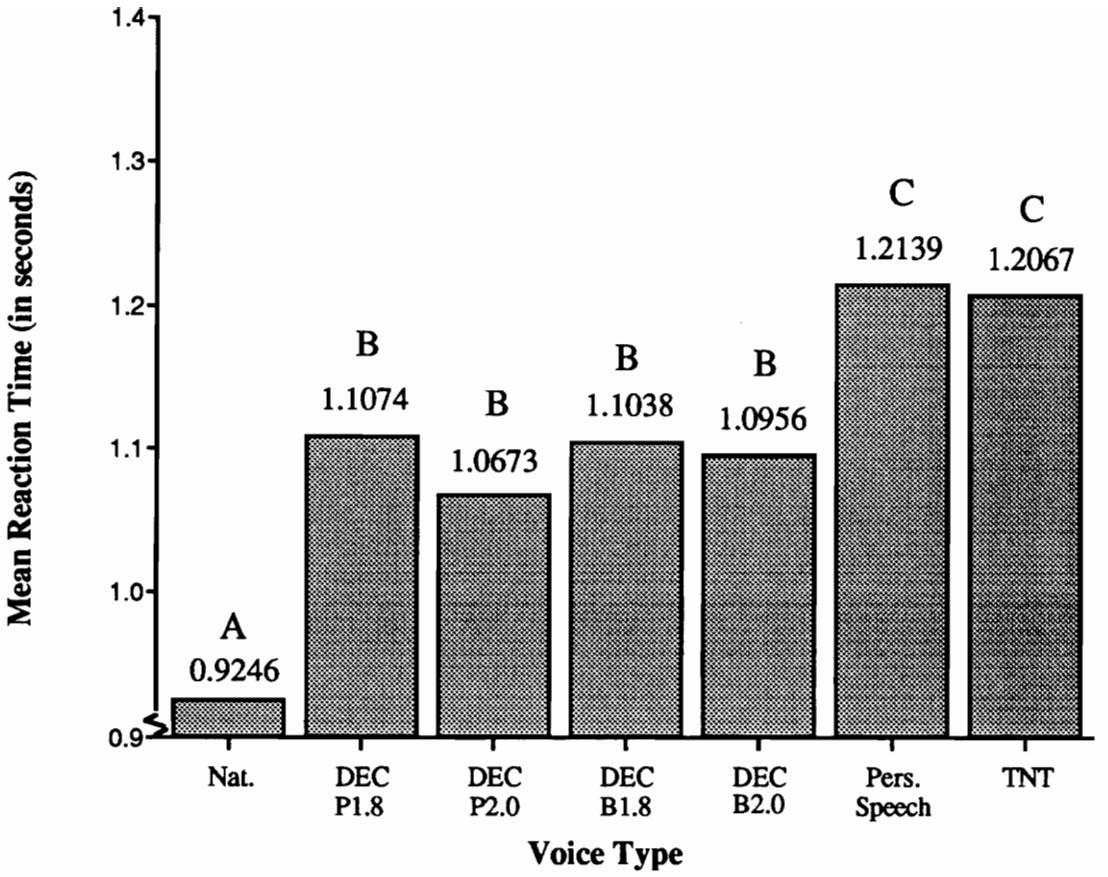


Figure 6. Reaction time means for voice type (excluding medium frequency). Reaction times for voice types with different letters were found to be significant at $p < 0.05$.

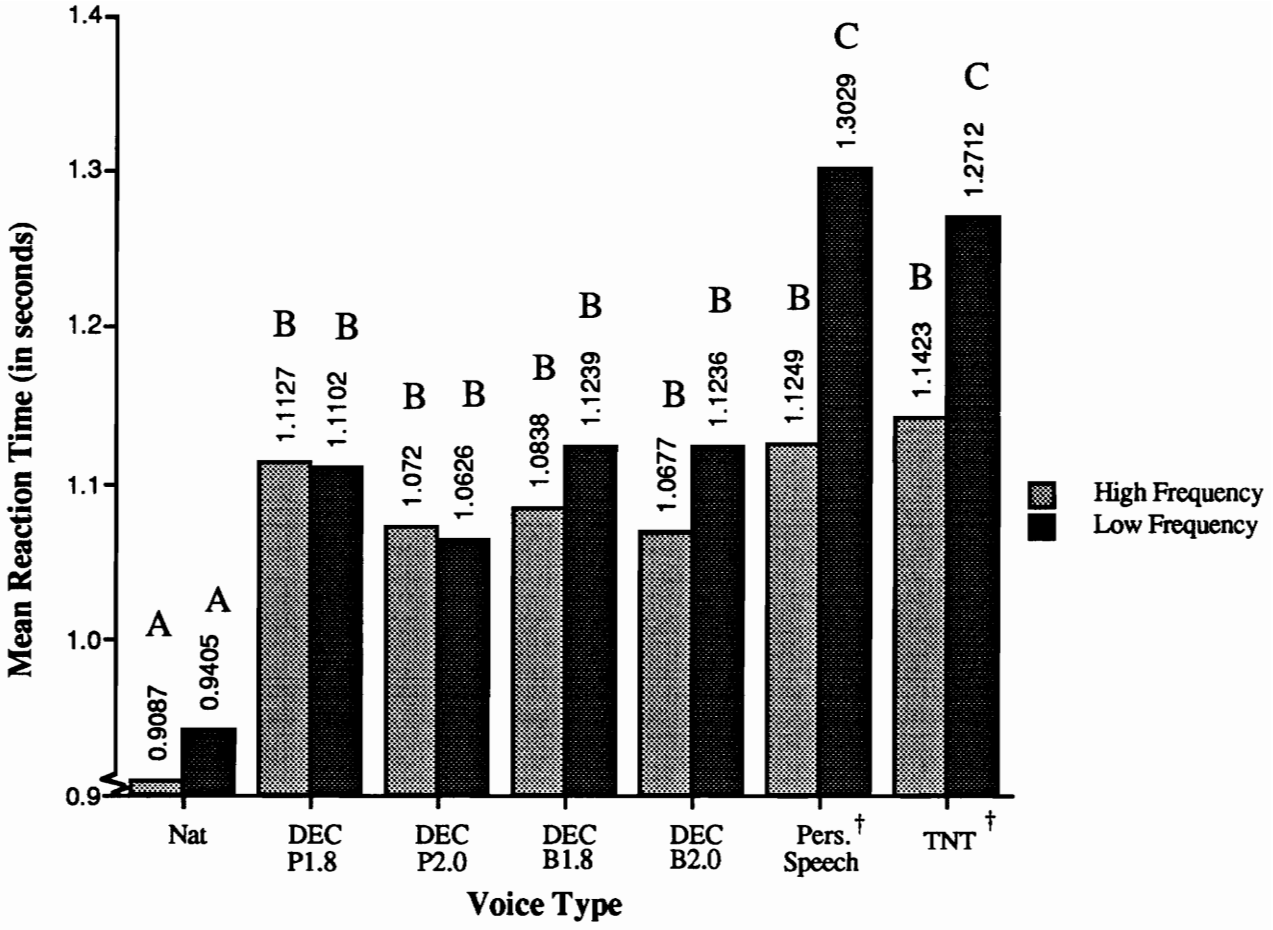


Figure 7. Comparisons used to evaluate the interaction of frequency by voice type. Reaction times for bars with different letters were found to be significant at $p < 0.05$.

† Voice type shows frequency effects.

TABLE 5

ANOVA Summary Table for Errors (Including Medium Frequency)

Source	*df*	*SS*	*F*	*p*	*Epsilon*
Between Subjects					
Subjects (S)	9	789.433			
Within Subjects					
Word Frequency(WF)	2	47.495	4.773	0.0304	0.82
WF x S	18	89.552			
Voice Type (VT)	6	2713.333	69.507	0.0001	0.55
VT x S	54	351.333			
WF x VT	12	29.038	0.959	0.959	0.36
WF x VT x S	108	272.581			
Total	209	4292.762			

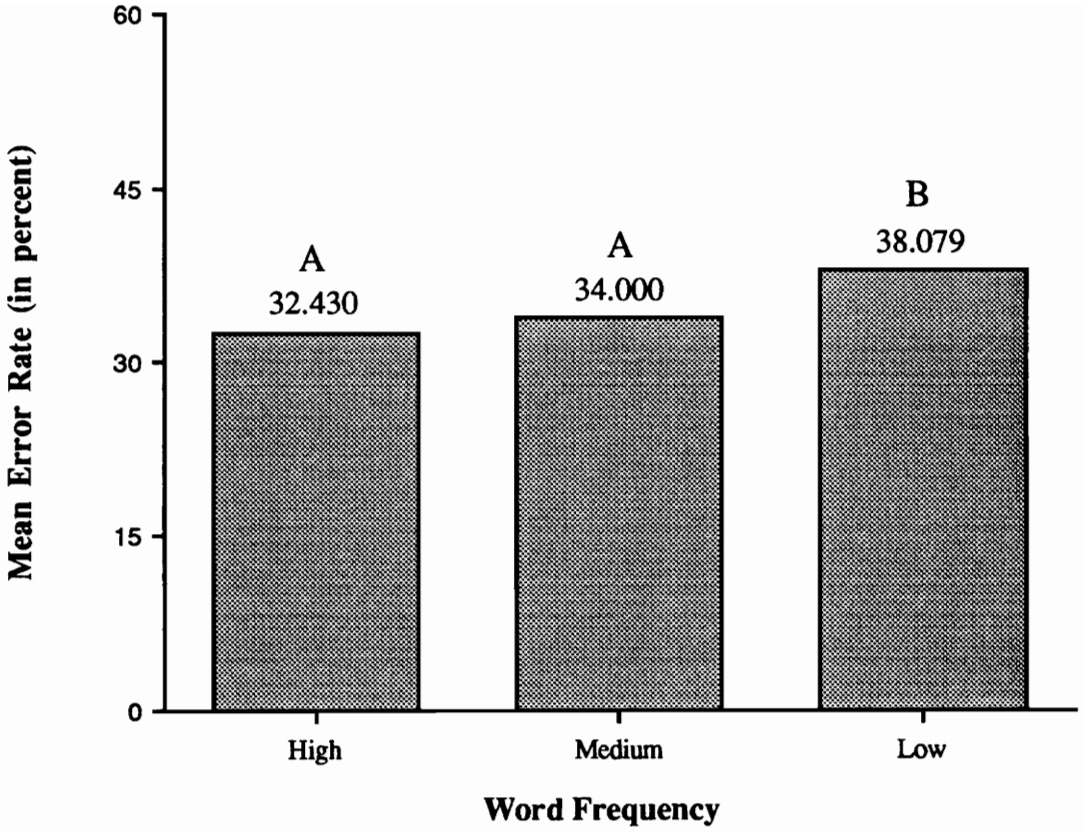


Figure 8. Error rate means for word frequency. Error rates for word frequencies with different letters were found to be significant at $p < 0.05$.

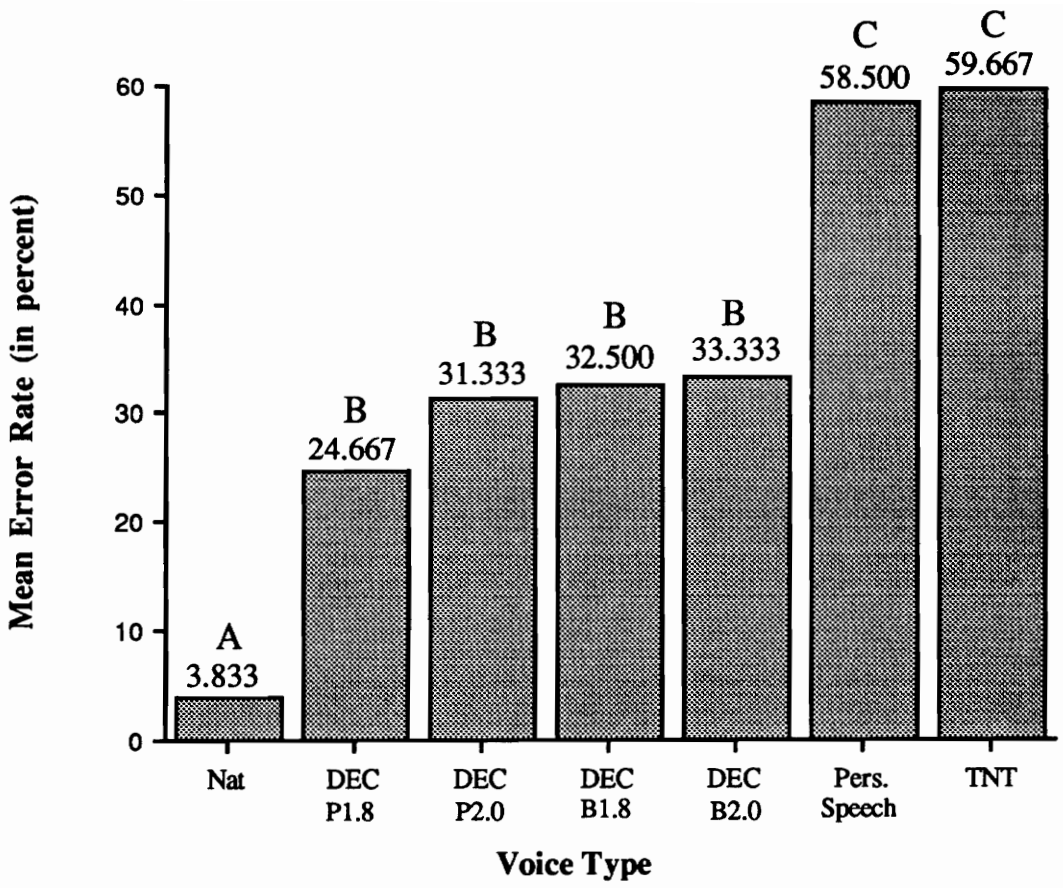


Figure 9. Error rate means for voice type (including medium frequency). Error rates for voice types with different letters were found to be significant at $p < 0.05$.

respectively. The word frequency by voice type interaction was not significant.

A Newman-Keuls post hoc analysis showed that the error rates observed for the high frequency and medium frequency conditions were significantly lower than those observed for the low frequency condition. There was no difference between the error rates for the high frequency and medium frequency conditions.

The summary table for the error analysis excluding medium frequency, is reported in Table 6. The main effects of word frequency and voice type were significant. Mean error rates for each voice type are shown in Figure 10. The word frequency by voice type interaction was not significant.

Results from Newman-Keuls analysis showed that for the main effect of voice type, the error rate for the natural voice was significantly lower than all of the other voices. This analysis also showed that the error rates of the DECtalk voices did not differ from each other and that the Personal Speech and the Type'n'Talk did not differ from each other. However, the results showed that the error rates observed for the DECtalk voices did differ from the error rates observed for the Personal Speech and the Type'n'Talk voices.

TABLE 6

ANOVA Summary Table for Errors (Excluding Medium Frequency)

Source	*df*	*SS*	*F*	*p*	*Epsilon*
Between Subjects					
Subjects (S)	9	564.864			
Within Subjects					
Word Frequency(WF)	1	44.579	6.153	0.0350	1.0
WF x S	9	65.207			
Voice Type (VT)	6	1648.300	55.983	0.0010	0.59
VT x S	54	264.986			
WF x VT	6	8.471	0.470	0.7047	0.50
WF x VT x S	54	162.243			
Total	139	2758.650			

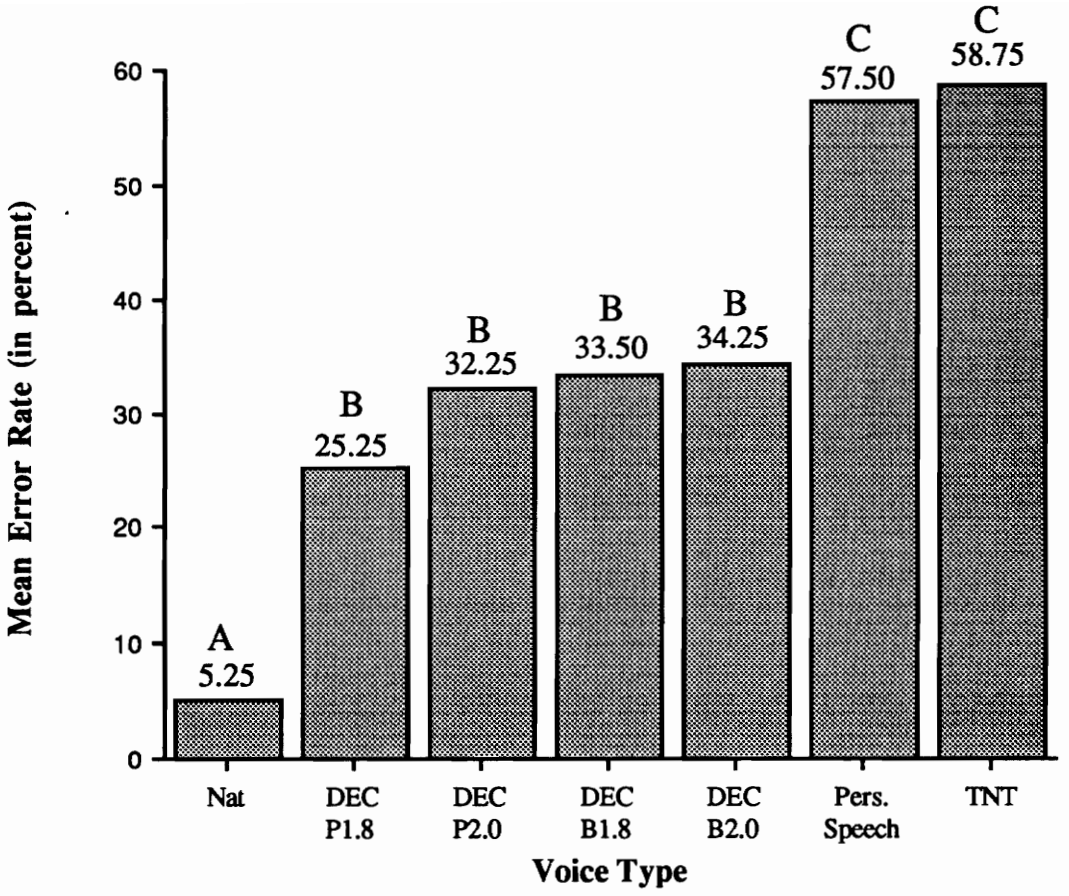


Figure 10. Error rate means for voice type (excluding medium frequency). Error rates for voice types with different letters were found to be significant at $p < 0.05$.

DISCUSSION

Main Effects

Frequency

The results show that word frequency significantly affected subject reaction time. Post hoc analysis showed that the differences observed were between high frequency words and low frequency words, and between medium frequency words and low frequency words. These results suggest that the use of more commonly used words can lead to reductions in time needed to recognize words spoken by synthetic speech displays or even digitized human speech displays. The results are also consistent with previous studies involving auditory lexical decision tasks (e.g. Connine *et al.*, 1990; Luce, 1986).

The practical significance of reaction time advantages of 0.06 second is highly application dependent. The trade off for using high versus low frequency words must be considered for only those applications where time differences of that size would impact on performance. Cockpit applications, for example, can place serious response demands upon the user. In contrast, regular voice response telephone applications would not place such demands on its users.

Error rates were also affected by word frequency. The post hoc analysis showed that error rates observed for the high frequency

words and medium frequency words were significantly lower than those observed for low frequency words. Again, the results indicate that the use of words that have a higher frequency of occurrence can lead to improved performance with speech displays.

Voice Type

The main effect of Voice Type was also found to be significant, for both reaction time and errors. Results from the post hoc analysis showed that the natural voice condition was significantly better than all of the other conditions in terms of reaction time, and that the DECtalk voices (Perfect Paul 1.8, Perfect Paul 2.0, Beautiful Betty 1.8, and Beautiful Betty 2.0) were significantly better than the Personal Speech and the Type'n'Talk voices. In addition, the DECtalk voices were not found to differ significantly from each other in terms of reaction time, nor were the Personal Speech and Type'n'Talk voices found to differ significantly from each other. Once again, the practical significance of applying these results is application dependent.

Newman-Keuls analysis showed that the same pattern found for reaction time was also found for error rate. The error rate for the natural voice was significantly lower than all of the other voices. The error rates of the DECtalk voices did not differ significantly from each other, but did differ significantly from the Personal Speech and the Type'n'Talk. The results also showed that the Personal Speech and the Type'n'Talk did not differ significantly from each other.

The results from the voice type analysis can be used to separate the voices into categories, as was done in the Greene *et al.* (1986) study. The overall rank order was the same in this study as in the Greene *et al.* (1986) study, with the DECtalk voices being better than the Type'n'Talk. However, error rates were about 10-15% higher in this study.

One explanation for the higher error rates is that, in this study, subjects were asked to respond quickly and may have unintentionally pressed an incorrect key. Also, subjects did not have the chance to change their response once it was made, nor were they required to transcribe the words. With additional time embedded in the task, as in the procedure utilized by Greene *et al.* (1986), the error rates might have been lower.

Another explanation stems from the differences between the tasks. Greene *et al.* (1986) used a Modified Rhyme test, while this study used a lexical decision task. The goal in the lexical decision task is to decide whether a stimulus is a word or a non-word. The goal in the Modified Rhyme test is to transcribe the word that was presented. In the Modified Rhyme test the subject can assume the stimulus is a word and concentrate on comprehension, however, the subject in the lexical decision task must try to tease apart the non-word "noise" from the low quality "noise" of the synthesizer. The lexical decision task can be thought of as a type of signal detection task, where words are the signal and non-words are the noise. Given

this analogy, it is plausible that the higher error rates are due to the increased amount of noise, specifically non-words, in the lexical decision task versus the Modified Rhyme test.

Word Frequency by Voice Type Interaction

The word frequency by voice type interaction was not found to be significant for the mean error rate analysis or the three frequency reaction time analysis. However, the interaction between word frequency and voice type was significant for reaction time when the difference between the high and low frequency conditions were highlighted by removing the medium frequency condition.

Newman-Keuls analysis showed that there was no difference in mean reaction time between high and low frequency words for the DECtalk. One explanation is that the DECtalk includes many exception rules. This may "level out" the quality so that both regular and exception words are processed equally. In other words, the advantage in processing regular words is lost because the quality of the exception words is higher than normally expected while the quality of the regular words has remained the same. Another explanation is that a higher digitizing rate is needed to bring out the frequency effects of this synthesizer. This would also explain the absence of frequency effects for the natural speech condition.

The fact that no frequency effects were found for the DECtalk synthesizers has positive and negative implications for speech system designers. The fact that reaction times do not increase

significantly when low frequency words are used, allows for the use of a wider range of words for the system vocabulary when using the DECtalk. However, this result suggests that humans are not processing the synthetic speech the same way they process natural speech. Therefore, it is necessary to take caution when applying results obtained for natural speech to synthetic speech displays.

The Personal Speech and Type'n'Talk did show a significant increase in reaction times for low frequency words as compared to high frequency words. One explanation is that these synthesizers do not include an adequate number of exception rules. If this is the case, the exception words will have longer processing times because of the mispronunciations. This would make the spread between the reaction times for high frequency and low frequency words larger.

The cost of the synthesizer often determines the quality. The Greene *et al.*, (1986) study indicated that, in fact, when it comes to speech synthesizers, "you get what you pay for" (Greene *et al.*, 1986, p. 105). The results of the present study are in agreement with this statement. It can also be assumed that deletion of exception rules is one way to lower the costs associated with a synthetic speech generator. The results of this study, then, suggest that it would be desirable to limit the use of low frequency words when using a low quality synthesizer and to use high frequency synonyms whenever possible. However, the time advantages associated with higher frequency words may not be significant for some applications.

The Activation-Verification Model

The results of this study are consistent with the Paap *et al.*, (1986) Activation-Verification (AV) Model of word recognition. Before discussing the results of the present study, a brief review of the AV Model is in order.

According to the AV model (Paap *et al.*, 1986), the word recognition process begins with the activation of phoneme units by phonemic feature detectors from the speech input. Next, lexical entries are activated, and those activated past some threshold value are placed in a verification list. Verification of one of the items in the list is achieved by sequential comparison of each item in the list and a stored representation of the stimulus. Items are verified in descending order of frequency. Verification of a word candidate can result in a match or a mismatch. A match occurs when the degree of fit between a lexical item and the stored representation of the stimulus exceeds criterion, a mismatch occurs when the criterion is not exceeded.

If the quality of the input signal is near perfect, only a small number of items will be included in the verification list. As the quality of speech synthesizers decreases, the number of activated features increases because the phoneme units are not uniquely identifiable. This increase results in an increase of lexical entries in the verification list, which in turn results in an increase in the amount of time needed to complete the verification process.

For example, if the stimulus were THIRST, and spoken by a low quality speech synthesizer, the phoneme units /θ/ and /f/ might both be activated, resulting in a verification list that includes the words FIRST and THIRST. Since FIRST has a higher frequency of occurrence, it will be checked before THIRST. If a high quality synthesizer or speech display had generated the signal, however, it is possible that only the /θ/ phoneme unit would be activated and the verification process would be shorter because of the decrease in items in the verification list. In the most extreme case, the criterion is not exceeded for any item in the verification list. This results in a mismatch and possibly an error.

Thus, according to the AV Model (Paap *et al.*, 1986) and the results of this study, it may be the case that making phonemes more uniquely identifiable, may lead to a reduction in the number of phoneme units activated by the feature detectors. This would in turn result in a shorter verification list and ultimately reduce the number of errors and reduce the amount of time needed to recognize words generated by synthetic speech systems.

Future Research

There are many more research questions that remain to be answered. One question involves the perception of synthetic speech in noisy environments. Information about how perception of synthetic speech in noise (e.g. different levels and different types) differs from perception of natural speech in noise could give insight

as to how to design synthetic speech systems in order to improve perception in noise.

Many research studies have identified factors which should be considered when designing speech displays. It is known that factors such as rate of speech, frequency of occurrence of words, segment size, and repetition have optimal values or ranges. Future research studies should concentrate on determining which of the known factors are the most useful for improving quality of synthetic speech. That is, it would be useful to know which factor or factors have the greatest influence on quality. This information would assist designers in making cost/benefit analyses.

Finally, research designed to determine how to make synthesized phonemes more distinguishable from one another should be conducted. This research should strive to determine the important factors in discriminating synthetic phonemes from one another. It should include examination of factors important for natural speech phoneme discrimination such as voice onset time, frequency spectrum characteristics, distinctive features, and formant transitions. The research should be directed at identifying the factors that are relatively unaffected by the absence of continuous speech cues such as coarticulation and context-conditioned variation. Finally, it must identify the factors that can be successfully included in text-to-speech systems given current technology. Similarity in processing of synthesized words with altered phonemes can then be compared to processing of synthetic words with current phonemes

and natural speech using the auditory lexical decision task and the methodology described in this thesis.

Conclusions

Summary of Results

The results of this study can be summarized as follows:

- The use of low frequency words with low quality synthetic speech generators, such as the Personal Speech and Type'n'Talk can increase the time needed for a human to recognize the output.

- Error rates for the Personal Speech and Type'n'Talk were significantly higher than the error rates for the DECtalk voices.

- Reaction times for the DECtalk voices were significantly lower than the reaction times for the Personal Speech and the Type'n'Talk.

- Reaction times and error rates were higher for the synthetic speech voices (DECtalk voices and Votrax voices) than for the natural voice.

- Increases in the error rates and the reaction times for low quality speech can be attributed to the drop in efficiency of phonemic feature detectors, resulting in a longer verification list.

Guidelines

The results imply the following guidelines for developing speech displays.

- Use of an auditory lexical decision task can be used to determine the relative quality among several speech synthesizers and natural speech.

- Including the use of low frequency words in the vocabulary of high quality speech displays should not affect time needed to recognize the output, nor should it result in an increase in errors.

- Replacing low frequency words with high frequency synonyms for low quality synthesizers, should be attempted whenever possible. Especially in time critical environments such as cockpits, or emergency alert systems.

- Cost/Benefit analysis of system requirements should be performed when deciding whether it is worth while to use the more expensive synthesizers or whether the less expensive models would be adequate for the application.

- Cost/Benefit analysis should be performed to decide whether response time decrements of a fraction of a second are critical to

an application before applying the type of results reported in this research.

REFERENCES

Carroll, D. W. (1986). *Psychology of language.* Monterey, CA: Brooks/Cole.

Connine, C. M., Mullenix, J., Shernoff, E., and Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(6), 1084-1096.

Davis, H., and Kranz, F. W. (1964). The international standard reference zero for pure-tone audiometers and it relation to the evaluation of impairment of hearing. *Journal of Speech and Hearing Research, 7*, 7-16.

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope, 58*, 955-991.

Fairbanks, G. (1958). Test of phonemic differentiation: The rhyme test. *Journal of the Acoustical Society of America, 30*(7), 596-600.

Francis, W. N., and Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Boston, MA Houghton Mifflin.

Frauenfelder, U. H., and Tyler, L. K. (1987). The process of spoken word recognition: An introduction. *Cognition, 25*, 1-20.

Greene, B. G., Logan, J. S., and Pisoni, D. B. (1986). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. *Behavior Research Methods, Instruments, and Computers, 18,* 100-107.

Greenhouse, S. W., and Geisser, S. (1959). *On methods in the analysis of profile data. Psychometrika, 24*(2), 95-112.

Herlong, D. W., and Williges, B. H. (1988). Designing speech displays for telephone information systems. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp.215-218). Santa Monica, CA: Human Factors Society.

House, A. S., Williams, C. E., Hecker, M. H. L., and Kryter, K. D. (1965). Articulation-testing methods : Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America, 37*(1), 158-166.

Jenkins, J. A. (1989). *Utilization of the naming task to manipulate the lexical and aural pathways of the extended activation-verification model.* Unpublished master's thesis, New Mexico State University, Las Cruces, NM.

Logan, J. S., Pisoni, D. B., and Greene, B. G. (1985). Measuring the segmental intelligibility of synthetic speech: Results from eight text-to-speech systems. *Research on Speech Perception,* Progress Report No. 11, Indiana University, 3-31.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon.* Unpublished doctoral dissertation, Indiana University, Bloomington, IN.

Merva, M. A. (1987). Effects of speech rate, message repetition, and information placement on synthesized speech intelligibility. Unpublished masters thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Miller, G. A., and Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27,* 338-352.

Morrill, J. C. (1986). Hearing measurement. In E. H. Berger, W. D. Ward, J. C Morrill, L. H. Royster (Eds.), *Noise and hearing conservation manual.* Akron, OH: American Industrial Hygiene Association.

Nusbaum, H. C., Schwab, E. C., and Pisoni, D. B. (1984). Subjective evaluation of synthetic speech: Measuring preference, naturalness, and acceptability. *Research on Speech Perception,* Progress Report No. 10, Indiana University, 391-408.

O'Shaughnessy, D. (1987). *Speech communication: human and machine.* Reading, MA: Addison-Wesley.

Paap, K. R., Newsome, S. L., McDonald, J. E., and Schvaneveldt, R. W. (1982). An activation-verification model for letter and word recognition: The word superiority effect. *Psychological Review, 89*(5), 573-594.

Paap, K. R., McDonald, J. E., Schvaneveldt, R. W., and Noel, R. W. (1986). Frequency and pronounceability in visually presented naming and lexical-decision tasks. In M. Coltheart (Ed.), *Attention & Performance, XII.* New York: Academic Press.

Pisoni, D. B. (1979). Some measures of intelligibility and comprehension. *Research on Speech Perception*, Progress Report No. 5, Indiana University, 3-47.

Rothauser, E. H., Urbanek, G. E., and Pachl, W. P. (1971). A comparison of preference measurement methods. *Journal of the Acoustical Society of America, 49*(4), 1297-1308.

Rubenstein, H., Garfield, L., and Milliken, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9,* 487-494.

Schvaneveldt, R. W., and McDonald J. E. (1981). Semantic context and the encoding of words: Evidence for two modes of stimulus analysis. *Journal of Experimental Psychology: Human Perception and Performance, 7*(3), 637-687.

Seidenberg, M. S., Waters, G. S., Barnes, M. A., and Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior, 23,* 383-404.

Simpson, C. A. and Marchionda-Frost, K., (1984). Synthesized speech rate and pitch effects on intelligibility of warning messages for pilots. *Human Factors, 26*(5), 509-517.

Simpson, C. A., McCauley, M. E., Roland, E. F., Ruth, J. C., and Williges, B. H., (1985). System design for speech recognition and generation. In R. M. Baecker and W. A. S. Buxton (Eds.) *Human computer interaction: a multidisciplinary approach.* San Mateo, CA: Morgan Kaufmann.

Slowiaczek, L. M. and Nusbaum, H. C. (1985). Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors, 27*(6), 701-712.

Streeter, L. A. (1988). Applying speech synthesis to user interfaces. In M. Helander (Ed.) *Handbook of human-computer interaction.* North Holland: Elsevier.

Thorndike, E. L., and Lorge, I. (1944). *The teacher's book of 30,000 words.* New York: Columbia University Press.

Voiers, W. D. (1977). Diagnostic evaluation of speech intelligibility. In M. E. Hawley (Ed.) *Speech intelligibility and speaker recognition*, Vol. 2, Benchmark Papers in Acoustics, Stroudsburg, PA: Dowden, Huchinson and Ross.

Appendix A

Stimuli Used in the Study

High Frequency Words and their associated frequency count†

Word	Frequency	Word	Frequency
WORK	496	LONG	541
YEAR	1661	BEGIN	583
CASE	503	DOWN	698
GROUP	512	FIRST	1031
GIVE	1264	THOSE	864
SHOW	640	PLACE	584
LIFE	772	CHURCH	451
THINK	982	TAKE	1575
HOLD	509	BOTH	731
VERY	703	EVEN	997

† Frequency count is number of occurrences per million words.

Medium Frequency Words and their associated frequency count†

Word	Frequency	Word	Frequency
HEAVY	110	ISSUE	200
VOLUME	182	FACTOR	176
DINNER	100	GOAL	100
THEORY	150	CHOOSE	177
THROW	150	MOTOR	108
NEWS	101	PAST	100
THAT	136	HEALTH	105
YARD	100	MUCH	212
PROVE	156	CLUB	178
TRIP	109	BASE	102

† Frequency count is number of occurrences per million words.

Low Frequency Words and their associated frequency count†

Word	Frequency	Word	Frequency
LATHE	2	ABIDE	14
DRAPE	9	POTENT	9
KEEN	11	THRIVE	11
VERIFY	11	ESCORT	11
WEAKEN	15	GOSSIP	13
DOING	10	FABLE	4
PUNISH	14	WOOL	13
SHRINK	12	IMPAIR	11
HEARTH	4	ROBE	10
TUNE	15	GRIEF	10

† Frequency count is number of occurrences per million words.

Non-word stimuli

MOKERN	CABET	PANGLE	AKLETE
BRASHEN	CINTER	BALLON	ROVINE
ZALACE	AUSTENE	TURNID	RIVEK
FLANT	TAURIC	THAUPE	VERTA
COGEST	CHIPE	TERMAC	DRIST
NOVER	GALIDE	THOGAL	BLICK
WONK	LUSK	LEGAR	NARVIS
INDICE	INPOIT	DOVAGE	IMPID
YONKER	MUND	LANT	DELSE
HELK	CLANGE	TAVOT	KLAFT
SMOAL	BINK	SCUND	MASIC
PRAUSE	ANVID	DREP	LANINE
COLICE	VALON	JAVIS	FROVE
VENEX	VOLUDE	GRAFE	SINT
PALON	SCOAT	ZANBER	ORBID

APPENDIX B

Instructions for the Auditory Lexical Decision Task

Instructions

In this experiment, your task will be to determine whether a letter string, presented through headphones, is a word or a non-word.

The experiment will consist of 70 practice trials and 840 experimental trials and will follow the procedure outlined below. If you have any questions about the procedures, do not hesitate to ask the experimenter for further explanation.

1. The experimenter will start the experiment when you are ready.
2. Once the trial has begun, you will see a warning dot on the screen in front of you. This dot indicates that a letter string (stimulus) will be spoken through the headphones.
3. Once you see the warning dot, direct your attention to listening for the stimulus.
4. Once you have heard the stimulus, your task is to indicate as **quickly** and as **accurately** as possible whether the stimulus was a word or a non-word. You indicate your choice by pressing the appropriate key on the keyboard in front of you. Take a minute and look at the keys. The key marked "**NONWORD**" is in the place where the "Z" key usually is, and the key marked "**WORD**" is in the place where the "?" and "/" usually are. If you can not find

the keys or have any questions about the procedure to this point, stop and ask the experimenter for assistance.

5. There will be a 2 second interval between trials, then the warning dot will be presented again and the next trial will have begun.

6. This procedure will continue for approximately four minutes. At that point, you will be given a one minute break.

APPENDIX C

SUBJECT INFORMED CONSENT DOCUMENT

Human-Computer Interaction Laboratory
533 Whittemore Hall

PARTICIPANT'S STATEMENT OF INFORMED CONSENT

You are asked to participate in a study to investigate human word recognition. In the experiment you will hear a letter string through the head phones. Your task is to decide whether or not this letter string is a word. Once you have decided, you are asked to quickly indicate your decision by pressing the key labeled "WORD" if you feel the sequence presented is in fact a word or by pressing the key labeled "NONWORD" if you feel that the sequence presented is not a word. More detailed instructions will be presented to you before the session begins. The study will take approximately 50 minutes. You will be paid $10.00 for your participation.

All information collected in the experiment will be held in strict confidence. We will use the information for statistical and summary purposes only, and will make certain that your name is not associated with your records.

To the best of our knowledge, there are no physical or psychological risks associated with the procedures in our study. The only known discomfort to which you will be exposed is possible fatigue resulting from the length of the experiment. However, you will be permitted to take rest breaks when needed.

As a participant in this study, you have certain rights. These rights will now be explained to you, and you will be asked for your signature, indicating that you consent to participation in this research.

1. You have the right to stop the experiment in which you are participation at any time if you feel that it is not agreeable to you.

2. You have the right to see your data and to withdraw it from the experiment if you feel that you should.

3. You have the right to be informed of the results of the overall experiment. If you wish to receive a summary of the results, please indicate you address (three months hence) with your signature. A summary will be sent to you. If you should then like further information, please contact the Human-Computer Interfaces Laboratory and a full report will be made available to you.

4. You have the right to call either **Reni L. Jenkins**, the experimenter, at 953-0192 or **Dr. E. R. Stout**, Institutional Review Board, at 231-5281, with your concerns about any aspect of the experiment.

The faculty and graduate students involved greatly appreciate your help as a voluntary participant. If you have any questions about the experiment or your rights as a participant, please do not hesitate to ask. We will do our best to answer them, subject only to the constraint that we do not want to pre-bias the experimental results.

Your signature on this form indicates that you have read your rights as a participant as stated above and that you consent to participation. If you include your printed name and address below, a summary of the experimental results will be sent to you.

Signature _____

Address (3 months hence):

_____

_____

_____

VITA

RENI LEE JENKINS

4408 D Emerald Forest Dr. Home: (919) 544-2046
Durham, NC 27713 Office: (919) 991-8641

EDUCATION

Master of Science, Industrial Engineering-Human Factors, Virginia Polytechnic Institute and State University, Blacksburg, Virginia. August 1990 to January 1992. GPA = 3.6/4.0
Thesis: "The Use of the Auditory Lexical Decision Task as a Method for Assessing the Relative Quality of Synthetic Speech", under Dr. R. Williges.

Bachelor of Arts, Psychology, New Mexico State University, Las Cruces, New Mexico. 1983-1988. GPA = 3.5/4.0

TECHNICAL EXPERIENCE

Spring 1992: Member of the Scientific Staff, Residential Services Planning, Bell Northern Research, Research Triangle Park, NC. Responsibilities include user interface design of telephone applications, research and writing of user interface guidelines for development of telephone applications, user testing of documentation and development tools.

Summer 1990: Staff scientist at Ashton-Tate Corporation, Torrance, CA. Responsibilities included design and development of graphical user interface prototypes, and user testing.

HONORS AND PROFESSIONAL SOCIETIES

Ford Foundation Predoctoral Fellowship Recipient, 1989-1991.
Human Factors Society Student Affiliate.

Reni Lee Jenkins

Birthdate: October 19, 1964