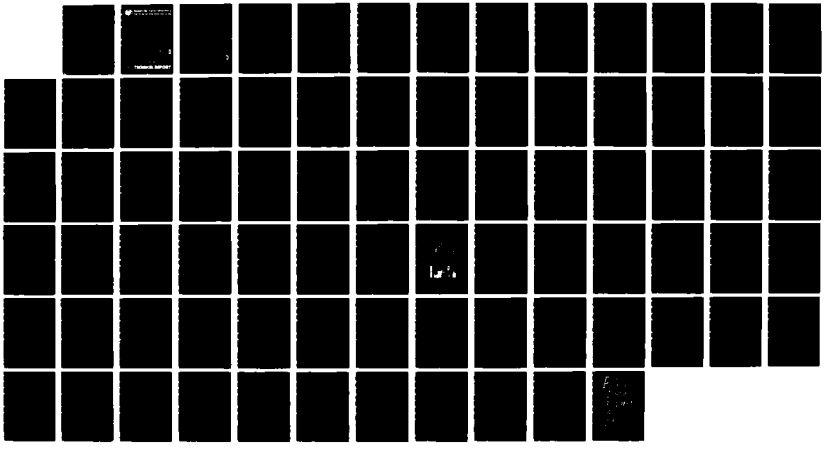
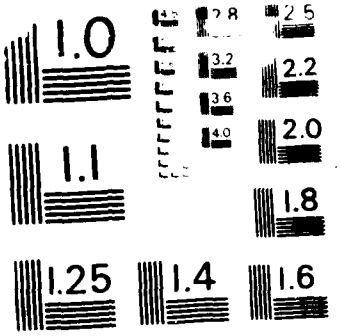
AD-A190 432 AN EXPERIMENTAL DETERMINATION OF THE INTELLIGIBILITY OF 1/1
TWO DIFFERENT SPE (U) PENNSYLVANIA STATE UNIV
UNIVERSITY PARK APPLIED RESEARCH LAB
UNCLASSIFIED R DEPAOLIS ET AL. DEC 87 ARL/PSU/TR-87-012 F/G 25/4 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

(4)

Applied Research Laboratory
The Pennsylvania State University

AN EXPERIMENTAL DETERMINATION OF THE INTELLIGIBILITY OF TWO DIFFERENT SPEECH SYNTHESIZERS IN NOISE

by

R. A. DePaolis and C. P. Janota

ARL:PSU

TECHNICAL REPORT

4

The Pennsylvania State University
APPLIED RESEARCH LABORATORY
P.O. Box 30
State College, PA 16804

AN EXPERIMENTAL DETERMINATION OF THE INTELLIGIBILITY OF TWO DIFFERENT SPEECH SYNTHESIZERS IN NOISE

by

R. A. DePaolis and C. P. Janota

Technical Report No. TR 87-012
December 1987

Supported by:
Naval Sea Systems Command

L. R. Hettche
Applied Research Laboratory

Approved for public release; distribution unlimited

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT (A) Unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) TR-87-00	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Applied Research Laboratory The Penna. State University 6b. OFFICE SYMBOL (If applicable) ARL	7a. NAME OF MONITORING ORGANIZATION Naval Sea Systems Command Department of the Navy
6c. ADDRESS (City, State, and ZIP Code) P. O. Box 30 State College, PA 16804	7b. ADDRESS (City, State, and ZIP Code) Washington, DC 20362
8a. NAME OF FUNDING / SPONSORING ORGANIZATION Naval Sea Systems Command 8b. OFFICE SYMBOL (If applicable) NAVSEA	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N-00024-85-C-6041
8c. ADDRESS (City, State, and ZIP Code) Department of the Navy Washington, DC 20362	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
AN EXPERIMENTAL DETERMINATION OF THE INTELLIGIBILITY OF TWO DIFFERENT SPEECH SYNTHESIZERS IN NOISE (unclassified)

12. PERSONAL AUTHOR(S)
R. A. DePaolis and C. P. Janota

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
MS Thesis	FROM _____ TO _____	December 1987	76

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Speech intelligibility, Speech recognition, speech synthesis

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Synthetic speech has become commonplace within society. One cause of this proliferation is the availability of varied inexpensive synthetic speech systems to meet almost any application from those in industry to those in communicative disorders. The ability to choose the most effective communication system is an increasingly important consideration as the role of synthetic speech in society grows.

This study examined two predominant inexpensive methods of synthesizing speech: formant synthesis(FS) and linear prediction coding(LPC). A pilot study indicated that upon first presentation of noncontextual material that FS was significant more understandable than LCP (α=0.01).

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code) 814/355-6344 / 22c. OFFICE SYMBOL ARL/PSU

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

An experiment using two groups of six subjects trained in either FS or LPC was conducted. Three sets of 66 contextual sentences from the revised SPIN test were presented at 80 dBC SPL at three noise levels of signal to babble (0, +5 and +10) to each subject.

The results fall into two categories: the training data and the subsequent test data. Whereas the training data indicate the eventual equality of mean percent correct word scores for the two synthesizers without noise, the test data indicate the superior performance of LPC with interfering noise (α-0.016). The effect of the interfering noise is studied as a cause and the significance of this study to current research in synthetic speech is discussed.

ABSTRACT

Synthetic speech has become commonplace within society. One cause of this proliferation is the availability of varied inexpensive synthetic speech systems to meet almost any application from those in industry to those in communicative disorders. The ability to choose the most effective communication system is an increasingly important consideration as the role of synthetic speech in society grows.

This study examined two predominant inexpensive methods of synthesizing speech: formant synthesis(FS) and linear prediction coding(LPC). A pilot study indicated that upon first presentation of noncontextual material that FS was significantly more understandable than LPC (α=0.01).

An experiment using two groups of six subjects trained in either FS or LPC was conducted. Three sets of 66 contextual sentences from the revised SPIN test were presented at 80 dBC SPL at three noise levels of signal to babble (0, +5 and +10) to each subject.

The results fall into two categories: the training data and the subsequent test data. Whereas the training data indicate the eventual equality of mean percent correct word scores for the two synthesizers without noise, the test data indicate the superior performance of LPC with interfering noise (α =0.016). The effect of the interfering noise is studied as a cause and the significance of this study to current research in synthetic speech is discussed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

The author acknowledges the most important contribution of Dr. Claus P. Janota, his thesis advisor. The knowledge and advice of this person were instrumental towards every phase of this research.

The author would like to thank Dr. Alan D. Stuart and Dr. Carter L. Ackerman for their reading of this manuscript and constructive comments.

Additional thanks are warranted to The Pennsylvania State University Speech and Hearing Clinic, specifically Patti Hergenreder, for the speedy hearing tests administered, and to Dr. Tom Frank for the use of his noise tapes.

This work was done at the Applied Research Laboratory of The Pennsylvania State University under contract with the Naval Sea Systems Command. Their support is gratefully acknowledged.

CHAPTER 1

INTRODUCTION

1.1 Statement of the Problem

The availability of inexpensive synthetic speech has improved markedly in recent years. The technology advances in integrated circuitry have laid the basis for a speech synthesizer on a chip. Pertinent applications have increased greatly due to reduced costs. Previous research has typically focused upon classifying the differences between natural and synthetic speech sources (Nye and Gaitenby, 1974; Clark, 1983 and Clark et al., 1985) or the effect of different speech parameters on intelligibility (Hart and Simpson, 1976; Simpson, 1976; Slowiaczek and Nusbaum, 1985; Schwab et al., 1985; Clark et al., 1985 and Simpson and Frost, 1984). Very little work has been done to evaluate the differences attributed to the use of dissimilar technology to synthesize speech (Keeler et al., 1976).

There are two major methods used to produce low cost text-to-speech synthetic speech: linear predictive coding (LPC) and formant synthesis (FS). The former models the vocal tract as at least ten equal cross sections while the latter reproduces the relationships between the formants. Currently there are no criteria for determining which method is most desirable in a given environment. This research was carried out to better define the differences between LPC and FS. An experimental approach to this problem is outlined, and the results of the experiment using trained University students are presented.

1.2 Importance of the Problem

There is no situation which better illustrates the need to evaluate synthesizers better than the application of synthetic speech sources to aid the visually and communicatively disabled. An understandable synthesizer can be used to outfit a computer so that it can be used by a blind person. A vocally disabled person can communicate with the use of a programmable unit which can speak previously entered messages. There are also many uses in industry that could be of importance, such as its use as a plant warning system (Dalton et al., 1983) or in the cockpit of an airplane to warn the pilot of a problem without forcing him to divert his attention (Simpson, 1976 and Hart and Simpson, 1976). More uses for synthetic speech are inevitable, once it is proven that flexible inexpensive synthesizers can be used effectively to communicate.

1.3 Research Objectives

The main goal of this thesis is to provide some criteria for classifying the effectiveness of the two major inexpensive synthetic speech technologies in a realistic communications environment. Since most of the applications of inexpensive synthetic speech would be with trained listeners, the experiment was carried out with a trained subject pool. Also, multi-talker noise was used because it provides a more realistic approximate of typical communicative environments. Finally any application of synthetic speech would involve conveying a meaningful message and thus contextual test items were used.

Chapter 2

NATURAL AND SYNTHETIC SPEECH COMPARISONS

2.1 General

This chapter reviews the process of synthesizing speech and the unique problems associated with evaluating this process. The need to evaluate synthetic speech is a result of its growing use in many fields with applications ranging from consumer products, aids for the visually impaired, industrial uses like plant production warning systems and its value to the vocally disabled to provide an effective means of communication. The varied uses and listening environments pose many questions relating intelligibility to purpose. What type of synthesizer is favorable in the presence of a certain type of noise? Is the processing task complex enough to hinder other related tasks? Would a synthesizer be effective with naive listeners, or do listeners need to be trained? These and other questions are basic to any communication system and because of the unique features of synthetic speech the answers cannot be found in the literature for natural speech but rather in the relatively new research on synthetic speech.

2.2 Methods of Synthesizing Speech

The types of speech synthesizers currently on the market provide a plethora of choices to those who wish to design an effective communication system. Therefore, a brief review of these methods is in order.

Natural speech has traditionally been analyzed on the phonetic level. This approach provides an advantage when

synthesizing speech by rule because the English language, with its thousands of words, can be reduced to a handful of basic sounds. Unfortunately there are many problems which arise when attempting to pronounce a word phonemically. First, unless the text is keyed into the processing system as phonemes rather than unmodified text, a large algorithm is required to determine which phoneme is the correct one. The many anomalies of spelling serve to illustrate the complexity of such an algorithm. Second, given a good algorithm, there is still a need to distinguish which allophone of a certain phoneme is appropriate. For example, the /t/ in tip and pit demonstrate that coarticulation is an important acoustic parameter in the intelligibility of speech. Finally, given a system with both the knowledge of syntax and grammar such that it is capable of identifying the correct phoneme and variation thereof, there is still a need to specify the acoustic properties of the transitions between phonemes or the resulting speech has robot like qualities and very little of the prosody of natural speech. Different speech synthesizers approach the concatenation of sounds in different ways and not surprisingly the price of a system is an indication of the sophistication of the methods used.

The above problems can be solved at the expense of flexibility and cost by utilizing a system based on synthesis by analysis. In this scheme the device first must analyze natural speech and then store the important parameters to be replayed as required. The result is a system which is qualitatively better than all but the most sophisticated synthesis by rule devices, but

a system that can only store as much speech as it can fit into its memory. The memory consideration limits applications considerably because a modest data rate of 2500 bits/second would fill up 4K of memory in about 15 seconds. Alternately, synthesis by rule systems have unlimited text-to-speech capability. Ultimately the application will dictate which system is appropriate since the size and variability of the message set control the choice of synthesis by rule or by analysis.

Another consideration is how price reflects both quality and intelligibility. There is a direct relationship between a listeners preference and the subsequent intelligibility of some of the text-to-speech systems on the market and thus an untrained listeners subjective judgements are important (Logan and Pisoni, 1986). A study using the extremely high quality MITALK text-to-speech synthesizer by Pisoni and Hunnicutt (1980) proved that for untrained listeners contextual synthetic speech can approach the intelligibility of natural speech. There is also some evidence to suggest that there are at least three distinct classes of speech with respect to segmental intelligibility and sentence verification speed (Green and Logan, 1986; Manous et al., 1986). These classes are natural speech, high quality synthetic speech and low to moderate quality synthetic speech.

The technology which is employed to realize these text-to-speech systems is based upon techniques which are quite straightforward, although expanded upon in some of the more expensive synthesizers. These are formant synthesis (FS) and linear predictive coding (LPC). LPC is a mathematical model of the vocal

tract. Typically the model comprises a minimum of ten equal length cross sections of varying diameter. Using an analysis of an actual speech wave, the reflection coefficients for each sections are calculated and thus speech can be generated by varying the diameter of each cross section relative to the actual movement of the vocal tract during the production of the reproduced speech sound. Typically the coefficients generated by this mathematical model are expanded through an all pole filter. FS approaches the problem by assuming that most of the information needed to understand speech is contained within the first three formants of a phoneme (some systems use five formants). Two sources are expanded through a bank of filters representing both the resonances and the antiresonances(formants) of the vocal tract. A noise source is used to produce sounds like fricatives while the second which pulses like the glottis, produces sounds such as vowels. Although these two methods are not the only ones used, they are employed in the majority of cases, most notably in less expensive text-to-speech systems.

2.3 Review of Test Material

While examining any communication system, one must always be wary of the test material used to evaluate it. A poorly constructed test may emphasize elements of communication, such as word frequency, which are not of interest. The test material for any experiment must therefore be selected very carefully.

Intelligibility has traditionally been approached on a phonetic level. To this end, phonetic tests such as the Rhyme test or the Harvard phonetically balanced word lists have been

developed and tested extensively. Lehiste and Peterson, 1959; Fairbanks, 1958; House et al., 1964 and Pollack, 1958 are just a sampling of this literature. These tests present the listener with lists of isolated words, typically monosyllabic forms such as consonant-vowel-consonant, which are scored at both the word and phonetic level. This data yields information about phonetic confusion and relative intelligibility of each phoneme. These types of tests do not closely model the complexity of the actual listening task. Synthetic speech is most useful in applications where it conveys a meaningful message and thus factors such as prosody and context cannot be ignored.

Kalikow et al. (1977) addressed this problem by developing a set of sentences which he called the Speech Perception in Noise(SPIN) test. In this research the major objective was "to produce a measure that would assess utilization of linguistic-situational information of speech in comparison with utilization of acoustic phonetic information" (Kalikow et al., 1977, p. 1339). They accomplished this by developing 10 forms of 50 sentences, half of the form contextual and half non-contextual. The sentences were designed to elicit a one word response, the last word in the sentence. The sentences were all five to eight words, each word six to eight syllables. The key word to be identified was a monosyllabic noun with a frequency count of 5 to 150 per million words. The predictability of the key words in these contextual sentences was determined by presenting those sentences to subjects without the key word and asking the subjects to write down the word most likely to occur. Once developed, the forms

were tested for equivalence and phonetic balance.

This type of test provides the material best suited for the evaluation of a synthetic communication system. The use of contextual material provides a more realistic approximate of the types of processing required to understand speech. Further research has been performed to validate the SPIN test. Since the test was designed such that the administering of any one form could be compared with that of another, form equivalence is important. Morgan et al. (1981) examined all 10 forms and found equivalence between 7 of them. This is not surprising since two of the three forms they found unsuitable were also deleted by Kalikow et al. (1977) in their final list of sentences. Another feature of the SPIN test is its attempt to determine how well a listener uses contextual information. The difference score is the contextual item score minus the non-contextual item score. It ideally yields some information about the listeners use of context. Owen (1981) studied the relationship of the difference score to syntactic skills, semantic skills, I.Q., hearing loss and signal to noise ratio. He found the difference score related mostly to the subjects hearing and the signal to noise ratio and thus not effective as a measure of a subject's use of contextual information. Given these facts it seems prudent to temper the use of the SPIN test with a knowledge of its shortcomings.

Bilger (1985) attempted to refine the SPIN test as a test instrument by carefully sifting through the test items. Using the same procedure as Kalikow et al. (1977), he revised the SPIN test into eight equivalent forms which he renamed the revised SPIN

test. Like the SPIN test, the revised SPIN test closely models everyday communication and is ideal for use with synthetic speech. To date, neither the SPIN test nor the revised SPIN test have been used with synthetic speech but other more traditional tests have yielded a wealth of information about the unique characteristics of synthetic speech. A review of this research and the important results obtained follows.

2.4 Unique Features of Synthetic Speech

For many years speech scientists have used forms of synthetic speech to gain a greater understanding of natural speech. In the 1950's the Haskins Laboratory used hand painted spectrograms, reproduced by modulated light beams, to study such things as transitional cues for consonants (Delattre et al., 1955). The extreme ease with which the spectra of synthetic speech can be altered make it an excellent tool for the investigation of the acoustic properties of speech. This same flexibility poses many questions about the most effective use of synthetic speech.

A comparison of natural and synthetic speech at the phonetic level will serve as a starting point towards an appreciation of the unique problems associated with evaluating synthetic speech. Vowels and the attribute of voicing in general show performance approaching that of natural speech (Clark, 1983 and Keeler et al., 1976). Consonants do not fair so well. While synthetic consonants exhibited a sharp steady decline from favorable to unfavorable noise conditions, natural consonants were resistant to masking down to 0 dB signal to noise ratio and then decreased in intelligibility (Clark, 1983). This could prove important to the

application of synthetic speech in a noisy environment. Different studies have documented different error rates for identical consonants and thus the particular characteristic of the synthesizer used in the study determines how well each consonant performs. The one common factor of all synthesizers is the consistently poorer performance of the synthetic consonants versus either natural consonants and vowels or synthetic vowels.

The prosodic features of speech provide acoustic cues which extend over greater than phonetic length. These are intonation, stress and rhythm. The effect of pitch contour has been studied by assuming that if the speech rate of a sentence is increased, the listener is forced to rely more on the pitch contour. Using simple meaningful and nonmeaningful sentences with and without pitch contours at different speech rates, the slower meaningful sentences were significantly better understood than the faster meaningless sentences, regardless of the pitch contour (Slowiaczek and Nusbaum, 1985). The use of more complex sentences using the same criteria produced a slight increase in intelligibility due to the addition of pitch contour, but the additional factors of processing effort, memory and speech rate serve to confound the significance of the finding. The relationship between pitch and intelligibility can also be examined by varying the fundamental frequency for, although pitch is a subjective measure of speech, it depends heavily upon frequency. The use of three separate fundamental frequencies (70,90 and 120 Hz) produces no significant change in intelligibility (Simpson and Frost, 1984).

Speech rate is easily adjusted on most synthesizers and the

potential to produce a more efficient message without a corresponding decrease in intelligibility has been studied. Increasing speech rates from 128 to 156 to 178 words per minute did not produce a decrease in intelligibility and did produce faser response times (Simpson and Frost, 1984). The need for messages which are quickly and easily understood is especially important to those who wish to use synthetic speech as cockpit warning messages. In this context the effects of linguistic redundancy on processing effort and response time were examined. The use of sentences as opposed to two keyword format produced more intelligible messages and faster response times. The two keyword format was also associated with increased mental demands (Hart and Simpson, 1976 and Simpson, 1976). Undoubtedly, the unnaturalness of the speech makes it difficult to understand a short phrase as opposed to a sentence in which the subject becomes accustomed to the voice. There are also extra clues to the message in the sentence format.

One important effect which must be considered when studying synthetic speech is the learning effect since the majority of research to date has documented some type of learning effect. The performance of untrained listeners identifying low quality synthetic speech steadily increased over a period of ten days with no indication that performance had peaked (Schwab et al., 1985). Even the intelligibility of the high quality MITALK synthesizer, a device which approaches the performance of natural speech, improves slightly with increased exposure (Pisoni and Hunnicutt, 1980). Identical studies with natural speech do not yield such a

marked learning effect (Schwab et al., 1985). The one obvious cause of this is the less consistent acoustic cues of phoneme-based systems, which attempt to paste phonemes together with no regard for the surrounding phonemes. The variance of natural versus synthetic speech serves to illustrate this point. While the variance of listeners intelligibility scores for natural speech even out to a constant upon successive presentations of test material, the variance for synthetic speech does not show any trend for improvement (Clark et al., 1985).

The presence of a dramatic learning effect highlights the need to separate learning improvements from test results. The lack of consistent acoustic cues which causes learning effects illustrates the importance of choosing the correct test material, since ultimately a communication system will be judged by its performance relative to well-chosen test material.

CHAPTER 3

INTELLIGIBILITY MEASUREMENTS OF SYNTHETIC SPEECH

3.1 Introduction

This chapter outlines the design of an intelligibility experiment comparing two speech synthesizers. The information in the preceding chapter was considered along with the results of two pilot studies to develop an effective approach to the problem. An explanation of the experimental objectives begins this chapter. A description of the equipment used to carry out these objectives follows with an outline of all calibration techniques used. Finally, a description of the subjects used and the safeguards employed to ensure their safety will be outlined.

3.2 Experimental Objectives

The main point of this experiment is to compare the predominant technologies used to synthesize speech. If the term predominant was not somehow qualified this would be a formidable task indeed because there are hundreds of different approaches to the same problem. The qualification made here is to only consider those methods which are low in cost and thus can be used in applications where high cost would be prohibitive, such as its use as an aid to the visually impaired or vocally disabled. An inexpensive synthesizer which is highly intelligible would also pave the way for more creative uses of synthetic speech like consumer products and industry related functions. The hidden cost of editing a usable vocabulary is also to be considered and thus a synthesis by analysis system, although of high quality,

is not flexible enough to be used in the above-mentioned application. Thus, only text-to-speech systems will be considered.

Another restriction placed upon this study is that the results be derived in a realistic but repeatable setting. To achieve this, the design of the experiment was approached with consideration to an actual application of speech synthesis. This affected decisions of the use of noise, test material and subject training. Ultimately the study should be used not only for a better understanding of the intelligibility of synthetic speech, but also be useful to those implementing a communication system. The experimental environment should thus model the actual listening task as closely as is practical. The experimental objective is not just a comparison of the technologies, but rather an intelligibility comparison of the technologies with a structured listening task which very closely models a realistic application.

3.3 Choosing Representative Synthesizers

The task of choosing representative speech synthesizers to be used in the experiment required the identification of the predominant technologies used to produce low cost text-to-speech synthetic speech. These methods are formant synthesis(FS) and linear predictive coding(LPC), both of which are outlined in section 2.2. The actual choice of synthesizers was done using the results of the following pilot study.

The pilot study compared five different speech processing systems as outlined in table 3.1. An inexpensive modified delta modulation digital recording(DR) scheme was investigated in the

pilot study. It was hoped that the digitizer might provide an upper limit for synthetic speech intelligibility and provide results comparative to a similar study (Dalton, 1983).

Table 3.1 Speech Processors Used in Pilot Study

Device	Type	Description	Price($)
Echo (GP)	LPC	Stand alone peripheral	180*
Echo (PC)	LPC	IBM PC circuit board	100
Votrax (Type'n talk)	FS	Stand alone peripheral	240*
Intex (Talker)	FS	Stand alone peripheral	360
Mimic digitizer	DR	Stand alone peripheral	170

*slected for study

Two sets of fifty noncontextual sentences were presented to ten untrained subjects. Each subject listened to twenty sentences from each speech system and was asked to identify the last word in the sentence. A 5x5 Latin square design was employed to reduce order of presentation effects. The results are presented in table 3.2. Semivowels are not averaged into either the consonants or the vowels. As was noted earlier in section 2.4 the intelligibility of a particular class of consonants is a function of the synthesizer used. Even though both the Votrax and the Intex system are FS technology, they differ greatly in percent nasals correctly identified. The low word scores recorded are a result of using

untrained subjects who have not had enough exposure to the systems to attune themselves to the peculiar acoustic characteristics.

Table 3.2 Percent Correct Scores for Five Speech Systems

Device	Word	Nasals	Plosives	Fricatives	Vowels
Echo GP(LPC)	31	61	52	73	72
Echo PC(LPC)	29	61	44	76	64
Votrax(FS)	40	64	62	80	72
Intex(FS)	43	47	66	87	64
Mimic(DR)	17	53	43	42	49

A three way statistical test comparing the methods was computed and FS found more intelligible than LPC which in turn was more understandable than the digital recorder (α =0.1). At this point the digital recorder was removed from the experiment due to the apparent failure to perform its intended purpose. It was not possible to statistically compare the Echo GP to the Echo PC and the Votrax to the Intex due to the small differences in mean scores and thus other criteria was employed. Since the performance of the two synthesizers in each class was comparable, the Echo GP and the Votrax were chosen due to their equivalent cost. The Echo GP was also qualitatively superior to the Echo PC, the PC sounding more mechanical and harsh.

The groundwork for the design of an intelligibility experiment was thus complete. Two comparable priced synthesizers using two different technologies to implement a phoneme based text-to-speech system were chosen. An inspection of a similar study will serve to illustrate the important design parameters.

3.4 A Comparative Study

Dalton et al. (1983) were faced with a similar task in an industrial setting. The problem was to determine the most effective system to be used as an alarm system for a batch process plant. The study compared a phoneme based text-to-speech synthesizer and a 32 kbit per second digitizer. The following hypothesis were tested:

1. No difference in intelligibility.
2. No difference between keywords and sentences.
3. No difference due to familiarity.
4. No difference in operator performance.

The subjects were split into two groups of eight, four hearing phrases and four hearing sentences of plant messages. The groups were then tested using new messages couched in phraseology used by the operators. The messages were first learned and then presented in the presence of plant noise. The results show that the digitizer is initially more intelligible than the Votrax synthesizer, but once accustomed to the Votrax the performance of the two was equal. This highlights the importance of separating learning effects from performance data with either a balanced experimental design or with trained subjects. Another major finding study was that listeners of synthetic speech were unable to use the information from one message set to help in the understanding of another. Dalton reached this conclusion by comparing the results of the initial presentation of two different message sets and noting that the synthesizer decreased in performance by three percent while the digitizer increased by four

percent. This general conclusion of the nature of the intelligibility of two types of speech processors is seemingly confounded by learning effects and differing complexity of the two message sets. This serves to illustrate the need for clearly defined variables because, although this was an excellent industrial study which accomplished its objective of choosing the best warning system for a batch processing plant, it is not easily transferable to more specific environments.

The need for a more balanced message set is another point of interest. The 3 and 4 percent differences noted above could easily be due to one class of consonant which is uncharacteristically stressed. The results in table 3.2, from the initial pilot study, documenting the differences of percent correct identification of consonants for different synthesizers stresses the need for carefully constructed test items. The contextual content of each message also must be controlled, as in this study the differing complexity of the message sets prove to make comparisons between the results of each message set untenable.

Dalton's study as a whole, highlights all of the important considerations in an investigation of synthetic speech. Most importantly it espouses an approach of clearly defining the objectives to the point where they are directly assessable by the experimental results.

3.5 Subject Training

The reduction of learning effects in the test results has been shown to be necessary and therefore a trained subject pool

was used. Schwab et al. (1985) studied this effect using varied material to force the subject to learn the acoustic cues present and not the specific test used. This same approach was utilized to train the subjects for this experiment with certain modifications relevant to the purpose of actually training the subject as opposed to measuring the performance of each subject at each presentation. Each subject was trained on only one synthesizer to avoid the effects of learning one synthesizer transfering to the second synthesizer. A second pilot study was conducted to further examine learning effects and to determine an estimate of the number of training sessions needed to avoid dramatic learning effect tainting the test results. Two subjects were exposed to the Votrax synthesizer for four consecutive Fridays. The test material for each session was 25 contextual and 25 noncontextual sentences from the SPIN test (Kalikow et al., 1977) presented without noise. The percent word scores are presented in table 3.3.

Table 3.3 Results From Learning Effects Pilot Study

Session #	1	2	3	4
Low Context	40	52	50	66
High Context	74	78	66	88

Based upon the results in table 3.3 four training sessions were thought to be sufficient. Although it is not realistic to assume learning effects will stop after four training sessions (Scwab et al., 1985), a balanced test design should be able to

factor out the greatly reduced learning effects. A detailed description of each of the five sections of one of the four training days follows. Each training day was identical in structure, although all of the items used were different.

Section #1: Harvard sentences (A). This set of material consisted of fifty phonetically balanced contextual sentences (IEEE, 1969). The syntactic structure varies around five key words. The task in this section was to identify one of the key words randomly chosen. A list of the sentences excluding the one word was supplied the subject and a set of instructions preceding this section asked the subject to identify all words aurally as well as visually.

Section #2: Prose passage. This second set of material consisted of a prose passage selected from a basic English text. The subject was supplied with the text and instructed to read along with the synthesizer to become accustomed to the unique features of the synthetic voice. This section was included to attune the listeners to the unique prosodic features of the synthesizer and to force the listener to abandon the normal method of discerning natural speech sounds.

Section #3: Haskins sentences. The third set of materials consisted of fifty syntactically normal but semantically anomalous sentences developed at Haskins Laboratories (Nye and Gaitenby, 1974). Each sentence

had four key monosyllabic high frequency of occurrence words in a structure of the order: (The (adjective) (noun) (verb, past tense) the (noun).) The subjects task in this section was to identify one of the key words randomly chosen. A list of the sentences excluding the key word was supplied the subject and a set of instructions preceding this section asked the subject to identify all words aurally as well as visually. This section was included to force the listener to identify words in a realistic sentence structure without word identification cues based on the meaning of the utterance.

Section #4: Harvard Sentences (B). The fourth set of materials consisted of fifty phonetically balanced sentences as in part one. In this section the subjects were not supplied with any written clues but were instructed that all of the sentences were meaningful. The subjects were then asked to identify the last word in each sentence. This section was very similar to the task during the actual testing on the fifth day and thus served to prepare the subject. It also tested the intelligibility of the synthesizer progressively for meaningful phrases.

Section #5: MRT lists. The final section consisted of 12 sets of six monosyllabic words taken from the modified rhyme test (House et al., 1965).

This test was used to determine the intelligibility of consonants and to identify possible areas to be investigated in the final analysis.

A sample session of the training days is presented in appendix A. The purpose of the four training sessions was varied. The exposure of the subjects to the synthetic speech should reduce the learning effects significantly. It was also hoped that the variance would be reduced allowing for a smaller subject pool for the final experiment. This is also a realistic approach since any application of synthetic speech would require similar training methods.

Another approach to the problem of learning effects is to completely randomize the experiment into eleven blocks of 8x8 Graeco-Latin squares in which each subject hears each synthesizer at each noise level. The one great advantage of this design is that it eliminates the difference due to subjects. The one great disadvantage is that the resulting experiment does not model an actual application of synthetic speech in any way. This approach was not felt appropriate for this type of study.

3.6 Design of a Balanced Test

The design of the actual test began by choosing the revised SPIN test as the test items (Bilger, 1985). The choice was dictated by the extensive research done on the original SPIN test (Kalikow et al., 1977; Morgan et al., 1981 and Owen, 1981) and the eventual refinement of that test by Bilger (1985). These tests are fully described in section 2.3. The revised SPIN test incorporates the features stressed in section 3.3. Briefly reviewed, it has

controlled word predictability, phonetic balance among forms and contextual phrases. The revised SPIN test also has noncontextual phrases but these were not used due to potential confusion between context and noncontext test results. This alteration of the revised SPIN test, i.e. using the contextual sentences from one form with the contextual sentences from its pair form, raises the question of equivalence among the altered forms. Phonetically the new forms remain balanced because the pair forms contain the same keywords in context and out of context. The equivalence of the forms in general is not as crucial since different presentations of the different forms will not be compared against each other as was originally intended for the revised SPIN test.

The choice of noise was influenced by the desire to compare results with previous studies using the SPIN test with natural speech and the desire to model an actual applications environment. Previous studies used babble at signal to noise ratios varying from -5 dB to +10 dB (Kalikow et al., 1977 and Owen, 1981). Babble is appropriate because it provides a confusion element which approximates speech in a crowded room. The specific multi-talker babble used in this experiment consisted of the simultaneous mixing of 8 male and 12 female voices such that none of the voices could be understood. A more detailed description of the noise is contained in the appendix of Frank and Craig (1984). Upon inspection of previous studies using natural speech and the SPIN test with the knowledge that the degradation of synthetic speech is more rapid in the presence of noise (Clark, 1983), the signal to noise ratios of 0, +5 and +10 dB were chosen.

Since three noise levels were used with 200 contextual test items, the revised SPIN test forms were split into three sets of 50 phonetically balanced items with the remaining 50 items divided up to yield three sets of 66 item forms. The subsequent analysis of the collected data will necessarily reflect the distinction between the 50 core items and the 16 added items since no attempt was made to balance the three forms with respect to the added items.

To determine the needed sample size, estimates of the variance and expected difference between the population means were needed. An estimate of 10.25% for the variance was obtained from the original pilot study outlined in section 3.3. From this same study the difference in means was estimated at 9% (table 3.2). The estimated sample size for $\alpha = .01$ and $\beta = .1$ is thus six assuming a student t-distribution for the data (Ostle, 1963).

A block design was chosen for the presentation of the material to control variability due to presentation order, subject differences and form equivalence. The experiment was divided into four blocks of a 3x3 Graeco Latin square. A sample block is presented in table 3.4.

Table 3.4 Graeco Latin Square

A1	B2	C3
B3	C1	A2
C2	A3	B1

This block would be presented to 3 subjects on one synthesizer.

The Latin letters represent the three SPIN test forms used, the numbers represent the three noise levels. The analysis of a 3x3 Graeco Latin square does reveal the relative sources of variability but one of its shortcomings is that the degrees of freedom for error is zero and thus no statistical inferences can be drawn about the blocked variables. This means that analysis of variance cannot be used for analysis of the data nor can any formal test be performed to test for interactions between the blocked variables.

3.7 Equipment and Equipment Calibration

The two synthesizers used in this experiment have some variable features and thus a description of the measurable parameters is needed. Both synthesizers were set at a speech rate of 125 words per minute. Because speech rate was linked to pitch on both of the synthesizers, they were not set to identical fundamental frequencies. A frequency analysis of both synthesizers sustaining the vowel /e/ (weed) yielded fundamental frequencies of 116 Hz for the Echo versus 128 Hz for the Votrax. This difference should not be significant (Slowiaczek and Nusbaum, 1985 and Simpson and Frost, 1984).

Both synthesizers were controlled by an AT&T model 6300 personal computer via the RS-232 interface. The signals were fed directly to a Crown 700 tape recorder. All taped material was edited to correct mispronunciations by the synthesizers. During the training sessions the recorded synthesized voice was played back through an amplifier to a mixing box which fed the signal to both ears of a pair of Pioneer SE-550 circumaural earphones. Since

it was not possible to directly calibrate the circumaural headphones a loudness balance technique was used. The procedure involved a pink noise source played through a loudspeaker in an Industrial Acoustics Company model 40 soundproof room. The sound pressure level(SPL) was measured at the ear facing the speaker by a calibrated B&K type 2209 sound level meter. The other ear was covered by the uncalibrated phone and three people with known normal hearing(500, 1K, 2K, 4K, 8k at 20 dB or less) were instructed to match the level of the loudspeaker to that of the phone for both ears with the phone level starting both high and low and then approaching the calibrated speaker. The results from the three people were then averaged to obtain a calibration voltage. The voltage was measured at the output of the mixing box by a Ballantine model 303-1 slow averaging voltmeter. The calibration is not exactly accurate due to variability in distance between the ear and the phone with different listeners but the calibration does not need to be precise for the following reasons:

a. The loudness of the signals is sufficiently high that they are well above the auditory threshold where the ear is linear.

b. The grossest errors will occur at very high and very low frequencies, limitations which do not affect the intelligibility of synthetic speech.

The training sessions were conducted in a quiet classroom after working hours or in a soundproof booth. All test sessions were conducted in the sound proof room. The mixing of the noise

and signal was performed with the variable attenuators as depicted in figure 3.1. Initially both the signal and the noise were calibrated to 80 dBC (the C weighted scale was used as opposed to linear measurement because of extraneous very high frequencies present in the sound booth.). The variable attenuators were then used to control the signal to noise ratios and to maintain an overall SPL of 80 dBC. The calibration of the speech signal was not easy due to its non-periodicity and dramatic transients. Therefore, the phrase "worker-enter," containing both dramatic transients and dips in energy, was used to calibrate the speech.

3.8 Subject Information

Prior to the implementation of any experimentation, approval was obtained from the Office for the Protection of Human Subjects so as to adhere to the University's policies and institutional assurance with the United States Department of HHS regarding the use of human subjects. A detailed prospectus of the procedures regarding the experiment was supplied the office, as well as those procedures used to ensure subject safety. Subjects were instructed in the details of the study in written form and verbally. Questions regarding their part in the experiment were encouraged. All relevant data pertaining to subject safety is included in appendix B.

Overall 13 university students ranging in age from 18 to 30 years of age were used. One subject's data was removed from the experiment due to a calibration problem discovered after the test session. The 12 remaining subjects included 4 males and 8 females. All subjects were screened by The Pennsylvania State

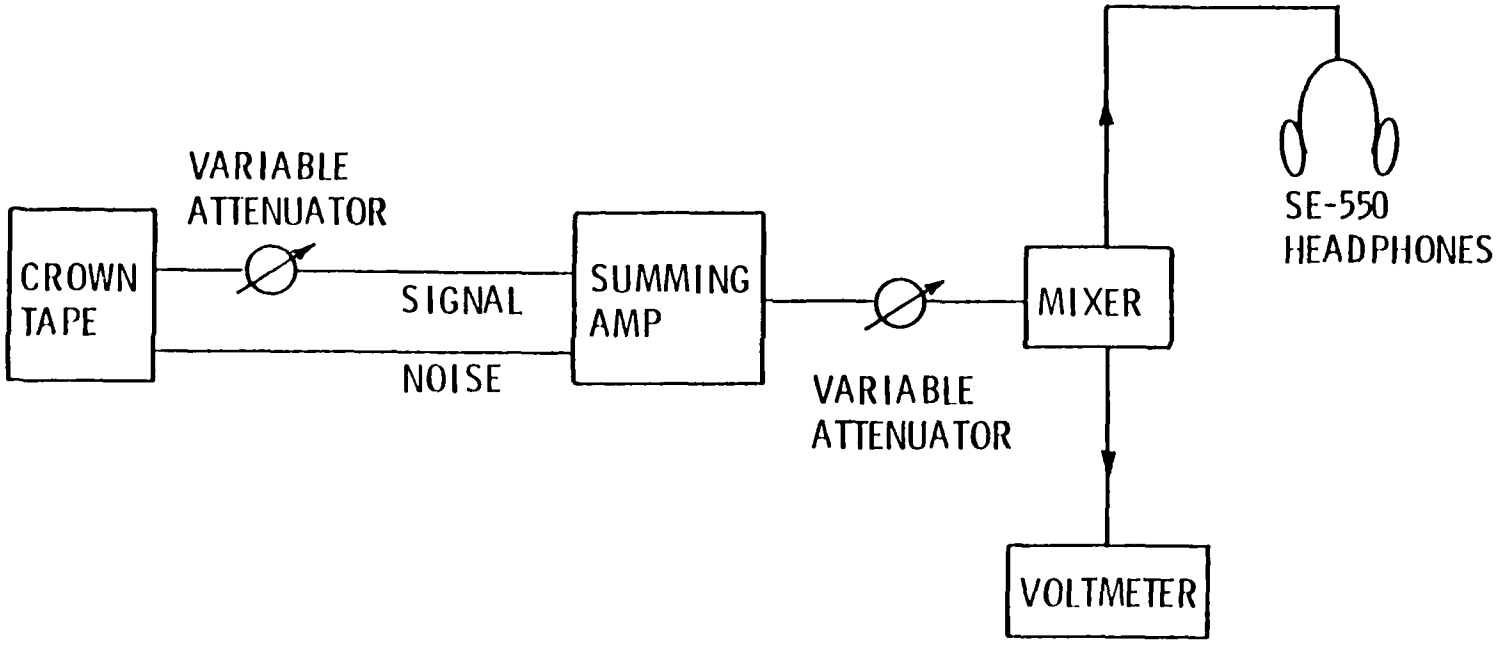


Figure 3.1 Block Diagram of Equipment for Test Sessions

University Speech and Hearing Clinic and were found free of any hearing defects (250, 500, 1K, 2K, 4K and 8K at 20 dB or less from the threshhold of hearing).

CHAPTER 4

RESULTS AND DISCUSSION

4.1 General

The results can be split into two sections, the training data and the test results. The training data was approached from the perspective of its purpose and therefore the progress of the subjects scores was assessed with a special emphasis placed upon the day to day variability. The test data was examined from a statistical perspective and inferences drawn about the two synthesizers. Finally, the results were examined in light of other research and then the possible causes for the results discussed.

4.2 Training Session Results

The purpose of the training session was to reduce the experimental error due to learning effects. It was also hoped that the 12 subjects would become increasingly attuned to the peculiarities of synthetic speech and thus decrease the variability due to subjects. The need for the training sessions is indicated by table 4.1. The use of written clues in the four training sessions seems to have been effective since a comparison of mean percent correct scores with Schwab et al. (1985) in figure 4.1 shows a more dramatic improvement over a shorter period of time for similar test material (Harvard(B) sentences). The plotted scores from this study are an average of both the Echo and Votrax synthesizers. This difference must be tempered with the difference in difficulty of the two tasks, the previous research requiring the subjects to identify five words in each sentence as opposed to simply aurally identifying all but the last word in the

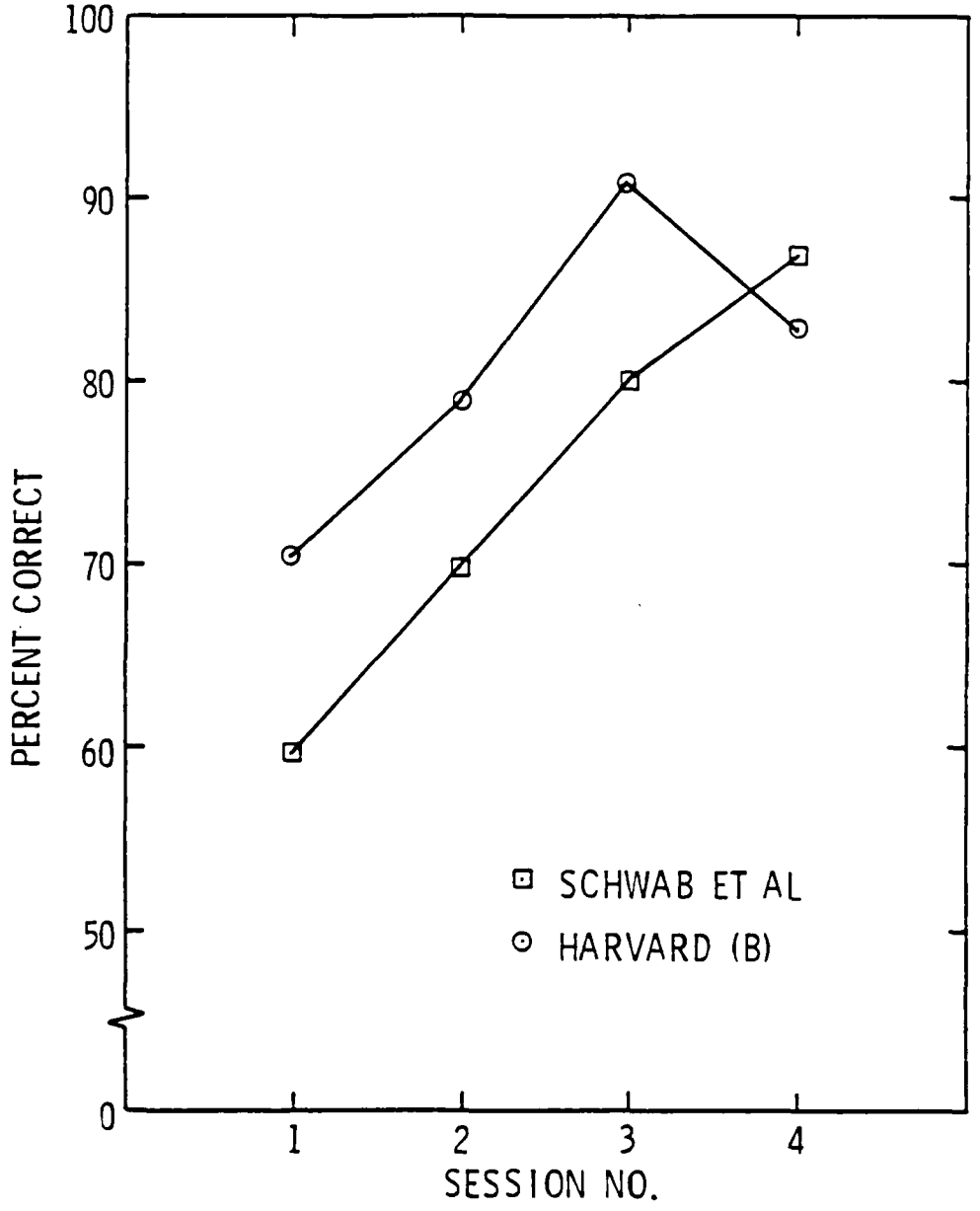


Figure 4.1 Training Sessions Comparison

sentence. It is obvious from the data in table 4.1 that no plateau of performance has been reached by the subjects but the dramatic improvements witnessed from day one to day three in all the sections, which could severely taint test results, need to be avoided.

Table 4.1 Training Session Mean Percent Correct

		Hvd(A)	Haskins	Hvd(B)	MRT
Echo	day 1	71.0	49.3	69.0	56.9
	day 2	83.0	64.6	79.0	70.8
	day 3	90.6	63.6	89.6	81.9
	day 4	90.0	69.3	85.3	79.1
Votrax	day 1	76.6	58.6	71.6	66.6
	day 2	84.0	70.3	79.3	79.2
	day 3	87.3	62.0	92.3	90.2
	day 4	87.6	67.0	81.0	86.1

Variability of percent correct scores from day to day is plotted in figures 4.2 thru 4.5. It is clear from these results that the variance shows no trend of becoming stable unlike that of natural speech (Clark et al., 1985). The hope that training the subjects would reduce the variability in the experiment between subjects was thus unfounded and the use of a conservative estimate for the variance when estimating sample size in synthetic speech studies is advisable. A statistical comparison of the two synthesizers during the training sessions was performed using the Wilcoxon rank sum test (Ott, L., 1984). The Harvard(A), Harvard(B) and Haskins sections were tested for each of the four days. All results excluding day one of the Haskins section displayed insufficient evidence to claim a difference in means (α=.0471).

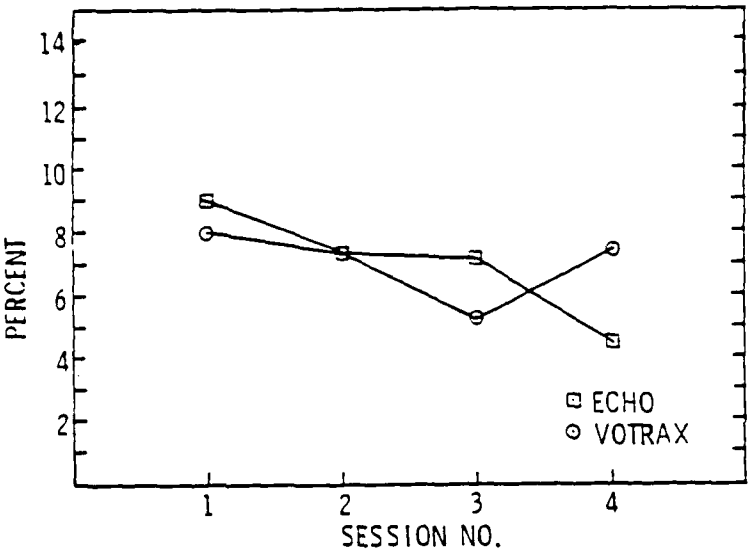


Figure 4.2 Harvard(A) Std. Dev. of % Correct

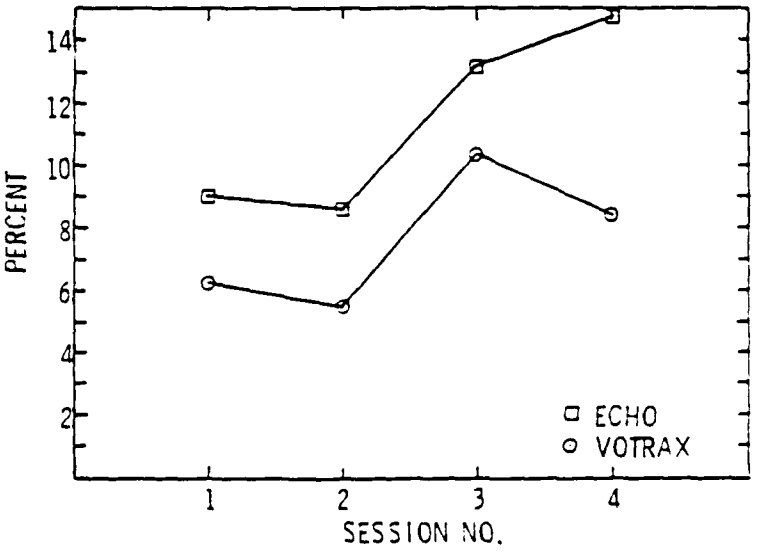


Figure 4.3 Haskins Std. Dev. of % Correct

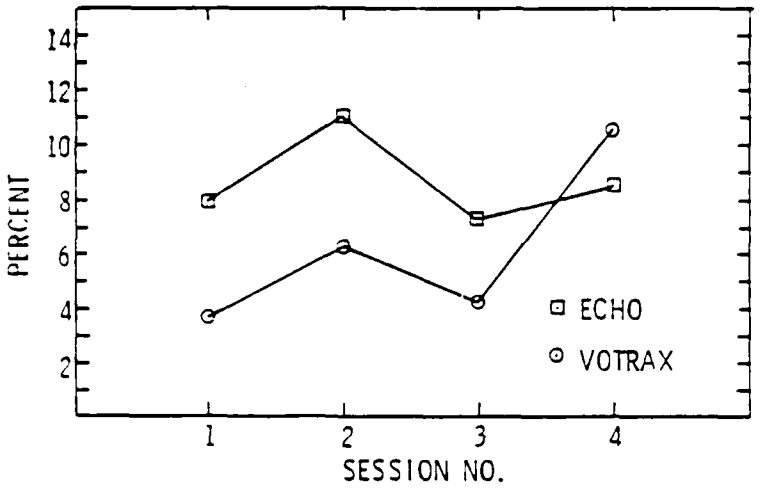


Figure 4.4 Harvard(B) Std. Dev. of % Correct

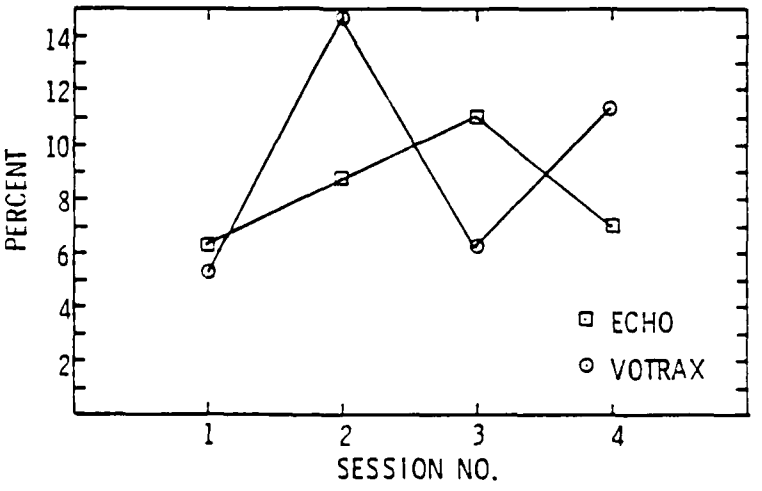


Figure 4.5 MRT Std. Dev. of % Correct

On day one of the Haskins section the mean of the Votrax synthesizer was significantly higher than that of the Echo ($p=.405$). This result agrees with the original pilot study which found that naive subjects understood noncontextual material presented without interfering noise better with the FS synthesizer that with an LPC. It is interesting to note that for contextual material there was no difference and that once accustomed to the synthesizers there was no difference in scores for noncontextual material.

4.3 Test Results

The test results are plotted in figure 4.6. The natural speech data was taken from Kalikow et al., 1977. The means and variances are presented in table 4.2.

An analysis of variance was performed on the four 3x3 Graeco-Latin squares (Montgomery, 1976). The sources of variability are presented in table 4.3. These values represent the percent of the total variability which is accounted for by the respective items, i.e., forms, order, subjects and noise. The

Table 4.2 Test Results in Percent

	Votrax		Echo	
SNR	Mean	St. Dev.	Mean	St. Dev.
0 dB	10.2	7.1	27.8	9.6
5 dB	18.5	12.3	39.0	12.7
10 dB	30.0	6.1	53.8	8.7

The difference in forms contribute very little and thus the questions raised earlier about the use of altered SPIN tests seems

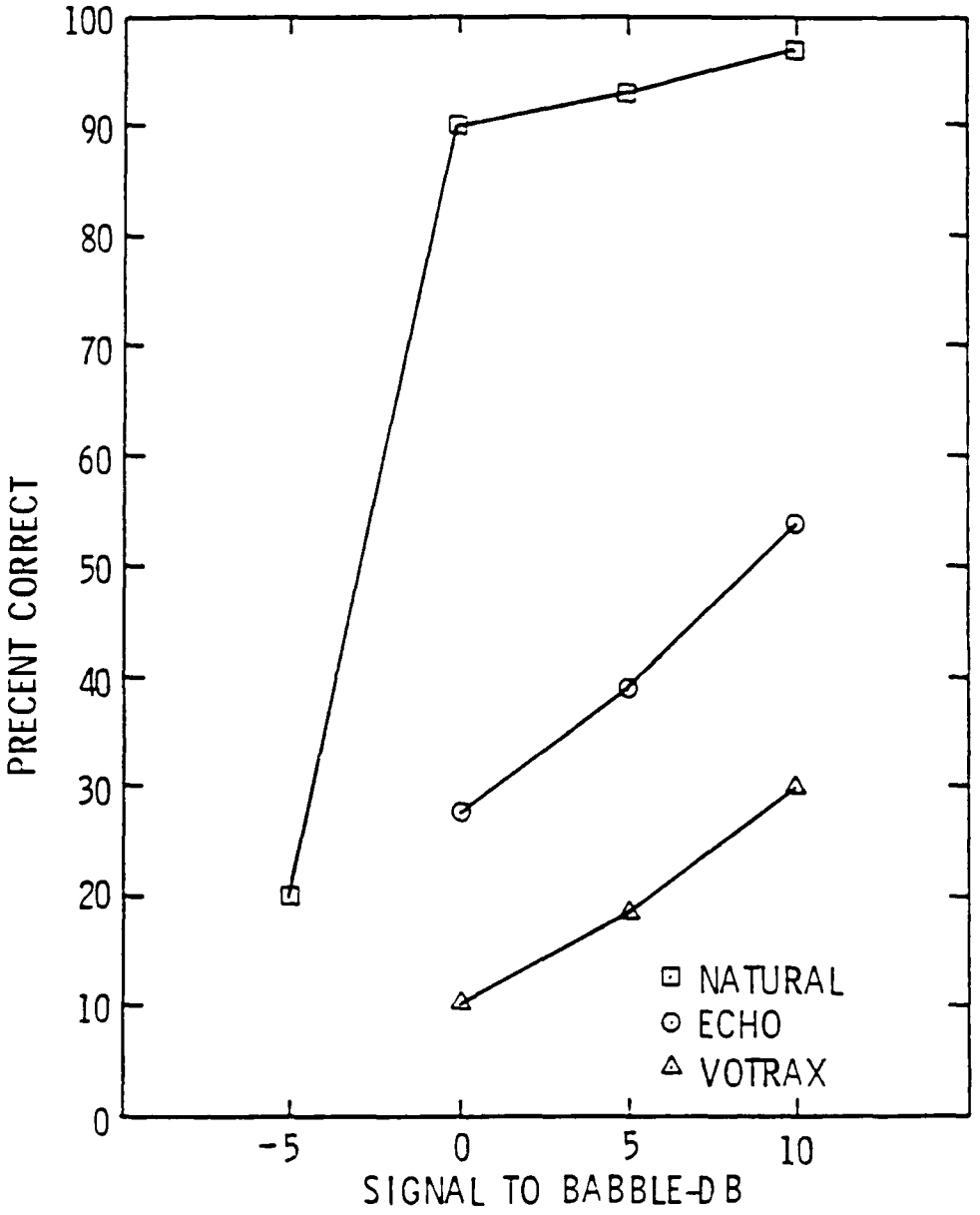


Figure 4.6 Test Results

to have had no effect upon the outcome of the experiment. Unfortunately the use of a 3x3 square does not lend itself to any statistical testing since the degrees of freedom for error is zero. Instead, the Wilcoxon signed rank sum test was performed on the data for the three noise levels (Ott, L., 1984). The Wilcoxon signed rank sum test was used because the data for each synthesizer are paired and also because the assumptions of a normally distributed data and

Table 4.3 Sources of Variability in Percent

	Echo	Votrax
Forms	4.3	4.7
Order	11.7	7.7
Subjects	22.0	34.9
Noise	61.5	53.2

equal variance are less stringent. This is an important consideration when sample sizes are small and data sets are very sensitive to outliers. The LPC synthesizer's mean percent word correct score was significantly more understandable than the FS at all three noise levels (α=0.016). This is a surprising result since the two synthesizers were comparable before the introduction of interfering noise. An investigation of the spectra of the noise and the two synthesizers may provide an explanation.

4.4 Discussion and Conclusions

A sonograph was used to investigate the acoustical properties of the two synthesizers. Figure 4.7 is spectrograph of the word /skirt/and the word /splash/ by both the synthesizers.

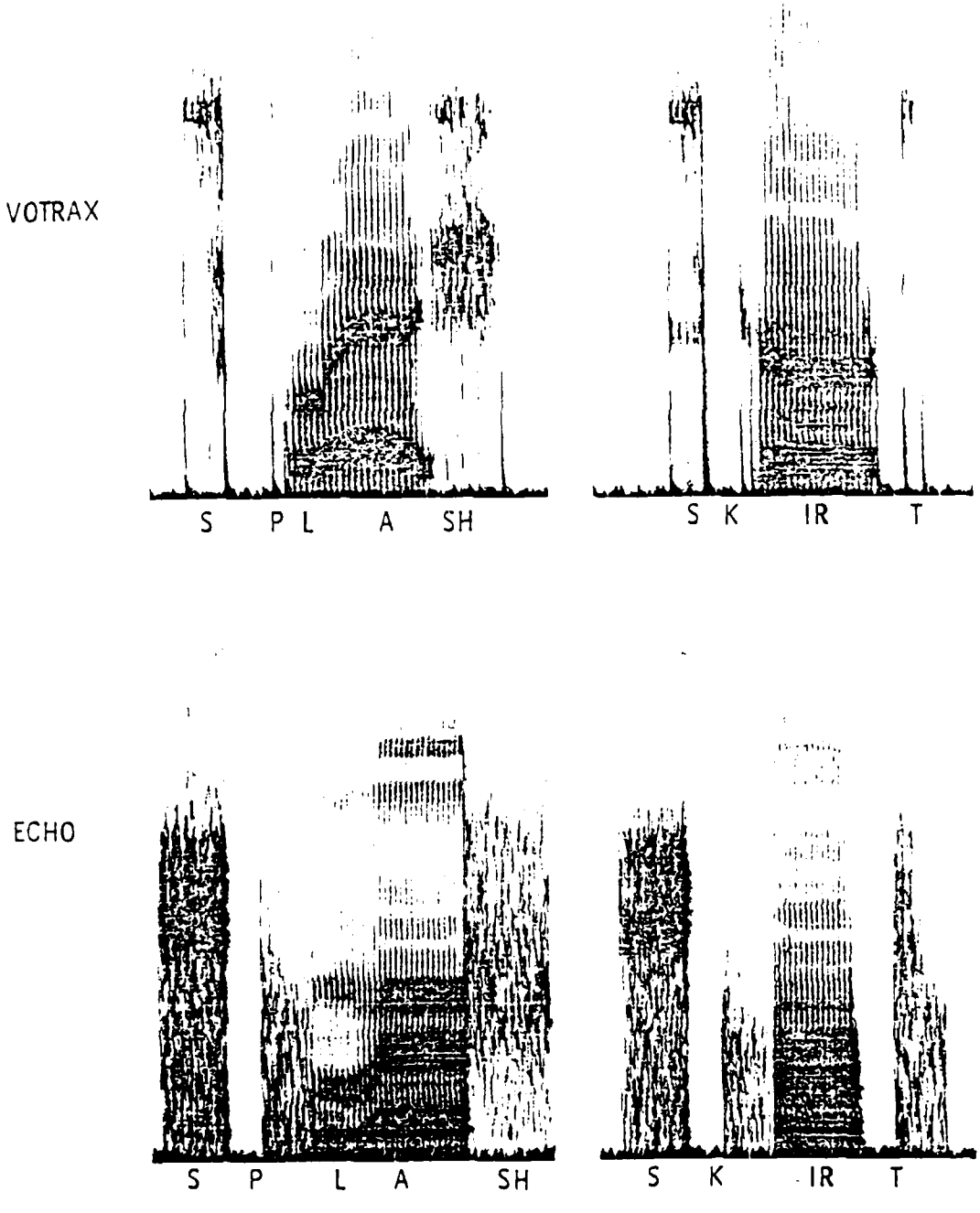


Figure 4.7 Sonograph Analysis

The deficiency of the FS consonants is immediately noticeable, in particular the small amount of energy to model the turbulent air flow of the fricative /s/ of both splash and skirt. The stops of the FS system are also produced with less robust qualities. Both synthesizers provide the brief gap in energy imitating the build of intraoral pressure, but the burst of energy which follows must be of sufficient intensity to be understood in the presence of noise. An analysis of the phoneme /s/ was undertaken with the test results for ten of the twelve subjects (one subjects data was lost and thus its pair subject had to be excluded also) to determine whether the lack of energy for this fricative noticed in the spectrograph had a significant effect on the test results. Of 101 total /s/'s presented at +10 dB, the Echo subjects correctly identified 93 while the Votrax subjects only identified 23. More importantly, of the 99 possible words which contained at least one /s/, the Votrax subjects identified 23 compared to the 70 identified by the Echo subjects. It is not hard to see the correlation between the phonemes with low energy in their spectra and the phonemes which were effectively masked by the noise. Interestingly, the Pilot study which was conducted with naive listeners without noise yielded a much different picture. The Echo subjects identified 73% of the fricatives while the Votrax subjects identified 80%. A tally of the number of /s/'s correctly identified on the fourth training day for the Harvard(A) section, which was similar in all aspects to the material used in the test sessions, yielded 99 out of 108 for the Votrax versus 101 out of 108 for the Echo. This is proof that the noise effectively masked

the /s/'s of the Votrax. This masking is not as abrupt as that of natural speech which resists masking down to a SNR of 0 dB and then its intelligibility steadily declines (Clark, 1983 and Kalikow et al., 1977). Synthetic speech seems to be steadily masked from at least an SNR of +10 dB down as is indicated from figure 4.6 and from research done by Clark, 1983. Accordingly the choice of synthesizers for a noisy environment cannot be taken lightly.

This study shows that the Echo synthesizer produces a more intelligible system than the Votrax. This implies that an LPC system is more intelligible than an FS system. Although this type of generalization may seem to bold, it is useful to realize that most inexpensive synthesizers have a common chip set. For example, the Votrax and the Intex formant synthesizers (both used in the initial pilot study) use the Votrax SC-01 chip which translates a 6 bit phoneme code and a 2 bit pitch code into the spectral parameters which adjust its internal electronic model of the vocal tract to synthesize speech. Therefore, any differences are due to the implementation of the supporting hardware.

This leads to an interesting observation about the Votrax synthesizer. Whereas from figure 4.7, the fricatives have less energy than the vowels, it seems a simple and inexpensive way to improve the intelligibility of the Votrax in noise would be to explore ways to more forcefully model the turbulent airflow of fricatives. The data presented upon the intelligibility of the Votrax's /s/'s indicates that this would be a worthwhile modification.

A previous study by Schwab et al., 1985, demonstrated that until the technology for high quality synthetic speech becomes less expensive, low quality inexpensive speech technology can be used effectively as a communication system. This study took that result a step further and qualified the word technology. It is apparent from this study that LPC(Echo) produces a more intelligible sytstem than the FS(Votrax) in the presence of babble. Future studies might examine the nature and extent of the noise masking as well as the masking effect of different kinds of noise.

Also, there are many questions remaining regarding the effect of speech parameters on intelligibility. Although the LPC device was found more intelligible than the FS, the harsh mechanical sound of the LPC synthesizer might not be a tolerable human factor. Finally, speech parameters such as speech rate, pitch contour and pitch have been examined without interfering noise and it is certainly not a sound practice to assume that in different types of noise, these parameters would have identical effects on intelligibility.

REFERENCES

Bilger, R. C. (1985). "Speech Recognition Test Development," ASHA, 14, 2.

Clark J. E. (1983). "Intelligibility Comparison for Two Synthetic and One Natural Speech Source," J. Phonet. 11, 80-93.

Clark J. E., Dermody, P. and Palethorpe, S. (1985). "Cue Enhancement by Stimulus Repetition: Natural and Synthetic Speech Comparisons," JASA, 78(2), 458-462.

Dalton, B., Umbers, I. and Townshend, K. (1983). "An Evaluation of Two Speech Synthesizers in an Industrial Context," Pp. 28, Warren Spring Labs., Stevanage England. PB 84-201318(No. 19, 1984).

Delattre, P. C., Liberman, A. M. and Cooper, F. S. (1955). "Acoustic Loci and Transitional Cues for Consonants," JASA, 27, 769-773.

Fairbanks, G. (1958). "Tests of Phonemic Differentiation: The Rhyme Test," JASA, 30(7), 596-600.

Frank, T. and Craig, C. H. (1984). "Comparison of the Auditec and Rintelmann Recordings of the NU-6," J. of Speech and Hear. Disorders, 49, 267-271.

Greene, B. G. and Logan, J. S. (1986). "Segmental Intelligibility of Synthetic Speech Produced by Eight text-to-speech Systems," JASA, Suppl. 1, 79, S24.

Hart, S. G. and Simpson, C. A. (1976). "Effects of Linguistic Redundancy of Synthesized Cockpit Warning Message Comprehension and Concurrent Time Estimation," Proc. 12th Ann. Conf. Manual Control, Champaign, 1976. Moffett Field, CA: NASA TMX-73, 170 May.

House, A. S., Williams, C. E., Hecker, M. H. L. and Kryter, K. D., (1974). "Articulation Testing Methods: Consonantal Differentiation with a Closed-Response Set," JASA, 37(1), 158-166.

IEEE. (1969). "IEEE Recommended Practice for Speech Quality Measurements," (IEEE No. 297).

Kalikow, D. N., Stevens, K. N. and Eliot, L. L. (1977). "Development of a Test of Speech Intelligibility in Noise using Sentence Materials with Controlled Word Predictability," JASA, 61(5), 1337-1351.

Keeler, L. O., Clement, G. L., Strong, W. J. and Palmer, E. P. (1976). "Two Preliminary Studies of the Intelligibility of Predictor Coeficients and Formant-Coded Speech," IEEE Trans. Acoust., Speech, Signal Process., ASSP-24, 429-432.

Lehiste, I. and Peterson, G. E. (1959). "Linguistic Considerations in the Study of Speech Intelligibility," JASA, 37(1), 280-286.

Logan, S and Pisoni, D. B. (1986). "Preference Judgements Comparing Different Synthetic Voices," JASA, Suppl 1, 79, s24.

Luce, P. A., Feustel, T. C. and Pisoni, D. B. (1983). "Capacity Demands in Short Term Memory for Synthetic and Natural Speech," Human Factors, 25, 17-32.

Manous, L. M., Pisoni, D. B., Dedina, J. and Nusbaum, H. C. (1986). "Comprehension of Natural and Synthetic Speech Using a Sentence Verification Task," JASA, Suppl. 1, 79, s24.

Montgomery, D. C. (1976). "Design and Analysis of Experiments," John Wiley & Sons, New York NY.

Morgan, D. F., Kamm, C. A. and Velde, J. M. (1981). "Form Equivalence of the Speech Perception in Noise (SPIN) Test," JASA, 69(6), 1791-1798.

Nye, P. W. and Gaitenby, J. H. (1974). "The Intelligibility of Synthetic Monosyllabic Words in Short, Syntactically Normal Sentences," Haskins Laboratories Status Report on Speech Research, 38, 169-190, New Haven, CT: Haskins Laboratories.

Ostle, B. (1963). "Statistics in Research," Iowa State Univ. Press, 552-553.

Ott, L. (1984). "An Introduction to Statistical Methods and Data Analysis," PWS Publishers, Boston MA.

Owen, J. H. (1981). "Influence of Acoustical and Lingusitic Factors on the SPIN Test Difference Score," JASA, 70(3), 678-682.

Poll, I., Rubenstein, H. and Decker, L. (1959). "Intelligibility of Known and Unknown Message Sets," JASA, 31(3), 273-279.

Pisoni, D. B. (1981). "Speeded Classification of Natural and Synthetic Speech in a Lexical Decision Task," JASA, Suppl. 1, 70, s98.

Pisoni, D. B. and Hunnicutt, S. (1980). "Perceptuall Evaluation of MITALK: The MIT Unrestricted Text-to-speech System," ICASSP, 572-575.

Schwab, E. C., Nusbaum, H. C. and Pisoni, D. B. (1985). "Some Effects of Training on the Perception of Synthetic Speech," Human Factors, 27, 395-408.

Simpson, C. A. (1976). "Effects of Linguistic Redundancy on Pilots Comprehension of Synthesized Speech," Proc. 12th Ann. Conf. Manual Control, Champaign, 1976. Moffett Field, CA: NASA TMX-73, 170, May.

Simpson, C. A. and Frost, K. M. (1984). "Synthesized Speech Rate and Pitch Effects on Intelligibility of Warning Messages for Pilots," Human Factors, 26(5), 509-517.

Slowiaczek, L. and Nusbaum, H. C. (1985). "Effects of Speech Rate and Pitch Contour on the Perception of Synthetic Speech," Human Factors, 27(6), 701-712.

APPENDIX A

SAMPLE TRAINING SESSION

HARVARD SENTENCES(A)

In this section you will be presented with 50 sentences. The sentences will be provided in written form excepting one word which you will be required to identify. Take care to identify all words so that your chance of identifying the unknown will be greater. Please print all responses in the blank space provided.

Example The carpenter used a _____ and nails.

If you have any questions please ask them now.

1. The _____ air passed through the tent.
2. The crooked maze failed to _____ the mouse.
3. _____ fast leads to wrong sums.
4. The show was a flop from the very _____.
5. A saw is a _____ used for making boards.
6. The _____ moved on well oiled wheels.
7. March the _____ past the next hill.
8. A cup of _____ makes sweet fudge.
9. Place a rosebud _____ the porch steps.
10. Both lost their _____ in the raging storm.
11. We _____ of the side show in the circus.
12. Use a pencil to write the first _____.
13. He ran _____ way to the hardware store.
14. The clock _____ to mark the third period.
15. A _____ creek cut across the field.
16. Cars and _____ stalled in snow drifts.
17. The _____ of china hit the floor with a crash.
18. This is a grand season for _____ on the road.
19. The dune rose from the edge of the _____.
20. Those _____ were the cue for the actor to leave.
21. A yacht slid around the _____ into the bay.
22. The two met while _____ on the sand.
23. The ink stain dried on the finished _____.
24. The _____ town was seized without a fight.
25. The lease ran out in _____ weeks.

26. A _____ squirrel makes a nice pet.

27. The horn of the car woke the sleeping _____.

28. The _____ beat strongly and with firm strokes.

29. The _____ was worn in a thin silver ring.

30. The fruit _____ was cut in thick slices.

31. The navy attacked the big _____ force.

32. See the cat _____ at the scared mouse.

33. There are more than two factors _____.

34. The hat brim was _____ and too droopy.

35. The lawyer tried to lose his _____.

36. The _____ curled around the fence post.

37. Cut the pie into large _____.

38. Men strive but seldom get _____.

39. Always _____ the barn door tight.

40. He lay prone and hardly moved a _____.

41. The slush lay deep along the _____.

42. A wisp of cloud _____ in the blue air.

43. A pound of _____ costs more than eggs.

44. The fin was sharp and _____ the clear water.

45. The _____ seems dull and quite stupid.

46. Bail the _____ to stop it from sinking.

47. The _____ ended in late June that year.

48. A tusk is used to make costly _____.

49. Ten pins were _____ in order.

50. The bill was _____ every third week.

PROSE PASSAGE

In this section you will be asked to listen to a selection of prose from an accomplished writer. All you are required to do is to follow along with the supplied text and become accustomed to the unique characteristics of the synthesized voice. If you have any questions please ask them now.

PROSE ONE

One of the most interesting and characteristic features of democracy is of course, the difficulty of defining it. And this difficulty has been compounded in the United States, where we have been giving new meanings to almost everything. It is, therefore, especially easy for anyone to say that democracy in America has failed.

"Democracy," according to political scientists, usually describes a form of government by the people, either directly or through their elected representatives. But I prefer to describe a democratic society as one which is governed by a spirit of equality and dominated by the desire to equalize, to give everything to everybody. In the United States the characteristic wealth and skills and know-how and optimism of our country have dominated this quest.

My first and overshadowing proposition is that our problems arise not so much form our failures as from our successes. Of course no success is complete; only death is final. But we have probably come closer to attaining our professed objectives than any other society of comparable size and extent, and it is from this that our peculiarly American problems arise.

The use of technology to democratize our daily life has given a quite new shape to our hopes. In this final chapter I will explore some of the consequences of democracy, nor for government but for experience. What are the consequences for everybody every day of this effort to democratize life in America? And especially the consequences of our fantastic success in industry and technology and in invention.

There have been at least four of these consequences. I begin with what I call attenuation, which means the thinning out or the flattening of experience. We might call this the democratizing of experience. It might otherwise be described as the decline of poignancy. One of the consequences of our success in technology, of our wealth, has been the removal of distinctions, not just between people but between everything and everything else, between every place and every other place, between every time and every other time. For example, television removes the distinction between being here and being there. and the same kind of process, of thinning out, of removing distinctions, has appeared in one area after another of our lives.

For instance, in the seasons. One of the great unheralded achievements of American civilization was the rise of transportation and refrigeration, the development of techniques of canning and preserving meat, vegetables, and fruits in such a way that it became possible to enjoy strawberries in winter, to enjoy fresh meat at seasons when the meat was not slaughtered, to thin out the difference between the diet of winter and the diet of summer. There are many unsung heroic stories in this effort.

One of them, for example, was the saga of Gustavus Swift in Chicago. In order to make fresh meat available at a relatively low price to people all over the country, it was necessary to be able to transport it from the West, where the cattle was raised, to the

Eastern markets and the cities where population was concentrated. Gustavus Swift found the railroad companies unwilling to manufacture refrigerator cars. They were afraid that, if refrigeration was developed, the cattle would be butchered in the West and then transported in a more concentrated form than when the cattle had to be carried live. The obvious consequence, they believed, would be to reduce the amount of freight. So they refused to develop the refrigerator car. Gustavus Swift went ahead and developed it, only to find that he had more cars than he had use for. The price of fresh meat went down in the Eastern cities, and Gustavus Swift had refrigerator cars on his hands. He then sent agents to the South and to other parts of the country, and tried to encourage people to raise products which had to be carried in refrigerator cars. One of the consequences of this was the development of certain strains of vegetables and fruits, especially of fruit, which would travel well. And Georgia became famous for the peaches which were grown partly as a result of Swift's efforts to encourage people to raise something that he could carry in his refrigerator cars.

HASKINS SENTENCES

In this section of the session you will be presented with 50 sentences. These sentences are unique, in that, although they are grammatically sound, they make no sense whatsoever. You will be given the sentences in printed form excepting one word, which you will be asked to identify. Take care to identify all words so that your chance of correctly identifying the unknown word will be greater. Please print all responses neatly in the blank space provided.

Example The small _____ fell the hill.

If you have any questions please ask them now.

1. The wrong shot lead the _____.
2. The _____ top ran the spring.
3. The great car _____ the milk.
4. The old _____ cost the blood.
5. The _____ are sent the cow.
6. The low walk _____ the hat.
7. The _____ paint said the land.
8. The _____ bank felt the bag.
9. The seat _____ grew the chain.
10. The _____ dog caused the shoe.
11. The last _____ fire the nose.
12. The young _____ saw the rose.
13. The gold rain _____ the wing.
14. The chance _____ laid the year.
15. The white bow had the _____.
16. The near stone thought the _____.
17. The end _____ held the press.
18. The deep head _____ the cent.
19. The next _____ sold the room.
20. The full leg shut the _____.
21. The _____ meat caught the shade.
22. The fine lip tired the _____.
23. The _____ can lost the men.
24. The dead _____ armed the bird.
25. The fast point _____ the word.

26. The mean _____ made the game.

27. The clean book _____ the ship.

28. The red _____ said the yard.

29. The late _____ aged the boat.

30. The large group _____ the judge.

31. The _____ knee got the shout.

32. The least _____ caught the dance.

33. The _____ week did the page.

34. The _____ cold stood the plant.

35. The _____ air heard the field.

36. The far _____ tried the wood.

37. The high sea _____ the box.

38. The blue _____ broke the branch.

39. The _____ feet asked the egg.

40. The _____ horse brought the hill.

41. The strong rock _____ the ball.

42. The _____ neck ran the wife.

43. The dry door paid the _____.

44. The child _____ spread the school.

45. The brown post _____ the ring.

46. The clear back _____ the fish.

47. The round _____ came the well.

48. The good _____ set the hair.

49. The bright guide knew the _____.

50. The hot nest gave the _____.

HARVARD SENTENCES(B)

In this section you will be presented with 50 sentences. You are required to identify the last word in each sentence. You will have no written clues, but keep in mind that the sentences are meaningful and thus you will have many aural clues. Please print your responses in the blank spaces provided. If you have any questions please ask them now.

1 _____
2 _____
3 _____
4 _____
5 _____
6 _____
7 _____
8 _____
9 _____
10 _____
11 _____
12 _____
13 _____
14 _____
15 _____
16 _____
17 _____
18 _____
19 _____
20 _____
21 _____
22 _____
23 _____
24 _____
25 _____

26 _____
27 _____
28 _____
29 _____
30 _____
31 _____
32 _____
33 _____
34 _____
35 _____
36 _____
37 _____
38 _____
39 _____
40 _____
41 _____
42 _____
43 _____
44 _____
45 _____
46 _____
47 _____
48 _____
49 _____
50 _____

MODIFIED RHYME TEST

In this section you will be presented with a spoken word and six possible choices. The word will be spoken in the form "Number one is _____." For example you will hear:

"Number one is mean."

Your choices would be:

seen teen mean wean lean queen

You would then circle the word which you feel to be the one spoken. For example:

seen teen mean wean lean queen

If you have any questions ask them now.

MODIFIED RHYME TEST ONE

1	late	lake	lay	lace	lane	lame
2	bean	beach	beat	beam	bead	beak
3	peel	reel	feel	heel	keel	eel
4	nest	vest	west	test	best	rest
5	seep	seen	seethe	seed	seem	seek
6	cut	cub	cuff	cup	cud	cuss
7	dig	dip	did	dim	dill	din
8	ten	pen	den	hen	then	men
9	sun	nun	gun	fun	bun	run
10	way	may	say	gay	day	pay
11	book	took	shook	cook	hook	look
12	peace	peas	peak	peal	peach	peat

APPENDIX B

PROTECTION OF HUMAN SUBJECTS

PILOT STUDY

1. The production of inexpensive synthetic speech equipment has made possible many new applications within industry and on the consumer market. Previous research has focused upon identifying the difference between natural and synthetic speech sources. This study is desigred to investigate the difference between the various methods used to produce speech. Sentences generated by linear predictive coding, formant synthesis and digitization will be presented randomly to volunteer subjects. The data will then be analyzed to det-ermine the perceptual dissimilarities.

2. I am a student of the Graduate program in Acoustics. This project is being supervised by Dr. Claus P. Janota, Assistant Professor in Acoustics. Dr. Janota has continually been involved in human subjects research for the past ten years.

3. The subject population will consist of males and females above the age of 18 with normal hearing.

4. All subjects will be volunteers.

5. Subjects will be seated in a soundproof room. Instructions will be presented to them in written form as well as in the form of the five various methods of computer generated speech described in step six. Basically, each subject will be presented with a sentence generated by one of the five methods and will be asked to identify the last word in the sentence. All material will be presented aurally, in acoustic free field from recordings. The subjects will be told to guess as to the answer if necessary. The identified words will be written on a form given to them by the experimenter. Before the actual experiment each subject will be presented with a small sample of representative stimulus such that learning effects will be minimized. The subjects will then be informed that the experiment is to begin. The actual experiment will contain 50 sentences, 10 using each synthesis device. The response form contains 50 blank spaces for the subjects to record the identified word.

6. The equipment involved includes the following commercial speech synthesis means:

Intex talker & Votrax type'n talk(Formant synthesis)
Echo PC & Echo GP(Linear predictive coding)
Mimic Speech Processor(Speech digitizer)

The devices are controlled by an AT&T model 6300 personal computer. All test material is then recorded onto a Crown series 700 reel to reel recorder and played back into an Industrial Acoustics Company model 40 soundproof room. The playback levels are such that the sound pressure level(SPL) does not exceed a comfortable listening level.(reference one)Sound level will be measured periodically using standard sound-level metering equipment to insure that unacceptably high levels cannot occur. The experimenter, Rory DePaolis, and his advisor will have sole

access to the data. Mr. DePaolis is a research assistant at the Applied Research Laboratory working towards his M.S. Dr. Janota's qualifications are described above.

7. Informed consent will be obtained explicitly via the attached informed consent form. The consent form and all other pertinent information will be reviewed with each subject such that any questions about the form itself, or the experiment as a whole can be addressed. These actions will take place at the experimenters office at the Applied Science Building.

8. The most serious potential risk is the exposure of a subject to sound levels exceeding a safe limit. The exposure duration and permitted sound pressure levels(SPL) are such that no adverse effects are known.

9. N.A.

10. The benefit to each subject will be an interesting exposure to the commercially available methods of speech generation. On the societal level, a better understanding of the generation of synthetic speech stands to be gained.

11. To avoid the possibility of mistakenly exposing subjects to SPL's exceeding a safe limit the output levels of the Crown recorder are set such that the maximum levels correspond to 85 dBA, free field, inside the sound booth. (ref. 1) To prevent disclosure of a subjects test results all test forms will be coded such that only the experimenter will know which test form corresponds to which subject.

13. Thurlow, Willard R. (1971). "Audition," in Experimental Psychology, edited by Kling, J. W. and Riggs. L. A. (Holt Rinehart and Winston), pp. 223-259.

PILOT STUDY

PURPOSE OF STUDY

The study you are going to participate in is an investigation of some popular methods of speech generation. The development of inexpensive speech synthesis eqipment has opened up many industrial and consumer applications but currently very little work has been done to examine the perceptual differences of these methods. The current study will address this issue and determine some of the basic differences between the various methods of producing speech.

PROCEDURE

You will be presented with 3 popular methods of producing speech, one hardware based, one software based and one method of digitizing speech. The speech will be presented in the form of sentences of which you will be asked to write down the last word of the sentence. It is important that you attempt to identify the last word in all of the sentences. Your written responses will be scored phonetically so that, for example, if you incorrectly identified the word 'push' as 'bush' you would still be scored as having correctly identified the /u/ and the /sh/. This data will then be analyzed and a conclusion reached as to any significant difference between the three speech production methods.

TIME REQUIRED

After a short familiarization period you will be asked to identify 50 words. The amount of time required of you from start to finish will not exceed 60 minutes.

POTENTIAL RISKS

There is a very slight possibility that you could be exposed to sound levels that might cause discomfort. The eventuality is unlikely.

CONTACT PERSON

In the event that you have any questions regarding this study, please feel free to contact:

Rory DePaolis, Research Assistant
Applied Research Laboratory
P.O. Box 30
State College, PA 16801

_____ _____
Volunteers Signature Date

_____ _____
Investigators signature Date

EXPERIMENT

1. The production of inexpensive synthetic speech equipment has made possible many new applications within industry and on the consumer market. Previous research has focused upon identifying the difference between natural and synthetic speech sources. This study is designed to investigate the difference between two synthetic methods used to produce speech. Sentences generated by linear predictive coding and formant synthesis will be presented randomly in the presence of four different noise levels to volunteer subjects. The data will then be analyzed to determine the perceptual dissimilarities.

2. I am a student of the Graduate program in Acoustics. This project is being supervised by Dr. Claus P. Janota, Assistant Professor in Acoustics. Dr. Janota has continually been involved in human subjects research for the past ten years.

3. The subject population will consist of males and females above the age of 18 with normal hearing.

4. All subjects will be volunteers.

5. Subjects will be seated in a soundproof room. Instructions will be presented to them in written form as well as in the form of a detailed explanation by the experimenter. Each subject will be presented with a sentence generated by one of the two methods of speech synthesis and will be asked to identify the last word in the sentence. All material will be presented aurally, in acoustic free field from recordings. The subjects will be told to guess as to the answer if necessary. The identified words will be written on a form given to them by the experimenter. The actual experiment will contain 50 sentences forms, two to four forms per test period. The response form contains 50 blank spaces for the subjects to record the identified word.

6. The equipment involved includes the following commercial speech synthesis means:

Votrax type'n talk(Formant synthesis)
Echo GP(Linear predictive coding)

The devices are controlled by an AT&T model 6300 personal computer. All test material is then recorded onto a Crown series 700 reel to reel recorder and played back into an Industrial Acoustics Company model 40 soundproof room. The playback levels are such that the sound pressure level(SPL) does not exceed a comfortable listening level.(reference one) Sound level will be measured periodically using standard sound-level metering equipment to insure that unacceptably high levels cannot occur. The experimenter, Rory DePaolis, and his advisor will have sole access to the data. Mr. DePaolis is a research assistant at the Applied Research Laboratory working towards his M.S. Dr. Janota's qualifications are described above.

7. Informed consent will be obtained explicitly via the attached informed consent form. The consent form and all other pertinent information will be reviewed with each subject such that any questions about the form itself, or the experiment as a whole can be addressed. These actions will take place at the experimenters office at the Applied Science Building.

8. The most serious potential risk is the exposure of a subject to sound levels exceeding a safe limit. The exposure duration and permitted sound pressure levels(SPL) are such that no adverse effects are known.

9. N.A.

10. The benefit to each subject will be an interesting exposure to the commercially available methods of speech generation. On the societal level, a better understanding of the generation of synthetic speech stands to be gained.

11. To avoid the possibility of mistakenly exposing subjects to SPL's exceeding a safe limit the output levels of the Crown recorder are set such that the maximum levels correspond to 85 dBA, free field, inside the sound booth. (ref. 1) To prevent disclosure of a subjects test results all test forms will be coded such that only the experimenter will know which test form corresponds to which subject.

13. Thurlow, Willard R. (1971). "Audition," in Experimental Psychology, edited by Kling, J. W. and Riggs, L. A. (Holt Rinehart and Winston), pp. 223-259.

EXPERIMENT

PURPOSE OF STUDY

The study you are going to participate in is an investigation of two popular methods of speech generation. The development of inexpensive speech synthesis eqipment has opened up many industrial and consumer applications but currently very little work has been done to examine the perceptual differences of these methods. The current study will address this issue and determine some of the basic differences between the two methods of producing speech.

PROCEDURE

You will be presented with two popular methods of producing speech, one hardware based and one software. The speech will be presented in the form of sentences of which you will be asked to write down the last word of the sentence. It is important that you attempt to identify the last word in all of the sentences. Your written responses will be scored phonetically so that, for example, if you incorrectly identified the word 'push' as 'bush' you would still be scored as having correctly identified the /u/ and the /sh/. This data will then be analyzed and a conclusion reached as to any significant difference between the two speech production methods.

TIME REQUIRED

You will be asked to participate in a maximum of five sessions which will last between 30 and 60 minutes. You will be paid a sum of . The amount of time required of you will not exceed five hours.

POTENTIAL RISKS

There is a very slight possibility that you could be exposed to sound levels that might cause discomfort. The eventuality is unlikely.

CONTACT PERSON

In the event that you have any questions regarding this study, please feel free to contact:

Rory DePaolis, Research Assistant
Applied Research Laboratory
P.O. Box 30
State College, PA 16801

_____ _____
Volunteers Signature Date

_____ _____
Investigators signature Date

Title of Investigation: The Perceptuall dissimilarities of Speech Production Methods

Investigator: Rory DePaolis

Date:

This is to certify that I, _____, hereby agree to participate as a volunteer in a scientific study as an authorized part of the education and research program of The Pennsylvania State University under the supervision of Rory DePaolis.

The study and my part in the study have been explained to me by Rory DePaolis, and I understand his explanation. A copy of the procedures of this study and a brief description of any risks and discomforts have been provided to me and has been discussed in detail with me.

I have been given the opportunity to ask whatever questions I may have had and all such inquiries have been answered to my satisfaction.

I understand that I am free to deny any answers to specific items or questions in interviews or questionnaires.

I understand that any data or answers to questions will remain confidential with regard to my identity.

I understand that, in the event of physical injury resulting from this investigation, neither financial compen-sation nor free medical treatment is provided for such phys-ical injury, and that further information on this policy is available from the Vice President for Research and Dean of the Graduate School, 114 Kern Graduate Building(865-6331).

I FURTHER UNDERSTAND THAT I AM FREE TO WITHDRAW MY CONSENT AND TERMINATE MY PARTICIPATION AT ANY TIME.

_____	_____	_____
Date	Date of Birth	Subject's Signature

I, the undersigned, have defined and fully explained the investigation to the above subject.

_____	_____
Date	Investigator's Signature

APPENDIX C

GLOSSARY OF TERMS

Allophone--variations of a phoneme

Coarticulation--the influence of surrounding phonemes upon the phoneme which is being produced.(Ex. of backward coarticulation, napkin)

Formants--resonant frequencies of the speech wave which reflect how the vocal tract is modified to produce sounds.

Fricatives--consonants in which the air is partially obstructed in the oral cavity.

Nasals--consonants in which the air passes through the nasal cavity.

Phoneme--a unit of spoken language which signals semantic distinctiveness.

Plosive--consonants in which the air is completely stopped in the oral cavity.

Semivowels--consonants whose production is similar to that of vowels, i.e., /l/ and /r/.

END

DATE

FILMED

4-88

DTIC