Comparing the Single-Word Intelligibility of Two Speech Synthesizers for Small Computers

Paula Sue Cochran
Charlottesville, Virginia

B.A., College of Wooster, 1976
Diploma in Linguistics, University of Cambridge, 1977
M.A., Ohio University, 1978

A Dissertation Presented to the Graduate
Faculty of the University of Virginia
in Candidacy for the Degree of
Doctor of Philosophy

Department of Education

University of Virginia

May 1986

Glen Bull
Raymy. Stoudt, Jr.

Abstract

Previous research on the intelligibility of synthesized speech has placed emphasis on the segmental intelligibility (rather than word or sentence intelligibility) of expensive and sophisticated synthesis systems. There is a need for more information about the intelligibility of low-to-moderately priced speech synthesizers because they are the most likely to be widely purchased for clinical and educational use.

The purpose of the present study was to compare the word intelligibility of two such synthesizers for small computers, the Votrax Personal Speech System (PSS) and the Echo GP (General Purpose). A multiple-choice word identification task was used in a two-part study in which 48 young adults served as listeners. Groups of subjects in Part I completed one trial listening to taped natural speech followed by one trial with each synthesizer. Subjects in Part II listened to the taped human speech followed by two trials with the same synthesizer.

Under the quiet listening conditions used for this study, taped human speech was 30% more intelligible than the Votrax PSS, and 53% more intelligible than the Echo GP. A statistically significant difference in word intelligibility was observed between the synthesizers, with the Votrax PSS being 18% more intelligible. Listeners who heard human

speech followed by two different synthesizers performed comparably to those who heard the more likely clinical combination of human speech followed by just one synthesizer.

The observed difference between these speech synthesizers is likely to be most noticeable in clinical applications in which other contextual cues are minimal, or in which listeners are unlikely or unable to take advantage of such cues. In considering the factors bearing on the purchase of speech synthesizers for such applications, clinicians are encouraged to increase the priority they give to intelligibility.

Acceptance of the Dissertation

The dissertation entitled "Comparing the Single-Word Intelligibility of Two Speech Synthesizers for Small Computers" is accepted by the Graduate Faculty of the School of Arts and Sciences, and the Department of Education, of the University of Virginia, in partial fulfillment of the requirements of the degree of Doctor of Philosophy.

Advisory Committee

Glen Bull
Dr. Glen Bull
Chairman

James Deese
Dr. James Deese

Jerry Short
Dr. Jerry Short

Ralph Stoudt
Dr. Ralph Stoudt

Dr. George Theodoridis

4/28/86
Date

Acknowledgements

I would like to thank my advisor, Glen Bull, for sharing his vision of "the big picture" and teaching me to look for my own. I would also like to acknowledge the helpful suggestions and supportive attitude of the rest of my advisory committee. I am grateful to Chris Halpin and Tom Lough for technical assistance.

PSC

Contents

Chapter Four: Procedures and Results

Chapter Five: Interpretation of Results

Chapter Six: Discussion and Conclusions

List of Figures

List of Tables

Chapter One
Speech Synthesis in Communication Disorders

Clinical Applications of Microcomputers

The clinical applications of microcomputers are becoming increasingly apparent and increasingly refined. Clinicians and teachers have discovered that microcomputers can be powerful instructional tools. Current applications for microcomputers in communication disorders range from analysis of diagnostic data to direct use of the computer with clients. Many clinical applications make use of artificial speech.

The clinical use of microcomputers by professionals in the field of communication disorders is becoming more common, even outside of the university/training environment. A survey by the American Speech-Language-Hearing Association (ASHA) indicates that the availability and use of microcomputers in this field has increased dramatically in the past few years. For example, the percentage of ASHA members reporting the use of computers at work increased from 23.8% in 1982 to 47.2% in 1985 (ASHA Omnibus Survey, reported by Hyman, 1985).

Microcomputers can play a part in the treatment of a surprisingly wide range of communicative disabilities. Computer-based assessment and/or therapy techniques are already available for cognitive rehabilitation, motor speech, voice, and fluency disorders, augmentative communication, developmental speech and language disorders, etc.. An important feature associated with many of these applications is artificial speech.

As has been observed elsewhere, special educators and clinicians have been especially quick to recognize the potential instructional and rehabilitative value of "talking" computer programs (Ginther, 1983). Until fairly recently, however, few commercially available educational or clinical software programs made use of either digitized or synthesized speech.

Even an informal examination of software and instructional materials catalogs suggests that within the past 2 or 3 years, there has been a remarkable change in the availability of software which makes use of synthesized speech. A likely factor in this change has been the advent of readily available, affordable speech synthesizers for popular microcomputers. Several already-assembled speech synthesizers are now available in the $150 to $400 price range, making them a financial investment comparable to a dot-matrix printer. Kits for building speech synthesizers are available for even less. This easy-to-use and widely available technology can trace its history back through a long line of dedicated scientists and intriguing artificial speech mechanisms.

Artificial Speech: A Glance Backward

I. Synthesized speech vs. digitized speech

The "talking machines" which consumers are most likely to interact with today, such as talking cash registers or soda machines, make use of digitized rather than synthesized speech. The speech in these machines sounds realistic because digitized speech begins with natural human speech. The natural speech sample is recorded and stored by a computer. This process requires that the computer "sample" the speech signal, and record numbers which reflect changes in the speech wave over time. Speech of high quality can be produced when these digits are converted back into an analog signal and then amplified. Among other factors, the quality depends on how many samples per second were taken from the original speech signal.

Although digitization makes it possible to incorporate high-quality speech into various computer-controlled applications, it has some disadvantages. The most obvious is that digitized speech requires the original human speaker -- all speech must be planned and recorded in advance. In addition, the storage of even small samples of digitized speech requires large amounts of computer memory.

Synthesized speech, in contrast, begins with an artificial sound source. The vocabulary need not be limited in content to previously prepared words, or in size by the memory limitations of a computer. There are several techniques for synthesizing speech, which will be briefly reviewed below. The search for a way to create artificial speech which could be systematically manipulated and analyzed began long before computers were available to assist in the process.

II. Models of synthesized speech

Early devices for producing artificial speech attempted to mechanically reproduce the human vocal tract and articulators (Borden and Harris, 1984; Flanagan, 1972). For example, in the late 1700's Wolfgang von Kempelen of Vienna constructed a speaking machine which produced vowel and consonant sounds as it was manipulated with two hands. According to Flanagan (1972), "It used a bellows to supply air to a reed which, in turn, excited a single, hand-varied resonator that produced voiced sounds. Consonants, including nasals, were simulated by four separate constricted passages, controlled by the fingers of the other hand" (p. 1381). A reproduction of this machine, built in Edinburgh in the late 1800's, may have influenced Alexander Graham Bell in his own attempts to produce a mechanical analog to the vocal tract (Flanangan, 1972, p. 1381).

Soon electrical devices replaced mechanical ones in the effort to reliably produce speech sounds. Major advances came with the development of the sound spectograph, which produced a frequency and intensity analysis of speech over time. Spectographic analysis of human speech provided more information about the acoustic properties of speech for use in speech synthesis. Formant-based speech synthesizers, such as the Pattern Playback synthesizer constructed by Franklin Cooper at Haskins Laboratories, soon became valuable tools in the study of speech perception. The Pattern Playback produced speech by converting formant-like patterns painted on acetate film loops into acoustic signals (Borden and Harris, 1984, p. 22). For the Pattern Playback and other formant-based systems, the parameters for speech synthesis were specified in terms of formants or the acoustic characteristics of speech, rather than in the mechanical terms of articulators.

Speech synthesizers used in conjunction with computers have dramatically increased the speed and complexity of possible speech signal manipulations. Some computer-based synthesizers operate from formant-based information in mathematical (rather than graphic) form. These include, for example, the MITalk - 79, the Votrax Type 'n Talk, and the Votrax Personal Speech System (PSS) (Greene, Logan, and Pisoni, in press). Mathematical algorithms are used to convert information from letter-to-phoneme rules into estimated formant values.

Linear predictive coding (LPC), is an alternative method for generating speech. It has been used in speech synthesizers such as the Echo +, Echo PC, Digitalker, and the IBM PCjr Speech Attachment. In the process of LPC, mathematical formuli model the filtering and resonating characteristics of the vocal tract, in response to regular text input or special phonemic code (Greene, Logan, and Pisoni, in press). Although these devices are not mechanically operated like those of von Kempelen and Bell, they generate speech by predicting how speech sounds would be produced and filtered in a particular vocal tract.

III. Text-to-speech

Speech synthesizers can be programmed to produce speech in various ways. Many synthesizers can be programmed directly, using a computer programming language such as BASIC. Frequently, specialized codes must be used, especially if any adjustments of rate or intonation are desired. Some synthesizers can be programmed "phonetically," by using symbols which represent individual phonemes of a natural language. This might be desirable, for example, when the speech accompanies a software program with a specified, pre-programmed vocabulary. Fine tuning the pronunciation of that set of words could be done in advance by a software designer.

For other applications, however, complex programming requirements can have the practical result of a limiting the vocabulary and utility of the synthesizer. That is, users with limited expertise and time might not master the programming skills required to make the speech synthesizer meet their needs. Such a reduction in flexibility would cancel out some of the advantage of synthesized speech over digitized speech.

Some speech synthesizers can function alternatively as text-to-speech systems. This means that they can take standard orthographic text as input, and translate it into speech. The text must be such that it can be converted to standard ASCII code by the computer. The synthesizer receives this code from the computer, and then converts the letters into speech sounds using a set of rules. The way in which the speech sound information is "synthesized" into speech -- either through formant-based rules or LPC, depends on the synthesizer. This capability to translate text into speech is important in many clinical applications of speech synthesizers. It means that special programming skills are not necessarily required of the user, and an unlimited vocabulary of words, phrases, or sentences is possible. An obvious difficulty, however, is the fact that standard orthographic text contains almost no information about the prosodic features of language. Text-to-speech systems must therefore produce speech without the benefit of the subtleties of intonation, pause, and stress, which often convey the true meaning of an utterance.

Clinical/educational Applications of Synthesized Speech

As the number of clinicians who have experience in the direct use of synthesized speech increases, so does the number of new clinical and educational applications. Until recently, the clinical use of synthesized speech might have been considered an "advanced" skill, requiring more experience and training than most clinicians possessed. With the advent of menu driven software and already-assembled synthesizers, this is no longer the case. Often it is the desire to make use of speech output with a particular client which encourages clinicians, special educators, or parents of handicapped children, to learn about computers. The most prominent uses of text-to-speech synthesizers to enhance communication or the development of communication skills, are described below.

I. Synthesized speech as part of an augmentative or alternative system of communication for disabled children or adults.

There are many approaches to facilitating communication when regular speech is not possible. Systems which require minimal special equipment and have been used frequently in the past include sign language, picture books/boards, and communication boards. Communication boards usually consist of pictures, words, letters, or other symbols arranged systematically on a board in front of the user. The "speaker" indicates his or her message through direct selection with a pointer or eye-gaze system, or through scanning. When the scanning method is used with one of these devices, the "speaker" and "listener" must gradually eliminate the possibilities from a set of hierarchically arranged choices (food--> breakfast--> cereal--> Cheerios).

Although many severely speech-impaired children and adults have learned to communicate successfully using these methods, they have several limitations. Two of these limitations are directly addressed by systems which make use of synthesized speech. The first pertains to proximity; with traditional methods of augmentative communication, the speaker and listener generally have to be quite close to one another. When a communication board is in use, both must be positioned so that they can see the board. The communication partner must give the board and "speaker" constant visual attention and verbal feedback. Related to this is the second limitation, which is the difficulty of communication between an augmented communicator and a novice partner. Communication through special devices tends to be threatening and uncomfortable for inexperienced "listeners." Routine exchanges are often difficult, especially in public.

Although it is not a cure-all, the addition of synthesized speech to augmentative devices can help make communication more effective. Listeners need only be within earshot, as they would have to be for regular speech, and thus conversing with individuals and in groups is more efficient and natural. Although synthesized speech may require careful listening, and varies in its intelligibility, most naive listeners can become accustomed to it more quickly than they could learn to decode messages through a scanning system or sign language.

The augmentative systems which make use of synthesized speech may or may not incorporate a personal computer. There are many self-contained devices which have speech output, covering a wide range of cost and complexity. For example, on the simple and inexpensive end of the continuum lies the Vocaid. The Vocaid is a small, easy to use communication board especially appropriate for addressing temporary communication needs. It is used widely in hospitals or with adult clients who are in the process of learning an alternative mode of communication, such as esophageal speech or sign language. The Vocaid has a pre-programmed vocabulary and comes with several overlays arranged by topic. The user presses the space which contains the picture, symbol, or word which he wants to express, and the Vocaid says the message.

The Minspeak and its smaller related systems can be found on the expensive and complex end of the continuum of self-contained communication devices with speech-output. The Minspeak, costing about $6000, is used primarily for diagnostic purposes. Either of two related devices, the

Light Talker or the Touch Talker, could be programmed to operate according to the optimal conditions identified through the use of the Minspeak. The Light Talker and Touch Talker cost about $2500 each. This family of dedicated communication aids can be programmed to meet individual requirements for vocabulary and complexity of operation. For example, each might be activated through a one-step direct-selection method (like the Vocaid) or a more sophisticated, layered system of topics and phrases. Any symbol set (pictures, letters, words, Blissymbols, etc.) or combination of symbols can be used to interface the user with the Minspeak devices.

Currently, the Minspeak devices are considered among the best choices for severely physically handicapped but cognitively alert individuals, who need an augmentative system which can continue to grow with them. Recently the Prenke-Romich corporation has made it possible to interface the Minspeak devices with a personal computer for some applications.

For a cost comparable to that of many sophisticated, dedicated communication devices, a personal computer with a speech synthesizer can be purchased. This is not the place for an exhaustive discussion of the pros and cons, but there are some primary reasons for acquiring this combination of devices rather than (or occasionally in addition to) a dedicated communication aid. These include:

- o Flexibility
- o Alternative Access
- o Alternative Uses.

Flexibility is the most important criterion in the comparison of dedicated communication aids to specially-configured personal computers. A computer-based system can often be tailored to match the important vocabulary and selection-method features of dedicated devices. In addition, the power and speed of a personal computer can be used to enhance these features.

Physical access to a computer-based communication system can usually be accomplished in a way which is comparable to or better than access to a traditional communication board. This might involve, for example, an alternative keyboard and pointer, or an electronic switch. It is also important to note that computer-compatible response devices have been developed which make use of advanced technology such as infrared and light-sensitive mechanisms and eye-gaze detectors. These devices along with the development of voice-recognition systems make computer access possible for a population of severely handicapped persons for whom other communication aids could not work.

In addition to a wide range of configurations and access methods, personal computers present a handicapped user with a range of possible uses beyond spoken language. Many dedicated communication devices include speech and printed output, but not the various other alternatives which a computer makes possible. For example, the same selection method which allows a disabled person to access synthesized speech and communicate with his family and friends, could be used to communicate with people all over the country via a modem. A computer/modem combination could connect him to an unlimited number of data bases or electronic bulletin boards. A word processor or a spreadsheet program could be used for personal business. A handicapped child could operate educational software, making choices using a selection method compatible with his everyday communication system rather than learning yet another "signalling" technique for educational and recreational activities.

While computer-based systems have their share of disadvantages (e.g., portability), clearly the use of synthesized speech via a personal computer is a desirable option in many instances. Thus one of the most important uses of a speech synthesizer for a small computer is in cases where it becomes the voice of a communicatively impaired person.

II. Synthesized speech as a model for potentially oral communicators.

As mentioned above, there has been an increase in the number of clinical and educational software packages for microcomputers which make use of synthesized speech. Some of these programs have been designed especially for use with speech/language handicapped children (e.g., First Words by Laureate Learning Systems, 1983; Exploratory Play by PEAL Software, 1985). For most of this software, the purpose of adding speech is to allow it to be used with non-readers or beginning readers. In some cases, however, the speech output of the program is particularly intended to serve as a model and a stimulus for oral speech/language development (Meyers, 1986). In this software the speech output is an essential feature, not just an enhancement.

Meyers (1984, 1986) has been at the forefront of this use of computers and synthesized speech. Her work with young non-oral children suggests that the synthesized speech itself, when under the control of the child, can serve as a motivation for verbal communication. She has used this method to facilitate language development in children with Down's syndrome and other language disorders.

III. Synthesized speech in interactive simulations and language activities.

Many of the synthesizers available for personal computers can operate in the text-to-speech mode. That is, they can "say" text as it is typed at the keyboard and printed on the screen, without any intermediate steps. Some of these synthesizers, such as the Votrax PSS and the Echo GP, can be attached to the printer card of the computer as if they were a printer. Information (text) can then be "spoken" with no more difficulty than if it were being printed.

This capability of speech synthesizers can be exploited by teachers and clinicians who are already using certain computer programming languages. For example, the computer language Logo is now widely available in public schools, and is often being taught to children who are pre-readers or beginning readers. It is particularly easy to make Logo "communicate" with a speech synthesizer like the ones mentioned above (Bull and Cochran, 1985; Wissick and Young, 1985). For the most part, then, the Logo activities which are now happening "quietly" in classrooms all over the country, could easily become even more interactive (and enticing) with the enhancement of synthesized speech.

Recent approaches to language therapy encourage clinicians to pay closer attention to the pragmatic aspects of communication (context, intent, shared knowledge, etc.). Goals related to conversation skills and clarification strategies, for example, are appearing on intervention plans in recognition of the importance of facilitating appropriate pragmatic language skills in speech/ language disordered children. In order to address these goals, however, the clinician is faced with the task of providing an ongoing, realistic context for conversation with her client (Snow, Midkiff-Borunda, Small, and Proctor, 1984). Many times, the strange and restricted environment of a therapy room does not help in this effort. In addition, children who have become accustomed to clinical and educational encounters with adults often adopt the "learner" role automatically. They may wait passively for instructions and assume that they will have minimal input into the choice of activities (Ripich and Panagos, 1985).

Cochran and Bull (1985) suggest that computer-based activities can serve to make children more active participants in language therapy. They propose a 3-way model of interaction, in which the child and clinician use the computer together to accomplish a shared goal. Although a variety of educational software packages may be used to develop an ongoing computer-based context for therapy, some Logo activities seem to be especially well suited to this purpose (Cochran and Bull, 1985; Bull, Lough, and Cochran, in press).

Logo activities can be designed to provide feedback to the child and clinician, so that they can see immediately whether or not the instructions they gave were clear and complete. Often, they may find that what they told the computer to do and what they really wanted it to do, did not match. This provides opportunities to discuss options and practice clarification. Although this can happen without including speech output in the Logo activity, speech is an especially desirable enhancement with young or cognitively delayed children.

With motorically handicapped children, it is sometimes desirable to include robots or other three dimensional objects (such as floor turtles) in computer activities. These children can use instructions to the computer to manipulate the robots and move them through space. The robots (through Logo and a speech synthesizer) can provide not only visual feedback, but also verbal feedback to the child.

IV. Synthesized speech and teaching specific academic skills: reading and writing.

Software designed specifically to provide or enhance instruction in reading and writing has now begun to incorporate synthesized speech. The _Writing to Read_ (Martin, 1982) and _Listen to Learn_ (Rosegrant and Cooper, 1985) software developed and distributed by IBM are examples of widely field-tested software packages which make use of speech output. In the case of this software, speech is produced by a special speech card inside the computer, or the Echo PC (identical to the Echo GP except for cable).

Listen to Learn combines synthesized speech with a word processing program. Such software is being used successfully not only in regular classrooms, but also with multiply-handicapped pre-schoolers, school-aged learning-disabled students (Rosegrant, 1984, 1986a, 1986b) and in cognitive rehabilitation with brain-injured adults.

Having a way to "hear" what they write is empowering for beginning readers and writers and also for language-impaired readers and writers. With a program like Listen to Learn, students can ask the computer to "read" to them upon demand, over and over if necessary. The combination of synthesized speech with word-processing facilitates not only early literacy skills in young handicapped children, but also improved speech and language performance (Rosegrant, 1984). Because a word-processing program is open-ended, unlike tutorial or drill-and-practice software, activities can be personalized. Vocabulary and language structures appropriate for the special needs of individual children or adults can be incorporated. In addition, as in the Logo activities described above, the handicapped user experiences control of the activity.

V. Synthesized speech as a tool for the exploration of language.

Although basic literacy skills may be the most obvious educational application of synthesized speech, especially in connection with a word-processing program, other objectives are possible as well. Either through a word-processor or a programming language such as Logo, learners can experiment with and explore language independently. Spelling patterns and word modifications (such as pre-fixes and suffixes) are examples of language structures which can be manipulated. Synthesized speech can provide auditory feedback to the student, resulting in an interactive learning activity.

Chapter Summary

The ready availability of speech synthesizers and appropriate software encourages clinicians and educators to assess the effects that the addition of aural language has on computer-based instruction (e.g., Miller, 1984). Many professionals are called upon to make recommendations concerning the allocation of funds for the software and hardware required for "talking" computer programs. There are several factors to be considered in choosing or recommending such a system, and in evaluating the effect of adding speech to instruction. Ironically, one important but frequently overlooked factor in both cases is the intelligibility of the speech produced by the synthesizer under consideration. Improved intelligibility was among the objectives which motivated changes in the technology of artificial speech production over the past 50 years. These changes also reflect an interdisciplinary approach to the study of speech production and analysis. New technologies have resulted in text-to-speech synthesizers which can make important contributions to remediating and compensating communication impairment. The most prominent uses of synthesized speech to enhance communication and the development of communication skills, were described.

The clinical and educational applications of synthesized speech described above involve a wide variety of users, settings, and objectives. In many instances the users of this technology are persons with impaired and/or yet-to-be acquired language skills. How does the quality of the speech output in these applications affect the success of the user? How intelligible is the speech produced by the synthesizers most commonly used in schools and clinics? Although to a large extent these questions remain unanswered, the information available thus far is reviewed in Chapter Two.

Chapter Two
A Review of Related Literature

Interdisciplinary History

Mones E. Hawley (1977), editor of a collection of key papers in speech intelligibility, presents an insightful history of this interdisciplinary field. He points out that although interest in speech and speech perception is as old as speech communication itself, scientific inquiry into the conditions which hamper or enhance speech transmission did not really begin until after the invention of the telephone. From that time till the present, according to Hawley, three primary movements in the field have occurred. The first was spurred on by physicists and lasted from before World War I until half way through World War II. By then, the development of the vacuum tube amplifier had made possible the analysis and improvement of speech transmitted through noise.

Hawley associates the second wave of new information with the development of another piece of equipment, and a different group of researchers. From World War II to the early 1960's, psychologists making use of tape recorders made major contributions to the study of the human perception of speech. During this time, speech audiometry became a reliable technique in hearing evaluation.

Starting in the late 1950's, computer engineers began to make their contributions to the field. They made use of digital computers for controlling experiments, producing speech signals, analyzing speech signals, and analyzing resulting data. Since 1977, when Hawley's observations were made, the computer-based exploratation of speech intelligibility has predictably increased. Interest in speech intelligibility and the assessment of speech transmission developed in parallel with synthesized speech. Before reviewing some of the main issues and findings in speech intelligibility studies, it is important to make some semantic distinctions.

Defining Intelligibility

As in nearly every aspect of the study of language and linguistics, terms in the study of speech intelligibility have frequently been defined and re-defined, mis-used and misunderstood. Terms which occur frequently in the literature in reference to similar, if not identical, behaviors, include: articulation, comprehension, discrimination, identification, intelligibility, perception, and recognition. The roles of speakers and listeners have been confused within this group of terms, as have the objectives of measurement: whether the variable of interest pertains to the quality of the signal or to the integrity of the receiver.

In a frequently cited about the influence of test materials on speech intelligibility measurement, Miller, Heise, and Lichten (1951) are careful to distinguish between speech _audibility_ and _intelligibility_, but deliberately blur any distinction between _intelligibility_ and _discriminability_. In explaining their position, they use several of the culprit words listed above:

> The crux of the difference is that intelligibility involves a complex discrimination and identification, whereas audibility is simply a discrimination of presence or absence. It seems reasonable, therefore, to call a speech unit intelligible when _it is possible for an average listener with normal hearing to distinguish it from a set of alternative units_ (their emphasis, p. 331).

For their purposes, a speech unit can be any combination of "vocal noises," such as phonemes, syllables, words, or phrases. They are clear about excluding listener interpretation from intelligibility, stating that their "definition reduces intelligibility to discriminability, and avoids the questions of semantic rules and meaning" (p. 331).

While the avoidance of semantics as a factor in the intelligibility of speech seems to simplify matters, it is not an acceptable solution for linguists such as Lehiste and Peterson (1959). Reminding readers that intelligibility is an aspect of the process of communication, Lehiste and

Peterson are convinced that "intelligibility differs from recognizability" (p. 280). They suggest that although a sound or speech event may be recognized as a phoneme, phonemes themselves are "signals" rather than "symbols". They prefer to ascribe the characteristic of intelligibility only to linguistic units which can carry meaning: morphemes, words, and larger units.

There is considerable precedent, however, for thinking of the intelligibility of a speaker or system in terms of individual speech sounds -- Alexander Graham Bell described the "intelligibility" of his first telephone as follows: "Indeed, as a general rule, the articulation was unintelligible except when familiar sentences were employed. . . . The elementary sounds of the English language were uttered successively into one of the telephones and its effects noted in the other. Consonantal sounds, with the exception of L and M, were unrecognizable. Vowel sounds in most cases were distinct " (Hawley, 1977, p. 2).

For his own editorial purposes, Hawley (1977) defined intelligibility as "the recognizability of a speech stimulus (a basic speech sound, word, sentence) and the response to it by repeating it, writing it down, choosing it from alternatives offered, or stating that the listener recognized it" (p. 2). He does not use the term "speech discrimination" separately. His is closer to the view of

Miller, Heise, and Lichten (1951) than to Lehiste and Peterson (1959).

Owens and Schubert (1968), however, make a useful distinction between speech discrimination and intelligibility. They suggest that speech discrimination be used when the purpose is to refer to an aspect of listening ability, or, more specifically, impaired listening ability. They use speech intelligibility to refer to the effectiveness of a system which connects unimpaired speakers and listeners.

In the present document, the Owens and Schubert (1968) convention will be slightly modified. Speech discrimination will refer to the assessment of a listener's ability to identify speech sounds and combinations of sounds. Intelligibility will refer to the ability of a speaker (or system) to produce the speech sounds of a language, individually or in combinations, in such a way that they can be understood by other speakers of that language. Listeners may indicate that the message was intelligible by responding in a number of ways, the most common of which are repeating the message, writing it down, or choosing it from an array of alternatives.

Intelligibility studies of the past can be divided into groups according to the "speaker" being assessed. For example, intelligibility has been a diagnostic (and prognostic) indicator for various communication disorders such as dysarthria, hearing impairment, or cleft palate. Some of the issues in intelligibility testing have a history in the study of these disorders, and will be discussed below. Some issues related to the design and implementation of intelligibility studies come from extensive work on assessing communication between normal speakers and listeners under adverse conditions (such as noise). As was indicated above, even before Alexander Graham Bell, researchers were interested the intelligibility of speech produced or transmitted by non-human mechanisms. Highlights from this body of literature will also be discussed below, especially as they relate to factors known to affect the measurement of speech intelligibility. The intelligibility studies related to the current generation of speech synthesizers for microcomputers will be discussed with emphasis on findings related to the synthesizers used in the present study.

Factors Which Influence Intelligibility

The factors which influence the intelligibility of speech include:

- Listening conditions (including equipment, physical setting and acoustic conditions)
- The task (how the listener is expected to signal understanding).

The influence of the task and listening conditions upon natural speech intelligibility measurement have been the subject of much research and must be considered in the design of synthesized speech studies.

I. The task

Speech intelligibility tasks vary according to the unit of speech which is used for stimuli (individual sounds, syllables, words, phrases, etc.) and the mode of response required. Response modes (repetition vs. writedown, for example) vary in their complexity for listeners and also for researchers who must then score or otherwise evaluate responses. An ideal task for speech intelligibility testing, as for any other research, minimizes the influence of extraneous variables on subject performance, and minimizes the likelihood of scorer error.

The tendency toward scorer bias in talkback response tasks has been observed by several investigators comparing methods of speech discrimination testing (Merrell and Atkinson, 1965; Nelson and Chaiklin, 1970). In a typical talkback task, a single word is presented (either via live voice or tape) and the subject is asked to repeat the word. One disadvantage of this task is that scorers may have a tendency to judge questionable responses as being correct rather than incorrect. Comparing talkback vs. writedown tasks, and inexperienced vs. experienced examiners, Nelson and Chaiklin (1970) concluded that even with experienced examiners there was a likelihood that scoring bias would occur.

On the basis of scoring reliability, then, it would seem that a task which requires a listener to write down what he thought he heard would be preferable to a talkback task. Writedown tasks can be in the form of open-response tasks, such as fill-in-the-blanks, or closed-response formats, such as multiple-choice. As Black (1957) points out, a multiple-choice format may have some experimental advantages over open-response tasks. He suggests that the single most important advantage of a multiple-choice form is ease and reliability of scoring. According to Black (1957), the results of writedown tasks depend more on the linguistic sophistication of both the subject and the examiner, than do results of multiple-choice tasks. Written responses must be analyzed carefully and with some degree of phonetic understanding on the part of the scorer. Because more judgement is required, there is more room for error.

In a comparison of an open-response and multiple-choice form of an intelligibility test, Black (1957) reported higher absolute scores (12 percentage points) with the multiple-choice form. This is as would be expected, and is in agreement with the observation of Miller, Heise, and Lichten (1951) that as the range of alternatives increases, so does the amount of information required for identification of an individual item. Black (1957) found that although absolute scores were higher for the multiple-choice task, the change in scores as a result of change in signal level (64 to 24 db re .0002 dyne/cm2) yielded the same slope as for the writedown task. He also found no difference between the writedown test and the multiple-choice test in split-half reliability. Black (1957) points out that it is easier to give directions for an open-response test, but that multiple-choice tests are faster to administer and easier to score. In addition, a multiple-choice format "makes possible the study of confusion characteristics among the fixed population of words" (p. 224).

No matter what the response mode or test format is, the content and context of the speech material tested is also an important factor in intelligibility testing. This was dramatically demonstrated by Miller, Heise, and Lichten (1951). Figure 2-1 is taken from their paper describing intelligibility as a function of the context of test materials. It illustrates the relative intelligibility of digits, words in sentences, and nonsense syllables (all in writedown tasks) presented to the same subjects under the same listening conditions.

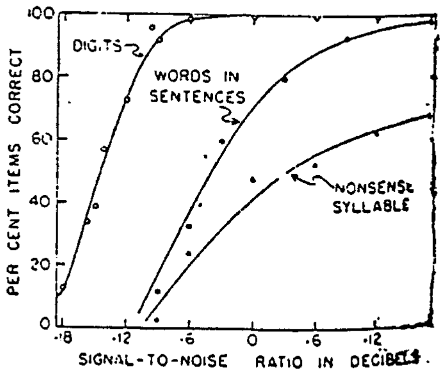


Figure 2-1. Effect of task and contextual information on intelligibility. Note. From "The Intelligibility of Speech as a Function of Test Materials" by G.A. Miller, G.A. Heise, and W. Lichten, 1951, Journal of Experimental Psychology, 41, p. 330).

Their results led Miller, Heise, and Lichten (1951) to conclude that the "most important variable producing the differences [in intelligibility] is the range of possible alternatives from which a test item is selected" (p. 331). They point out that listeners can accurately guess at digits with less information than is needed to identify a nonsense syllable, for which knowing one phoneme is no help in predicting what the other(s) might be. A similar narrowing of the possibilities occurs when words are tested in the context of meaningful sentences instead of isolation; the syntactic rules of English make the range of possible alternatives smaller, and so scores are higher (Miller, Heise, and Lichten, 1951; Theodoridis and Schoeny, 1982; Theodoridis, Schoeny, and Anne, 1985; Trual and Black, 1965).

Speech discrimination testing is an integral part of audiometric assessment. The objective of such tests is to examine a listener's ability to discriminate among similar sounds or among words that contain similar sounds (Newby, 1964). There have been several sets of speech discrimination materials developed for audiometric purposes such as the evaluation of auditory discrimination ability in noise or hearing aid selection. So-called phonetically balanced (PB) word lists are most often used as the basis for these speech materials.

In clinical settings probably the most widely used PB lists are those that were initially developed at the Psycho-Acoustic Laboratory at Harvard University (Egan, 1948) and subsequently revised at the Central Institute for the Deaf (CID) (Hirsh, Davis, Silverman, Eldert, and Benson, 1952). These lists are commonly referred to as the CID W-22 lists. There are 24 equivalent lists, each list consisting of 50 monosyllabic words chosen so that together they present each speech sound with the same frequency as it occurs in English. There have been many discussions of the desirable and undesirable features of these lists, and many attempts to devise improved alternatives (e.g., Black, 1952, 1957, 1968; Bull, Ruth, and Schoeny, 1979; Fairbanks, 1958; Giolas and Epstein, 1963; Lehiste and Peterson, 1959; Rintelman, Schumaier, and Jetty, 1974). For audiometric purposes, the lists are most often used in a task which requires the listener to either write the word in a blank or say it back to the examiner (see Nelson and Chaiklin, 1970, for a comparison of the these two methods).

The CID W-22 lists, as well as other materials especially designed for auditory discrimination testing, have frequently been used as test materials in the evaluation of speech intelligibility. They have been used to measure the intelligibility of a wide variety of speakers, including normal speakers (e.g., Giolas and Epstein, 1963) hearing impaired speakers (e.g., Sitler, Schiavetti, and Metz, 1983), and speech synthesizers (e.g., Chial, 1976; Schwab, Nusbaum, and Pisoni, 1985). Similarly, the Modified Rhyme Test (MRT) (Fairbanks, 1958; Kreul, et al., 1968) has been used interchangeably as a speech discrimination test for hearing impaired listeners (e.g., Northern and Hattler, 1974) and as the "de facto" test of segmental intelligibility for synthesized speech (Pisoni, Nusbaum, and Greene, 1985).

Even using speech discrimination materials for their intended purpose, however, is not foolproof. Northern and Hattler (1974) compared the performance of a hundred hearing impaired and 25 normal hearing subjects on four popular speech discrimination tests, including the the CID W-22 words, Egan's (1948) original lists (labelled PAL PB-50 words), the MRT, and a sentence identification task (Speaks and Jerger, 1965). Their review of the literature and the results of their study lead them to conclude that "Speech discrimination measurement in audiologic clinics suffers appreciably from a lack of standardized test materials and procedures" (p. 33). They identified problems similar to those which will be discussed below in regard to intelligibility testing.

It is important not to assume that the same speech materials and tasks are equally appropriate and effective for assessing speech discrimination and intelligibility. The possible interaction between materials, the speaker, and the listener, must be considered. For example, some sounds and sound combinations have been found to enhance the intelligibility of words (Black, 1952). In an analysis of 3697 moderately intelligible words from the Thorndike lists, Black (1952) found that [s] blends such as [sl], [st], [sp], and [sm] tended to make words more intelligible to normal listeners. On the other hand, most /s/ blends tend to be acquired late in speech development, and are more likely to be produced incorrectly by young children than many other single consonants or consonant blends. Thus the structural characteristics which facilitate auditory discrimination of words do not necessarily facilitate their articulation.

Black (1952) identified a relationship between several characteristics of words and their intelligibility. He identified eighteen sounds which tended to enhance the intelligibility of words, and seven sounds ([c, ou, p, f, th, h, l]) which apparently deterred word intelligibility. Since the issue of phonetic and/or phonemic balance has been central to many of the discussions of the PAL PB-50 and CID W-22 lists (e.g., Lehiste and Peterson, 1959) the possible unequal influence of some sounds is particularly interesting.

Black found that words which occur more frequently in English were more intelligible, even among generally familiar words. Also, words with many sounds and words with more than one syllable tended to be more intelligible than words with few sounds or words with one syllable. Interestingly, however, single syllable words with few sounds tended to have higher familiarity ratings than more complex words. Black concluded that, "Thus two contrary influences, word familiarity and word complexity, appear to operate in the auditory recognition of a word somewhat independently of the phonetic content. The prediction of word intelligibility from phonetic content alone, as may be feasible in the instance of nonsense syllables, becomes virtually impossible" (1952, p.417). These observations confirm the importance of the content of materials in intelligibility testing. It seems likely that even well-designed materials will not be equally sensitive to influences on listener performance (discrimination) and influences on speaker performance (articulation).

II. Listening Conditions

The most common variable in the listening conditions of intelligibility studies is noise. The presence of background noise has been frequently manipulated in assessing the intelligibility of speech in a variety of tasks (Black, 1952, 1957, 1968; Black and Agnello, 1964; Greene, Logan, and Pisoni, in press; Theodoridis, Schoeny, and Anne, 1985).

Other factors which have been shown to be important in the listening conditions of intelligibility testing include signal level (Black, 1957; Theodoridis, Schoeny, and Anne, 1985), sound field vs. presentation via headphones, and the equipment or recordings used for stimulus presentation (Black, 1957; Giolas and Epstein, 1963).

Intelligibility of Synthesized Speech

I. Influences on listener performance

For the past several years, researchers at the Speech Research Laboratory at Indiana University have been studying the perception of "synthesis-by-rule" synthetic speech (Greene, Logan, and Pisoni, in press; Luce, Feustel, and Pisoni, 1983; Pisoni, Nusbaum, and Greene, 1985; Salasoo and Pisoni, 1985; Schwab, Nusbaum, and Pisoni, 1985; Slowiaczek and Nusbaum, 1983). Although there are other individual researchers working in the area of intelligibility of synthesized speech (e.g., Chial, 1976; Clark, 1983; Klatt, 1983; Kraat and Levinson, 1984; and Voiers, 1977, 1984), no other person or group has established a line of research which systematically addresses the intelligibility of synthesized speech.

The Indiana University group has collected and published behavioral data based on responses to eight different text-to-speech systems (Greene, Logan, and Pisoni, in press). The synthesizers evaluated thus far represent a wide range in cost and availability, from proto-types of systems costing many thousands of dollars to commonly available systems costing a few hundred dollars. To some extent, their collection is also a chronological representation of what has been developed and improved in text-to-speech systems in the last few years. The synthesizers used in their research include: Berkley Systems Works, DECtalk V1.8 (Digital Equipment Corporation), Echo (Street Electronics), Infovox SA 101, MITalk-79, Prose 2000 V3.0 (Speech Plus), TSI Prototype-1 of the Prose 2000 (Telesensory Systems, Inc.), and the Type 'n 'Talk (Votrax, Inc.).

Some of the factors known to affect the results of natural speech intelligibility studies were discussed in the previous section. Task complexity, linguistic context, and listening conditions are also considered to be constraints on the performance of human observers in synthesized speech studies (Greene, Logan, and Pisoni, in press). Greene, Logan, and Pisoni (in press) identify three more factors which may have particular implications for assessment of speech synthesizers:

- Short-term memory limitations of human listeners
- Signal characteristics of the synthesized speech
- Previous experience of listeners.

The results of two Indiana studies (Luce, Feustel and Pisoni, 1983) have implications pertaining to short-term memory and the processing of synthesized speech. The first study involved the visual presentation of digits followed by the auditory presentation of either natural or synthesized words. The examiners found that subjects recalled more of the digits and more of the words presented in the natural speech condition. In addition, they observed an interaction between the length of the digit lists recalled perfectly, and the type of speech presented (as the length of digit lists increased, significantly fewer subjects who listened to synthesized words were able to recall all the digits). They suggest that these results demonstrate that "synthetic speech requires more short-term memory capacity than natural speech" (Pisoni, Nusbaum, and Greene, 1985, p. 1672).

In a related study, Luce, Feustel, and Pisoni (1983) found an interaction between subjects' ability to recall synthesized or natural words in lists, and the position of words in the lists. Words heard first in synthesized lists tended to be recalled less accurately than words heard first in natural lists. This suggests that synthesized words heard later were interfering with the active rehearsal of earlier words. The investigators point out that the increased processing demands created by synthesized speech "may place important perceptual and cognitive limitations on the use of of voice response systems in high information load conditions or severe environments" (Pisoni, Nusbaum, and Greene, 1985, p. 1673).

The signal characteristics of synthesized speech are considered to be another constraint on listener performance. The acoustic-phonetic and prosodic characteristics of synthesized speech are unlike the acoustic-phonetic and prosodic characteristics of natural speech (Pisoni, Nusbaum, and Greene, 1985). Natural speech reflects the constraints imposed by linguistic rules and the acoustic properties of the vocal tract. Synthesized speech "is an impoverished signal representing phonetic distinctions with only a limited subset of the acoustic properties used to convey phonetic information in natural speech" (p. 1667). Results of intelligibility studies which suggest some possible distinctions between the acoustic-phonetic features of natural and synthesized speech will be discussed below in more detail.

Besides short-term memory contraints and speech signal characteristics, another factor which may particularly effect the intelligibility of synthesized speech is previous listener experience or training. Some research suggests that relatively little practice can improve listeners' scores in response to synthesized speech (Chial, 1976; Schwab, Nusbaum and Pisoni, 1985). This will be discussed in more detail below.

II. Natural vs. Synthesized Speech

The intelligibility of synthesized speech can be evaluated at several levels, depending on the linguistic unit of interest. Studies designed to measure intelligibility at the phoneme level most often make use of the Modified Rhyme Test (MRT), originally developed by Fairbanks (1958) and then modified to include a more complete set of phonemes by House, Williams, Hecker, and Kryter (1965). The MRT requires the subject to listen to a consonant- vowel-consonant (CVC) word and then choose it from a set of six alternatives which differ by a single consonant phoneme. In this form it is a multiple-choice, closed-response task, although it has also been administered in an open-response form (Greene, Logan, and Pisoni, in press).

In their comparison of natural speech and the eight synthesizers listed earlier, the researchers at Indiana University have obtained baseline data using the MRT (Greene, Logan, and Pisoni, in press; Pisoni, Nusbaum, and Greene, 1985). Figure 2-2 shows results based on both the closed and open response format of the MRT.

<< Insert Figure 2-2 About Here >>

Figure 2-2. MRT error rates for closed and open formats, for eight speech synthesizers and natural speech. Note. From "Perception of Synthetic Speech Produced Automatically by Rule: Intelligibility of Eight Text-to-Speech Systems" by B.C. Greene, J.S. Logan, and D.B. Pisoni, in press, Behavior Research Methods, Instruments, and Computers.

The error rate was lowest for natural speech, being less than 1% in the closed response task, and about 3% when subjects were asked to write down the word they heard in an open format. Results for the eight synthesizers range from the DECtalk v1.8 (voice called "Paul") with a closed response error rate of a little less than 4%, to the Echo with a closed response error rate of about 36%.

Pisoni, Nusbuam and Greene (1985) point out that even in the open response format, natural speech was more intelligible (97.2 % correct) than any of the synthesizers were with a closed format (DECtalk v1.8, Paul: 96.7 % correct). It can also be seen from Figure 2-2 that the effects of an open-response format are greater for synthesizers which already had high error rates for the restricted response task. For example, the difference between the closed and open format error rates for the DECtalk v1.8 (Paul) was about 10% (from 4% to 14% incorrect). For the Echo synthesizer, however, the difference was about 40% (from 36% to 76% incorrect). The authors suggest that this indicates that when speech is less intelligible, listeners rely more heavily on contextual cues, or, in this case, response-set constraints. This is in good agreement with recent findings pertaining to the measurement of the contribution of context in speech perception studies (Theodoridis and Schoeny, 1982; Theodoridis, Schoeny, and Anne, 1985). The evaluation of these synthesizers compared to natural speech indicates that, as measured by the MRT, natural speech is still more intelligible than even the best synthesized speech.

III. Decoding synthesized speech

Pisoni, Nusbaum, and Greene (1985) describe two hypotheses which have been postulated to account for the greater difficulty which listeners have in "encoding synthetic speech." Throughout this otherwise excellent article, the authors use the word "encode" to describe the listener's ability to identify, process, or recognize various aspects of a speech signal. This is precisely what is usually meant by the opposite word, "decode", in much of the literature pertaining to communication disorders. It is a more transparent use of the word "encode" to refer to what a message sender does, and "decode" to refer to what a message receiver does. Therefore, throughout the remaining discussion, "decode" will be used to refer to the way in which listeners perceive and interpret speech signals.

Listening to synthesized speech can be considered comparable to listening to natural speech in noise (Clark, 1983). This hypothesis would suggest that the acoustic-phonetic cues which are present in natural speech but obscured in noisy listening conditions, are similarly present but obscured in synthesized speech.

A second hypothesis, preferred by Pisoni, Nusbaum, and Greene (1985), proports that synthesized speech is not like degraded natural speech: "By this account, synthetic speech is fundamentally different from natural speech in both degree and kind because many of the important criterial acoustic cues are either poorly represented or not represented at all" (Pisoni, Nusbaum and Greene, 1985, p. 1670). Pisoni, Nusbaum and Greene, 1985, describe the results of a study designed to investigate these hypotheses (Nusbaum, Dedina, and Pisoni, 1984).

It was predicted that if listening to synthesized speech is like listening to "noisy" natural speech, patterns of errors would be similar in response to both natural speech in noise and to synthesized speech. Consonant-vowel (CV) syllables were used as stimuli. Results were obtained for natural speech at several signal-to-noise (S/N) ratios, and were compared to the DECtalk v1.8, the Speech Plus Prose-2000 v2.1, and the Votrax Type'n'Talk. Errors made in response to each synthesizer were compared to errors obtained in response to natural speech presented at a S/N ratio which produced a comparable overall score. The error patterns were different between the synthesized speech and the natural speech, and also between the synthesizers. For example, although listeners rarely confused /b/ and /r/ when listening to natural speech (even in the poorest condition, -10 dB S/N), this confusion accounted for 100% of the errors made in the identification of /r/ phoneme in response to DECtalk v1.8 (Pisoni, Nusbaum, and Greene, 1985, p. 1671).

Several other differences in error patterns were observed between natural speech in noise and synthesized speech. Errors for stop consonants, however, were similar for the Votrax Type'n'Talk and natural speech in noise. Some consonants which have acoustic-phonetic similarity were easily confused with the Type'n'Talk, as was the case with natural speech in noise. However, results for liquids and glides (r,l,w,j) were different. For natural speech, relatively few glides or liquids were confused with stop consonants, which are acoustically dissimilar. This confusion, however, accounted for the largest number of errors for glides and liquids with the Votrax.

The results from this study suggest that some consonant identification errors in response to synthesized speech may be due to the similarity of the acoustic-phonetic characterstics of those consonants, as happens with natural speech in noise. Other errors suggest that listeners are responding to miscues present in the synthesized speech signal. These errors cannot be predicted by a hypothesis which considers synthesized speech to be comparable to natural speech in noise (Pisoni, Nusbaum, and Greene, 1985).

IV. Practice Effects and Synthesized Speech

Several investigators have been interested in the effects of listener training on the intelligibility of synthesized speech (Chial, 1976; Nusbaum and Schwab, 1983; Schwab, Nusbaum, and Pisoni, 1985). Interestingly, these studies have all used versions of a Votrax speech synthesizer. On the whole, they suggest that the performance of normal listeners in response to synthesized speech can be significantly improved with practice.

As part of a series of experiments involving a Votrax VI, Chial (1976) evaluated learning effects across eight trials in which human and synthesized presentations of test lists were alternated. Stimulus items were drawn from random orderings of the CID W-22 phonetically balanced word list 1, presented monoaurally via headphones. Chial observed an improved performance (10-20% improvement) for synthesized lists as a result of practice, but not for human speech (1%).

Schwab, Nusbaum, and Pisoni (1985) were particularly interested in controlling for practice effects, in order to distinguish between the effects of listening to synthesized speech and the effects of greater familiarity with the task. They wanted to know whether changes in scores were attributable to improvements in the ability of subjects to process synthesized speech. Test materials included the MRT, words from PB lists, and both meaningful and anomolous sentences. The Votrax Type'n'Talk synthesizer was used for the synthesized speech tasks. According to the experimenters, the Votrax Type'n'Talk was chosen "primarily because of the poor quality of its segmental synthesis." Since low scores could be expected initially, they hoped to prevent ceiling effects from obscuring training effects. Their experiment involved three groups of subjects who were pre-tested on Day 1 and post-tested on Day 10 with synthesized speech stimuli. In the intervening days, one group received practice listening to synthesized speech, one received identical natural speech practice, and one group received no training.

Performance improved dramatically only for the group which had received synthetic speech training. In addition, their post-test scores were significantly higher than those for both other groups. The experimenters found that six months later, with no further contact with synthesized speech, their subjects had retained the training. They concluded that 1) subjects had acquired detailed information about the rule system being used to generate the synthesized speech, and that 2) human listeners can modify (and improve) the perceptual strategies they use to decode even poor-quality synthetic speech, with relatively little training (Schwab, Nusbaum, and Pisoni, 1985).

V. Language Impaired Listeners

Very few reported studies have compared the intelligibility of natural speech to synthesized speech with any population other than normal adults (Rentschler, 1985; Sevik and Romski, 1985). Sevik and Romski (1985) reported the results of an experiment involving nonspeaking severely retarded individuals who were learning new vocabulary symbols in a non-verbal communication system. A Votrax speech synthesizer was used, and subjects did learn new symbols in response to it. Further details of procedures and results are not available.

Forty-seven language impaired children served as subjects in a comparison of the Votrax (model not reported) speech synthesizer and live human voice (Rentschler, 1985). These school-aged children completed the _Goldman-Fristoe-Woodcock Test of Auditory Discrimination_ (GFW), (Goldman, Fristoe, and Woodcock, 1970) two times. The GFW requires listeners to match the target word to a picture, choosing from an array of 4 in each picture plate (e.g., sack, shack, tack, stack). They completed it the first time in response to live voice and, and then later in response to the Votrax speech synthesizer. The children obtained a mean number of errors of 7.21 for natural speech and 13.08 in response to the Votrax.

Comparing Speech Synthesizers

As Pisoni, Nusbaum, and Greene (1985) point out, in the past there have been very few studies dealing with the intelligibility of synthesized speech or the technical issues and problems which surround such research (p. 1666). This is in spite of the increasing need for a systematic and reliable method of evaluating the quality of speech synthesizers. Speech synthesis systems are becoming easier to obtain and easier to use; at the same time, some of their features are becoming more complex and sophisticated. For example, the ability to synthesize more than one voice or more than one language is now available in some new systems (Pisoni, Nusbaum, and Greene, 1985).

At the present time it is difficult to describe the comparative performance of a new system or measure the value of changes to the way speech is synthesized. This difficulty is due in part to the necessity of depending on human listeners for feedback. As Pisoni, Nusbaum, and Greene (1985) point out, "Unfortunately, there is no existing method for automating the assessment of synthetic speech quality . . . The perception of speech depends on the human listener as much as it does on the attributes of the acoustic signal itself and the system of rules used to generate the signal" (p. 1665). In addition to the need for objective assessment in controlled laboratory conditions, there is a need for information on the comparative performance of speech synthesizers in less ideal surroundings.

Despite the obstacles inherent in the task, some researchers have endeavored to compare the quality of speech synthesizers on the basis of intelligibility (Clark, 1983; Greene, Logan, and Pisoni, in press; Kraat and Levinson, 1984). This work will be discussed particularly as it relates to the intelligibility of two brands of synthesizers commonly used for clinical and educational applications: Votrax and Echo.

The synthesizers under discussion can operate in a text-to-speech mode. As was explained in Chapter One, text-to-speech means that the synthesizer can convert normal text, in the form of ASCII code, to synthesized speech. This allows for the synthesis of an unlimited number of sounds, words, and phrases. According to Pisoni, Nusbaum, and Greene (1985), there are three components of a text-to-speech system which might directly affect the intelligibility of speech output:

1) the spelling-to-sound rules
   (sometimes called letter-to-sound rules)

2) the derivation and implementation of suprasegmentals
   (pitch, intonation, timing, stress, etc.)

3) the phonetic implementation rules which convert the
   internal representation of phonemes or allophones (from 1
   above) into a speech signal.

Except for some variation in the internal amplifiers and speakers which present the final speech signal to the listener, most of the technical differences in speech synthesizers are accounted for by these three components.

The Echo text-to-speech system converts text into allophonic control codes using an algorithm developed at the Naval Research Laboratory (cited in Morris, 1979). In the Echo used for the Indiana University studies, these codes were then converted to speech using linear predictive coding (LPC) by a TMS-5200 chip (Greene, Logan, and Pisoni, in press). Currently, the Echo II (Apple compatible) Echo-GP (general purpose), and Echo-PC (personal computer: IBM compatible) synthesizers make use of a Texas Instruments' TMS 5220 chip (Chial, 1985).

The Votrax Type'n'Talk converts text into phoneme control codes through a text-to-speech translator module. These codes are converted to speech using a formant synthesis technique, by the Votrax SC-01 phoneme synthesizer chip (Greene, Logan, and Pisoni, in press). The Votrax Personal Speech System (PSS), uses a Votrax SC-01A chip (Chial, 1985). The Votrax systems can be accessed by most microcomputers, either through a serial or parallel interface.

Most of the results from the Indiana University studies described above are based on the Modified Rhyme Test (MRT). It is important to point out again that the MRT in its usual closed-response format evaluates segmental intelligibility. It should also be mentioned that, like most materials used to evaluate synthesized speech, it was not designed for this purpose. Target sounds are tested in either the initial or final position of monosyllabic, CVC words. The intelligibility of single phonemes is tested in the context of words with minimal phonemic contrasts (got/dot/cot, etc.). Therefore, it can be argued that the MRT data used to compare speech synthesizers reflects differences at the level of individual speech sounds rather than words.

Some information about how the Votrax Type'n'Talk and the Echo compare with natural speech and with more sophisticated (and expensive) synthesizers was presented earlier (see Figure 2-2). The error rates observed in the closed response format were 27.44% for the Type'n'Talk and 35.56% for the Echo (Table 2, Greene, Logan, and Pisoni, in press). There was about 8% difference in the segmental intelligibility scores of the Echo and the Type'n'Talk. Together, these two synthesizers make up the "low-quality synthetic speech" group in the Indiana University studies. The ranking of the synthesizers by intelligibility reflects, according to Greene, Logan, and Pisoni, "the adequacy of the phonetic implementation rules used in the individual text-to-speech systems which in turn is directly related to the amount of speech knowledge incorporated into each system "(in press).

Illustrative Experiment: Votrax PSS vs. Echo II

I. Summary of the Study

Although Pisoni, Nusbaum, and Greene (1985) describe some synthesizer comparisons using tasks other than the MRT, such as a word-in-sentences identification task and a paragraph comprehension task, apparently neither the Type'n'Talk nor the Echo were used in this aspect of their work. A task other than the MRT, however, was used by Kraat and Levinson (1984) to compare the Votrax PSS and the Echo II. These two synthesizers were chosen because they are commonly used in augmentative communication systems. The study was designed to compare these two synthesizers based on three measures: intelligibility of sentences, the relative effects of pause time on sentence intelligibility, and the number of alternate spellings required of the user entering text, in order to compensate for mispronunciation of frequently used words.

In the first part of the Kraat and Levinson (1984) study, twenty adults aged 20-65 served as listeners in a sentence intelligibility task. This test consisted of 64 sentences, each eight words in length, selected from the the Assessment of Intelligibility of Dysarthric Speech (Yorkston and Beukelman, 1981). Sixteen sentences were assigned to each of four testing conditions: the Votrax PSS (normal rate), Echo II (normal rate), Votrax PSS (with pauses), and the Echo II (with pauses). In the pause conditions, a 2.5 second pause was inserted between each word of each sentence. In the normal rate condition, word boundaries were marked with the usual single press of the space bar during keyboard entry. Standard spellings were used for the text of the 64 sentences, except in cases where this resulted in a vowel substitution or syllable addition or deletion. In these instances, the spelling of stimulus words was altered in such a way that normal pronunciation was approximated.

Sentences were presented in an open sound field, with a 20 second delay between sentences. Subjects were asked to write what they heard after each sentence presentation. Conditions were randomized for order of presentation. The results indicated that the Votrax PSS in the normal and pause conditions was significantly more intelligible than the Echo II. In the normal rate condition, average scores were 45.7% for the Echo and 70.4% for the Votrax. In the pause condition, average scores were 81.1% and 84.3%, respectively.

In the second part of the Kraat and Levinson (1984) study, five graduate students in speech pathology were asked to judge each synthesizer's pronunciation of frequently occuring words. The 1500 words tested consisted of the 1000 most frequently occuring words in English according to

Thorndike and Lorge (1944), as well as the 500 words most freqently used by persons using augmentative communication devices (Beukelman, Yorkston, Poblete, and Naranjo, 1984). Listeners were asked to judge whether or not the synthesized version of each word (normal spelling) produced a vowel substitution, or added or deleted a syllable.

The results from this part of the study are summarized in Table 2-1.

	Thorndike/Lorge List	Beukelman, et al. List
Echo II	175/1000	55/500
Votrax PSS	45/1000	36/500

Table 2-1. Number of frequently occuring words mispronounced, according to 5 judges. Data from Kraat and Levinson (1984).

The Votrax PSS was judged to produce fewer incorrect pronunciations for each list. It is surprising that on the whole, so many frequently occuring words were judged to be correctly produced by both synthesizers.

This information puts an additional perspective on the impression of poor intelligibility created by the MRT data presented earlier (Note: the Votrax and Echo synthesizers used by Kraat and Levinson were later models). It should be pointed out again that judges in the Kraat and Levinson study were listening for vowel distortions and syllable alterations. Neither of these types of errors are likely factors in the MRT task, which emphasizes initial and final consonant discrimination in single syllable words.

According to Kraat and Levinson (1984), their results suggest that the Votrax PSS is superior overall, but that the Echo II performs comparably when pauses are added between words in sentences. There are several possible confounding factors in their study, which makes it necessary to interpret their results with extreme caution. These factors will be discussed in detail, because they are illustrative of some of the difficulties encountered in research of this type.

II. Threats to validity due to the task and the subjects

There is a need for information about the intelligibility of synthesized speech at the sentence level, and the experimenters drew their test materials from a source which was designed to assess speaker intelligibility. However, as was discussed earlier, open-response formats depend heavily on the linguistic sophistication of the listener (Black, 1957). The verbal and written language competency of the twenty subjects (having a 45 year age range) in the Kraat and Levinson study is likely to have varied widely. In addition, the difficulties of reliably scoring even single-word open-response tests have been suggested in the literature. It is possible that these difficulties were resolved in this study, but the details of scoring and reliability were not reported.

III. Threats to validity due to procedures

1) The test sentences were randomly assigned to the four listening conditions.

Results from other intelligibility studies (Theodoridis, Cochran, and Bull, unpublished data, Greene, Logan, and Pisoni, in press; Pisoni, Nusbaum, and Greene, 1985; Renschler, 1985) suggest that some speech synthesizers distort some phonemes or phoneme combinations more than others. Pisoni, Nusbaum, and Greene (1985) have suggested that listeners are "miscued" by some incorrect or inadequate information present in synthetic speech signals. This means that even speech materials which have been shown to be "equivalent" for natural speech, may be selectively distorted when translated into synthetic speech. In the Kraat and Levinson study, there is no way to tell that the sentences assigned to the Votrax and the Echo were equivalent in listening difficulty. That is, groups of sentences may have contained unequal numbers of phonemes which were particularly susceptible to synthetic distortion or miscuing.

2) When poor text-to-speech conversions occurred, alternate spellings of stimulus words were used.

Data from the second part of the Kraat and Levinson study confirm that using alternate spellings in the sentence intelligibility task is likely to have had an effect which was not equal for both synthesizers. More alternative spellings were probably necessary in sentences used with the Echo II, although this information was not reported. In any case, using alternative spellings as input to a text-to-speech system serves to compensate for some of the characteristics which are under investigation.

3) Conditions were randomly presented to the subjects.

It is not clear from the account available (Kraat and Levinson, 1984) whether subjects were tested individually or as a group, although group testing seems the more likely. Either way, random ordering of conditions, rather than counterbalancing them between groups or individuals, has unpredictable consequences in this study. Practice effects and contextual effects may be disguised. Since other studies (e.g., Schwab, Nusbuam, and Pisoni, 1985) indicate that skills used for decoding synthesized speech may be especially sensitive to practice or training effects, the order of presentation is critical when more than one synthesizer or condition is presented to the same subjects.

IV. Conclusions

Conclusions based on the Kraat and Levinson study must be made with caution. The information from the second part of the study, about relative pronunciation of frequently occuring words, may be more valid than information from the sentence intelligibility experiment. However, since the judges were not assessing overall correctness of pronunciation, but rather the presence of particular distortions, this data should be interpreted with care also.

Chapter Summary

The relatively few studies which have evaluated the intelligibility of synthesized speech have done so primarily under ideal laboratory conditions using normal adult listeners. A review of the literature suggests the following:

1) As yet, the segmental intelligibility of even the best synthesized speech is significantly worse than natural speech. A relatively few consonants, however, account for most segmental confusions on the part of normal adult listeners.

2) Some errors in the perception of synthesized speech, especially poor quality synthesized speech, are similar to errors which occur in the perception of natural speech

in noise. The prevalence of errors which do not match this pattern, however, suggests the presence of "miscues" in some synthetic speech signals.

3) There are no standardized methods or materials developed particularly for assessing the intelligibility of synthesized speech. Most studies have made use of test materials originally designed to reflect inadequacies in listener perception (speech discrimination) rather than speaker performance (articulation, or speech intelligibility).

4) Studies to date have based evaluations of synthetic speech intelligibility primarily on data which reflect contrasts in single phonemes in the context of monosyllabic words.

5) The listening task is an important factor in the assessment of the intelligibility of synthesized speech; contextual cues influence performance more as the quality of speech decreases.

6) The synthetic speech decoding skills of normal listeners may be susceptible to change as a result of even brief periods of exposure and/or training.

7) In general, the intelligibility of speech synthesizers is correlated with their price. This may become less predictable, however, as the variety of special features is extended and multiple synthesizers within a given price range become available.

8) Evidence suggests that compared to natural speech, decoding of synthetic speech places more cognitive and short-term memory demands on listeners. This implies that low-cost/ poor quality synthetic speech should only be used in applications in which there is little competing stimuli, and other task requirements are not severe.

Educational and clinical applications of synthesized speech, however, tend to involve situations that are cognitively demanding and in which auditory, visual, and tactile stimuli are competing for attention. Because cost will be a major factor in most school and clinic purchase decisions, synthesizers at the lower end of the price range will most often be considered. More information about the intelligibility of low-cost speech synthesizers is needed, particularly at levels beyond individual speech sounds in monosyllabic words.

Chapter Three
Experimental Design

A review of the literature pertaining to the intelligibility of synthesized speech reveals that there is still a lack of information about the intelligibility of low-cost, commonly used synthesizers for small computers. Information about the segmental intelligibility of early models of some relatively inexpensive synthesizers (such as the Echo and the Votrax Type'n'Talk) is available. Very little information about word, sentence, or contextual intelligibility has been established. It would not be expected that every model of synthesizer would be evaluated in the literature. However, the intelligibiliy of at least those which are likely to be widely purchased for clinical and educational use should be assessed.

The literature also reveals several factors which should be considered in the design of this type of research. These factors include:

- The task and test materials
- Listening conditions
- Practice and contextual effects.

This chapter discusses these factors specifically as they relate to the design of the present study.

Research Questions

The questions addressed by the present study are as follows:

1) Is the intelligibility of words different for the Votrax PSS and the Echo GP, when operated in the text-to-speech mode?
2) Does listening to more than one type of speech synthesizer influence the performance of subjects in intelligibility testing?
3) To what extent is the word intelligibility of human speech different from that of the Votrax PSS and Echo GP speech synthesizers under identical listening conditions?

Equipment: Choice of Synthesizers

There are several reasons which make it important to obtain objective data pertaining to the intelligibility of the Votrax Personal Speech System (PSS) and the Echo GP (general purpose). These specific synthesizers have been chosen because:

1) The Votrax PSS and the Echo GP have not been objectively compared before.
2) Of the various available models of Votrax and Echo synthesizers, these are the only ones which can be used with several different computers.
3) The clinical and educational software already developed suggests that they are likely to be available and in

demand for a long time.

4) The Votrax PSS and Echo GP are representative of the low end of the cost continuum, and are therefore most likely to be considered for purchase in clinics and schools.

Some of the research previously discussed in Chapter Two pertained to either a Votrax (Chial, 1976; Pisoni, Nusbaum, and Greene, 1985; Rentschler, 1985; Schwab, Nusbaum, and Pisoni, 1985) or an Echo speech synthesizer, or both (Greene, Logan, and Pisoni, in press; Kraat and Levinson, 1984; Wilson, 1986). In most instances, these researchers made use of earlier models of these synthesizers, which used different microprocessor chips to synthesize speech. There is no reported data from which a direct comparison of the Votrax PSS and Echo GP could be made.

Such a comparison is of interest because these two particular synthesizers have in common their ability to be interfaced to more than one kind of computer (Apple, IBM, etc.). Unlike a speech card which fits into a slot within the computer, these synthesizers are both self-contained. They are accessed by the user through a serial port of the computer (like a printer would be). This flexibility facilitates a direct comparison, since both synthesizers can be tested using the same computer and the same program for input. It also makes these synthesizers a good value for buyers who need speech output for more than one kind of computer.

Both of these speech synthesizers are becoming established products for clinical and special educational use (Wilson, 1986). There is enough available software already to suggest that they are and will continue to be in demand. To some extent, this may be because the Votrax PSS and the Echo GP represent the low end of the cost-and-sophistication continuum of speech synthesizers. A Votrax PSS costs approximately $400, and an Echo GP costs approximately $180. Because they are within the price range of clinic and school budgets, they are incorporated into clinical and educational software packages instead of more expensive synthesizers.

Some direct comparison of the output from the Votrax PSS and Echo GP is possible through the use of sound spectography. A spectogram represents a mapping of a speech wave over time. From it, the length of some sound segments, formant frequencies, and pitch or fundamental freqency, can be estimated. Spectograms made from a sample of speech from a Votrax PSS, an Echo GP, and an adult male speaker were compared. Broadband spectograms were produced, with a bandwidth of 500 Hz and a frequency range of 1000 to 8000 kHz. The utterance used was randomly chosen from the Black (1985) word lists, and consisted of these three words pronounced without special pauses between them: akin, bowl, steward. The following observations were made:

1) Total utterance time: 1.60 sec., Votrax PSS
                         1.25 sec., Adult male
                         1.60 sec., Echo GP

   The total utterance time measurements were somewhat surprising in that listeners frequently report that the Echo GP sounds like it is talking fast, especially in contrast with the Votrax.

2) Consonant sounds: The Echo GP produced plosive consonants (/k, b, t/) which were longer in total time, but included more silence and less energy in the upper frequencies than did the Votrax GP. The fricative /s/ as produced by the Echo GP consisted of a concentration of energy from 3000 to 5000 kHz only, in contrast to the human production of /s/, in which energy is concentrated from 4000 to 7000 kHz and continues with less intensity through 8000 kHz. The Votrax /s/ consisted primarily of energy from 3000 to 5000 kHz, but some energy was present in the signal through 8000 kHz.

Although no conclusions can be made on the basis of such a small sample of speech, clearly there are differences between the speech signals produced by these synthesizers. It is important to find out how these differences are reflected the intelligibility of the speech produced.

Task and Test Materials

In Chapter Two some of the issues surrounding the task and test materials in intelligibility testing were discussed. It was pointed out that it has sometimes been assumed that the same speech materials and tasks that are appropriate for assessing speech discrimination are equally effective for assessing speech intelligibility. Evidence of some of the specific characteristics of words which can either facilitate or deter their intelligibility (but not necessarily their "utterability") were also discussed (Black, 1952). The materials and tasks most frequently used thus far in assessing the intelligibility of synthesized speech have also been described above.

No validated materials have been designed especially for assessing the intelligibility of synthesized words. In fact, there are relatively few tests or word lists even for natural speech which have taken into account the many variables which can influence listener performance (such as word familiarity, and word complexity). An exception is the set of multiple-choice intelligibility tests developed by Black and Haagen during the 1950's (Black, 1957, 1968, 1971, 1985; Black and Haagen, 1963).

The multiple-choice intelligibility tests, re-named Word Discrimination in a recent new edition (Black, 1985), consist of eight different test forms, each divided into twelve shorter lists. In Forms A and B each short list, called a Speaker list, consists of 24 items. In Forms C and D, developed at a different time, each Speaker list consists of 27 items. The Speaker lists within a Form are equivalent in mean intelligibility and the variance of scores of items (Black, 1968).

These tests require the listener to identify the word he hears from four possible responses on a printed answer sheet. Foils are not necessarily words which are minimal phonemic contrasts, however, as in the Modified Rhyme Test (MRT). Rather, the three error responses in the Black lists consist of words which were the most frequently written error responses for each item, when it was administered in an open-response format (Black, 1957). In other words, foils consist of words which are known to be often confused with the target word in good listening conditions, regardless of their phonetic structure. Table 3-1 presents sample targets and their foils taken from Form C, Speaker list 1.

popular	nurse	GET
POPLAR	first	gap
hopper	birth	guess
opera	BURST	guest

Table 3-1. Sample items from Form C, Speaker list 1, Black's multiple choice intelligibility tests (Black, 1957, 1985). Target words are in all caps.

One of the things which distinguishes the Black word lists from other speech discrimination or intelligibility materials is the systematic way in which test items were originally devised and validated. This process has been described elsewhere in some detail (Black, 1957, 1968; Black and Haagen, 1963). It will be briefly reviewed here.

The process of devising equivalent multiple-choice intelligibility tests began with the 10,000 most frequently used words in English as identified by Thorndike ratings (Thorndike, 1944). Proper nouns, homonyms, and homographs were removed, and an intelligibility rating in both quiet and noise was obtained for each word. In order to obtain discriminating test items, only words which were 15-85 per cent intelligible were used from this point on. These 3500 words formed the initial pool from which test items were later constructed (Black, 1957).

As described above, foils for test items were determined by the frequency with which they occurred as errors in response to the target word in quiet and in noise. When these groups of frequently confused words were compiled into test lists, Black found that some error responses were much more likely to occur than others. For example, Table 3-2 shows the percentages for each possible response chosen by listeners for 3 items in Speaker list 1 of Form C. It reflects the choices made by 132 listeners in response to 12 human speakers in 110 db of noise (Black, 1957, p. 223).

Item	%	Item	%	Item	%
popular	35.1	nurse	2.3	GET	56.5
POPLAR	61.8	first	12.2	gap	3.8
hopper	1.5	birth	10.7	guess	27.5
opera	1.5	BURST	74.8	guest	12.3

Table 3-2. Proportions of responses to each word for sample items from Form C, Speaker list 1, Black's multiple choice intelligibility tests (data from Black, 1957, Table 9, p. 223). Target words are in all caps.

In his discussion of the data from which the examples in Table 3-2 are taken, Black points out that the confusion values of certain response choices are dependent, on listening conditions. The relative frequency with which a certain choice is made, therefore, would likely change according to the type and degree of listening adversity. This tendency is also indicated by differences which exist between Forms A and B of the test, compared to Forms C and D. The foils for Forms A and B were identified as a result of responses to test words spoken over carbon microphones in the presence of simulated aircraft noise (Black and Haagen, 1963). Forms C and D, on the other hand, were derived from words originally spoken and recorded in quiet, and reproduced to panels of listeners in noise (Black, 1957). Black points out the differences in the forms that likely resulted from the original listening conditions. For example, in Forms C and D, the words containing /s/ tended to be more intelligible than others, whereas the presence of /s/ in a word in Forms A and B tended to decrease its intelligibility (Black, 1968).

A number of characteristics of Black's tests have been studied, including the intelligibility functions of noise level, signal level, single and multiple distortions, distance between speaker and listener, clarity of answer sheet, and absolute scores compared to PAL PB-50 lists (Black, 1957, 1968). Another issue related to phonetic structure of test items came to light during a study comparing Forms A, B, C, and D.

According to Black, items and foils in Forms C and D received ratings from listeners which suggested that they sounded less "alike" than did items and foils in Forms A and

B (Black, 1968). The difference between target words and foils was then quantified in terms of phonemic distinctions, and compared to listener ratings of the response sets. Sameness ratings for Forms A and B did correlate significantly with phonemic disparity values, although this was not the case for Forms C and D. As Black suggests, this may have implications for rhyme tests of intelligibility in which it is assumed that response sets which have minimal phonemic differences actually sound alike. According to Black, it had been reported that rhyming response sets were judged to sound alike. In Black's study, however when sameness ratings and phonemic disparity values were correlated with intelligibility results, a low correlation was observed.

In other words, actual performance for individual items on Black's multiple choice tests did not necessarily correspond to listener judgements of sameness, or quantifiable phonemic differences between items and foils. Again, this suggests an interaction between listener performance and test materials in intelligibility testing which involves many factors. Of these, simple phonetic structure may not be the most important. The influence of word complexity, familiarity, and specific listening conditions, cannot be overlooked.

Because of the integrity of the Black word lists, and the overall suitability of the task for intelligibility testing, it is appropriate to try out these materials (Black, 1985) in the collection of data pertaining to the intelligibility of synthesized words. A review of the literature reveals that thus far, this has not been done.

Contextual Influence

In the Kraat and Levinson (1984) study discussed in detail in Chapter 2, all subjects listened to both the synthesizers under evaluation. There are some experimental advantages to a within-subjects design such as the one used in their study. Individuals, and therefore groups, are likely to vary in their performance of intelligibility tasks, and the effects of that variation are controlled by a within-subjects design. Assuming other reasonable precautions were taken, it would allow experimenters to conclude that differences between measures were due to differences in the synthesizers rather than in groups. Indeed, this is what Kraat and Levinson suggest. If different groups each listened to different synthesizers, it would be more difficult to dismiss the possibility that observed differences were due merely to basic differences between the groups of subjects rather than the synthesizers.

At first glance, then, a within-subjects design seems to have advantages for a study comparing the intelligibility of more than one synthesizer. The results from other kinds of perceptual studies, however, indicate that there are some potential dangers inherent in within-subjects designs (Poulton, 1973, 1982). Poulton (1982) describes one such undesirable consequence, which he calls asymmetric transfer. He postulates that in some within-subjects studies, subjects adopt or learn a performance strategy during one task which influences their performance in a subsequent task. Usually, Poulton contends, the transfer of a strategy to another task is most likely to occur and go unreported (and possibly unsuspected) when the strategy is unobtrusive, and different treatment conditions are interleaved randomly within a block of trials.

Poulton (1982) presents several examples from the literature in auditory and visual perception and processing to support his theory of asymmetric transfer. For example, he recounts a visual scanning experiment in which subjects were asked to scan for one, five or ten possible targets. The experimenters found that subjects were able to scan for one of 10 targets as quickly and accurately as they were able to scan for one or five targets. They suggest that this supports a theory of parallel processing of visual features. Poulton, however, contends that subjects may have continued to scan for 10 targets even when instructed to scan for a smaller set. In this way, a strategy adopted during an early task may have been inappropriately maintained during later trials.

Poulton's theory of asymmetric transfer has implications for assessing the intelligibility of synthesized speech. Pisoni, Nusbaum, and Greene (1985) present evidence which suggests that synthesized speech may send "miscues" to listeners due to missing or poorly synthesized information in the speech signal. Their argument was discussed in Chapter 2, as well as the research which indicates that subjects who receive training or practice listening to synthesized speech are likely to improve their skills fairly quickly. It is conceivable, therefore, that listeners could learn to decode the particular features of one speech synthesizer, and persist in use of a learned decoding strategy even when listening to a different synthesizer. Although within-subjects designs may help control for variance between groups, the possibility that listening to more than one synthesizer may cause additional unaccounted for effects must be considered.

Practice and Order Effects

The possible effects on performance due to practice have been well documented in a wide range of intelligibility studies (e.g., Black, 1957; Schwab, Nusbaum, and Pisoni, 1985; Theodoridis, Schoeny, and Anne, 1985). Since even relatively short periods of practice may improve performance, especially for discrimination of synthesized speech, data used in the comparison of speech synthesizers should be collected under conditions which are controlled for practice and order effects. Presentation of test materials should be counterbalanced or equivalent forms should be used.

Statistical Design

The following experimental design was proposed to address the questions stated at the beginning of this chapter. The study included two parts. Listeners in both parts of the study listened to taped human speech first. The two synthesizers used for the study, the Votrax PSS and the Echo GP, constituted a within-subjects factor for Part I, and a between-subjects factor for Part II. This allowed for comparisons between the synthesizers and controlled for group variation in Part I, where all groups heard both synthesizers, and contextual influences in Part II, where each group heard one synthesizer only. There were four groups of 6 subjects each, in both parts of the study.

A multiple-choice word identification task was employed, to minimize variance due to linguistic sophistication of subjects and to facilitate speed and accuracy of scoring. Speaker lists 1-6 of Form C of the Black word lists (Black, 1985) were used as stimulus items, with 2 Speaker lists administered for each condition. Each group of subjects heard 2 Speaker lists delivered via natural speech, to gain experience with the task and to provide a means of comparing baseline skills across groups.

Lists used with each synthesizer were counterbalanced between groups, to avoid effects caused by non-equivalent lists. Even though the materials used have been shown to be equivalent for natural speech, they may not be for synthesized speech. In Part I, the order of presentation of the synthesizers was counterbalanced between groups, to control for possible practice effects. Table 3-3 summarizes the statistical design for Part I of the study, and Table 3-4 summarizes the statistical design for Part II.

<u>Summary of Part I: Contrasting Synthesizers</u>

Factor A: Type of Synthesizer
2 levels, within-subjects factor,
counterbalanced

A1: Votrax PSS
A2: Echo GP

Factor B: List Order
2 levels, between-subjects factor,
counterbalanced

B1: Form C, Lists 3 and 4 then 5 and 6 (Black, 1985).

B2: Form C, Lists 5 and 6 then 3 and 4 (Black, 1985).

Factor C: Sythesizer Order
2 levels, between-subjects factor,
counterbalanced; groups randomly assigned to levels of C

C1: Votrax then Echo
C2: Echo then Votrax

Groups 1-4		A1 Votrax	A2 Echo
B1 C1	n = 6	54 items	54 items
B2 C1	n = 6	54 items	54 items
B1 C2	n = 6	54 items	54 items
B2 C2	n = 6	54 items	54 items
Total	n = 48		

Table 3-3. Statistical design for Part I. Suitable for analysis of variance to look for main effects between levels of A, and interactions between A, B, and C.

Summary of Part II: One Synthesizer Per Group

Factor A: Type of Synthesizer
2 levels, between-subjects factor, counterbalanced; groups randomly assigned to levels of A

A1: Votrax PSS
A2: Echo GP

Factor B: List Order
2 levels, between-subjects factor, counterbalanced

B1: Form C, Lists 3 and 4 then 5 and 6 (Black, 1985).
B2: Form C, Lists 5 and 6 then 3 and 4 (Black, 1985).

Factor C: Trials
2 levels, within-subjects factor

C1: Trial 1
C2: Trial 2

		C1	C2
		Trial 1	Trial 2
Groups 5-8			
A1 B1	n = 6	54 items	54 items
A1 B2	n = 6	54 items	54 items
A2 B1	n = 6	54 items	54 items
A2 B2	n = 6	54 items	54 items
Total	n = 48		

Table 3-4. Statistical design for Part II. Suitable for analysis of variance to look for main effects between levels of A, and interactions between A, B, and C.

In addition, results for Trial 2 of each synthesizer in Part II can be compared to those obtained in Part I for each synthesizer when it was presented second in order. Differences in these scores would suggest possible presence of asymmetrical transfer, if scores for natural speech are comparable across groups.

Chapter Summary

Several important factors must be considered in the design of studies for the purpose of comparing the intelligibility of speech synthesizers. Two factors which may influence results but which have been overlooked in previous studies, received special emphasis. These factors are the potential contextual effects in within-subjects designs, and the importance of using test materials which have been constructed on the basis of criteria other than phonetic structure only. A design was presented for a study which would address the research questions posed at the beginning of the chapter. This proposed study accounts for the major issues raised, by making use of established intelligibility testing materials and a two-part, mixed experimental design.

Chapter Four
Procedures and Results

Preparation of Stimuli

Professional quality cassette taperecordings were made using the words from the lists for Speakers 1, 2, and 7 of Form C of the Word Discrimination (Black, 1985) word lists. A trained male speaker recorded the items, following the directions for speakers provided with the lists (Black, 1957). Three target words were read within one phrase unit. For example, the speaker would say:

"Number 1 grew modest vice"

using the same intonation as if it were a sentence. No deliberate pauses occurred between words, although about 3 seconds of response time occurred between presentations.

A computer program was written to allow the words from Speaker lists 3, 4, 5, and 6 of Form C to be sent to either the Votrax PSS or the Echo GP. The same program was used for both synthesizers, with no spelling changes or other modifications of the text of stimulus items. Presentation of each set of target words was exactly as above.

Subjects

Forty-eight young adults (6 M, 42 F) participated, in groups of six. All subjects passed a pure tone hearing screening test (25dB, HTL) and were native speakers of English. Subjects were not experienced in listening to synthesized speech.

Apparatus and Test Environment

All testing was done in a large classroom which had an ambient noise level of 57dB (B & K sound level meter, C scale) when empty. Subjects were seated at tables in two rows of three. The first row of tables was 3 feet from the speech presentation apparatus, and the second row of tables was 7 feet from the apparatus (see Figure 4-1).

A microcomputer, cassette tapeplayer, Votrax PSS, and ECHO GP were arranged on a table in front of the six seats. The speech signal from the tape player and both the Votrax PSS and the Echo GP was set at a level of 70dB (B & K sound level meter, C scale) as measured at the middle seat of row one. Level of presentation varied at the six seats, from 70db to 74dB for the Echo GP, and from 70dB to 64dB for the Votrax PSS. This variation was thought to be due to the difference in placement of the speakers on the two synthesizers.

Both synthesizers were in plain sight throughout testing, but were attached to a switch box so that subjects could not see which one was being used for a given trial.

Test Procedure

Subjects received verbal and written instructions explaining the listening task. The text of the written instructions (Appendix A) was adapted from Black (1957), and was tape recorded for verbal presentation. Subjects received the instruction sheet (Appendix A) and a practice response sheet (Appendix B) identical in format to the one used later for test items.

After the taped instructions were presented, questions were invited. Subjects were then asked to get ready to listen to Speaker 7, and the tape of list 7, Form C (described above) was played. After this practice list (27 items) was completed, questions were again invited. Then the instruction and practice sheets were laid aside, and each subject received a response form for Speaker lists 1-6, Form C (Appendix C). Subjects were then presented with Speaker lists 1 and 2 (54 items, total) via taped human speech. The six Speaker lists from Form C (Black, 1985) which were used for this study were combined to form 3 paired lists. That is, Speaker lists 1 and 2 were always presented consecutively and will be labelled List A (54 items) for the purposes of discussion. Speaker lists 3 and 4 were always presented consecutively, and are hereafter referred to as List B (54 items). Speaker lists 5 and 6 were always presented consecutively, and are hereafter referred to as List C (54 items).

I. Procedures for Part I

Four groups of subjects were randomly assigned to testing conditions, which varied according to what followed after the presentation of human speech. There were 4 test conditions in Part I, consisting of different presentation orders for both synthesizers and word lists (see Table 4-1).

Group	n	Human	Trial 1	Trial 2
Group 1	6	List A	Votrax, B	Echo, C
Group 2	6	List A	Votrax, C	Echo, B
Group 3	6	List A	Echo, B	Votrax, C
Group 4	6	List A	Echo, C	Votrax, B

Table 4-1. Order of presentation of synthesizers and word lists for Part I.

Each of the 4 groups of subjects in Part I was then presented with 54 words (List B or C) via each synthesizer, in the order assigned to the group as shown in Table 4-1.

II. Procedures for Part II

Each of the remaining four groups of subjects was randomly assigned to a testing condition, which varied according to what followed after presentation of human speech. There were 4 test conditions in Part II, consisting of different presentation orders of word lists with a single synthesizer for each group (see Table 4-2).

Group	n	Human	Trial 1	Trial 2
Group 5	6	List A	Votrax, B	Votrax C
Group 6	6	List A	Votrax, C	Votrax B
Group 7	6	List A	Echo, B	Echo C
Group 8	6	List A	Echo, C	Echo B

Table 4-2. Presentation of synthesizers and order of word lists for Part II.

Scoring

The responses sheets for each subject were scored and double-checked. Items for which no word was marked were counted as incorrect, as were items for which more than one word was marked. Scores for List A (Speaker lists 1 and 2), List B (Speaker lists 3 and 4), and List C (Speaker lists and 5 and 6) were calculated, resulting in three scores for each subject.

Statistical Analysis

The raw scores obtained in Part I were analyzed with a repeated measures analysis of variance utilizing a multivariate approach (O'Brien and Kaiser, 1985) and an analysis of covariance. Raw scores obtained in Part II were also analyzed using a standard analysis of variance. The computer program SPSS-X (Norusis, 1984) was used to complete these analyses.

Results: Part I

The mean raw scores and standard deviations for groups 1-4 for both synthesizers are presented in Table 4-3. The maximum possible score per list was 54. The combined mean score for human speech was 50.50, or 94%, with a standard deviation of 1.96 (less than two words missed). The combined mean score for the Echo GP was 32.29, or 60%, with a standard deviation of 4.63. The combined mean score for the Votrax PSS was 38.16, or 71%, with a standard deviation of 2.96. Thus there was an 11% difference in raw scores between the Votrax and the Echo.

Group #	n	Human	S.D.	Echo	S.D.	Votrax	S.D.
Group 1	6	50.00	1.96	35.00	4.34	39.12	1.47
Group 2	6	51.50	1.87	30.66	2.50	36.16	3.31
Group 3	6	50.33	2.07	29.50	4.68	38.83	4.17
Group 4	6	50.16	2.28	34.00	5.18	38.16	1.83
Total	24	50.50	1.96	32.29	4.63	38.08	2.96

Table 4-3. Raw score means and standard deviations for groups 1-4.

A repeated measures analysis of variance was computed for synthesizer scores for Groups 1 through 4. Table 4-4 is a summary table for the analysis of variance testing for the effects of synthesizer order, list order, or a combination of those factors. _F_ values were not significant at the 0.01 level.

Source of Variation	SS	DF	MS	F
WITHIN CELLS	353.917	20	17.696	
CONSTANT	59431.687	1	59431.687	3358.513
SYNTH.ORDER	.187	1	.187	.011
LIST ORDER	93.521	1	93.521	5.285
SYNTH.ORDER BY LIST ORDER	9.187	1	9.187	.519

Table 4-4. Summary table for analysis of variance: synthesizer order and list order, Part I.

Table 4-5 shows the results of an analysis of variance which tested for a main effect between synthesizers, and interactions between the synthesizers, list order and synthesizer order. A significant _F_ value was obtained for synthesizer effect (F = 43.69, $p < .01$).

Source of Variation	SS	DF	M	F
WITHIN CELLS	184.250	20	9.212	
SYNTHESIZER	402.521	1	402.521	43.693 *
SYNTH.ORDER BY SYNTH.	11.021	1	11.021	1.196
LIST ORDER BY SYNTH.	31.687	1	31.687	3.440
SYNTH.ORDER BY LIST ORDER BY SYNTHESIZER	11.021	1	11.021	1.196

* Significant beyond the .01 level.

Table 4-5. Summary table for analysis of variance: synthesizer effect and interactions, Part I.

An analysis of covariance was computed to confirm that scores for human speech were not a good predictor for synthesized speech scores. Human speech scores were used as a covariate for synthesizer scores for each group. Results were not statistically significant, suggesting no linear relationship (as expected). Recall that human speech scores did not vary much (the largest standard deviation observed was 2.28, for Group 4).

Results: Part II

The mean scores and standard deviations by group in Part II are presented in Table 4-6. The raw score means and standard deviations by trial for both synthesizers are summarized in Table 4-7.

Group #	n	Human	S.D.	Trial1	S.D.	Trial2	S.D.
				Votrax		Votrax	
Group 5	6	49.83	1.83	39.83	3.97	38.33	1.51
Group 6	6	49.16	1.17	37.50	4.04	40.83	3.31
Total	12	49.50	1.50	38.66	4.00	39.58	2.41
				Echo		Echo	
Group 7	6	50.50	1.64	30.33	5.99	36.33	5.31
Group 8	6	51.17	0.98	34.16	5.64	32.67	7.34
Total	12	50.83	1.31	32.25	5.81	34.50	6.33
Gr. 5-8	24	50.17	1.41				

Table 4-6. Group raw score means and standard deviations for Part II.

The combined mean for the Votrax in Trials 1 and 2 was 39.12, or 72%, with a standard deviation of 3.21 (Table 4-7). The combined mean for the Echo in Trials 1 and 2 was 33.38, or 62%, with a standard deviation of 6.07. Thus there was a 10% difference in scores between listeners who heard only the Votrax after human speech and listeners who heard only the Echo after human speech. As can be seen in Table 4-7, mean scores for both synthesizers in Trial 2 were slightly higher than in Trial 1.

Trial #	n	Votrax	S.D.	Echo	S.D.
Trial 1	12	38.66	4.00	32.25	5.81
Trial 2	12	39.58	2.41	34.50	6.33
Total	24	39.12	3.21	33.38	6.07

Table 4-7. Raw score means and standard deviations by trial, for all groups in Part II.

A analysis of variance (repeated measures) was computed for the synthesizer scores obtained in Part II. Table 4-8 shows the results of the test for the main effects (Trials 1 and 2 as a within-subjects factor).

Source of Variation	SS	DF	MS	F
WITHIN CELLS	749.167	20	37.458	
CONSTANT	63075.000	1	63075.000	1683.871
SYNTHESIZER	396.750	1	396.750	10.592 *
LIST ORDER	.083	1	.083	.002
SYNTHESIZER BY LIST ORDER	0	1	0	0

* Significant at the .01 level.

Table 4-8. Summary table for analysis of variance: differences between trials, Part II.

A significant F value of 10.6 ($p < .01$) was obtained for the synthesizer effect. This indicates that when Trials 1 and 2 are assumed to be equal, a significant source of variation between them is attributable to differences between synthesizers.

Another analysis of variance was computed considering each trial separately. The summary table for an analysis of variance for simple effects is found in Table 4-9. No significant effects were observed between trials, however a significant interaction between trials, synthesizer, and list order was found (F = 10.08, p < .005). This suggests that the interactions between synthesizers and list orders were different between trials in Part II.

Source of Variation	SS	DF	MS	F
WITHIN CELLS	226.167	20	11.308	
TRIAL (1 or 2)	30.083	1	30.083	2.660
SYNTHESIZER BY TRIAL	5.333	1	5.333	.472
LIST ORDER BY TRIAL	5.333	1	5.333	.472
SYNTHESIZER BY LIST ORDER BY TRIAL	114.083	1	114.083	10.088 *

* Significant beyond the .01 level.

Table 4-9. Summary table for analysis of variance: interactions between trials, synthesizers, and list order, Part II.

A closer examination of mean scores per list per synthesizer revealed that differences between synthesizers were greater for List B than List C (see Table 4-10). Highest scores for the Votrax occurred with List B, while highest scores for the Echo occurred with List C. Thus, when the interactions between list order and synthesizer were averaged across trials (Table 4-8), no significant effect was apparent. When effects were separated for each trial, the fact that the interactions between list order and synthesizers were different was revealed by the significant F value (Table 4-9).

Synth	Trial 1	List	Trial 2	List
Votrax	40	B	38	C
Votrax	38	C	41	B
Echo GP	30	B	36	C
Echo GP	34	C	33	B

Table 4-10. Rounded mean scores showing interaction between lists and synthesizers, Part II.

Results: Part I and Part II Compared

The similarity of the results obtained in Parts I and II can be seen in Figure 4-2. Figure 4-2 compares the scores obtained for human speech and both synthesizers in Part I and Part II, showing mean raw scores and percentages.

<<Insert Figure 4-2 about here >>

Figure 4-2. Plot of mean scores and percentages for human speech, Votrax PSS, and Echo GP, Parts I and II.

To address the question of possible contextual influence in Part I, the scores obtained for each synthesizer in the second trial of Part I were compared with scores obtained for the same synthesizer in Trial 2 of Part II. For the purpose of this analysis, then, Trial 2 became a between-subjects factor.

Source of Variation	SS	DF	MS	F
WITHIN CELLS	705.833	40	17.646	
CONSTANT	63438.021	1	63438.021	3595.071
SYNTHESIZER	346.687	1	346.687	19.647 *
PART (I or II)	22.687	1	22.687	1.286
PART BY SYNTHESIZER	1.021	1	1.021	.058
LIST ORDER	17.521	1	17.521	.993
LIST ORDER BY SYNTH	93.521	1	93.521	5.300
LIST ORDER BY PART	4.688	1	4.688	.266
LIST ORDER BY PART BY SYNTHESIZER	1.021	1	1.021	.058

* Significant beyond the .01 level.

Table 4-11. Summary table for analysis of variance, for Trial 2 across Parts I and II of the present study.

The results of an analysis of variance computed for Trial 2 across Parts I and II and including both synthesizers is summarized in Table 4-11. As was expected, a significant F for synthesizer effect was obtained (F = 19.65, p < .001).

Interactions in Trial 2 between either the Votrax or the Echo and factors such as list order for either Part I or II of the study were not significant. That is, no significant amount of variation observed in Trial 2 was attributable to factors or interactions between factors other than the synthesizers themselves (see analysis of variance summary table, Table 4-12).

Source of Variation	SS	DF	MS	F
WITHIN CELLS	705.833	40	17.646	
CONSTANT	63438.021	1	63438.021	3595.071
PART WITHIN VOTRAX	7.042	1	7.042	.399
PART WITHIN ECHO	16.667	1	16.667	.945
LIST ORDER	17.521	1	17.521	.993
SYNTHESIZER	346.687	1	346.687	19.647 *
LIST ORDER BY PART WITH VOTRAX	5.042	1	5.042	.286
LIST ORDER BY PART WITH ECHO	.667	1	.667	.038

* Significant beyond the .01 level.

Table 4-12. Summary table for analysis of variance of Trial 2, across Parts I and II, showing possible interactions.

These results suggest that scores obtained during the second trials of Part I were not significantly different from the scores obtained during the second trials of Part II, despite the fact that listeners in Part I had heard a different synthesizer during Trial 1. Differences between synthesizers remained significant, but overall scores in Parts I and II were similar, as was seen in Figure 4-2. The results of Parts I and II were combined to obtain overall means, standard deviations, and percentages for the Votrax PSS, the Echo GP, and human speech. These are presented in Table 4-13.

Part	%	Human	S.D.	%	Votrax	S.D.	%	Echo	S.D.
I	94	50.50	1.96	71	38.08	2.96	60	32.29	4.63
II	93	50.17	1.41	72	39.12	3.21	62	33.38	6.07
Total	93	50.33	1.69	72	38.60	3.09	61	32.83	5.35

Table 4-13. Overall means, standard deviations, and percentages for Parts I and II combined.

Overall results of this study indicate that taped human speech was 93% intelligible under these listening conditions. The words produced by the Votrax PSS were 72% intelligible, overall, compared to words produced by the Echo GP, which were 61% intelligible. When results from Parts I and II were combined, an 11% difference between the scores obtained for each synthesizer was observed. The implications of these findings will be discussed in Chapters Five and Six.

Chapter Five
Interpretation of Results

Natural vs. Synthesized Speech

I. Natural speech intelligibility

The average intelligibility for human speech across Parts I and II of the present study was high (93%) and very consistent (S.D. = 1.69). These scores were notably higher than those reported by Black (1957) for Form C under "quiet" listening conditions. In one study, Black reported percentage scores of 72, 70, 76, and 72 for the four lists used in the present study, and an overall average score for Form C in quiet of 72.7% (Black, 1957, p. 217). In another study evaluating the effects of distance from the speaker on listener performance, Black reported scores as high as 84.6% at a distance from the speaker of 6 feet, in "classroom quiet". "Classroom quiet" for Black's study involved background noise of 68db (General Radio sound level meter, C scale). His subjects also served as live "speakers" for the lists, and so signal level was not controlled (Black, 1957). Thus higher scores in the present study may be due to one or more of the following:

o Presentation of human speech with less background noise

o A controlled signal level

o Closer proximity to the "speaker" (taperecorder)

o A single, professionally trained male speaker.

II. Natural speech compared to the Votrax PSS and the Echo GP

The mean raw score for human words correctly identified was 50.33, or 93%. The mean across Parts I and II was 38.6 (72%) for the Votrax PSS and 32.8 (61%) for the Echo GP. The differences between human scores and the scores for each synthesizer can be converted to a percentage to indicate how much more intelligible natural speech was than either synthesizer. Thus, natural speech scores averaged 11.7 items or 30% better than the Votrax (11.7 / 38.6 = .304). Natural speech raw scores averaged 17.5 more than those for the Echo GP, for a converted percentage of 53% (17.5 / 32.83 = .533). In this study, then, taped human speech was 30% more intelligible than the Votrax PSS and 53% more intelligible than the Echo GP.

It has been suggested that specific errors produced in response to synthesized speech do not necessarily match the errors produced in response to "noisy" natural speech (Pisoni, Nusbaum, and Greene, 1985). Nevertheless, it is interesting to consider what listening conditions for human speech would result in scores comparable to those observed for the Votrax and Echo in the present study. A similar level of degradation of overall performance can be seen in the effects of either distance or noise on natural speech intelligibility. For example, Black (1957) reported an average score of 70.5% for Form C, when listeners were 21 feet from the speaker. (However, even at a distance of 39 feet, the farthest distance reported, scores averaged 65.9%.) The effects of background noise were also reported by Black (1957). For example, 106 db noise produced average scores of 72.2% and 114 db noise reduced scores to an average of 60.5%. This suggests another way of thinking about how the present results for the Votrax and the Echo compare to human speech. Scores for the Votrax were comparable to Black's scores for human speech in 106 db noise, and those for the Echo GP were comparable to human speech in 114 db noise.

Votrax PSS vs. Echo GP: Listening to One or Both

When subjects listened to human speech and then both synthesizers, words produced by the Votrax PSS were significantly more intelligible than those produced by the Echo GP (Part I). A comparable difference between synthesizers was also observed when listeners heard human speech followed by one or the other rather than both synthesizers (Part II). Table 5-1 shows how the difference in the average scores for Parts I and II was converted to a percentage which reflects how much more intelligible the Votrax was than the Echo.

	Votrax V	Echo E	Diff V - E	% Better (V - E) / E
Parts I and II	38.60	32.83	5.77	(5.77 / 32.83) = .1757

Table 5-1. Derivation of per cent by which Votrax PSS was more intelligible than the Echo GP.

Thus, the results of this study indicate that the Votrax PSS was 18% more intelligible than the Echo GP.

How These Results Compare to Other Studies

The data from the present study and previous studies should not be compared without noting that there were major differences in the tasks, subjects, and procedures used in previously reported studies. Even the speech synthesizers themselves were not identical. However, the similarities and differences in the present findings and previous related research are interesting and somewhat surprising.

It is of interest to look at the present findings in light of previously studies to compare the following:

- o Segmental intelligibility
- o Word intelligibility
- o Sentence intelligibility.

Greene, Logan, and Pisoni, (in press) reported results from the MRT for earlier models of the Echo and Votrax (summarized in Table 2-2). On the forced-choice format of the Modified Rhyme Test (MRT), average scores were about 65% and 73% for the Echo and Votrax Type'n'Talk, respectively. Thus although their results suggest slightly less difference in intelligibility between those models, overall scores for segmental intelligibility were remarkably similar to the word intelligibility scores of the present study. This is shown in Figure 5-1.

<< Insert Figure 5-1 about here. Comparison of segmental, word, and sentence intelligibility data for Votrax and Echo speech synthesizers>>

Although the MRT evaluates speech sounds in the context of whole CVC words, it is more a measure of segmental intelligibility than word intelligibility. Therefore, higher scores might have been expected on a word intelligibility test in which items contained more information: more sounds, more syllables, more consonant blends, etc.. Scores in the present study, however, were infact slightly lower than reported MRT scores (see Figure 5-1).

Likewise, sentences usually provide even more cues for word identification. Thus sentence intelligibility scores would be expected to be higher than those for single words (e.g., see Figure 2-1). However, word intelligibility results for the Votrax PSS in the present study were very similar to sentence intelligibility results reported by

Kraat and Levinson (1984) (see Figure 5-1). The word intelligibility of the Echo (61%) was higher than reported sentence intelligibility (45%), when no adjustments for rate were made in the sentences.

For natural speech, much research has considered the task of predicting the intelligibility of contextual speech from single-word data (e.g., Giolas and Epstein, 1963; Schiavetti, Sitler, Metz, and Houde, 1984). Usually, scores based on larger linguistic units are higher (recall Figure 2-1). Thus we might predict that word intelligibility scores for synthesized speech would be improved in a sentence intelligibility task. Available data for the synthesizers in question, however, does not confirm this prediction. Several possibilities could account for this departure on the part of synthesized speech from what we might predict for natural speech. The relevant factors include:

- Speaker vs. listener assessment
- Inadequate text-to-speech translation
- Limitations of written vs. spoken language

Again it becomes important to distinguish between testing the listener (speech discrimination) and testing the speaker (speech intelligibility). In studies which concern the comparison of low-information (sounds or words) to high-information (sentences or discourse) tasks, this distinction has often been blurred. Thus, although a formula for predicting "contextual speech intelligibility" from single word data for natural speech has been derived (Schiavetti, et al., 1984), it may be a better predictor for listener performance than for speaker performance. That is, with more contextual cues, a normal listener is likely to do a better job of decoding. It does not follow that the complex task of producing (articulating) speech is easier in larger, more meaningful units, than in simpler, shorter units.

The child with a speech disorder makes a useful example. Any working clinician can describe a case such as this: a typical word identification articulation test is administered to a child. Results suggest, perhaps, a mild speech delay. When the child is asked to tell about his favorite toy, however, the clinician can barely understand one word in five. Results from a single-word performance task, in this case, do not predict the severity of the problem, or the unintelligibility of a contextual speech sample.

Obviously, a child and a modern speech synthesizer do not produce speech in the same manner. The factors which contribute to the child's difficulty producing intelligible ongoing speech are not necessarily paralleled by the inadequacies of a synthesizer, so it is best to halt the analogy here. The point is that when speech is deficient in some aspect at the word level, it is likely to be that bad or worse at sentence and discourse levels. When so-called "intelligibility" scores go up for sentences, it is more likely a function of listener improvement than speaker improvement. The production of ongoing speech in real time is a complex process. It requires the careful synchronization of intonation, stress, and pause, along with production of complex phoneme combinations. The subtleties of this process are not entirely captured in the formuli used by even the most sophisticated speech synthesizers.

The prosodic features of oral language, such as intonation and stress, have been shown to be critical to listener comprehension many times. Examples of sentences which cannot be interpreted without prosodic cues abound in the psycholinguistics literature. Written language, however, contains minimal information about the prosodic characteristics of a phrase or sentence that would be present if it were spoken. Thus it should be acknowledged that text-to-speech systems must make use of input that is incomplete in this dimension. There is little reason, therefore, to expect that text-to-speech synthesizers would produce ongoing speech more intelligibly than they produce single words.

On the contrary, word boundaries and syllabic stress patterns may be so distorted as to make intelligibility worse for sentences than for words, when poorly synthesized.

This is probably what happened in the Kraat and Levinson (1984) study discussed previously, in which sentence intelligibility for the Echo was poor (45%) before two-second pauses were inserted between each word in the sentences. Inserting a pause of this length changes this back to a word identification task, with the advantage of semantic expectations.

Evidence of Contextual Effects

There was the possibility in Part I of the present study that listening to two different synthesizers may have influenced subject performance. That is, subjects may have adopted a decoding strategy while listening to one synthesizer, which they continued to use while listening to the second. Also, just the contrast between the synthesizers may have effected performance. For this reason, subjects in Part II listened to one synthesizer only. Listening to a single synthesizer in the context of human speech is also more likely in clinical situations than is listening to two different synthesizers in close succession.

By looking at data for Trial 2 across both studies, it was possible to see whether scores for either synthesizer were effected by context (practice with the same or another synthesizer in Trial 1). Scores for Trial 2 in Part II were slightly higher than in Part I, for both synthesizers (see

Figure 4-1). Statistical analysis revealed however, that differences between Part I and Part II were not significant. This suggests that in this instance, performance was not influenced by listening to more than one kind of speech synthesizer.

Practice Effects

Their performance on the natural speech task suggests that the four groups in Part I were similar to each other and to groups in Part II in basic ability. Scores in Part II were 1% higher for the Votrax and 2% higher for the Echo, but were on the whole very similar to those in Part I, as was seen in Figure 4-2. Although scores improved slightly from Trial 1 to Trial 2, no significant difference between scores in Trial 1 and Trial 2 was observed for either synthesizer. This indicates that the amount of practice from one trial (54 items) was not enough to significantly effect the listeners' abilities to understand either the Votrax or the Echo.

Chapter Summary

Natural speech intelligibility scores in the present study were higher than those previously reported for this experimental task. Several factors may have contributed to high scores, including advantageous listening conditions and speaker expertise. Word intelligibility scores for the two speech synthesizers evaluated were significantly poorer than for human speech, as was expected. The synthesizers were significantly different in word intelligibility, the Votrax PSS being 18% more intelligible than the Echo GP. In the present study, the effects of practice were not significant. In addition, listeners who heard both synthesizers performed comparably to listeners who heard only one.

Present findings are in general agreement with previous comparisons of similar synthesizers. Comparing the present results with those of other studies suggests the following:

1) Evaluations of the intelligibility of Votrax synthesizers (the Type'n'Talk and PSS) have resulted in remarkably consistent findings. Segmental, word, and sentence intelligibility scores hover around 70%.
2) The intelligibility of Echo synthesizers (Echo, Echo II, Echo GP) is poorer and varies more in comparison with the Votrax. Segmental and word intelligibility scores range from 60-65%.
3) The relationship between word intelligibility and sentence intelligiblity may not be as predictable for synthesized speech as it is for natural speech.

Chapter Six
Discussion and Conclusions

For several reasons it was important to obtain an objective comparison of the intelligibility of the specific speech synthesizers used in this study. These reasons include the following:

1) The Votrax PSS and Echo GP have not been objectively compared before.
2) Of the various models of Votrax and Echo synthesizers, these are the most flexible because they can be used with several brands of computers.
3) The clinical and educational software already developed for each suggests that they are likely to be available and in demand for a long time.
4) Votrax and Echo synthesizers are representative of the low end of the cost continuum, and are therefore most likely to be considered for purchase in clinics and schools.

The results of the present comparison of the Votrax PSS and the Echo GP have implications for clinical applications, purchasing decisions, and future intelligibility studies.

Summary of Major Findings

I. Taped human words were:
30% more intelligible than Votrax
53% more intelligible than Echo

The Echo GP and the Votrax PSS were significantly less intelligible than natural speech under the same listening conditions. This is important information for teachers and clinicians who may not realize how much more intelligible human speech is than low-cost synthesized speech. The speech output of these synthesizers was difficult to understand even for normal adults. Clinicians should be reminded that adding a speech synthesizer to a learning situation or a communication system is not comparable to adding a human voice.

II. Votrax PSS was 18% more intelligible than the Echo GP

Scores for the Votrax PSS averaged 72% (39 / 54) correct, compared to 61% (33 / 54) correct for the Echo GP. The average difference in raw score was 6, which converts to an 18% advantage in word intelligibility for the Votrax PSS (6 / 33 = .181). This advantage was observed whether listeners heard both synthesizers during trials or just one. Besides being statistically significant, this degree of difference in intelligibility is likely to be clinically significant. For some clinical applications, this may justify purchase of a Votrax PSS rather than an Echo GP, in spite of higher cost.

III. No evidence of practice or contextual effects

Listeners did not significantly improve their scores during the second of two back-to-back trials. This amount of listener practice therefore, was not enough to effect the intelligibility of either synthesizer. In most clinical or educational settings, users would listen to a speech synthesizer in the context of natural speech. It would be fairly unusual for more than one kind of synthesizer to be presented in a single clinical activity. In the present study, listeners who heard human speech followed by just one synthesizer performed comparably to listeners who heard human speech followed by two different synthesizers.

Implications for Clinical Applications

I. Deterents and enhancements to intelligibility

Deterents to the intelligibility of speech, whether natural or synthesized, include:

o Adverse listening conditions

o Impaired or less sophisticated linguistic skills.

In most clinical situations, more adverse listening conditions will be present in the form of more background noise and more competing stimuli than were present in the ideal conditions of this study. In addition, listeners will often be speech and/or language disordered. For these reasons it may be that the 61% and 72% intelligibility scores obtained for the Echo and Votrax respectively, represent high estimates of their word intelligibility in some clinical situations.

Enhancements to intelligibility include:

- o Communicative context
- o Listener practice or training.

Just as many clinical situations will have conditions which adversely effect intelligibility, most clinical applications will include practice and a communicative context which are likely to enhance the intelligibility of synthesized speech.

It would be a mistake, however, for clinicians to assume that situational cues will compensate for poorly synthesized speech. Although evidence pertaining directly to synthesized speech is still scant, previous research suggests that as speech becomes less intelligibile, normal listeners rely more heavily on contextual information for understanding. Thus even when synthesized speech is used in a clinical application which has the benefits of linguistic and pragmatic cues, basic speech intelligibility may still be crucial. This is because clinical populations are the ones least likely to notice and take advantage of the linguistic and pragmatic cues that normal listeners would use to compensate for poorly synthesized speech.

The combination of intelligibility deterrents and enhancements that are likely to occur in clinical situations complicates generalization of the present findings to clinical applications. The intelligibility of synthesized speech does not necessarily vary as predictably as natural speech. For example, the sentence intelligibility of the Echo and Votrax may not be as predictably related to word intelligibility as it is for natural speech (see discussion in Chapter Five). It cannot be assumed that the 18% advantage in intelligibility of the Votrax would be maintained in clinical situations. Nevertheless it seems likely that the advantage of the Votrax PSS over the Echo GP would be especially noticeable in some clinical applications. These include clinical applications in which other contextual cues to aid understanding are minimized.

II. Clinical applications most sensitive to intelligibility differences

Several of the clinical and educational applications of synthesized speech that were described in Chapter One could involve the presentation of speech in single words. For example, simple communication boards are often designed to convey one-word messages. The listener is then required to make some assumptions about what the single word might mean. A listener might guess that "cup" means "I want my cup." Of course it could mean anything from "I want a cup of coffee," to "Your cup is going to fall off the table!" Familiarity and communicative context help listeners make the appropriate interpretation. But if the single-word message was unintelligible to begin with, communication is less likely to be smooth.

Speech-impaired children and adults are hardest to understand when the topic of conversation is unknown and other visual or contextual cues are absent. It follows that situations most likely to demonstrate the limitations of a less intelligible synthesizer would be those which incorporate minimal contextual cues and minimal redundancy. Examples of such situations include:

- Communication board applications which do not have accompanying printed output or pictures.
- Applications in which language-handicapped users create their own input for the synthesizer, as with <u>Listen to Learn</u> or other talking word processing systems.
- Presentation of vocabulary or spelling words outside of a meaningful context.
- Applications in which listeners are linguistically impaired, or otherwise unlikely to take advantage of contextual cues (such as young blind children).

These are the kinds of clinical/educational applications of synthesized speech in which the advantage in intelligibility of the Votrax PSS compared to the Echo GP would be most apparent.

The Cost of Intelligibility

I. Factors in choosing a speech synthesizer

There are several factors which should be considered before the purchase of a speech synthesizer for clinical or educational use. Important considerations include:

- o Cost
- o Computer compatibility
- o Software compatibility
- o Ease of use (programmability)
- o Special features (pitch, volume, rate control)
- o Intelligibility

These are the factors that should be considered. Many purchasers, however, are not sufficiently aware of their relative importance.

It is unlikely that intelligibility has very often been a deciding factor in the purchase of speech synthesizers for small computers. The results of intelligibility studies conducted thus far (mostly with very expensive synthesizers) have not been published in journals commonly read by working clinicians and educators. In order to make a decision based on intelligibility, purchasers outside of university settings would have to have access to more than one synthesizer already, and then make a choice based on their own impressions or anecdotal information from fellow users. In the absence of more objective information, clinicians are likely to rely on what they read in software catalogs.

II. Influences from the marketplace

Clinicians and teachers (especially ones who are just starting to develop expertise in computer applications) are likely to follow a decision-making path such as this:

- Assume if software appears in several catalogs, it is probably better than other software.
- Choose a software program that is compatible with the computer already purchased.
- Buy the speech synthesizer that is compatible with that particular software (and perhaps even marketed with it).

Thus the purchase of a speech synthesizer is more likely to be based on how frequently it appears in software catalogs, than on the quality of speech it produces. The first three purchasing factors listed above may get due consideration, but the last three, including intelligibility, often do not.

It is not necessarily the case that clinicians are oblivious to intelligibility, or rather the lack of intelligibility of commonly used synthesizers. However, in the absence of research which might suggest otherwise, they depend on what they hear from software developers and distributors. At least two university-affiliated speech/language professionals who have developed their own popular clinical software (Meyers, 1986; Wilson, 1986) have claimed the superiority of the speech produced by an Echo speech synthesizer. Meyers (1986) commented that the Echo was chosen to accompany her software because is was easier for young language-impaired children to understand than other synthesizers. Wilson (1986) suggested that the Echo is superior because its speech is more human-sounding and less robotic than that of a Votrax. Thus far, there is no information in the literature or in the present study that would support these claims.

III. Justifying the purchase of greater intelligibility

The retail price of a Votrax PSS is about $400. An Echo GP costs about $180. However, the higher price of the Votrax does not necessarily reflect the cost of greater intelligibility. That is, the features which make the Votrax PSS more expensive may not be related to intelligibility at all. For example, there are several structural differences which could contribute to a higher production cost. These include:

- A metal rather than plastic case
- An external speaker at both ends, rather than a single one
- An internal rather than external power supply
- Both a serial and parallel interface, rather than serial only
- A computer chip for synthesizing music as well as a speech

Another factor related to price may be that the primary market targetted by the manufacturer of the Votrax seems to be business rather than education. The Echo synthesizers, on the other hand, are sold almost exclusively for use in educational settings, by a company with an expressed interest in applications for special populations. The pricing and marketing strategies for these two synthesizers may be as different as pricing and marketing are for business vs. educational software. Clearly, many of the factors which influence the price of speech synthesizers for small computers may be irrelevant to speech quality. People who order equipment for schools and clinics are as unlikely to be aware of these factors as they are unaware of differences in intelligibility.

When two synthesizers are equally software/hardware-compatible for a particular application, cost becomes the determining factor, often without regard for intelligibility. A clinic or school system considering the purchase of several synthesizers is likely to buy the cheapest ones when other factors are perceived to be equal, or nearly equal. Thus even a school system having successful experience with a Votrax PSS, might choose to purchase three or four additional Echo GP synthesizers rather than two additional Votrax machines (C. Wissick, Albemarle County Schools, personal communication, March 13, 1986). An obvious justification would be having more schools equipped with a speech synthesizer; and up till now, there has been almost no objective evidence to support the purchase of the more expensive product.

It is interesting to consider how much money intelligibility would be worth to a prospective buyer. If more information about intelligibility became widely available, clinicians and educators might begin to give it more consideration in their choice of synthesizers. In the present study, the Votrax PSS was 18% more intelligible than the Echo GP. At this time, a Votrax PSS costs a little more than twice as much as an Echo GP. Possible reasons for price differences regardless of intelligibility were outlined above. When factors such as quality of construction and optional features are not crucial (and often they are not), a clinician is still faced with deciding when the difference in price is worth the difference in intelligibility.

Some price/intelligibility combinations would be obvious consumer choices. For example, choosing between a $2000 synthesizer that was 18% more intelligible than a $200 speech synthesizer would be an easy decision in most circumstances. That is, the advantage in intelligibility would not be perceived as being worth ten times the price, even if the $2000 were available. At the other extreme, even a slight advantage in intelligibility would probably be worth some price difference. Thus a $60.00 synthesizer which was 5% more intelligible would likely be purchased in preference to a $50.00 synthesizer. Decisions about price and intelligibility differences which are less extreme become more difficult.

To some extent, a clinician's choice may be influenced by the application for which the synthesizer is primarily intended. Less intelligible synthesized speech is more likely to be noticed in applications in which minimal contextual cues are available to assist in comprehension, or in which the listener is unable to make use of those cues. For these applications at least, the additional expense of more intelligible synthetic speech may be justified.

Methodological Factors to Be Considered in Future Intelligibility Studies

I. Assumptions Which Should be Avoided

The results of the present study suggest two assumptions that should be avoided in future studies of the intelligibility of synthesized speech. First, it should not be assumed that speech materials which are equivalent for natural speech will be equivalent for various speech synthesizers. The present study made use of a task and materials especially designed for intelligibility testing, but not previously used with synthesized speech. Word-lists were counterbalanced to avoid the confounding of results by non-equivalent lists. Results suggest that speech materials which are equivalent for natural speech may not be equivalent for synthesized speech. Counterbalancing (rather than randomizing) speech materials to be used to compare speech synthesizers is recommended.

Secondly, the relationship between single word and sentence intelligibility may not be as predictable for synthesized speech as it is for natural speech. Especially when comparatively low segment or word intelligibility is found, assumptions about sentence intelligibility are not warranted (see discussion in Chapter 5).

II. Task considerations

Important features of the task and materials used to assess the intelligibility of synthesized speech include:

- o Ease of administration
- o Ease of scoring
- o Consistency of subject performance
- o Control for linguistic considerations (word familiarity)
- o Multiple forms, sufficient number of items

The present study made use of a Black's multiple choice speech intelligibility lists (Black, 1957, 1985) which were especially designed for intelligiblity testing, but not previously used with synthesized speech. This task is well-suited to group administration, and makes it possible to collect a large number of individual responses in a minimal amount of time. For example, participants in the present study were able to listen to directions (5 minutes), complete a practice list of 27 items, and respond to 162 test items in about 30 minutes. This included time for the examiner to change stimuli between trials. In contrast, during a recent pilot study using CID W-22 words in a writedown task, it took 35-40 minutes to obtain 124 responses per subject, including practice. The ease with which the Black tests were administered and scored, and the overall consistency of subject performance suggests that these lists meet the first three criteria mentioned above.

At the time of their development (see discussion in Chapter Three) the target words in these lists had frequency of occurence ratings comparable to or better than the words in the Harvard PB word lists (Black, 1957). Since this has not been as important a consideration in the development of some materials previously used with synthesized speech (such as the Modified Rhyme Test), the Black lists probably meet this criterion as well or better than alternative available materials.

Although it cannot be assumed that any or all forms of the Black lists are equivalent for synthesized speech, there is a sufficient pool of forms and lists to counterbalance, so that even large numbers of responses could be obtained. The closed- response format of this task and the content of these materials endorse the use of the Black lists for future synthetic speech intelligibility studies.

III. Between- vs. within-subjects design

In the present two-part study, a within-subjects study was replicated by study in which the primary independent variable became a between-subjects factor. This replication addressed concerns about potential contextual effects of listening to multiple synthesizers on the one hand, and concerns about potential between-group variability on the other. Although no evidence of contextual influence was found, it cannot be assumed that this would be the case for any other pair (or more) of synthesizers.

The consistency of performance between groups within Part I and Part II and across Parts I and II suggests that a between-subjects design would be suitable for extending the present study to include more synthesizers. Results for natural speech could be obtained for each additional group, and could signal unusual group variability if it should occur.

IV. New directions

There are many ways in which the present study could be extended to answer further questions about the intelligibility of synthesized speech with special reference to clinical applications. Four alternatives of particular interest would involve the assessment of:

- More synthesizers in the low-to-moderate price range
- Practice effects
- Sentence intelligibility
- Intelligibility with language-impaired listeners and other clinical populations

Information such as that resulting from the present study should be available for the majority of low-to-moderately priced speech synthesizers for small computers. Previous research assessing the intelligibility of synthesized speech placed an emphasis on segmental intelligibility. The primary objective was to obtain detailed information about individual speech sounds and sound combinations, which might provide specific information to developers at the forefront of speech technology. Information about the general adequacy or inadequacy of the text-to-speech translation of a variety of synthesizers was gleaned, but with emphasis on the high end of the cost-and-sophistication continuum. This is understandable, since even a few years ago it would have been difficult to predict the current availability of personal computers and the subsequent demand for low-cost synthesized speech. Since this demand has increased dramatically and is likely to continue, now there is a need for more information about synthesizers which are most likely to be widely purchased for clinical and educational use.

It is especially important to find out more about the effects of listener training and practice on the intelligibility of synthesized speech. Although some research has indicated that the ability to decode synthesized speech may be quite responsive to practice, more information about specific synthesizers is called for. For example, it would be of clinical value to know whether listener practice is sufficient to compensate for the differences in intelligibility between two speech synthesizers. If practice causes a less expensive synthesizer to become as intelligible as a more expensive one, such information would have bearing on the price vs. intelligibility issues discussed above.

Since the word intelligibility of some text-to-speech synthesizers may not be related to sentence intelligibility as predictably as it is for natural speech, it would be of value to know more about this relationship. Many clinical applications incorporate synthesized speech at the sentence level. Word-intelligibility scores can form the basis for some comparisons, but sentence-intelligibility information is also imperative.

Future studies should also test the extent to which differences observed between synthesizers under controlled laboratory conditions are also observed in the less predictable and more demanding environment of the clinic or classroom. Information is needed about the way in which speech and language handicaps effect listeners' abilities to decode synthesized speech. It is not known to what extent communicative context contributes to the intelligibility of synthesized speech in clinical situations. Thus far, there is no way to tell whether the intelligibility scores obtained in the present study represent high or low estimates of intelligibility in clinical situations. The results of the this study could be used as a baseline for future experiments designed to address these questions.

Summary

Under the same listening conditions, taped human speech was 30% more intelligible than the Votrax PSS and 53% more intelligible than the Echo GP. A statistically significant difference in word intelligibility between the Votrax PSS and the Echo GP was observed, with the Votrax PSS being 18% more intelligible. Listeners who heard human speech followed by two different synthesizers performed comparably to those who heard the more likely clinical combination of human speech followed by just one synthesizer.

The observed difference between synthesizers is likely to be most noticeable in clinical applications in which other contextual cues are minimal, or in which listeners are unlikely or unable to take advantage of such cues. In considering the important factors bearing on the purchase of a speech synthesizer for such applications, clinicians are encouraged to increase the priority they give to intelligibility.

References

Beukelman, D.R., Yorkston, K.M., Poblete, M, & Naranjo, C. (1984). Frequency of word occurrence in communication samples produced by adult communication aid users. *Journal of Speech and Hearing Disorders*, *49*, 360-367.

Black, J.W. (1952). Accompaniments of word intelligibility. *Journal of Speech and Hearing Disorders*, *17*,409-418.

Black, J.W. (1957). Multiple-choice intelligibility tests. *Journal of Speech and Hearing Disorders*, *22* (2), 213-235.

Black, J.W. (1968). Responses to multiple-choice intelligibility tests. *Journal of Speech and Hearing Research*, *11*, 453-466.

Black, J.W. (1985). *Word discrimination*. Danville, IL: Interstate Printers & Publishers.

Black, J.W., & Agnello, J.A. (1964). The prediction of the effects of combined deterrents to intelligibility. *Journal of Auditory Research*, *4*, 277-284.

Black, J.W. & Haagen, C.H. (1963). Multiple-choice intelligibility tests, Form A and B. *Journal of Speech and Hearing Disorders*, *28* (1), 77-86.

Borden, G.J., & Harris, K.S. (1984). *Speech science primer: Physiology, acoustics, and perception of speech.* Baltimore, MD: Williams & Wilkins.

Bull, G. & Cochran, P. (1985). Creating tools for clinicians and teachers. *Journal for Computer Users in Speech and Hearing*, *1* (1), 45-49.

Bull, G., Lough, T., & Cochran, P. (in press). Logo and exceptional individuals. In J. Lindsey (Ed.) *Computers and Exceptional Individuals*. Columbus, OH: Charles Merrill.

Chial, M.R. (1976). Evaluation of a synthetic talker for speech intelligibility testing. *Journal of the Acoustic Society of America*, *59*, S15.

Chial, M.R. (1984, November). Comparison of commercial speech synthesizers for small computers. Paper presented at the annual convention of the American Speech-Language-Hearing Association, San Fransisco.

Chial, M.R. (1985, January). Tutorial on speech synthesis. Paper presented at the 1985 American Speech-Language-Hearing Foundation Computer Conference, New Orleans, LA.

Clark, J.E. (1983). Intelligibility comparisons for two synthetic and one natural speech source. Journal of Phonetics, 11, 37-49.

Cochran, P., & Bull, G. (1985, November). Creating a shared context: Using a computer in language therapy. Paper presented at the annual convention of the American-Speech-Language-Hearing Association, Washington, D.C..

Cook, T.D., & Campbell, D.T. (1979). Quasi-experimentation: Design & analysis issues for field settings. Boston: Houghton Mifflin.

Dillon, H. (1983). The effect of test difficulty on the sensitivity of speech discrimination tests. Journal of the Acoustical Society of America, 73, 336-344.

Echo (1982). User's Manual. Carpinteria, CA: Street Electronics.

Egan, J. P. (1948). Articulation testing methods. Laryngoscope, 58(9), 955-991.

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

Flanagan, J.L. (1972). Voices of men and machines. Journal of the Acoustical Society of America, 51, 1375-1387.

Ginther, D.W. (October, 1983). Micro's are talking back in the classroom: The promise of speech technology in education. T.H.E. Journal, pp. 105-107.

Giolas, T.G. & Epstein, A (1963). Comparative intelligibility of word lists and continuous discourse. Journal of Speech and Hearing Research, 6, 349-358.

Greene, B.C., Logan, J.S., & Pisoni, D.B. (in press). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. Behavior Research Methods, Instruments, and Computers.

Hawley, M.E. (Ed.) (1977). Speech intelligibility and speaker recognition. Stroudsburg, PA: Dowden, Hutchinson, & Ross, Inc.

Hyman, C. (1985, November). Computer usage in the speech-language pathology and audiology profession. Poster session presented at the annual convention of the American Speech-Language-Hearing Association, Washington, D.C..

Kraat, A., & Levinson, E. (1984, October). Intelligibility of two speech synthesizers used in augmentative communication devices for the severely speech impaired. Paper presented at the Third International Conference on Augmentative and Alternative Communication, M.I.T., Cambridge, MA.

Laureate Learning Systems (1983). First Words. Burlington, VT: Laureate Learning Systems, Inc.

Lehiste, I. & Peterson, G. (1959). Linguistic considerations in the study of speech intelligibility. Journal of the Acoustic Society of America, 31, 280-286.

Luce, P.A., Feustel, T.C., & Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural word lists. Human Factors, 25, 17-32.

Merrell, H.B., & Atkinson, C.J. (1965). The effects of selected variables upon discrimination scores. Journal of Auditory Research, 5, 285-292.

Meyers, L.F. (1984). Unique contributions of microcomputers to language intervention with handicapped children. Seminars in Speech and Language, 5 (1), 23-33.

Meyers, L.F. (1986, January). Using computers with synthesized speech output to facilitate language development in young children. Paper presented at the 1986 American Speech-Language-Hearing Foundation Computer Conference, Orlando, FL.

Miller, J.M. (1984). The effects of voice synthesis on the acquisition of Bliss symbols by nonvocal motorically impaired and intact mentally retarded persons. Unpublished doctoral dissertation, School of Education, University of Virginia.

Miller, G.A., Heise, G.A., & Lichten, W. (1951). The intelligibility of speech as a function of test materials. Journal of Experimental Psychology, 41, 329-336.

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

Nelson, D.A., & Chaiklin, J.B. (1970). Writedown versus talkback scoring and scoring bias in speech discrimination testing. Journal of Speech and Hearing Research, 13, 645-654.

Norusis, M.J. (1985). SPSSx advanced statistics guide. New York: McGraw-Hill.

Nusbaum, H.C., Dedina, M.J., & Pisoni, D.B. (1984). Perceptual confusions of consonants in natural and synthetic CV syllables. Speech Research Laboratory Technical Note 84-02, Bloomington, IN: Indiana University.

O'Brien, R.G., & Kaiser, M.K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. Psychological Bulletin, 97 (2), 316-333.

Owens, E., & Schubert, E.D. (1968). The development of constant items for speech discrimination testing. Journal of Speech and Hearing Research, 11, 656-667.

PEAL Software (1985). Exploratory Play. Duarte, CA: PEAL Software, Inc.

Pisoni, D.B., Nusbaum, H.C., & Greene, B.G. (1985). Perception of synthetic speech generated by rule. Proceedings of the IEEE, 73 (11), 1665-1675.

Poulton, E.C. (1973). Unwanted range effects from using within-subjects experimental designs. Psychological Bulletin, 80, 113-121.

Poulton, E.C. (1982). Effects of one strategy on another in the within-subjects designs of cognitive psychology. Psychological Bulletin, 91(3), 673-690.

Rentschler, G.J. (1985, November). *Computer synthesized speech: Intelligibility and the language-impaired child.* Paper presented at the annual convention of the American Speech-Language-Hearing Association, Washington, D.C..

Rintelman, W.S., Beasley, D.S., Mosher, N.A., & Mosher, R.A. (1974). Repeated measures of speech discrimination with normal listeners: Counterbalancing vs randomization. *Journal of Auditory Research, Supplement 2*, pp. 18-20.

Ripich, D.N., & Panagos, J.M. (1985). Accessing children's knowledge of sociolinguistic rules for speech therapy lessons. *Journal of Speech and Hearing Disorders, 50*, 335-346.

Rosegrant, T.J. (1984). Use of microprocessors to remediate speech through literacy. In W. H. Perkins (Ed.) *Current therapy of communication disorders: Language handicaps in children* (pp. 57-62). New York: Thieme-Stratton.

Rosegrant, T. (Winter, 1986a). Listening to `at risk' children learning. *Educators' Report*, pp. 10-11.

Rosegrant, T.J. (1986b, January). *Using microcomputers to help the language-impaired write.* Paper presented at the 1986 American Speech-Language-Hearing Foundation Computer Conference, Orlando, FL.

Rosegrant, T., & Cooper, W. (1985). *Listen to Learn*. Boca Raton: IBM, Inc.

Schiavetti, N., Sitler, R.W., Metz, D.E., & Houde, R.A. (1984). Prediction of contextual speech intelligibility from isolated word intelligibility measures. *Journal of Speech and Hearing Research*, *27* (4), 623-6.

Schwab, E., Nusbaum, H., & Pisoni, D. (1985). Effects of training on the perception of synthesized speech. *Human Factors*, *27*(4), 365-408.

Sevik, R.A. & Romski, M.A. (1985, November). *Comprehension of synthesized speech by nonspeaking severely retarded individuals.* Paper presented at the annual convention of the American Speech-Language-Hearing Association, Washington, D.C..

Snow, C. Midkiff-Borunda, S., Small, A., & Proctor, A. (1984). Therapy as social interaction: analyzing the contexts for language remediation. *Topics In Language Disorders,* *4* (4), 72-85.

Stalker, J.L., Hawk, A.M., & Smaldino, J.J. (1982). The intellgibility and acceptability of speech produced by five different electronic artificial larynx devices. Journal of Communication Disorders, 15, 299-307.

Theodoridis, G.C. & Schoeny, Z.G. (1982). A procedure for measuring the minimum information required for identification of a word in context. Kybernetes, 11, 183-188.

Theodoridis, G.C., Schoeny, Z.G., & Anne, A. (1985). Measuring the contribution of printed context information to acoustical word recognition by normal subjects. Audiology, 24, 104-116.

Thorndike, E.L., & Lorge, I. (1944). The teacher's word book of 30,000 words. New York: Teacher's College, Columbia University.

Trual, G.N. & Black, J.W. (1965). The effect of context on aural perception of words. Journal of Speech and Hearing Research, 8, 363-369.

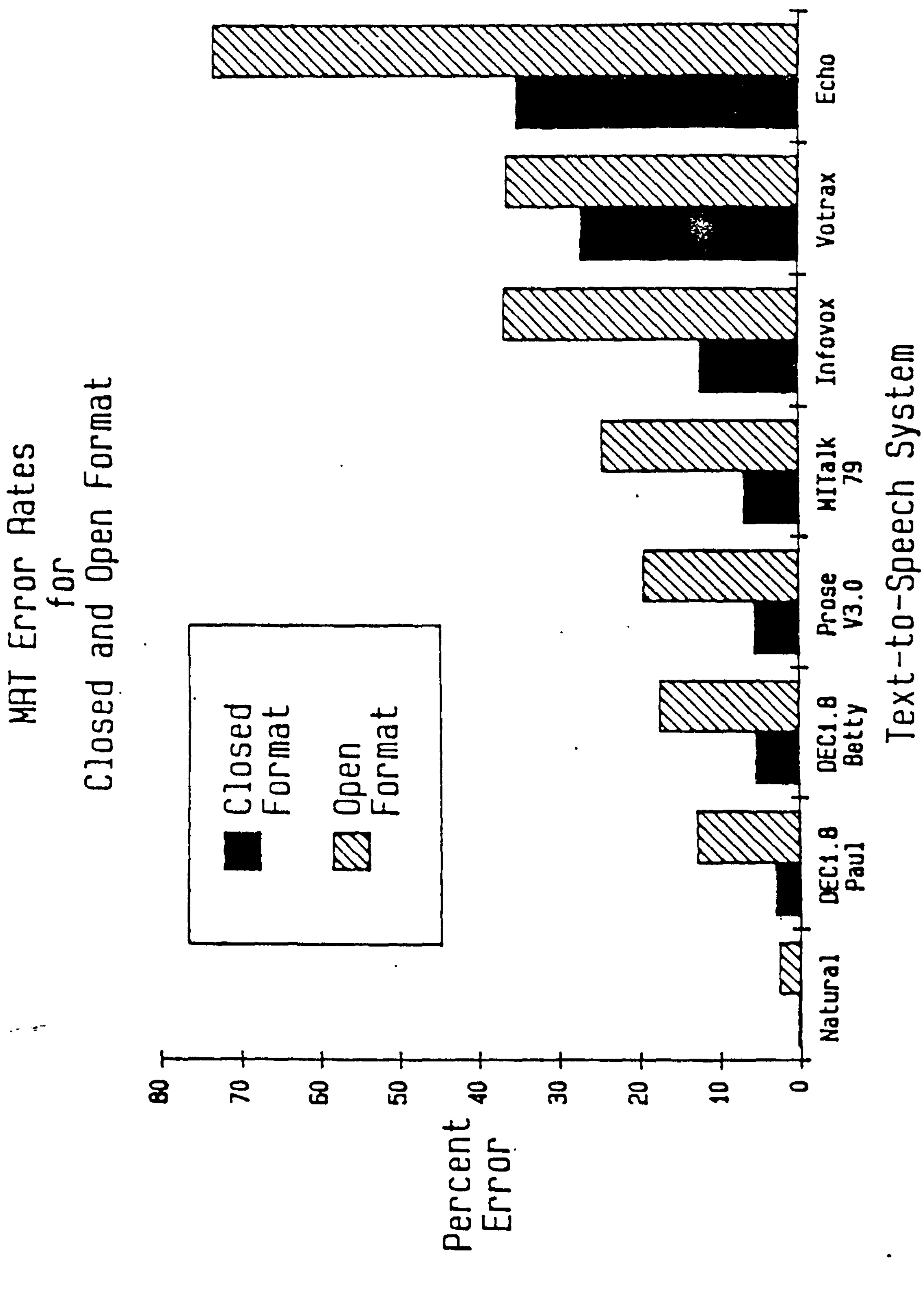
Wilson, M.S. (1986, January). Comparison of synthesized speech for clinical software. Paper presented at the 1986 American Speech-Language-Hearing Foundation Computer Conference, Orlando, FL.

Wilson, M.S. & Fox, B.J. (1983). Microcomputers: A clinical aid. In H. Winitz (Ed.) Treating language disorders: For clinicians by clinicians (pp. 248-255). Baltimore, MD: University Park Press.

Wissick, C., & Young, K. (1985). Say it with Logo: Logo and the speech synthesizer as tools for communication. In M. Gergen & D. Hagen (Eds.), Computer technology for the handicapped: Proceedings from the 1984 Closing the Gap Conference (pp. 94-96). Henderson, MN: Closing the Gap.

Yorkston, K.M., & Beukelman, D.R. (1981). Assessment of intelligibility of dysarthric speech. Seattle, WA: C.C. Publications.

Figure 2-2


MRT Error Rates
for
Closed and Open Format
Closed Format
Open Format
Percent Error
80
70
60
50
40
30
20
10
0
Natural
DEC1.8 Paul
DEC1.8 Betty
Prose V3.0
MITalk 79
Infovox
Votrax
Echo
Text-to-Speech System

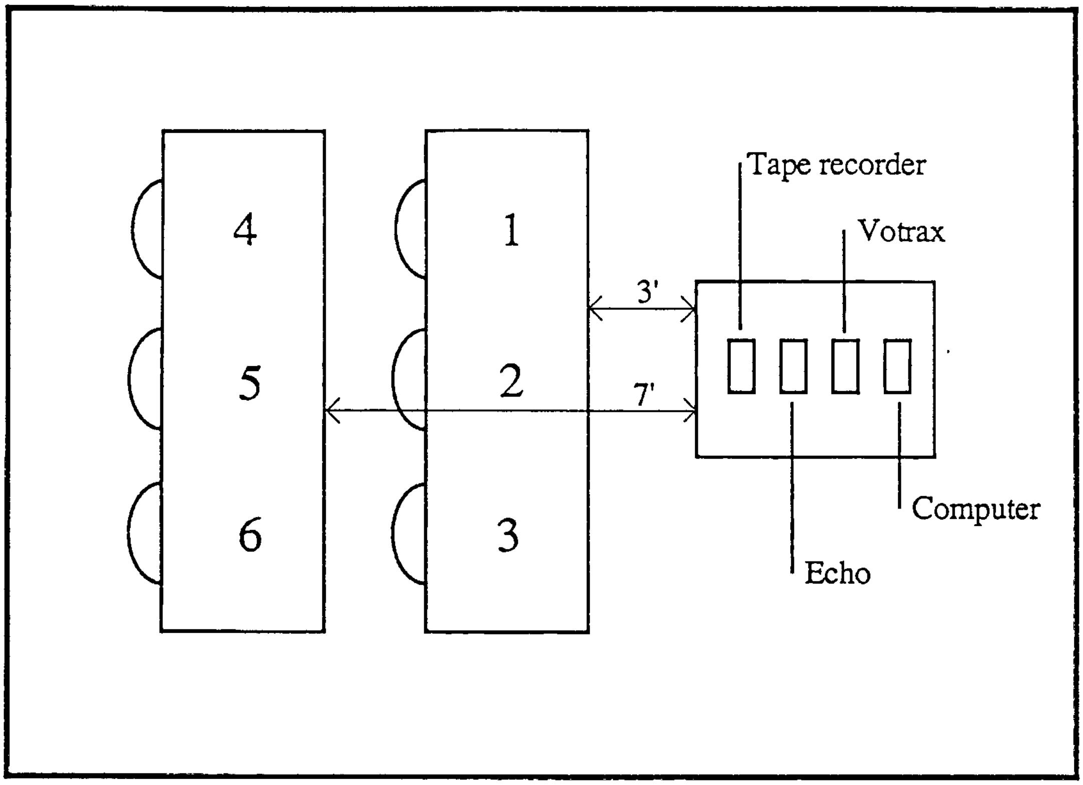
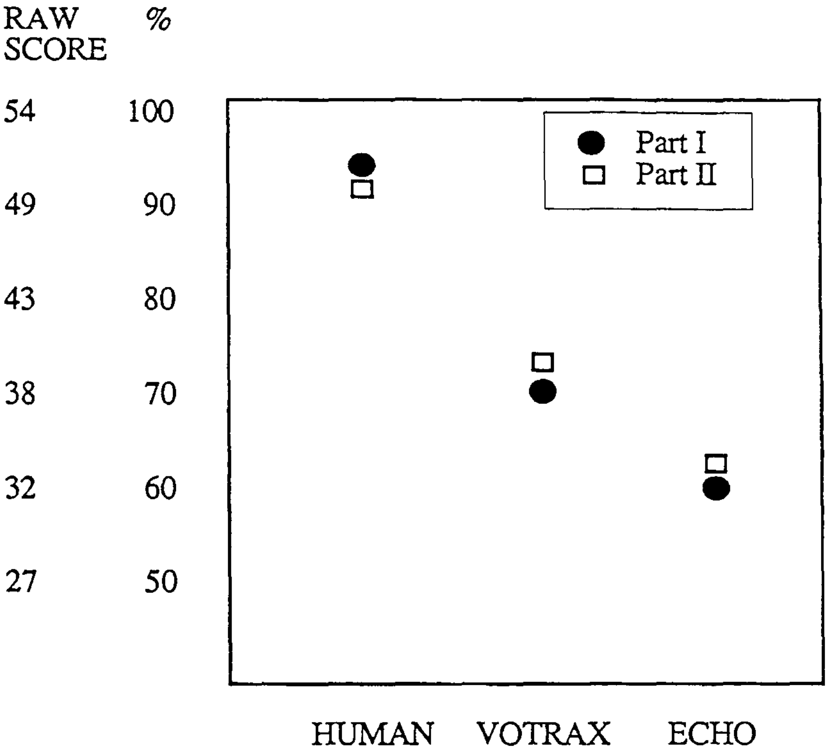


Figure 4-1. Room arrangement for the present study.



	HUMAN	VOTRAX	ECHO
Part I	51	38	32
Part II	50	39	33

Figure 4-2. Mean scores for human speech, Votrax, and Echo, Parts I and II.

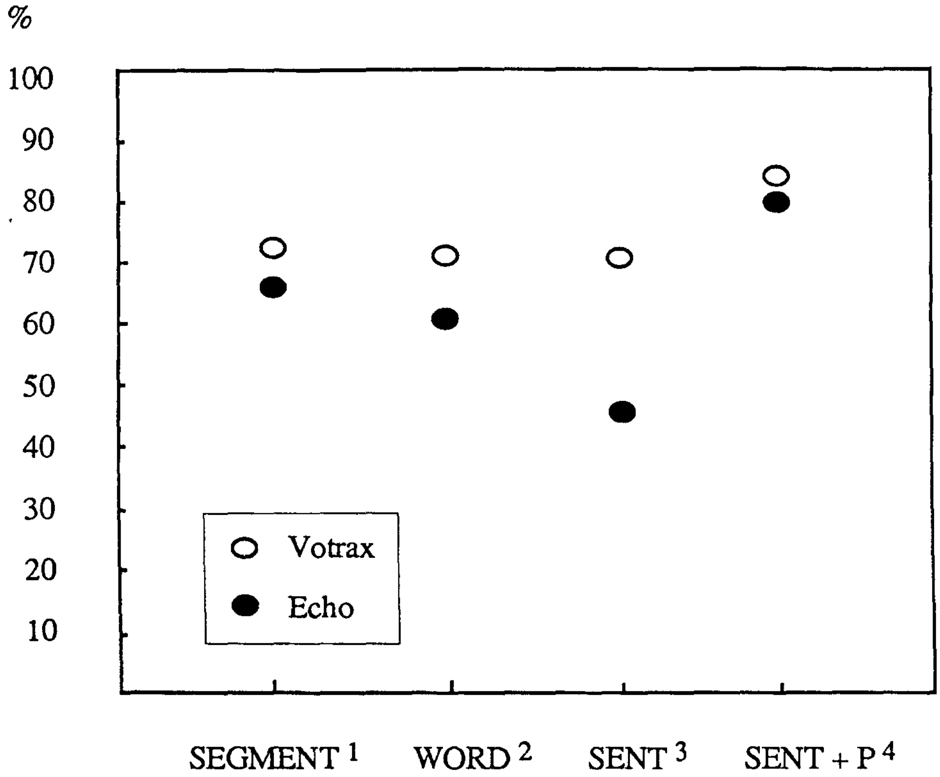


Figure 5-1. Comparison of segmental, word, and sentence intelligibility.

[1] Greene, Logan, and Pisoni, in press.
[2] Present study.
[3] Kraat and Levinson, 1984 (no pause data).
[4] Kraat and Levinson, 1984 (with 2 sec. pause between words).

Appendix A

Instructions to Subjects

You will be helping me test the intelligibility of computer-generated speech. Now look at the answer form, and find the section for Speaker 7 in the upper left corner.

You are going to hear a series of groups of three words. You will hear:

"I am Speaker 7; I say again, I am Speaker 7.

(pause) Number 1 province worst sledge

(pause) Number 2 grow wearer staunch

(pause) Number 3 zephyr swam grandsire "

You will notice that for each word I read there are four possible choices on the answer sheet in the section for Speaker 7. You heard me say, "Number 1 province worst sledge." The first word after Number 1 was <u>province</u> and appears in the first group of four words. The second word <u>worst</u> is found in the second group of four words of <u>Number 1</u> and the third word, <u>sledge</u>, you will find in the third group of four words of <u>Number 1</u>.

Your job is to draw a line <u>through</u> the word that you hear, making one mark in each group of four words. Erasures are permitted. Remember, draw a line <u>through</u> the words you hear, or think you hear. Try not to leave any blank.

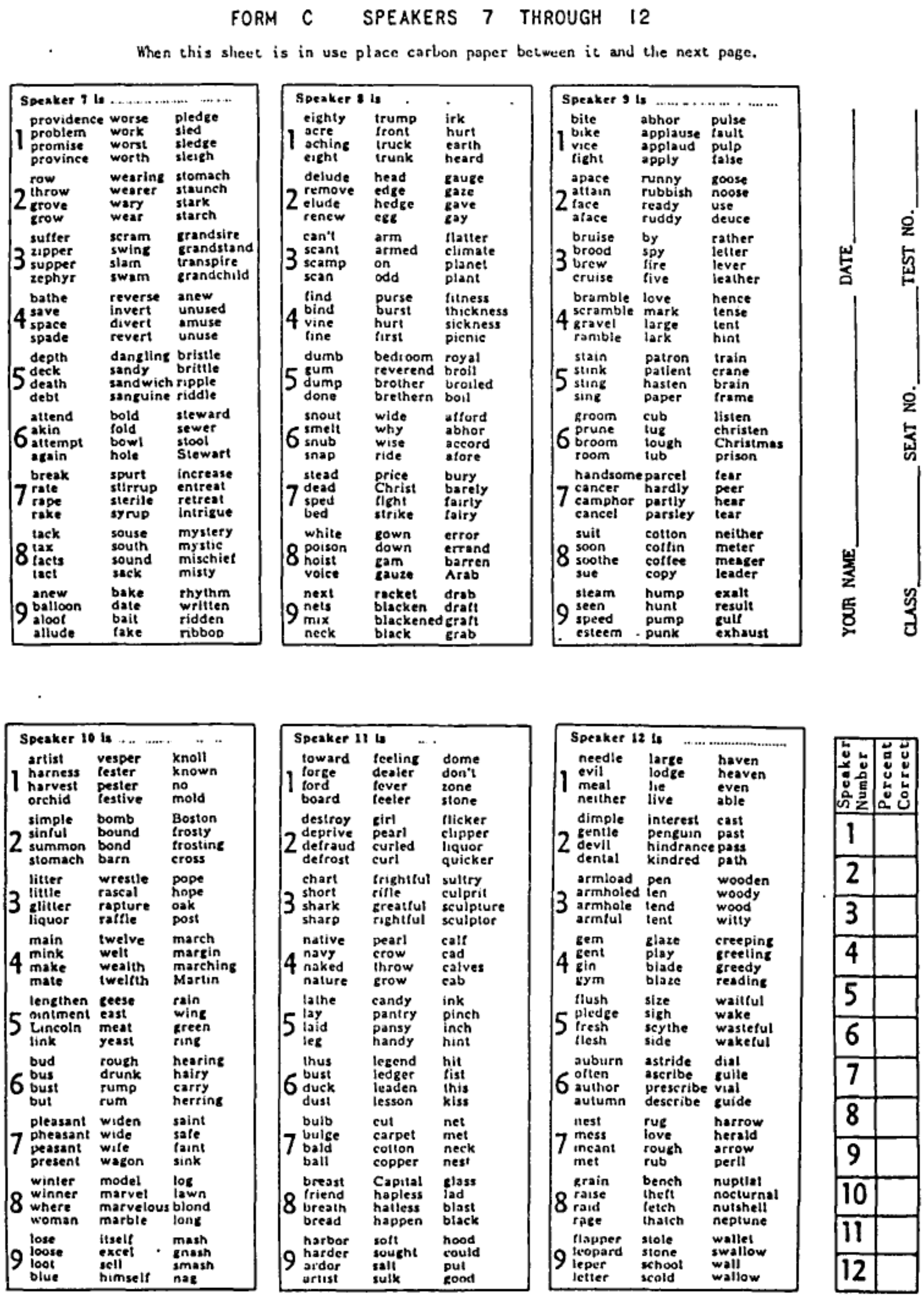
For example, for Number 2, you would draw a line through the words <u>grow</u>, <u>wearer</u>, and <u>staunch</u>.

Are there any questions?

We will do the list for Speaker 7 for practice, and then you will have another chance to ask questions.

Get ready to listen to Speaker 7.

Appendix B



FORM C SPEAKERS 7 THROUGH 12

When this sheet is in use place carbon paper between it and the next page.

Speaker 7 is ....

1	providence	worse	pledge
	problem	work	sled
	promise	worst	sledge
	province	worth	sleigh
2	row	wearing	stomach
	throw	wearer	staunch
	grove	wary	stark
	grow	wear	starch
3	suffer	scram	grandsire
	zipper	swing	grandstand
	supper	slam	transpire
	zephyr	swam	grandchild
4	bathe	reverse	anew
	save	invert	unused
	space	divert	amuse
	spade	revert	unuse
5	depth	dangling	bristle
	deck	sandy	brittle
	death	sandwich	ripple
	debt	sanguine	riddle
6	attend	bold	steward
	akin	fold	sewer
	attempt	bowl	stool
	again	hole	Stewart
7	break	spurt	increase
	rate	stirrup	entreat
	rape	sterile	retreat
	rake	syrup	intrigue
8	tack	souse	mystery
	tax	south	mystic
	facts	sound	mischief
	tact	sack	misty
9	anew	bake	rhythm
	balloon	date	written
	aloof	bait	ridden
	allude	fake	ribbon

Speaker 8 is . . .

1	eighty	trump	irk
	acre	front	hurt
	aching	truck	earth
	eight	trunk	heard
2	delude	head	gauge
	remove	edge	gaze
	elude	hedge	gave
	renew	egg	gay
3	can't	arm	flatter
	scant	armed	climate
	scamp	on	planet
	scan	odd	plant
4	find	purse	fitness
	bind	burst	thickness
	vine	hurt	sickness
	fine	first	picnic
5	dumb	bedroom	royal
	gum	reverend	broil
	dump	brother	broiled
	done	brethern	boil
6	snout	wide	afford
	smelt	why	abhor
	snub	wise	accord
	snap	ride	afore
7	stead	price	bury
	dead	Christ	barely
	sped	fight	fairly
	bed	strike	fairy
8	white	gown	error
	poison	down	errand
	hoist	gam	barren
	voice	gauze	Arab
9	next	racket	drab
	nets	blacken	draft
	mix	blackened	graft
	neck	black	grab

Speaker 9 is ....

1	bite	abhor	pulse
	bike	applause	fault
	vice	applaud	pulp
	fight	apply	false
2	apace	runny	goose
	attain	rubbish	noose
	face	ready	use
	aface	ruddy	deuce
3	bruise	by	rather
	brood	spy	letter
	brew	fire	lever
	cruise	five	leather
4	bramble	love	hence
	scramble	mark	tense
	gravel	large	tent
	ramble	lark	hint
5	stain	patron	train
	stink	patient	crane
	sting	hasten	brain
	sing	paper	frame
6	groom	cub	listen
	prune	tug	christen
	broom	tough	Christmas
	room	tub	prison
7	handsome	parcel	fear
	cancer	hardly	peer
	camphor	partly	hear
	cancel	parsley	tear
8	suit	cotton	neither
	soon	coffin	meter
	soothe	coffee	meager
	sue	copy	leader
9	steam	hump	exalt
	seen	hunt	result
	speed	pump	gulf
	esteem	punk	exhaust

YOUR NAME ______ DATE ______ TEST NO. ______

CLASS ______ SEAT NO. ______

Speaker 10 is ....

1	artist	vesper	knoll
	harness	fester	known
	harvest	pester	no
	orchid	festive	mold
2	simple	bomb	Boston
	sinful	bound	frosty
	summon	bond	frosting
	stomach	barn	cross
3	litter	wrestle	pope
	little	rascal	hope
	glitter	rapture	oak
	liquor	raffle	post
4	main	twelve	march
	mink	welt	margin
	make	wealth	marching
	mate	twelfth	Martin
5	lengthen	geese	rain
	ointment	east	wing
	Lincoln	meat	green
	link	yeast	ring
6	bud	rough	hearing
	bus	drunk	hairy
	bust	rump	carry
	but	rum	herring
7	pleasant	widen	saint
	pheasant	wide	safe
	peasant	wife	faint
	present	wagon	sink
8	winter	model	log
	winner	marvel	lawn
	where	marvelous	blond
	woman	marble	long
9	lose	itself	mash
	loose	excel	gnash
	loot	sell	smash
	blue	himself	nag

Speaker 11 is ...

1	toward	feeling	dome
	forge	dealer	don't
	ford	fever	zone
	board	feeler	stone
2	destroy	girl	flicker
	deprive	pearl	clipper
	defraud	curled	liquor
	defrost	curl	quicker
3	chart	frightful	sultry
	short	rifle	culprit
	shark	greatful	sculpture
	sharp	rightful	sculptor
4	native	pearl	calf
	navy	crow	cad
	naked	throw	calves
	nature	grow	cab
5	lathe	candy	ink
	lay	pantry	pinch
	laid	pansy	inch
	leg	handy	hint
6	thus	legend	hit
	bust	ledger	fist
	duck	leaden	this
	dust	lesson	kiss
7	bulb	cut	net
	bulge	carpet	met
	bald	cotton	neck
	ball	copper	nest
8	breast	Capital	glass
	friend	hapless	lad
	breath	hatless	blast
	bread	happen	black
9	harbor	soft	hood
	harder	sought	could
	ardor	salt	put
	artist	sulk	good

Speaker 12 is ....

1	needle	large	haven
	evil	lodge	heaven
	meal	lie	even
	neither	live	able
2	dimple	interest	cast
	gentle	penguin	past
	devil	hindrance	pass
	dental	kindred	path
3	armload	pen	wooden
	armholed	ten	woody
	armhole	tend	wood
	armful	tent	witty
4	gem	glaze	creeping
	gent	play	greeting
	gin	blade	greedy
	gym	blaze	reading
5	flush	size	waitful
	pledge	sigh	wake
	fresh	scythe	wasteful
	flesh	side	wakeful
6	auburn	astride	dial
	often	ascribe	guile
	author	prescribe	vial
	autumn	describe	guide
7	nest	rug	harrow
	mess	love	herald
	meant	rough	arrow
	met	rub	peril
8	grain	bench	nuptial
	raise	theft	nocturnal
	raid	fetch	nutshell
	rage	thatch	neptune
9	flapper	stole	wallet
	leopard	stone	swallow
	leper	school	wall
	letter	scold	wallow

Speaker Number	Percent Correct
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	

Appendix C

FORM C SPEAKERS I THROUGH 6

When this sheet is in use place carbon paper between it and the next page.

Speaker 1 ....

1	groove	modern	vice
	drew	moderate	fight
	crew	modesty	mice
	grew	modest	bite
2	say	forbade	chink
	stay	pervade	kink
	stayed	surveyed	check
	spade	survey	chin
3	stung	drunk	intent
	stun	grunt	intend
	sun	brunt	content
	stunned	runt	intense
4	quench	busy	wade
	went	physics	waves
	whence	physic	wave
	when	visit	way
5	pass	clearly	fine
	past	weary	find
	cast	quarry	sign
	task	query	kind
6	popular	nurse	get
	poplar	first	gap
	hopper	birth	guess
	opera	burst	guest
7	immense	named	only
	commence	name	woman
	emit	main	pullman
	cement	knave	omen
8	latter	last	swain
	ladder	lash	slain
	lattice	laugh	flame
	rabbit	glass	plain
9	crash	gold	pail
	crab	bowl	poor
	craft	cold	polo
	crack	bold	palace

Speaker 2 ....

1	ninety	drum	harrow
	nineteen	rung	peril
	nightly	rum	herald
	nine	run	arrow
2	ran	putter	need
	rank	tucker	lead
	rang	pocket	lean
	rag	pucker	leave
3	kick	see	depot
	tick	seed	people
	pick	siege	equal
	hick	seize	decoy
4	shower	earthen	bath
	scholar	earthly	bat
	sour	urban	bad
	scour	bourbon	back
5	berry	spring	listless
	carry	pray	mistress
	bearing	spray	restless
	very	spread	blissful
6	mouse	Saturn	fog
	mouth	sat	bar
	now	second	bog
	mount	satin	bug
7	quarter	felt	horrible
	fortress	belt	orchid
	portrait	dealt	orphan
	porter	bell	organ
8	heavy	did	dollar
	happen	live	jealous
	package	led	zealous
	happy	lid	develop
9	hamper	tendon	pond
	pamper	tender	on
	panther	pendant	hound
	pamphlet	pendulum	pawn

Speaker 3 ....

1	apply	gift	lamp
	supply	if	lance
	amply	hit	glance
	fly	it	land
2	bust	handle	free
	fuss	anvil	freeze
	but	amble	freed
	bus	ample	tree
3	airy	fed	laugh
	hairy	stead	glad
	arid	spend	lash
	carry	sped	flash
4	throw	low	rod
	froze	rose	brown
	prose	loathsome	brow
	probe	lonesome	proud
5	desk	stance	science
	depth	stand	silent
	dead	stamp	sound
	death	spent	silence
6	broke	code	begun
	growth	told	begot
	throat	cold	forgot
	wrote	coal	deduct
7	sister	hulk	mild
	system	halt	mile
	cistern	pulp	miles
	pistol	fault	mine
8	strike	limp	town
	spite	limb	townsman
	fight	lend	townsmen
	spike	lent	count
9	paid	cute	fell
	page	cunning	spell
	age	honey	felled
	haze	puny	bell

YOUR NAME ______________ DATE ______________

Speaker 4 ....

1	much	uplift	cypress
	mud	uproot	cipher
	month	approve	siphon
	monk	group	sightless
2	twelve	mind	blister
	well	mild	blissful
	dwell	mine	listful
	weld	line	wistful
3	wren	barter	found
	went	barker	crown
	rent	sparkle	cloud
	lent	parker	clown
4	guide	lively	love
	die	widen	lull
	died	wisely	low
	dive	widely	lag
5	stove	amiss	equipped
	sold	omit	acquit
	stole	amid	equip
	soul	emit	quit
6	reverse	sired	simple
	traverse	siren	dimple
	perverse	fire	pimple
	pervert	sire	temple
7	drove	warrant	dog
	stroke	one	gone
	strode	warm	don
	strove	warn	darn
8	fire	stale	evil
	hire	jail	easel
	tired	dale	measles
	tire	gale	needle
9	gaily	barn	lip
	fail	bark	lift
	daily	bought	lisp
	five	spark	list

Speaker 5 . .. .

1	crash	least	wouldn't
	drag	lease	wood
	trash	niece	wooden
	thrash	leaf	wooded
2	pillow	peg	loosely
	pillar	keg	gruesome
	killer	egg	loosen
	filler	pay	nuisance
3	lava	wait	hour
	loud	which	how
	lock	wake	howl
	robber	wig	owl
4	glad	fable	part
	lad	tablet	art
	laugh	habit	heart
	lag	cattle	arch
5	puncture	sigh	bake
	teacher	size	bait
	tincture	side	fate
	picture	scythe	faith
6	tempt	green	seller
	tense	cream	solemn
	tent	tree	solid
	hemp	creed	sullen
7	youth	allege	muster
	you	away	lusty
	use	allayed	bluster
	mute	allay	luster
8	light	birds	chat
	pike	bird	cap
	height	birth	check
	hike	burden	chap
9	devise	chaff	Ed
	defy	shaft	head
	divide	chap	add
	beside	shack	ebb

Speaker 6 .. ..

1	feel	fruit	pelvis
	deal	true	elder
	steal	troop	elbow
	veal	truth	eldest
2	tasty	sheep	add
	hasty	shield	ask
	hasten	she	as
	pastry	sheath	has
3	wrist	depth	fortune
	risk	death	fort
	rip	deaf	important
	list	guest	forty
4	shoe	defense	hamper
	choose	methinks	tamper
	too	repent	hampered
	chew	bethinks	hamburg
5	led	palace	stow
	red	palate	stole
	ledge	talent	stowed
	leg	pilot	stove
6	butter	heat	tick
	flutter	hate	chicken
	flood	paint	ticket
	fluttered	ink	picket
7	thumb	coy	auto
	from	toy	bottom
	come	tore	often
	sum	torque	autumn
8	bower	fast	sit
	borrow	fact	six
	flower	fat	sick
	power	that	sift
9	deceive	cars	heard
	precede	carve	verge
	concede	card	urge
	receive	car	herb