1991

Computer-generated speech training versus natural speech training at various task difficulty levels

James Michael Fillpot

COMPUTER-GENERATED SPEECH TRAINING VERSUS NATURAL SPEECH TRAINING AT VARIOUS TASK DIFFICULTY LEVELS

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

in

Psychology

by

James Michael Fillpot

September 1991

COMPUTER-GENERATED SPEECH TRAINING VERSUS NATURAL SPEECH TRAINING AT VARIOUS TASK DIFFICULTY LEVELS

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

by

James Michael Fillpot

September 1991

Approved by:

[redacted]

Janet L. Kottke, Chair, Psychology

3/15/91

Date

[redacted]

Vera J. Dunwoody, Psychology,
Chaffey College

[redacted]

James G. Rogers, Management

Abstract

The researcher conducted an experiment that examined performance degradation at varying task difficulty levels between subjects trained with either natural or computer-generated speech. The researcher hypothesized that: 1) in presenting a simple task, subjects trained via computer-generated speech would exhibit similar performance rates as subjects presented with natural speech training; 2) in presenting a moderately difficult task, subjects trained via natural speech would suffer minimal performance degradation compared to subjects presented with computer-generated speech training; 3) whether presented with natural or computer-generated speech training, all subjects would experience statistically significant performance degradation between moderately difficult and difficult task levels; and 4) while performance degradation would occur for both computer-generated and natural speech training as task difficulty increased, computer-generated speech trained subjects would exhibit statistically significantly greater performance degradation at the difficult task level than natural speech trained subjects. Data analyses supported the third and fourth stated hypotheses, where statistically significantly lower performance rates were observed among the group trained via computer-generated speech.

Table of Contents

Table of Contents (continued)

List of Tables

List of Figures

List of Appendices

Introduction

In examining the corporate world, one of the major (if not the primary) concerns is cost effectiveness. In both the public and private sectors of business and industry, administrators and managers are constantly examining the work environment in order to ascertain areas where costs can be reduced. As well as controlling the budget, responsible parties continually pursue avenues which may also improve the production ability of employees at the onset of tenure with the organization. In analyzing both of these facets of the work environment (cost effectiveness and increased production via rapid, trenchant training), an area which holds great promise in fulfilling these administrative expectations is computer-generated speech, specifically training employees via computer-generated speech.

Traditionally, employee training has been conducted employing natural speech as the main channel of interaction in information transmission. However, current research provides a great deal of support and hope for the future usage of computer-generated speech as an alternative training tool. Recently, Schwab, Nusbaum, and Pisoni (1985) demonstrated the effectiveness of computer-generated speech as a method for training employees. Schwab et al. presented five sets of stimulus material to nine subjects using synthetic speech as the mode of training. Synthetic speech was generated by the Votrax Type-'n-Talk system and "was chosen primarily because of the relatively poor quality of its segmental (i.e., consonant and vowel) synthesis" (Schwab, Nusbaum, and Pisoni, 1985, p. 398). The five sets of stimuli presented were: (1) 12 lists of 50 monosyllabic phonetically balanced words (PB lists); (2) four sets of 50 monosyllabic words taken from the Modified Rhyme Test (House, Williams, Hecker, and Kryster, 1965) (MRT lists); (3) 100 Harvard psychoacoustic sentences (Egan, 1948; IEEE, 1969) (Harvard sentences); (4) 100 syntactically normal but semantically anomalous sentences developed at Haskins Laboratories (Nye and Gaitenby, 1974) (Haskins Sentences); and (5) 39 prose passages taken from a variety of sources, including popular magazines and reading comprehension tests (Prose passages). The entire experiment lasted two working weeks (Monday through Friday), with approximately a one-hour session administered once per day. Seven subjects received identical testing, with the exception that the stimuli material was presented using natural speech (speech produced by a male talker). Ten subjects were in the control group and received no training, participating in the experiment on Day 1 (pretest) and Day 10 (posttest) only. At the conclusion of the study, Schwab et al. found that, while the performance level of the natural speech group was consistently higher than that of the synthetic speech group, subjects in the synthetic speech group showed a more consistent increase in their performance over the course of training than did the natural speech group. The lower performance level of the synthetic speech group was expected by the researchers, due to the quality of the computer-generated speech. In previous experiments, researchers have found similar results using lexical (base word) decision tasks with natural and synthetic speech. Pisoni (1981) found that performance improved for both types of speech within a one-hour session, while Slowiaczek and Pisoni (1982) found performance improvements for both types of speech after a five-day experiment. Greenspan, Nusbaum, and Pisoni (1985) also found that training subjects with synthetic speech improved their ability to recognize both words and sentences. Most encouraging, Schwab et al. noted:

> . . . based on a trend analysis, no evidence was obtained that these subjects had received asymptotic levels of performance even by the last day of training with the synthetic speech. Thus, it is entirely possible that further improvements could have been obtained if training had been carried out for a longer period of time (p. 404).

Audition as Training Modality

Existing research tends to support audition as the most effective method of training, at least in comparison to other sensory input modality. Computer-generated speech training in particular appears to provide a number of advantages as a training tool. Klapp, Kelly, and Netick (1987) examined hesitations in continuous tracking that involved a number of concurrent discrete tasks. All subjects were instructed to perform a visually guided pursuit tracking task with their right hand. The experimental group, however, was also instructed to perform a concurrent auditory reaction-timed task involving manual handle movement with their left hand. Since right-

and left-handed manual tasks involved the same forced function, any differences in tracking performance were expected to be attributable to the presence of the secondary stimuli (response time to the auditory signal). After controlling for active muscle freezing and muscle relaxation, Klapp et al. found a minimal right-hand rate of hesitation associable to the introduction of the secondary concurrent task. As Klapp et al. note, "The rate of hesitations decline with practice, and this improvement in right-hand performance was accompanied by an improvement in performance of the concurrent left-hand response" (p. 327). Similar research by Wickens, Mountford, and Schreiner (1981) has found that responses to speech displays could be time-shared with a visual/manual tracking task with no decrement in performance from the single- to the dual-task condition for speech displays. These findings indicate that, in tasks requiring multiple responses, cuing of one or more tasks via auditory stimulation creates minimal degradation of task performance.

Despite findings that audition might be a superior mechanism, Hofmann and Heimstra (1972) have discussed the widespread usage of visual displays in most man-machine systems, citing the frequency with which visual displays have been employed to convey information to the human operator. As these researchers note, overload problems in visual display training pose a serious problem. Alternative modes of training which are equal or better must be discovered. In investigating the effectiveness of auditory, cutaneous and visual feedback displays on compensatory tracking tasks, Hofmann and Heimstra found that, based on the two significantly independent dimensions of speed of response and goodness of performance, auditory feedback displays proved to be the most effective method utilized in compensatory tracking tasks.

Citing the advantages of speed of response and goodness of performance, Hakkinen and Williges (1984) studied the effects of presenting emergency messages to subjects in a simplified air traffic control task experiment via computer-generated speech. In a series of experiments, Hakkinen and Williges varied the presence of light and tone alerting cues and the presentation of non-critical messages (visual or auditory). Hakkinen and Williges found that, when computer-generated speech was used for multiple functions, a greater number of emergency messages were detected. In single function tasks, visual alerting cues were even found to present a detrimental effect, lengthening the response time to the perception of the message. Wickens's theory on multiple resources (1980) supports Hakkinen and Williges' findings. According to Wickens, the introduction of another input modality for secondary information will result in the operator incurring less mental workload than when using the same modality as that required in the primary task.

In two studies involving experimentation in a cockpit environment, further evidence has been forwarded to support the implementation of speech as an input modality. Hawkins, Reising, Lizza, and Beachy (1983) compared speech and pictorial displays. While a slight difference was found between these two input modality in terms of actual performance, a post-test questionnaire showed that subjects overwhelmingly preferred the speech display. Williamson and Curry (1984) presented a visual tracking task to subjects, introducing secondary systems status information either through speech, pictorial, or alphanumeric displays. Responses to the speech mode were found to be faster than responses to the pictorial or alphanumeric displays.

Evidence that secondary displays partially negate or decrease the effectiveness of computer-generated speech training has been presented by Luce, Feustel, and Pisoni (1983). Luce et al. compared recall performance for computer-generated and natural speech monosyllabic word lists. Luce et al. found that the visual display of digits of varying lengths prior to the presentation of the spoken word lists reduced the subjects' ability to retain words presented via computer-generated speech as the digit length increased. Thus, similar to the results reported by Hakkinen and Williges, single-function extraneous sensory inputs tend to decrease the effectiveness of the computer-generated speech message.

In examining the evidence, it appears that computer-generated speech training presents unique advantages, especially in practical tasks which require critical feedback and multiple-function ability. While the introduction of input modalities other than speech tend to degrade the primary task where speech is the input modality, the introduction of speech as the secondary input modality does not appear to conversely create performance deterioration in primary tasks involving other sensory input modalities.

Natural Speech Analysis

Although computer-generated speech presents a number of advantages as a training tool, the most critical drawback appears to be the inability of computer-generated speech to intelligibly mimic natural speech. Before discussing important considerations in the production and transmission of computer-generated speech, an analysis of natural speech must first be conducted in order to more readily understand the requirements of successful language transmission.

Sherwood (1979) presents a fine example of how natural speech sounds are generated. Over-pressure and lowered pressure in the lungs causes the vocal folds to blow apart and collapse back together, respectively. This process repeats itself periodically. However, as the vocal folds are blown apart, a puff of air is emitted into the mouth cavity. This puff of air creates a sound, an acoustics signal which can be varied depending on the position of the tongue, lips and jaw.

Phonemes.

In examining natural speech sounds, Van Gieson and Chapman (1968) point out that a speaker of English creates speech by combining about forty different classes of sounds. These sounds are known as phonemes. A phoneme is defined as the smallest meaningful unit in the sound system of a language. Simpson and Marchionda-Frost (1984) alter this definition somewhat, adding the caveat that a phoneme, ". . . if changed, will alter the meaning of the word (See Appendix A for a phoneme chart display)" (p. 509). For example, the sounds "t" and "d" in *tin* and *din* distinguish the two words. The vowels "a" and "e" in *tan* and *ten* also operate in this manner, creating a different sound for each word. Thus, as Van Gieson and Chapman state, "words are created by generating proper sequences of the various phonemes" (p. 31). Language is an extension of this principle, creating messages by properly sequencing combinations of words. However, the inclusion of segments of silence is equally important to the successful transmission of spoken messages. Unlike printed messages, spoken messages are made up of sequences of sounds which are continuous and may blend together. Thus, the major consideration in developing and analyzing spoken messages is the sequencing and combination of phonemes and segments of silence.

As complex as language creation may appear on the surface, it is aided a great deal by the fact that speech is highly redundant. Most words contain sounds and segments of silence which can be eliminated without reducing intelligibility. In natural speech, the speaker often eliminates these components of language without realizing it. The long-winded speaker who is running short of breath often strings words together without any pause. In a similar fashion, certain English dialects and regional accents eliminate phonemes in words without significantly affecting intelligibility.

Kent (1973) spent considerable time evaluating phoneme combinations in the creation of intelligible natural speech. Kent found that, in examining human imitation of computer-generated vowels, vowels are represented in memory as a continuous transformation of the acoustics signal or a representation of classification transformation. Once again, phonemes and segmentations of silence (or lack of) play an important role in the production of natural speech. Further research by Kent (1974) indicates that the recognizability of the phoneme also plays an important role in speech intelligibility. In presenting American English vowels and foreign language vowels to American subjects, Kent found that intelligibility was significantly greater for American English vowels. Kent attributed this finding to the American subjects' lack of familiarity with foreign vowels. Thus, ambiguity in the spoken message, primarily in phoneme recognition, reduced natural speech intelligibility.

de Haan and Schejerderup (1978) conducted research similar to Kent's, providing results that support Kent's findings. de Haan and Schejerderup examined the intelligibility of connected speech, or speech without segments of silence. These researchers found that, while intelligibility of compressed speech remained relatively high, comprehension test scores decreased as rate of speech compression increased. This finding is attributable primarily to the fact that compression increases the speed of speech and thus creates pitch distortion. Speech rate, pitch, frequency and other facets of message intelligibility will be discussed later in this paper.

Consonants.

In addition to the study of vowels and how they relate to phonemes and segmentations of silence in the production of natural speech, other researchers have also examined the role of consonants in natural speech generation. Yuchtman, Nusbaum, and Pisoni (1985) have found that, in perceiving consonants, spacing and structure play a vital role in intelligibility. Based on the indications of the aforementioned research, further support exists to substantiate the findings that phonemes and segments of silence are all important in intelligible natural speech generation.

Keeping in mind the combinations of phonemes and segments of silence which create words and spoken messages, various researchers have attempted to pinpoint accurate methods of measuring natural speech transmission quality through a variety of procedures. Steeneken and Houtgast (1980) note that natural speech transmission quality is often determined by the performance of speakers and listeners on intelligibility tests. While this approach has some advantages, the lack of well-trained speakers is a major drawback. Steeneken and Houtgast suggest that the channels through which speech is transmitted should be studied.

Channels of Speech Transmission.

In following this line of reasoning, research exists which supports the contention that optimum channels of speech transmission exist. Literature indicates that esophageal speech is widely and routinely preferred over laryngeal and artificially-generated speech. Hyman (1955) demonstrated in early research on natural speech production that esophageal speech was preferred when only auditory cues were available to the listeners. Crouse (1962) extended these findings, discovering that esophageal speech was preferred by both sophisticated and naive listeners when judgments could be based on both auditory and visual cues. A number of researchers (Arnold, 1960; Curry and Snidecor, 1961; DiCarlo, Amster, and Herer, 1956; and Gardener and Harris, 1961) agree that esophageal speech is more convenient and easily understood. More recent research (Clark and Stemple, 1978) supports these contentions. In testing listeners for preference rankings, Clark and Stemple found that esophageal speech was preferred over artificial laryngeal speech, normal laryngeal speech, and even pulmonary esophageal speech.

In viewing the research on natural speech generation, one can see that a few very basic concepts must be kept in mind. Natural speech is produced by phoneme combinations and segments of silence. Research which examines vowel and consonant identification and intelligibility lends support to this view. While phonemes and segments of silence can be eliminated or altered, research also indicates that at prescribed levels elimination or alteration has a seriously detrimental effect on intelligibility. Wholly separate from phoneme combinations and segments of silence, the mode or channel by which natural speech is presented also creates a great deal of difference in the intelligibility of natural speech. Esophageal speech is often preferred over other methods of speech generation. Keeping these facts about natural speech generation in mind, the researcher will now examine these requirements in the context of computer-generated speech. Specifically, the researcher will examine how computer-generated speech is created and, in light of the aforementioned research on natural speech production, discuss some considerations in synthetic speech generation.

Computer-Generated Speech Analysis

Currently, a number of computer-generated speech systems exist. However, these systems basically operate on the same principle. Going back to the aforementioned section on natural speech, Sherwood discussed how natural speech sounds were generated by puffs of air into the mouth cavity. These sounds were altered based on the position of the tongue, lips and jaw.

Most important to the reproduction of speech via computers, these sounds occur in a periodic yet repetitive manner. The segments of silence in between these sounds create breaks in the sounds produced. This is the fundamental cornerstone of computer-generated speech. The production of phoneme sounds and the segments of silence in between these sounds create waveshapes. Waveshapes can be reproduced by computers, thus in effect creating the base by which sounds and speech can be generated. However, a great deal more must be considered before computer-generated speech can come close to replicating natural speech.

Formant Frequencies.

The first aspect which must be examined in computer-generated speech is how computers, operating from the base of waveshape formation and recognition, replicate the functions of the tongue, jaw and lips. As Doddington and Schalk (1981) state, "a key element in recognizing the information in spoken sound is the distribution of energy with frequency" (p. 28). Particularly important are the energy peaks, or formant frequencies. A formant is usually described in terms of harmonic-oscillator resonances (alternating vibrations that are an integral multiple of the fundamental frequency). The size of the peaks of the waveshapes are determined in part by formant frequencies. In human speech, frequencies are created and altered by the tongue, jaw and lips. Because frequencies can be measured exactly, computers can be programmed in a number of ways to recognize and reproduce basic frequencies, thus altering the size of the waveshape. While formant frequency analyses become more complicated as phoneme combinations become more complex, this basic tenet continues to hold true. Difficulties for the computer to replicate human speech occur when various dimensions of speech are in constant transition (i.e., pitch, frequency, speed, noise amplitude, etc.). These points will be discussed in greater detail in a later section on message intelligibility.

Advances in technology have resulted in the creation of computers that are capable of extracting measurements from acoustic signals (Guillemin and Nguyen, 1984). The two primary methods of creating computer-generated speech are synthesized speech and digitized speech. Simpson, McCauley, Roland, Ruth and Williges (1985) define both of these methods of computer-generated speech:

> Synthesized speech refers to speech generated by rule, without the aid of an original human recording. The term digitized speech applies to human speech that was originally recorded digitally . . . another pair of terms used to describe these methods are synthesis by rule for speech synthesis and synthesis by analysis for digitized speech generation (p. 118)

Rule-Generated Speech.

As the definition of synthetic speech indicates, speech can be generated by rule. Ainsworth (1974) states, "in this method, each utterance of the vocabulary is stored as a sequence of numbers representing its phonetic transcript" (p. 493). The computer stores tables which enables it to create speech by selecting the parameter values necessary for synthesizing each utterance (Ainsworth, 1972). Yulsman (1983) offers further clarification, discussing how computers are programmed with basic phonemes, as well as rules of pronounciation and stress, from which it assembles words. Yuslman notes the enormous versatility affordable in synthesis by rule, since *any* word can be created and introduced. However, what is gained in flexibility is often lost in clarity. It is extremely difficult, if not impossible under the limits of current technology, to reduce all the permutations and inflections involved in pronunciation and speech down to a single, specific set of rules.

Digitized Speech Generation.

In synthesis by analysis, or digitized speech, the computer takes recorded samplings of the human voice and analyzes the sound wave at key intervals (usually every one-hundredth of a second). Key attributes such as predominant frequencies and energy levels (discussed in-depth later in this study) are extracted and stored. The computer is now capable of "mimicking" speech. Through a series of electrical impulses, the computer, through the use of filters, oscillators, and noise generators, creates sound. Since computers have a pre-created pattern to monitor and mimic, subtle nuances can be captured and stored, creating extremely lifelike voices. However, as is the case with synthesis by rule, drawbacks do exist. The actual vocabulary a computer can produce is limited to the words that have been programmed into it's memory. This type of programming requires an amazingly large amount of memory, reducing the number of words that can actually be reproduced. In storing a large number of words, the price becomes quite costly, prohibiting usage.

Gallant (1987) compares both types of computer-generated speech. In examining the pros and cons of both types of computer-generated speech, Gallant notes the "quantity vs. quality" issue as the major consideration in determining which computer-generated speech system is best suited for an individial or company's needs (p. 63). However, Gallant further states that advances in computer chip technology are allowing larger vocabulary lists to be committed to a computer's memory. A number of major computer-oriented corporations, led by Texas Instruments, are strong proponents of synthesis by analysis, a possible indication that any major breakthroughs or advances in technology will likely occur in synthesis by analysis before synthesis by rule.

While advantages and disadvantages exist for both types of computer-generated speech, for the purposes of this thesis only synthesis by analysis (digitized) speech will be examined and studied. The researcher chose to examine digitized speech over synthesis by rule speech for a number of reasons. Digitized speech systems can have an unlimited variety of different voices since they depend on human speakers for their vocabulary. Synthesis by rule speech does not depend on human speakers for new vocabulary; this limits synthetic speech systems to about six different voice types. Because it mimics human speech, digitized speech comes closer to replicating natural speech than synthesis by rule speech, a key factor in this study. Digitized speech can also usually be stored easier in a degraded form, thus making it more economical. A number of research studies (Flanagan, Johnston, and Upton, 1982; Campbell, 1974; and Schroeter and Sondhi, 1985) demonstrate the practical advantages of digitized speech.

Analog Speech.

An alternative method of computerized speech delivery is through analog speech. As Smith and Goodwin (1970) note, certain dimensions of speech occur naturally in an analog form. For example, frequency can be measured in cycles per second, for which the international reference is Hertz (Hz). Loudness is associated with the intensity of the sound and is expressed in terms of decibels (dB) (McCormick and Sanders, 1987). The point is that both Hertz and decibels, the measurements used to represent frequency and loudness, respectively, can be determined through equations. The result is a measurement which can be expressed as an integer. Smith and Goodwin demonstrate how, in this raw form, an analog computer (a computer that operates with a functional relationship among directly observable quantifications) can understand integers which represent certain levels of frequency, loudness, etc. and, through various means of amplification, reproduce a sound.

Although analog computers are capable of reproducing sounds, the intelligibility of such sound is low. Analog sound formation serves a more useful purpose in that it is the basis from which a number of computer-generated speech systems create digitized speech. As cited earlier, digitized speech refers to pre-recorded human speech which is usually converted first into analog (single integer) and then digitized (multiple numeric digits) form. Digitized speech is better than analog form speech in that the implementation of multiple digits allows for storing of strings of phonemes (known as framing), creating more natural sounding, intelligible speech.

Smoothtalker.

In terms of the type of synthesis by analysis (digitized) speech system to be used in this study, the researcher is limited by what is personally available. However, the research on one of the text-to-speech systems available to the researcher, Smoothtalker, supports the feasibility of implementing such a system. In Smoothtalker (produced by First Byte, Inc.), text is parsed using letter-to-sound rules which serve to generate control codes. These codes are then matched against prestored allophonic segments. The segments are concatenated together to produce a speech waveform. Logan, Greene, and Pisoni (1989) recently compared ten text-to-speech systems to natural speech.

Logan et al. found that the greatest influence on text-to-speech system intelligibility was the accoustic-phonetic knowledge present in the rules used in the formant synthesis system. However, in determining that the segmental intelligibility scores of the ten text-to-speech systems formed a continuum, Logan et al. found that Smoothtalker, while not fairing nearly as well as high-quality systems such as DECtalk or Prose, substantially outperformed text-to-speech systems such as Votrax and Echo. In examining overall error rates on the Modified Rhyme Test (MRT), Smoothtalker placed seventh out of the ten systems, with an overall error rate of 27.22. While experiencing a statistically significantly higher error rate than natural speech (0.53), Smoothtalker provides a good representation of computer-generated speech systems, serving as a middle-of-the-road example of such systems for general comparison purposes.

Intelligibility

To this point, the researcher has examined existing literature on training employees with either natural speech or computer-generated speech and how these two modes of speech are produced. In discussing the production of either natural or computer-generated speech, one would be remiss in neglecting to discuss speech intelligibility.

Intelligibility has been defined in a number of various ways. Early research (Woodsworth and Schlosberg, 1954; Foulke, 1965) examined reaction time, believing that low intelligibility resulted in increased choice reaction time (goodness of response). Later, Foulke and Sticht (1969) noted that intelligibility could be defined as, "the ability to repeat a word, phrase, or short sentence accurately (goodness of performance)" (p. 52). McCormick and Sanders (1987) term intelligibility, "the extent to which the transmitted message is understood by the listener" (p. 157). Simpson, McCauley, Roland, Ruth, and Williges (1985) claim the term intelligibility has a very precise meaning. Simpson et al. refer to intelligibility as, "the percentage of speech units correctly recognized by a human listener out of a set of such units" (p. 118). It is this later definition of intelligibility which will be employed for the purposes of this study.

While it is fairly easy to provide a general operational definition for intelligibility, a plethora of sub-components exist which, either individually or en masse, affect intelligibility. These main components include (but are not restricted to): time-compression; frequency; pitch; noise; sensation level; the properties of the message being transmitted (basically, the content of the message); the meaningfulness of the message; and the syntax or syntactical structure of the message being transmitted.

Time-Compression

As the research cited earlier mentions, phonemes and segments of silence can be removed without any apparent change in the pace of a computer-generated word or its intelligibility. Removing phonemes or segments of silence increases the speed of the message being presented. This is referred to as time-compressed speech. Foulke and Sticht also refer to time-compressed or accelerated speech as, "speech which has been reproduced in less than the original production time" (p. 50).

In natural speech, compression is created by the speaker increasing the pace of his or her presentation, or by taping the presentation and then replaying it at a different speed level. In computer-generated speech, one of two methods is basically employed. Usually, either segments of silence or phonemes in the presentation are removed or the entire message is compressed together, increasing other functions such as pitch and frequency. For now, however, only the effects of time-compression itself will be examined.

Two basic methods exist by which computer-generated time-compressed speech is studied. The first body of research examines the presentation of words (simple tasks) at various speeds. Some recent research exists which supports the contention that compression does not affect intelligibility on word recognition tasks. For example, seminal research conducted by Garvey (1953) on intelligibility of time-compressed speech, Garvey found that, at compression speeds of up to 2.5 times that of the original speech, intelligibility was minimally affected by compression (93.33% and higher intelligibility). However, at three times original speech speed, intelligibility dropped to 78.33%; at 3.5 times, 58%; and at four times, 40%. In examining the advantages of computer-generated speech training over natural speech training, Pisoni (1981) found that computer-generated speech was recognized faster than natural speech when it was compressed. As Pisoni notes, widespread advantages exist for "its (compressed speech) application in voice response systems used in applied settings." In presenting words at intervals of 1, 2, and 5 seconds per word, Luce, Feustel, and Pisoni (1983) found that the decrement in intelligibility of computer-generated speech did not increase at faster speech rates. In similar research, Simpson and Marchionda-Frost (1984) presented words at word/minute rates of 123, 156, and 178 to helicopter pilots. Simpson and Marchionda-Frost found that intelligibility did not decrease at faster rates. However, these researchers did find that the response time to messages at faster rates increased. This finding was attributed to the need for additional cognitive processing time at faster speech rates.

While the majority of research supports the contention that time-compression does not affect the intelligibility of speech, research findings to the contrary also exist. For example, Beasley, Schwimmer, and Rintelmann (1972b) presented time-compressed monosyllable words under five time-compression conditions, ranging from 30% to 70% in compression ratios. Research indicated that intelligibility was inversely related to time-compression ratio. de Haan (1977) presented research in which compression rates were also greatly altered. de Haan presented words at seven different rates, ranging from 203 words/minute to 408 words/minute. de Haan's findings: intelligibility decreased greatly as compression increased.

The second body of research examines the intelligibility of time-compressed sentences (difficult recognition tasks). While it appears from the literature surveyed that intelligibility does not decrease drastically when words are time-compressed, inverse findings are found when the intelligibility of time-compressed sentences is examined. Wingfield (1975) compressed sentence consisting of 10 English words to 80%, 70%, 60%, 50%, and 40% of normal playing time, corresponded to rates of 259, 296, 345, 414, and 518 words/minute, respectively. In comparison to the intelligibility of normal rate speech which was 87.5%, sentences compressed to 80%, 70%, 60%, 50%, and 40% of normal playing time displayed intelligibility rates of 80%, 66.6%, 69.3%, 27.8%, and 10%, respectively. As Winfield states:

> . . . the perceptual act is not a passive handling of the speech on a word-by-word basis . . . so long as there is some minimal intelligibility, subjects actively reconstruct the heard fragments so as to produce responses that are meaningful

Wingfield, Buttet, and Sandoval (1979) proffered further research which examined time-compression effects on sentences in both English and French. Wingfield et al. found that, in both English and French, as sentence compression increased (implementing the same rates as in Wingfield's previous study), intelligibility decreased. Slowiaczek and Nusbaum (1985) also found similar effects of time-compression on sentence intelligibility, noting that in slow sentences (150 words/minute) correct word identification was 86.7%, while in fast sentences (250 words/minute), intelligibility was down to 58.9%. Over and above these experiments, additional research exists (Beasley, Bratt, and Rintelmann, 1980; Winfield, Lombardi, and Sokol, 1984; Maarics and Williges, 1988) which suggests that the intelligibility of sentences decreases as compression rate increases. Foulke and Sticht attribute decreases of intelligibility of time-compressed sentence to the perceptual and cognitive processes of the listener. This will be discussed in greater detail in a later section on listener perception, abilities and requirements.

From the research on time-compression, one can see that an interesting pattern begins to emerge. It appears that, in low difficulty comprehension tasks (i.e., word list identification), compression does not affect the intelligibility of the message presented. However, in high difficulty comprehension tasks (i.e., sentence or prose passage identification), intelligibility decreases as compression of computer-generated speech increases.

Frequency

As was previously mentioned, a sound-generating source emits a series of waveshapes which affect the surrounding molecules, changing the surrounding air pressure. The magnitude of the waveshapes, how long it takes the waveshape to affect above normal and below normal changes in surrounding air pressure and return to a midline point, is called a cycle. Frequency refers to the number of cycles a sound makes per second. Frequency is expressed in terms of Hertz (Hz), which is equivalent to cycles per second. Different sounds produce a different number of Hertz. Middle C on the musical scale has a frequency of 256 Hz; an octave higher would produce 512 Hz. The human ear is sensitive to a wide range of frequencies, capable of hearing sounds in the 20 to 20,000 Hz range (McCormick and Sanders, 1987).

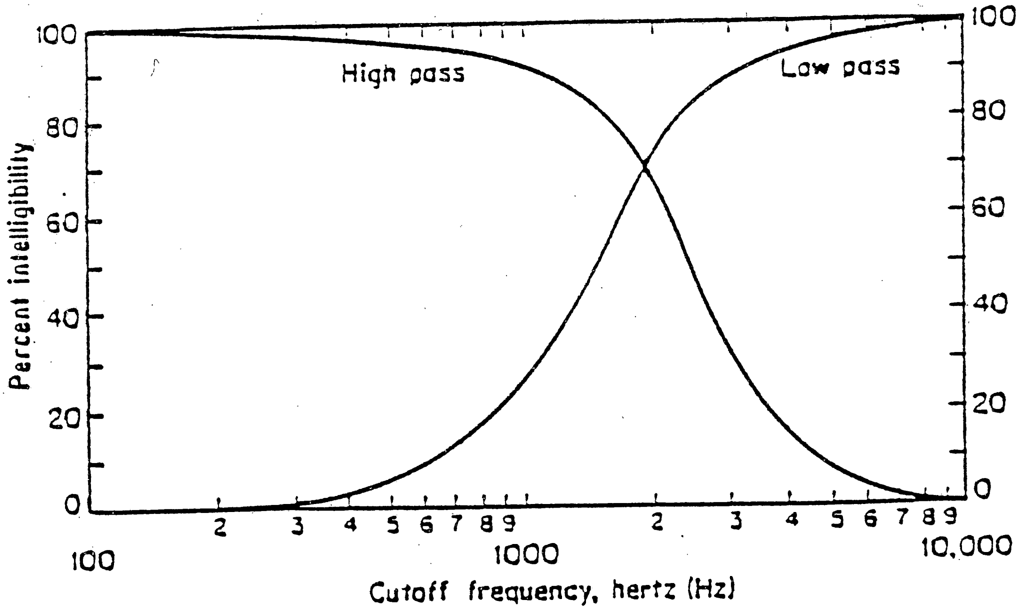
In addition to the normal frequency which a sound emits, frequency can also be altered by artificial means. Time-compression of speech can greatly alter frequency. In time-compressed speech, sound is condensed and the cycles per second ratio increases. In speech, increasing the frequency of a sound above (or below) its normal frequency range affects the intelligibility of the sound. Fortunately, filtering devices exist which allow the researcher to examine and, if desired, correct sounds altered by artificial means.

Speech is filtered by blocking out certain frequencies, thus permitting only selected frequencies to be transmitted. Frequency filtering devices are usually of two types: high-pass filters, or low-pass filters. High-pass filters eliminate frequencies below a preset level. Low-pass filters operate in the exact opposite fashion, removing frequencies above a preset level. Different filtering levels will affect speech in different ways. French and Steinberg

(1947) provide a solid depiction of how frequency filtering affects the intelligibility of speech (see Figure 1).

In early research, Giolas and Epstein (1963) demonstrate how intelligibility is affected by frequency variations created through filtering. Giolas and Epstein examined monsyllabic and phonetically balanced words, as well as representations of speech encountered in everyday situations. These

Figure 1. Effects on intelligibility of elimination of frequencies by the use of filters.



Note. From Human Factors Engineering and Design (p. 164) by E. J. McCormick and M. S. Sanders, 1987, New York, N.Y., McGraw-Hill Book Company. Copyright 1987 by McGraw-Hill Book Company.

speech samples were passed through seven low-pass frequency filtering conditions (no filtering, 2,040 cycles per second (cps), 1,560 cps, 1260 cps, 960 cps, 780 cps and 540 cps) with a 30 dB/octave frequency cut-off. The researchers found that, as frequency distortion caused by filtering increased, intelligibility decreased for both word lists and continuous discourse. Phonetically balanced words were found to be less intelligible than monosyllabic words as frequency distortion increased. Greater frequency distortions also increased error rates for continuous discourse.

Speaks and Jerger (1965) studied the effects of low-pass frequency filtering in the intelligibility of "real" sentences. Speaks and Jerger defined real sentences as sentences whose "meaning may be conveyed by only one or two key words" (p. 187). The key words for the sentences were chosen from a pool of the 1000 most common words as identified by the Thorndike-Lorge (1944) count. These sentences were taped and routed through a low-pass frequency filter with a cut-off frequency of 350 cps and an attenuation rate of 24 dB/octave. Speaks and Jerger found that, when a message was low-pass filtered in order to improve intelligibility, performance improved, especially when the amount of information transmitted was minimal.

Speaks (1967) extended his previous research on the intelligibility of filtered computer-generated speech when he examined the effects of low-pass and high-pass frequency bands on sentences intelligibility. As Speaks notes:

> French and Steinberg (1947) suggest that the most important frequencies for intelligibility of monosyllables occur between 1,500 and 2,500 Hz. The point of intersection of low- and high-pass functions indicates that frequencies above and below 1900 Hz. contribute equally to intelligibility (p. 289).

However, Speaks found that low-pass filtering appeared to be significantly more important to intelligibility than high-pass filtering. When the cut-off frequency was set at 1000 Hz, the level of correct responses was similar to that obtained when no filtering was used. Thus, the addition of frequencies above 1000 Hz appears negligible. High-pass frequency filtering must be extended down to 300 Hz before significant results are obtained, also indicating that high frequency energy is not as vital as low-frequency filtering. Interestingly, Speaks found that the low- and high-pass filtering intersection occurred at approximately 725 Hz, substantially below the findings of French and Steinberg.

Recent research focuses on the functional gain associated with frequency response. Functional gain can be defined as, "the difference between aided and unaided thresholds for third-octave bands of noise" (Pascoe, 1975, p. 6). Functional gain is important in that it reflects the true gain (over-correction response) produced by the listener. In examining functional gain and frequency response, Skinner (1980) presented five frequency levels to subjects with normal hearing and subjects with permanent noise-induced hearing loss above 1000 Hz. Hearing-impaired listeners have more difficulty identifying high-frequency speech sounds than low-frequency speech sounds. Skinner found that, compared to hearing-impaired listeners, nonhearing-impaired listeners experienced a 20 to 30 dB functional gain increase at each frequency level. Similar research (Owens and Schubert, 1968; Owens, Benedict, and Schubert, 1972; Pisoni and Koen, 1982; Bornstein, Randolph, Maxon, and Giolas, 1982) supports these findings.

In examining frequency response level and functional gain, some interesting themes begin to emerge. While, the human organism is capable of hearing a wide spectrum of sounds, optimum ranges exist for speech intelligibility. Early research focused on determining this range; however, contradictory findings continue to emerge. While speech sounds are affected differently by the elimination of various frequencies, it appears that the optimum range falls somewhere between 300-600 Hz and 4,000-4,5000 Hz (depending on the frequency filtering pass implemented). The intersection at which high or low-pass frequency filtering affects intelligibility the same is somewhere between 700-2000 Hz. Low-pass filtering appears to be more critical to speech intelligibility. Recent research examining functional gain indicates that, in addition to optimum frequency response levels, individual variations in frequency perception affect intelligibility. As frequency response levels increase, so do functional gain.

Pitch

The term pitch is used to refers to the highness or lowness of a tone.

As McCormick and Sanders note, "since high frequencies yield high-pitched sounds and low frequencies yield low-pitched tones, we tend to think of pitch and frequency as synonymous" (p. 126). However, a number of other factors come into play which allow researchers to differentiate between pitch and frequency. For example, one of the more predominant factors that influences the perception of pitch is the intensity of the tone. Intensity is associated with human sensation levels or loudness, which will be discussed in greater detail in a following section. For the purposes of discussing pitch, it is sufficient to state that when intensity increases, low-frequency tones (tones less than 1000 Hz) and high-frequency tones (tones greater than 3000 Hz) become lower and higher in pitch, respectively. A number of researchers have examined the effects of pitch modification on the intelligibility of speech. What follows is a representative selection of this research.

Rabiner (1977) notes that in examining pitch detection, one of the most robust and reliable methods of pitch detection is autocorrelation analysis. While autocorrelation analysis is a time consuming process, computations are made directly on the waveform. The autocorrelation computation is also easily amendable to digital hardware implementation, a feature which makes this method of pitch detection attractive for both natural speech and computer-generated speech analysis. Further, this method of pitch detection is basically insensitive to phase distortion, a noteworthy distinction considering the varying ranges of intelligibility associated with computer-generated speech. However, while autocorrelation analysis presents a number of advantages, there are still several problems associated with its use. In analyzing a section of speech, a number of autocorrelation peaks are created due to the formant structure. Thus, one problem is deciding which autocorrelation peak corresponds with the main pitch peak. A second problem is determining the period of time, the window, which is sufficient for analysis. Ideally, the analysis window should contain 2 to 3 pitch periods. For higher pitches the window should be short (5-20 ms); it should be longer (20-50 ms) for lower pitches. While autocorrelation analysis begins to provide researchers with a method for analyzing pitch, further methods of pitch manipulation and control must also be considered.

de Haan and Schjelderup (1978) describe existing instrumentation which allows speech rate to be varied with or without pitch. As speech is compressed, pitch usually increases as well. One such instrument is the AmBiChron pitch compensator. The AmBiChron pitch compensator digitizes speech, processing speech in a complicated procedure which allows pitch to be held constant despite changes in speech rate (Koch, 1974). While the error in pitch correction increases to 10% at 3.7 times normal speech, this level is still well within the range before intelligibility is seriously degraded (50% above normal pitch, according to Garvey, 1953). de Haan and Schjelderup found support for their hypothesis that pitch distortion reduced intelligibility. In holding pitch constant, intelligibility was limited only by the listener's ability to process verbal information. However, when pitch was not held constant, pitch distortion compounded this affect and further reduced intelligibility.

Simpson and Marchionda-Frost (1984) studied the effects of pitch variation on synthetic speech warning messages delivered to helicopter pilots. As Simpson and Marchionda-Frost note, "voice pitch provides a variety of cues at various linguistic levels of speech perception" (p. 510). Ofttimes speech comprehension is facilitated by pitch. Syllables that are stressed are higher in pitch and are usually longer in length. In carrying this research out to phrases and clauses, Sorenson and Cooper (1980) also found that phrases and clauses are marked by certain pitch contour variations. Other researchers (Cole and Jakimik, 1980; Larkey and Danly, 1983) support these contentions. In examining speech intelligibility in a 70-120 Hz pitch range, Simpson and Marchionda-Frost found that listeners consistently preferred a certain voice pitch (90-92 Hz). It appears that in regard to pitch level, a very rigid and narrow band range exists at which intelligibility is optimum.

Wolfe and Ratusnik (1988) have also demonstrated how vocal roughness may effect pitch perception. Wolfe and Ratusnik taped the speech of fifty-one individuals diagnosed as having various laryngeal disorders. Each individual recorded vowels /a/ and /i/ on a high-fidelity system, resulting in 102 one-second vowel samples. Wolfe and Ratsunik presented these vowel samples to speech/language graduate students, holding the pitch and loudness constant. After being provided with fifteen non-study samples of clarity/roughness, these students were asked to rate the vowel samples on a seven point scale (1, a clear voice or maximum clarity; 7 indicating maximum roughness or a severe quality disorder). In examining the results, Wolfe and Ratsunik discovered that, even with pitch and loudness held constant, the *perception* of these two variables was correlated with the vocal roughness of the taped presenter. In cases where the presenter's laryngeal dysfunction created a more pronounced vocal roughness, listeners experienced greater difficulty in correctly matching perceived pitch and loudness with the actual pitch and loudness created by the researchers. This study echoes other research findings that suggest the existence of a preferences for certain types of speech and vocal patterns.

With the exception of the research conducted by Simpson and Marchionda-Frost, the previous research cited has largely been conducted under conditions that implemented natural speech. Minimal research currently exists on the effects of pitch contour on synthetic speech perception, although pitch appears to play a vital role in the listener's preference for natural or synthetic speech (Nusbaum and Pisoni, 1985; Pisoni, 1981, 1982; Slowiaczek and Pisoni, 1982). In examining pitch contour, Slowiaczek and Nusbaum (1985) focused their research specifically on synthetic speech perception. Slowiaczek and Nusbaum hypothesized that pitch may predict upcoming information and aid in speech processing. Thus, in the absence of pitch, word perception should be impaired, especially in more complex and longer sentences. In analyzing their results, Slowiaczek and Nusbaum found evidence to support their hypothesis. A significant main effect was found for pitch contour. "Inflected pitch produced 75.1% correct word identification, and monotone pitch produced 70.4% correct identification . . . inflected pitch improved word identification for these more complex sentences" (Slowiaczek and Nusbaum, 1985, p. 708).

In summarizing this section on pitch, it appears that, while various methods exist for determining and controlling pitch (i.e., autocorrelation analysis, AmBiChron pitch compensator, etc.), no one method possesses any relative advantages over the other methods. However, for both natural and computer-generated speech, research overwhelmingly indicates that pitch contour creates a difference in listener intelligibility, especially in longer, more complex sentences, phrases and clauses. While pitch is often associated with time compression because of the effect time-compression has on it, pitch should be viewed and examined in a distinctly singular nature in order to more readily determine and control for its true effect on intelligibility.

Noise

As Burrows (1960) notes, noise can be considered as, "that auditory stimulus or stimuli bearing no informational relationship to the presence or completion of the immediate task" (p. 426). In examining noise, the key phrase "informational relationship," must always be kept in mind. In addition to noise that results from sounds that are not task-related, noise may also be generated by, ". . . task-related sounds that are informationally useless" (McCormick & Sanders, 1987, p. 426).

The effects of noise on performance has been examined by a plethora of researchers. Early seminal research conducted by Miller and Licklider (1950) focused on the disruption of speech intelligibility under three conditions: 1) interrupted speech in a quiet environment; 2) continuous speech masked by intermittent white noise; and 3) a combination of these two previous conditions (turning speech on as noise was turned off and vica versa). Miller and Licklider found that, in a quiet environment, speech remained intelligible until the frequency of the interruptions reached 100 per second; deterioration, although slight, continued between 200 and 2000 interruptions per second. In comparison, Miller and Licklider found that when only 10 white noise interruptions per second were introduced, intelligibility was reduced to 75%. In the third condition, speech and noise were alternated at a rate of 100 times per second. When noise was 18db more intense than the speech, intelligibility was reduced to 4%. At 215 alternations per second, speech became unintelligible. Miller and Licklider concluded that noise, especially loud noise, had serious effect on intelligibility. Intelligibility is also affected more when noise is introduced in the presence of speech, rather than when speech and noise are alternated. Additional early research (Hirsh, Reynolds, and Joseph, 1954; Speaks, 1967) supports these findings.

Hirsh, Reynolds, and Joseph (1954) noted that "the intelligibility of a word is a direct function of the number of syllables in the word and . . . the relation between intelligibilities of each word . . . is not the same . . . when the system is impaired by noise" (p. 530). Modern researchers have examined both word and sentence intelligibility under a wide variety of conditions and have discovered evidence to support this statement. Kalikow, Stevens and Elliot (1977) examined the intelligibility of key words in both low- and high-predictability sentences under various signal-to-noise ratios. As they hypothesized, Kalikow et al. found that key words were easier for subjects to predict in highly predictable sentences than in low-predictability sentences. However, in both high- and low-predictability sentences, the intelligibility of key words decreased as signal-to-noise ratio increased. Martin and Mussell (1979) found almost identical results in their research on the influence of pauses in the identification of key words in competing discourse. When speech noise was added to continuous discourse, subjects found it much more difficult to identify key words. Thus, it appears that word intelligibility decreases as signal-to-noise ratio increases, regardless of the understandability of the message context.

In addition to research which examines the effects of noise on natural speech intelligibility, a number of researchers have also examined the effects of noise on the intelligibility of computer-generated speech. Pisoni and Koen (1982) compared the intelligibility of computer-generated speech and natural speech at various signal-to-noise ratios. Pisoni and Koen mention both the abundance and paucity of literature which examines the effects of noise on natural speech and synthetic speech intelligibility, respectively. In comparing monosyllabic words produced either naturally or synthetically over a wide range of signal-to-noise ratios, Pisoni and Koen found that synthetic speech experienced a greater decrement in intelligibility than natural speech. This was especially true when an open free response format was employed. As Pisoni and Koen stated, "the increase in uncertainty affected recognition of the synthetic items more than the natural ones" (p. 94).

Yuchtman, Nusbaum and Pisoni (1985) forwarded two hypotheses as to why differences in intelligibility between natural and synthetic speech might occur when exposed to noise. One hypothesis is that synthetic speech is structurally equivalent to natural speech degraded by noise. An alternative hypothesis is that the acoustic-phonetic structure of synthetic speech is impoverished in comparison to natural speech, in that a minimal set of acoustic cues are used to implement phonetic segments. In analyzing synthetic speech and natural speech at several signal-to-noise ratios, Yuchtman, Nusbaum and Pisoni determined that greater support existed for the second hypothesis. "The properties of the perceptual spaces obtained for the synthetic consonants differed considerably from those obtained for the natural consonants" (p. 83).

In summarizing the effects of noise on the intelligibility of speech, it appears that the continuous presence of noise presents the most significant detrimental effect to intelligibility. Another factor affecting the intelligibility of speech is the signal-to-noise ratio. As the signal-to-noise ratio increases, the intelligibility of speech decreases. In further analyzing the effects of noise on intelligibility, it appears that, in the presence of noise, computer-generated speech suffers greater degradation than does natural speech. However, while single channel computer-generated speech appears to be lower in intelligibility than natural speech, Hansen and Clements (1985) have demonstrated that various enhancement procedures can improve computer-generated speech quality, even in the presence of noise. Thus, while natural speech currently appears to suffer less from exposure to noise than computer-generated speech, advancements in technology are leading to improvements in computer-generated speech intelligiblity, affording researchers greater control over a noisy environment.

Sensation Level/Loudness

As was previously mentioned in the discussion of pitch, loudness is associated with the human perception of sound intensity. Sound intensity is defined in terms of power per unit area (for example, watts per square meter

(W/m2). The most convenient and frequently used measurement of loudness is the decibel (dB), which is 1/10 of a bel. The number of bels that represent perceived loudness is based on a logarithm of the ratio of two sound intensities. Unfortunately, the power of a sound cannot be directly measured. What can be measured, what we express when we record loudness in decibels, is the pressure waves that a given sound emits that are above or below normal air pressure. This measurement, sound-pressure level (SPL), measures sound power directly proportional to the square of sound pressure.

Only recently has research in speech intelligibility examined optimum levels of loudness. Madell and Goldstein (1972) examined responses of subjects at nine sensation levels. Levels of loudness ranging from -10 to 70 dB were presented to normal hearing adults. Subjects were asked to judge the loudness of each sensation level after being presented 10 clicking sounds at a standard sensation level (30 dB). Madell and Goldstein found that, although subjects varied a great deal in their perception of sensation level, certain patterns emerged. As expected, no subjects were able to hear information presented at the -10 dB sensation level condition. Stimuli presented at the threshold sensation level (0 dB) was heard by subjects only half of the time. The most accurate perception of loudness occurred at the 50 dB sensation level.

Beasley, Schwimmer, and Rintelmann (1972) examined the effects of sensation level on the intelligibility of time-compressed monosyllables.

Beasley et al. presented time-compressed monosyllables to 96 young adults at four different sensation levels (8, 16, 24, and 32 dB). Words were compressed 30%, 40%, 50%, 60%, and 70%. Beasley et al. found that, under all conditions of time-compression, discrimination improved as sensation level increased. The greatest increase in intelligibility occurred between 8 and 16 dB, where intelligibility was found to increase by 2 - 3.5% per dB. Intelligibility was found to be highest at the 32 dB sensation level, where, with the exception 70% time-compression of speech, intelligibility of monosyllabic words was above 90%. This finding is consistent with Madell and Goldstein's results. Beasley, Bratt, and Rintelmann (1980) discovered similar results in their study of time-compressed sentential stimuli.

Konkle, Beasley and Bess (1977) have also examined the effects of sensation level on the intelligibility of time-compressed speech in relation to age. Konkle et al. administered the Northwestern University Auditory Test Number 6 (NU-6) to subjects ranging in age from 54 to 84 years old. Speech was compressed either 0%, 20%, 40%, or 60% and were presented at one of three sensation levels (24, 32, or 40 dB). Konkle et al. found a significant time-compression X sensation level interaction. As Konkle et al. note, "the mean scores for 24 dB SL were significantly lower than the scores for 32 and 40 dB SL, respectively, under the 0%, 20%, and 40% time-compression conditions" (p. 111) These results support the previously stated research. In addition to these findings, however, Konkle et al. also note a significant sensation level X age interaction. While significant differences were not obtained between subjects at the 32 and 40 dB sensation levels, significant differences were obtained between the 24 dB and 40 dB sensation levels for all age groups and between the 24 dB and 32 dB sensation levels for the three older age groups.

From this research on sensation level, a very basic statement concerning the effects of loudness can be made. Clearly, optimum sensation levels exist. This is especially true in regard to time-compressed speech, whether it be words or sentences. While a great deal of research examines sensation levels in the 20-50 dB range, this appears to be more than appropriate for examining speech intelligibility (see Appendix B, Peterson and Gross, 1972). Clearly, as Beasley, Schwimmer and Rintelmann (1972b) state, "the articulation functions are characterized by curvilinear progressions in which discrimination scores improve less with progressive increases in intensity, approaching an asymptote at 32-dB SL" (p. 344). Thus, in examining the true effects of any intelligibility measure, it appears that the sensation level should be set at a minimum of 30 dB.

Syntax/Syntactical Structure

Syntactical structure refers to the way words are put together in order to form phrases, clauses, or sentences. In examining the effects of syntactical structure on intelligibility, early research generally focused on the periodic interruption of speech (Miller and Licklider, 1950) or the number of syllables which a word contained (Hirsh, Reynolds, and Joseph, 1954). However, recent research has generally concentrated on a different tact, examining the manner in which altered intonation patterns effect the underlying syntactical structure.

Dooling (1974) focused on how the intelligibility of a message is effected by changes in rhythm and syntax. Subjects were presented with a varying number of consecutive sentences which contained the same grammatical structure. On a final experimental sentence, subjects were presented with a different sentence which contained either: 1) the same syntax and rhythm (SAME-SAME); 2) the same syntax but a different rhythm (SAME-DIFF); or 3) changes in both syntax and rhythm (DIFF-DIFF). In analyzing the results, Dooling determined that, while the effects of syntactical structure alone were not significant, changes in rhythm created a major reduction in intelligibility. As Dooling notes, these ". . . results point out the fundamental importance of rhythm in speech perception and suggest caution in attributing speech perception effects to syntax without controlling for rhythm" (p.255).

In his research, Wingfield (1975) examined normal intonation patterns and intonation patterns which conflicted with the underlying syntactical structure. Twenty specially constructed ten-word sentences were presented to subjects. Each sentence was specially constructed so that intonation patterns could be made to either agree or conflict with the underlying syntactic structure. Each subject heard four of the sentences in one of five time-compression ratio conditions (80%, 70%, 60%, 50%, or 40% compression of normal playing time). Wingfield found that, while an overall decrease in intelligibility was attributable to time-compression, sentences which contained intonation patterns anomalous with the underlying syntactic structure were significantly less intelligible than their intonation pattern correct counterparts. This was found to be true even under normal speech rate conditions. As Wingfield states, ". . . these results build a fairly clear picture of perceptual processing guided by syntactic analysis of the heard speech" (p. 103). Wingfield, Buttet, and Sandoval (1979) performed this same experiment with English and French speaking subjects and obtained the same results. Even across languages, it appears that sentences which contain intonation patterns anomalous with the underlying syntactic structure experience greater decrementation in intelligibility than do sentences which contain intonation patterns which do not conflict with the underlying syntactical structure.

Wingfield, Lombardi and Sokol (1984) have also examined the effects of syntactical vs. periodic segmentation on paragraph-length passages of time-compressed speech. Again, various intonation patterns were created. Passages were either presented in list intonation (monotone), in normal prosody (normal speech), or were electronically processed to produce speech devoid of pitch variation but otherwise normal. Passages were also compressed to either 65% or 50% of normal playing time. In analyzing the data, Wingfield et al. reported that, as expected, intellgibility scores decreased as time-compression increased. However, significant results were also obtained for both the type of segmentation which occurred and the intonation pattern implemented. Passages presented in list intonation were found to be significantly poorer in intelligibility, especially as speech rate increased. Additionally, periodic segmentation (segmentation which occurred randomly in the passage) was found to be less intelligible than syntactic segmentation (segmentation which corresponded to sentence and major clause boundaries).

From the research on syntactic structure, it appears that definite effects for intelligibility occur in relation to the type of intonation pattern implemented. Intonation patterns often provide clues to the words being presented. Under poor conditions (noise, time-compressed speech, etc.), this aspect becomes even more pronounced and important. When intonation is altered or anomalous to the message being presented, reductions in intelligibility are to be expected. Thus, in creating text of optimum intelligibility, care should be taken in order to ensure that intonation patterns are consistent with the underlying syntactical structure.

Meaningfulness of the Message (Content)

In addition to the aforementioned factors which influence the intelligibility of transmitted speech, a final variable which must be mentioned in any comprehensive discussion of intelligibility concerns the content of the message itself. If an individual is presented with a phrase or sentence which contains meaning or, due to the context of the situation, reduces the possibility of choices, the individual has an increased likelihood of correctly identifying key words or missing components of the message. For example, if an individual were to hear the phrase, "A rolling ______ gathers no moss," he or she would be much more likely to be able to provide the missing key word than if the individual were presented with a phrase such as, "On Wednesday he ______ ." Seminal research by Miller, Heise, and Lichten (1951) has demonstrated that, under adverse conditions, the content or meaningfulness of the message plays a vital role in enabling the listener to correctly identify what is being transmitted. Miller et al. forwarded the belief that distinct types of contexts existed which aided the listener in his or her understanding of the message:

> Three kinds of contexts are explored: (a) context supplied by the knowledge that the test item is one of a small vocabulary of items, (b) context supplied by the items that precede or follow a given item in a word or sentence, and (c) context supplied by the knowledge that the item is a repetition of the immediately preceding item (p. 329).

In order to test these three types of contexts, Miller et al. presented subjects with words from vocabularies of various sizes (2, 4, 8, 16, 32, and 256 words). Words were presented to subjects under various signal-to-noise ratios (ranging from -18 to 9 dB) as well. In analyzing the results of all three contexts, Miller et al. clearly demonstrated that the percent of key words correctly identified was strongly correlated with the size of the vocabulary implemented. Similar research (Hirsh, Reynolds, and Joseph, 1954; Lehiste and Peterson, 1959; Speaks and Jerger, 1965; Epstein, Giolas, and Owens, 1968) has supported this finding. Thus, it appears that in examining intelligibility, sentences are more intelligible than isolated words, and, in similar fashion, isolated words are more intelligible than syllables.

While the majority of research concerning intelligibility as a function of message context and meaningfulness was conducted some time ago, present day researchers are still examining the effects of message meaningfulness in practical applied settings. Representative of this fact is the research conducted by Slowiaczek and Nusbaum (1985). In examining settings of practical application, Slowiaczek and Nusbaum note how recent advances in technology have demanded transmission of more semantically correct messages. In presenting subjects with either semantically correct or anomalous sentences in conjunction with a number of other variables, Slowiaczek and Nusbaum determined that meaningfulness of the message was one of the more predominant moderating variables which influenced speech intelligibility. As Slowiaczek and Nusbaum comment, "the results suggest that in many applied situations the perception of the segmental information in the speech signal may be more critical to the intelligibility of . . . speech" (p. 704).

In summarizing the effects of meaningfulness and context of the message on intelligibility, it appears that two very distinct comments can be forwarded. The meaning or context of the message does play a major role in the intelligibility of the message. Intelligibility may be affected in one of three ways. First, it appears that limiting the range of items from which the listener must choose a specific item reduces ambiguity and increase intelligibility. Second, items which precede or follow a critical item aid in identifying that critical item and placing it in context. Finally, specified items which are repetitious of preceding items create familiarity, an effect which will also increase intelligibility.

The second main comment which may be made about the meaningfulness of a message concerns the size of the message itself. As was previously stated, sentences provide more context information than words, and words more than syllables. Increasing the size of the message being transmitted also is likely to increase intelligibility. Thus, whenever possible, words should be transmitted instead of syllables, sentences instead of words, and phrases or clauses instead of sentences. If only words or syllables are capable of being transmitted, it is more favorable if they contain more than one syllable and have meaning when used alone.

Perception (Listener Capabilities)

In discussing speech perception, one can easily view how modifications of the aforementioned variables would affect intelligibility. A slight adjustment in any of these variables may drastically alter the message, serving to reduce the intelligibility of the message being transmitted. However, even if the message is transmitted to the listener in such a manner that it arrives with intelligibility intact, moderating variables may still come into play. While the aforementioned variables which could have influenced intelligibility belonged to the message, the current section examines the role the listener may play in influencing the level of intelligibility.

One of the primary moderating variables that the listener brings to the situation which may affect intelligibility is short-term memory (STM) ability. A number of researchers (Neisser, 1967; Wickelgren, 1969; Liberman, 1970; Massaro, 1972) have forwarded data which suggest that phonemes are coded in short-term memory. These phonemes are coded by virtue of their distinctive features and are implemented to create syllables, which are also maintained in STM. However, researchers have demonstrated that phonemes which share similar distinctive features are often substituted for each other, especially when the listener attempts to recall these phonemes from STM (Wickelgren, 1965, 1966).

Citing this body of previous research, Cole (1973) acknowledges the role short-term memory plays in influencing intelligibility in his examination of the way subjects remember a series of syllables. Cole presented subjects with a series of consonant-vowel (CV) syllables. In order to ensure that the obtained results were a function of forgetting in STM and not misperception, Cole instructed a control group to press a response key as soon as they heard specific consonant or vowel phonemes. The results demonstrated that substitution errors were not generated by misperception. Cole then asked an experimental group of subjects to perform the same task after a 0.5 second delay. Cole found that the overall error rate for subjects after only a 0.5 second delay was 44% for consonant phonemes and 36% for vowel phonemes. Cole concluded that, while phonemes are coded independently of each other in STM, once they are forgotten they are grouped. Thus, even slight delays in recall may trigger the recalling of an incorrect substitute phoneme, affecting the intelligibility of at least that word, if not the entire message.

Pisoni (1981) has generated support for Cole's findings. In examining natural and synthetic words in a lexical decision task, Pisoni found that, while subjects responded faster (145 ms) to synthetic speech than natural speech and recognized words 140 ms faster than non-word stimuli, no interaction was found to occur between these two variables (signal type and classification response). Pisoni states:

> These results suggest that differences in perception between natural and synthetic speech lie at early stages of perceptual analysis in which the initial phonetic or segmental representation of the input signal is developed rather than at later stages of lexical access and search where these representations are examined or compared prior to execution of the observer's classification response.

In addition to delays in processing time affecting STM intelligibility, Luce, Feustel and Pisoni (1983) have also demonstrated how increased processing demands may affect STM intelligibility. Luce et al. presented subjects with synthetic and natural monosyllabic word lists. These words were presented at intervals of 1, 2, or 5 seconds per word. Luce et al. found that, for both natural and synthetic-produced words, recall ability increased as the subjects were allowed more time to comprehend each word. However, at each presentation rate, natural words were recalled significantly better than synthetic words. Luce et al. attribute this latter finding to the fact that encoding difficulties are more likely to be encountered in synthetically-produced speech, reducing the subjects' ability to rehearse, store, and recall words. In a second experiment, subjects were instructed to also recall and repeat digit strings of zero, three or six characters in length throughout presentation of the aforementioned word lists. Luce et al. concluded that, "synthetically produced word lists may interfere with the subjects' ability to maintain information in short-term memory" (p. 25).

In addition to short-term memory functions which serve to reduce the listener's ability to recall words and, hence, intelligibility, it appears that a natural preference exists for certain forms of speech as well. Clark and Stemple (1982) examined four different modes of speech: 1) pulmonary esophageal speech; 2) traditional esophageal speech; 3) artificial laryngeal speech; and 4) normal laryngeal speech. In each of these four speech modes, 10 synthetic sentences were presented to subjects in the presence of a competing background message at varying signal-to-noise ratios (0, -5, or -10 dB). Clark and Stemple's results indicate that, despite being the least intelligible of the four speech modes, in the two most difficult signal-to-noise ratio conditions (-5 and -10 dB), pulmonary esophageal speech was the speech mode which subjects preferred the most. While on the surface this finding may be difficult to explain, it lends credence to the notion that listeners do indeed prefer certain types of speech, regardless of intelligibility.

In their comparison of the perceptual and acoustic characteristics of tracheoesophageal and speech pathologist defined excellent esophageal speakers, Sedory, Hamlet, and Connor (1989) obtained results similar to Clark and Stemple. Sedory et al. taped both groups presenting a three sentence passage. In presenting the stimuli to ten normal-hearing subjects, Sedory et al. instructed subjects to select the sentence passage they preferred "... using their subjective impression ... which may be based on different aspects of speech, including intelligibility, voice quality, fluency, rate, naturalness, communicative effectiveness, or just which [they] would most like to listen to" (p. 210). While statistically significant results were not observed, all of the listeners stated that their selection was based upon the smoothness, clarity, and "more normal sounding" voice of the preferred speaker (p. 213).

Nusbaum, Greenspan and Pisoni (1985) have examined listener preference of natural speech vs. computer-generated speech. Nusbaum et al. instructed subjects to specify target syllables in one of three conditions: 1) targets and distractors (the text in which the target syllables appear) produced by the same human talker (N/N); 2) targets produced by a synthetic talker and distractors produced by a human talker (S/N); 3) targets produced by a synthetic talker and distractors produced by the same synthetic talker and a natural talker (S/N + S). Nusbaum et al. discovered that targets were highly intelligible in the S/N condition. Intelligibility was lower for target identification in the N/N condition, and much worse in the S/N + S condition. Nusbaum et al. attributed this finding to the distinctive mechanical sound of synthetic speech. Nusbaum et al. concluded:

> The distinctive mechanical sound of synthetic speech only appears to aid perception when there is just a single synthetic message among natural messages. When listeners must discriminate among synthetic messages, performance is significantly worse than when they must discriminate among natural messages.

While not the concern of this thesis, two additional listener moderating variables which affect intelligibility deserve brief mention.

Chronological Aging.

Konkle, Beasley and Bess (1977) present research which examines effects of chronological aging on the intelligibility of time-compressed speech. Presenting time-compressed speech (either 0%, 20%, 40%, or 60% that of normal speech) at various sensation levels (24, 32, and 40 dB) to subjects who ranges from 54 to 84 years of age, Konkle et al. discovered that, "older listeners exhibited marked difficulty in perceptually processing time-compressed speech" (p. 113). These results are in agreement with previous research findings (Calearo and Lazzaroni, 1957; Bocca and Calearo, 1963; de Quiros, 1964; Sticht and Gray, 1969; Schon, 1970; DiCarlo and Taub, 1972). Konkle et al. conclude that the intelligibility of time-compressed speech as a function of aging is related to changes in the central auditory processing system. Clearly, any experiment which involves the elderly must take into account the degenerative effects of aging on the auditory system.

Feedback.

Finally, it appears that feedback may influence an individual's ability to perceive the message being transmitted. Research by Loeb and Binford (1964) is exemplary of this fact. Forty-eight subjects were instructed to respond to occasional increases of a pulse sound. Loeb and Binford presented half of the subjects with feedback; the other half did not receive feedback. Loeb and Binford determined that subjects who received feedback made fewer false responses. In later sessions, false responses were reduced for both groups of subjects. This latter reduction in particular tends to suggest that feedback, even in the form of practice effect, increases the listener's confidence level of perceiving messages correctly. Feedback also serves to create familiarity with the message being transmitted, a factor which was previously mentioned as a variable which increases intelligibility.

Marics and Williges (1988) discovered similar findings in examining feedback presented via synthetic speech. In examining speech rate, message repetition, and location of information in a synthetic speech message, Marics and Williges presented nineteen naive subjects with eight different messages. Subjects were randomly presented the message to be repeated either once, twice, or three times. In examining the results, Marics and Williges found that among subjects who were exposed to the message twice, error rates dropped about 60% and response latency was about 50% faster than for subjects who were exposed to the message only once. Subjects who were presented the message two or three times also demonstrated improved transcription accuracy. When asked to type the message on a computer, subjects who were presented the message two or three times evidenced greater certainty about the accuracy of the message, a finding reflected in the actual accuracy of the transcribed message.

In reviewing the literature on listener capabilities and intelligibility, a few general statements can be forwarded. Clearly, it appears that the listener's short-term memory plays a vital role in determining the intelligibility of the message. Delays in processing time or overload of the short-term memory's processing ability appear to have the greatest affect on perceived intelligibility. Besides short-term memory errors in message perception, existing natural preferences for certain types of speech also affect intelligibility. This is especially true in natural speech/synthetic speech comparisons. While natural speech is generally preferred by listeners, in certain cases the distinctive nature of synthetic speech sets it out against background natural speech. This finding is largely due to the encoding preferences of the listener. Finally, while it is not within the scope of this experiment to examine their effects, researchers must also be aware of the influences of chronological aging and feedback as they relate to listener auditory deterioration and vigilance, respectively.

Types of Stimuli Presented

As one can see from examining the research cited, the intelligibility of the message being transmitted may be affected in any or all stages on its journey from the speaker to the listener. Whether it is formed naturally or artificially, the message may lose intelligibility when it is first created by the speaker. A number of factors such as time-compression, frequency, pitch, noise, loudness, syntactical structure, and content or meaningfulness of the message may also moderate the intelligibility of the message. Even if the message arrives to the listener unadulterated, the listener's own unique capabilities may affect his or her perception of the message, thus affecting intelligibility. While the present study acknowledges a number of factors which can and must be controlled if optimum intelligibility is to be maintained, a final factor must be discussed in order to fully exhaust all variables which might influence natural or computer-generated speech production, transmission and perception. Researchers must also examine the types of stimuli implemented in training.

The test materials implemented in speech generation and perception research have traditionally been classified into three distinct subgroups. These subgroups are: 1) syllables (consonants and/or vowels); 2) words (ranging from monosyllabic to polysyllabic words, created naturally or computer-generated); and 3) continuous discourse (sentences, phrases, clauses or paragraphs of speech). The various advantages and disadvantages of these subgroups will now be scrutinized in greater detail.

Perhaps the least implemented of the three test material stimuli mentioned, the presentation of syllables as a means of determining intelligibility does have certain unique advantages. Beasley, Schwimmer, and Rintelmann (1972b) favored this method of stimuli presentation in their experiment concerning the effects of time-compression on intelligibility. In examining time-compression, Beasley et al. note that word lists have been criticized as being too easy to be effective in differential diagnosis (Carhart, 1965). The use of sentences as the presentation stimuli was also rejected by Beasley et al., largely due to the fact that these researchers believed sentences provided contextual information and thus did not truly reflect the degree to which time-compression alone affected intelligibility.

Cole (1973) also favored the use of syllables in his study on perception and memory. Cole noted that consonant-vowel (CV) syllables most closely represented pure phonemes. As Cole notes, "an analysis of intrusion errors during a serial recall task revealed that consonant and vowel phonemes are coded by the same distinctive features in a variety of different CV syllables" (p. 37). Cole supports the arguments forwarded by Beasley et al. and Carhart, adding that experiments which implement either word lists or sentence stimuli are incapable of accurately analyzing phonemes. Cole comments that since phonemes are at the base of all speech, research which hopes to create new inroads into the understanding of speech intelligibility should focus on phonemes.

More recently, Yuchtman, Nusbaum and Pisoni (1985) discussed previous laboratory research which suggested that, "synthetic speech is less intelligible and more capacity demanding than natural speech" (p.83). Yuchtman et al. hypothesized that one reason this difference in intelligibility may exist is that, "the acoustic-phonetic structure of synthetic speech is impoverished in comparison to natural speech in that a minimal set of acoustic cues are used to implement phonetic segments" (p.83). Yuchtman et al. preferred presenting consonants as stimuli in their experiment, believing that neither word lists nor sentences were capable of accurately reflecting the real-life encoding which occurs, changing the input signal into segmental phonemic representations.

Some researchers prefer to present syllable stimuli because of the apparent advantages it presents in terms of accurately representing true experimental manipulations, phoneme segmentations, and real-life encoding processes. However, a far greater number of researchers employ word lists as a means of presenting stimuli to subjects. Pisoni and Koen (1982), for example, chose to use a word list. In examining the effects of noise on both synthetic and natural speech perception, Pisoni and Koen note that if syllables were presented, they might not be distinguishable from the noise itself. If one truly wishes to study speech, these researchers argue, stimuli which have "real-world" applications should be implemented. As well as being representative of stimuli encountered in the "real-world," it has also been argued that words create a truer sense of intelligibility than syllables. While syllable stimuli are basically transmitted as phonemic sounds, words are symbolic and have meaning. Thus, it may be easier to determine how much intelligibility has been affected if one can determine to what extent a word has lost its meaning.

Luce, Feustel, and Pisoni (1983) have examined the capacity demands in short-term memory, noting how the type of stimuli presented is affected. Luce et al. presented subjects with stimuli of varying length. Subjects were then instructed to recall these stimuli under various conditions. Luce et al. determined that the length of the stimuli significantly affected the subjects' ability to correctly recall, especially when stimuli were presented synthetically. Thus, it appears that increasing stimuli length has an effect on the processing demands in short-term memory. This points to the fact that effects on intelligibility may be more accurately measured by stimuli which has the ability to be varied in length.

In a slightly different vein, Benjafield and Muckenheim (1989) have demonstrated how the presentation of isolated words can afford the researcher greater control over what he or she intends to measure. In examining 1,046 words sampled from the *Oxford English Dictionary*, Benjafield and Muckenheim attempted to determine norms for each word on familiarity, imagery, concreteness, and goodness of fit. As Benjafield and Muckenhiem note:

> Such norms should be useful to researchers interested in sampling very uncommon or unfamiliar words, as well as quite common or familiar ones.... researchers particularly concerned with using a sample that is fairly representative of the range of words in the written language should find the database particularly valuable (p. 31).

In presenting the findings, Benjafield and Muckenheim demonstrate that certain words tend to have more of these qualities than other words with the four different measures. In presenting word lists as stimuli, certain words therefore are likely to evoke a more recognizable or pronounced response, based upon a number of dimensions. However, certain words that provoke a similar response can be identified.

Finally, Schiavetti, Sitler, Metz and Houde (1984) have demonstrated in their research that contextual intelligibility can be predicted from isolated words. Employing intricate formulae to examine the predictive ability of key isolated words, Schiavetti et al. examined four different sets of speech intelligibility data. Schiavetti et al. found that isolated words proved to be an excellent measure of contextual intelligibility. Other researchers (Duffy and Giolas, 1974; Kalikow, Stevens, and Elliot, 1977) have also conducted experiments which indicate that key words can have great predictive power of intelligibility in continuous discourse.

The final manner in which test material stimuli is usually presented is through continuous discourse (sentences, phrases, clauses and paragraphs). A number of advantages appear to exist for continuous discourse presentation. Early research by Speaks and Jerger (1965) indicated that informational content in a sentence could be controlled easier than in syllables or words, allowing the researcher more control over the manipulation of the stimuli to be presented. Dooling (1974) has also demonstrated the benefits that control over informational content of a sentence affords a researcher. Dooling was able to manipulate both syntax and rhythm in a series of sentences presented in noise. In particular, the only time Dooling was able to vary rhythm was when sentences were employed as the presentation stimuli.

Toscher and Rupp (1978) indicated that sentence stimuli may also present an advantage in that it enables researchers to study and analyze the occurrence of phenomena which might not be detectable when syllable or words are the method of stimuli presentation. Toscher and Rupp presented the Synthetic Sentence Identification Test (Speaks and Jerger, 1965) to groups of stutterers and nonstutterers to assess central auditory function. Toscher and Rupp concluded that differences did indeed occur between stutterers and nonstutterers in central auditory function. What is important about this study in regard to the present discussion is that this experiment would not have been possible if syllables or words were implemented. Syllables and words simply do not accurately measure the number of times a person stutters in normal speech. A stuttering subject is much more likely to recite syllables and words without stuttering, thus portraying a false picture of what is actually occurring.

In examining semantically congruent and incongruent word presentation, Lukatela, Carello, Kostic and Turvey (1988) present evidence for the depreciation of message coherence in non-sentence presentation conditions. In presenting word pairs that were either semantically congruent or incongruent to twenty-six subjects, Lukatela et al. noted a significantly higher level of message coherence when word pairs were congruent. This research indicates that, in situations were message coherence may be affected by incongruency or ambiguity, the presentation of contextual information may increase the listener's ability to correctly process the unfocused message.

Beasley, Bratt and Rintelmann (1980) have also noted instances when the use of sentences as the test material stimuli may be preferable. As Beasley et al. mention, "monosyllables have been studied relative to the assessment of central auditory disorders" (p. 722). However, in certain cases, such as cases which involve peripheral hearing loss, sentences may be more useful. In comparing monosyllable stimuli to sentential stimuli, Beasley et al. found that sentences could be controlled to the point where the effects of time-compression was approximately the same as that for monosyllables. Under these conditions, using sentences instead of monosyllables or word lists appears to be more practical.

Finally Slowiaczek and Nusbaum (1985) have determined that the actual length of the sentence itself may have a moderating effect on intelligibility. Slowiaczek and Nusbaum presented subjects with sentences that varied in a number of manners, one of these being length. In analyzing the results, Slowiaczek and Nusbaum discovered that large effects on intelligibility were attributable to sentence length. Words in short sentences (four words) were identified consistently better than words in long sentences (eight words).

In summarizing the relative advantages and disadvantages of each test material stimuli, it appears that each must be examined in the context of the experiment in order to determine which stimuli might be the most preferable presentation stimuli. Syllables are advantageous in that 1) certain intelligibility variables (e.g., time-compression) exert less influence; 2) syllables are more representative of pure phoneme segments; and, 3) syllables reflect real-life encoding processes. A disadvantage of syllable stimuli is that it might be undistinguishable from ambient noise.

Word list are advantageous in that an individual can usually distinguish words from noise. Word lists are also more representative of the "real world" than syllables. Words can also be predictive, allowing researchers to determine intelligibility in a sentence by examining a few key words. On the negative side, word stimuli have been criticized as being too easy to be effective in differential diagnosis and incapable of analyzing phoneme segments.

Sentence stimuli can be advantageous in that it allows the researcher greater control over the experiment, enabling the researcher to manipulate variables more easily. Certain experiments may also only be feasible when sentence stimuli are implemented. Both an advantage and a disadvantage is that sentence stimuli provides contextual information to subjects. This may aid or reduce the validity of an experiment, depending upon what the researcher hypothesizes and intends to examine. Finally, people are less capable of accurately analyzing phoneme segments when sentences are employed as the presentation stimuli.

Hypotheses.

Guided by the findings of the previously cited research, the researcher conducted an experiment which would study the rate of successful task performance across three task difficulty levels, based upon the type of training method employed (natural speech vs. synthetic speech). The researcher forwarded the following hypotheses: 1) In presenting simple tasks, subjects presented with computer-generated speech training will exhibit similar performance levels as subjects presented with natural speech training; 2) in presenting moderately difficult tasks, subjects presented with natural speech training will suffer minimal performance degradation, while subjects presented with computer-generated speech training will suffer significantly greater performance degradation. However, the performance levels between these two groups will not differ significantly; 3) whether presented with natural or computer-generated speech training, all subjects will experience statistically significant performance degradation between moderately difficult and difficult task levels; 4) while performance will decrease for both computer generated and natural speech training as task difficulty increases, computer-generated speech trained subjects will exhibit statistically significantly greater performance degradation at the difficult task level than natural speech trained subjects.

Method

Subjects

The subject pool consisted of Cerritos Community College students. Students who were enrolled in Introductory Psychology, Research Methods, and Physiological Psychology classes during the Spring 1990 semester were solicited to volunteer for the study. The researcher approached instructors who were teaching any of these three psychology classes and asked for permission to enter the classroom for approximately 20-30 minutes in order to obtain volunteers. During the course of the classroom presentation, the researcher provided a brief overview about the purpose of the study, taking care not to bias prospective volunteers about the expected or desired outcome. Students were informed that the study would require approximately 15 - 20 minutes of their time, where the experiment was located, and that they would be fully debriefed after participating in the study. As an added incentive to volunteer and participate in the experiment, subjects were told that the study would include a hearing test, and that they would be advised about the outcome of their individual hearing test. Arrangements were also made at this time to provide results to students who chose not to participate in the study but were interested in the results of the study once the experiment was completed. As the researcher discussed non-specific parameters of the experiment, a sign-up sheet was passed around the classroom, specifying numerous dates and times that students could reserve for participation. Dates and times were based upon class meeting times, with a one to two hour window for participation arranged immediately before or after the class. Appointments were set up in fifteen minute blocks. Before the researcher left the classroom, students were reminded to write down the appointment time that they had reserved, and once again where the experiment was located. Any remaining questions that did not influence or bias the experiment were also answered at this time.

Design

The researcher implemented a 2 x 3 repeated measures MANOVA design. The repeated measures design was of a Type A (between groups fixed, within groups fixed) nature, since the levels of both variables were intentionally, not randomly, selected. Two variables were introduced: 1) type of speech employed; and 2) level of task difficulty presented. Two levels of type of speech were employed: 1) natural speech; and 2) computer-generated (synthetic) speech. Natural speech refers to speech produced by a human speaker which is presented to the listener unadulterated. Computer-generated speech refers to speech created and presented by a computer. The exact details of the computer-generated speech system chosen for this particular experiment will be described in full detail when the researcher discusses the apparatus used for this experiment.

Three levels of task difficulty were presented to subjects in both the natural and synthetic speech presentation groups: 1) a simple task level; 2) a moderately difficult task level; and 3) a difficult task level. The three levels of task difficulty will be discussed in full detail when the researcher outlines the procedure of the experiment.

Apparatus

In conducting the experiment, the researcher employed the following apparatus: a Maico MA-19 Audiometer; a Macintosh Plus personal computer; the computer software package Smoothtalker2; a high-fidelity tape recorder; a headphone set; a slide projector with an automatic frame advancer; 80 color slides; and a room that provided a relatively noiseless environment.

The Maico MA-19 Audiometer was chosen as an auditory screening device primarily because of its ability to test subjects at specific frequency levels and at various senstion levels. The Maico MA-19 Audiometer is capable of testing frequencies at 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 3000 Hz, 4000 Hz, 6000 Hz, and 8000 Hz. Hearing threshold level can also be adjusted, with presentation levels available from 0 to 110 dB (ANSI) at 5 dB intervals. The built-in headphone set also provided unique advantages, allowing all subjects to receive independent left and right ear testing. The headphones have also been designed and tested to reduce ambient noise; satisfactory testing can be administered in an area where the ambient noise level is as high as 40 dB. The researcher was trained how to use the Maico MA-19 Audiometer by the Director of Cerritos College's Speech, Language, and Hearing Center and was able to test subjects' auditory ability himself.

The Macintosh Plus microcomputer was selected for a variety of reasons. The Macintosh Plus has 800K capability, a feature that ensured the computer-generated speech software package chosen for the study could be successfully implemented, regardless of the memory demands of the Smoothtalker 2 software package. The Macintosh Plus has a pre-existing outlet jack for a speaker/headphone, a feature that allowed the researcher to present computer-generated stimuli directly to subjects, effectively eliminating background (ambient) noise. The Macintosh Plus was also the most easily accessible quality caliber microcomputer for this experiment. Finally, in generalizing the results of this experiment to a "real-world" setting, the use of a microcomputer in computer-generated speech training most closely replicated the equipment which is commonly available in industry. In previously cited research by Logan, Greene, and Pisoni (1989) that tested the quality of Smoothtalker as a text-to-speech system, a Macintosh Plus was also the personal computer used, a precedent which provides further support for the implementation of this type of personal computer.

Smoothtalker2 was selected because of the relative advantages it offered in regard to ability to control for intelligibility factors previously cited. Smoothtalker2 is able to control for speed of speech (time-compression), pitch, tone (frequency), and volume (loudness). In addition to controlling for these intelligibility factors, Smoothtalker2 attempts to replicate natural speech in creating synthetic speech. Smoothtalker2 converts whatever text is entered into phonemes. Over 1000 English rules are also applied to incoming text. Thus, Smoothtalker2 encodes and accounts for stress, pitch, inflections and the like caused by punctuation. After converting the incoming text into phoneme building blocks and regulating effects caused by punctuation, Smoothtalker2 then proceeds to convert these encoded phonemes into speech. It is in this final form that computer-generated speech is presented to subjects.

While the Smoothtalker 2 software package will admirably handle the chore of computer-generated speech presentation, natural speech selection, because of its highly unreliable nature, requires thorough training. To account for this fact, the researcher solicited the aid of a colleague in presenting natural speech. The researcher spent approximately one hour discussing and rehearsing the natural speech stimuli that were presented to subjects by presenter. After listening to pre-recorded computer-generated speech presentations of the word lists, the natural speech presenter attempted to replicate the pitch, tone, and volume of the computer-generated speech stimuli, as well as the presentation time. When the researcher judged the natural speech stimuli equitable across all these variables to the computer-generated speech stimuli, the natural speech stimuli was recorded on a high-fidelity recording system to ensure consistent stimuli presentation to all subjects in the natural speech training group.

To control for ambient noise, the researcher employed a headphone set. With the proper audio jack, it was possible to plug the headphone set directly into the Macintosh Plus, allowing subjects to receive computer-generated speech stimuli directly from its source, free of extraneous noise.

Besides controlling for ambient noise, this method also allowed the researcher to place subjects in a position where they were not exposed to the stimuli being transmitted on the screen or the non-verbal behavior of the researcher (visual feedback). This also afforded the researcher greater control; any procedures which require manual computer keyboard operation by the researcher were easier to preform under this method.

Natural speech stimuli transmission also benefitted from the implementation of a headphone set. As with computer-generated speech stimuli, presentation of the natural speech stimuli via headphone set allowed the researcher to position subjects in a manner that prevented them from inadvertently receiving unintentional non-verbal cues created and presented by the researcher. Transmitting natural speech stimuli in this manner also served to control for ambient noise which might have confounded the actual affects attributable to natural speech stimuli transmission. This latter statement is true for computer-generated speech stimuli as well.

In order to randomize task difficulty level, a secondary task was presented simultaneously, involving indentifying a set color every *nth* slide. While more will be mentioned about this in the procedure section of this study, the apparatus involved required a slide projector and 80 color slides. The slides that were used were developed by a member of the Cerritos College Instructional Media Services area. Slides were a solid color: blue, red, green, or yellow. Twenty of each color were developed and provided to the researcher for use in the study.

A relatively sterile visual and auditory environment was essential to the success of the study. To this end, the researcher selected the Cerritos College Innovation Center. The Innovation Center holds the relative advantage over other possible experimental locations in that it is located away from possible sources of ambient noise, allowing the researcher freedom to control the experiment without outside pressures and constraints (e.g., beginning and ending experiments within certain time confines in order to work around scheduled classes or office personnel schedules), and provides at least face validity of additional test credibility (e.g., the impression to students that the experiment is occurring in a rigid academic setting and should thus be considered by the subject in an appropriate manner).

Procedure

As subjects arrived, the researcher greeted them at the entrance of the Innovation Center and escorted them into the experimental area. In order to maintain consistency and reliability, the greeting was rehearsed and standardized. Each subject was informed that the nature of the experiment concerned the perception of different forms of speech stimuli. Subjects were asked if they had any general questions before the experiment began. Questions that may have affected the outcome of the study were deferred until after the study was concluded. Subjects were also informed that any questions that might arise about the experiment at a future point in time could be answered by contacting the researcher at a telephone number provided to subjects. Subjects were also informed that they could contact the researcher at this same number to find out the results of the experiment once all data was collected and analyzed. When all general questions were answered, the researcher queried students as to whether they had any physical limitations that, knowing the non-specific parameters of the experiment, might limit their ability to participate in the experiment. Any severe impairments directly related to the experiment (e.g., medically diagnosed hearing loss or color blindness) resulted in a subject's dismissal from the study. In classes where participants were awarded extra-credit for volunteering for the study, subjects were informed that they would receive full-credit (as agreed upon in previous discussion with all instructors), regardless of whether or not they were able to participate.

Once subjects were greeted and comfortably seated, the first step of the experiment involved testing subjects to ensure that they possessed adequate auditory and visual ability. To test subjects' auditory ability, the aforementioned Maico MA-19 Audiometer was implemented. Subjects were seated facing a blank non-textured wall, with the audiometer located directly behind them. Subjects were then instructed that they would be receiving an auditory signal in either the left or right ear, but not in both ears at the same time. When the subjects thought that they heard the signal, they were instructed to raise the hand that represented the ear where they believed they heard the stimuli (e.g., "... if you hear the signal in your left ear, raise your left hand."). Subjects were asked to repeat the instructions to the researcher to ensure that they correctly understood the directions. Once directions were correctly repeated, the subjects were handed the headphones and instructed to place them so that they fit, securely, snugly, blocked background ambient noise, and were comfortable. The researcher also examined the placement of the headphones to further ensure that a uniformity of usage occurred and that all subjects had the headphones placed in a manner that created optimum transmission quality.

Once the subject indicated that the headphones were comfortably in place, the researcher turned on the audiometer and commenced auditory testing. During this phase of the study, subjects received an auditory signal in both the left and right ear at four different Hz levels: 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz. Auditory signals were randomized by ear and Hz level and presented to subjects in 10 second bursts. If a subject failed to correctly identify a signal in the proper ear after 10 seconds, the response would be recorded as incorrect. After further auditory testing was concluded, any subject who incorrectly identified a signal would be informed about the incorrect response and dismissed from the study. Before leaving the testing area, the subject would also be informed that the researcher, while fully trained on the audiometer, was a novice practioner. Results did not imply hearing loss or impairment, but further in-depth testing might be prudent. The researcher would then provide the subject(s) with a referral source on campus (Speech, Language, and Hearing Center) and follow-up with the subject and the referral source. Subjects who successfully recognized all auditory signals advanced to the visual testing stage.

In the visual testing stage, subjects were tested for color-blindness. To test for color-blindness, subjects were again placed so that they sat facing a blank, non-textured wall. As they sat facing the wall, subjects were instructed that they would be presented four colors, each of which they would later see in the study. Subjects were then instructed to verbally identify what color they believed they saw (e.g., "...if you believe you see a yellow color on the wall, please say, 'yellow' aloud."). Again, subjects were asked to repeat the instructions back to the researcher in order to ensure that they fully understood the task they were required to perform. After subjects indicated via their feedback of the instructions that they understood the directions given by the researcher, each color was projected individually onto the wall in front of them. Colors were presented at a uniform height on the wall and projected from a uniform distance. Each color was projected onto the wall until the subject responded, at which point the next color was projected. After each color was projected onto the wall, the subjects' response to the color was recorded. If at any point a subject incorrectly indetified a color, the incorrect response would be recorded. After completing testing, the subject would then be informed about the incorrect response and be dismissed from the study. Again, a referral service was available (the campus nurse) for any subject who failed to correctly identify a color.

In addition to subjects who failed to meet the criteria for hearing and visual ability, subjects who indicated that they had participated in hearing or speech studies or had more than a cursory exposure to synthetic speech stimuli were also dismissed from the experiment. This was included as a condition in order to control for any previous learning effect. While these prerequisites were implemented, all subjects successfully passed visual and auditory screening, and none of the subjects had been exposed to or participated in other speech or hearing studies.

After having been screened for lack of hearing impairment and color-blindness, subjects were randomly assigned to one of the two levels of speech stimuli, either the natural speech stimuli presentation group or the computer-generated speech stimuli presentation group. Regardless of the speech stimuli presentation group to which the subject was assigned, the directions and procedure for the experiment remained the same.

The actual experimental stage began with subjects remaining seated in the same direction as they faced while undergoing color-blindness screening. As they faced the blank, non-textured wall, the researcher began by explaining the subject's role in the next stage of the study. Subjects were told that they would be asked to perform two simultaneous tasks.

The primary task involved identifying word lists. The word lists consist of phonetically balanced (PB) words. Phonetically balanced words were chose because of their close approximation to everyday spoken English and their high comprehensibility. The word list was comprised of words rated high in familiarity and concreteness as identified by Benjafield and Muckenheim (1989). Benjafield and Muckenheim identify familiarity and concreteness in the following manner:

> ... words differ in their *familiarity* — that is, how commonly or frequently they have been experienced or how familiar they seem to be...words differ in the extent to which they refer to *concrete* objects, persons, places, or things that can be seen, heard felt, smelled , or tasted, as contrasted with *abstract* concepts that cannot be experienced by our senses (p.33).

Fifteen words were presented to subjects at each task difficulty level, for a total of 45 words presented to each subject. Words within each fifteen set group were of a similar familiarity and concreteness rating (± .50 on both familiarity and concreteness ratings; that is, on a scale of 1.00 (low) to 7.00 (high), all words presented were between 6.50 and 7.00 on both familiarity and concreteness). The three fifteen-item word lists, along with their familiarity and concreteness ratings, are presented in Appendix C. The highly stringent criteria for word inclusion was also expected to create a high degree of reliability between all three fifteen word set groups. Via a headphone set, subjects received either natural or synthetic speech words at four second intervals. Subjects in the natural speech stimuli presentation group received a pre-recorded taped list of words, presented by the researcher's trained colleague. Subjects in the computer-generated speech stimuli presentation group received the same word lists; however, subjects in this latter group received words created by the Smoothtalker2 computer-generated speech software program, transmitted through a Macintosh Plus personal computer. The quality of the headphone set allowed subjects to receive the stimuli in both ears, optimizing cognition of the word. Pilot testing with Cerritos College full-time classified staff who had submitted to the same auditory screening procedures revealed that a four second interval between words provided an optimum response period without allowing the subject too much time to formulate an "intelligent guess." After they heard a word, subjects were instructed to repeat the word that they believed they heard aloud (e.g., "...if you think that you hear the word 'dog,' I want you to say the word, 'dog' aloud."). When a subject repeated the word that he or she believed had been presented, the researcher recorded whether the word the subject said was correct or incorrect. Correct responses were identified as an exact duplication of the word presented to the subject within the four second period immediately proceeding word presentation. An incorrect response was recorded when the subject provided a different word (including approximations, e.g., "hat," instead of "cat") than the one presented, or when the subject failed to respond in the allocated four seconds. Late responses (those occuring after the allocated four second response period) were also recorded as incorrect, regardless of whether or not the response was indeed correct. Word presentation and response was repeated in this manner until a sequence of fifteen words were presented to subjects. Three fifteen item word lists were involved in the study. Specific word lists were affixed to specific task difficulty levels; that is, the same word list was always presented at a simple task difficulty level, a different fifteen item word list always presented at the moderately difficult task level, and so on.

Concurrent to word presentation and identification, subjects were also presented with a secondary task. As Knowles (1963) notes, a number of factors must be taken into consideration when presenting a secondary task. The secondary task, "should not physically interfere with, nor otherwise disrupt, primary task performance" (p. 156). Also, the secondary task should be simple. "The task should require very little learning and should show little inter-subject variability" (Knowles, 1963, p. 156). To meet these criteria, the secondary task consisted of subjects identifying a specific colored slide. This type of secondary task does not interfere with the primary task; nor does it create sensory channel overload. The visual stimuli, the flashing of a set color, also requires little (if any) learning and should demonstrate very little inter-subject variability. The visual stimuli was also presented randomly among a group of other flashing colors so that subjects would not attempt to respond based upon a non-existent perceived presentation pattern.

Prior to presenting the word list to subjects, subjects were also given a second set of directions. As subjects sat facing the blank wall, they were instructed that the color slides they had previously viewed for the color-blindness test would be randomly presented on the wall in front of them. While subjects were asked to attend to all slides regardless of color, they were instructed to pay particular attention to blue colored slides. Other colors presented were red, green and yellow. The choice of these colors reflects a large dichotomy in the color spectrum. In addition to being highly distinguishable from each other, color-blindness for any of these four colors is easily detectable. Increase in the attention demands of the secondary task acted as the variable that altered task difficulty. Responses for the secondary task were also recorded to further examine any degradation in secondary task performance that may have occurred.

While subjects started the experiment at different task difficulty levels, the specifics of the secondary task remained the same. A computer mouse was implemented as a dummy response button. The mouse was placed in front of the subject, offset to either the subject's left or right, depending upon the subject's self-reported handedness. The mouse cable was strung across the table in front of the subject and the loose end taped under the table, providing a further illusionary measure of computer control and sophistication to subjects. Subjects were instructed to place their hand to either side of the mouse. When the task difficulty level required a response, subjects were instructed to gently but firmly press the click-and-drag button on the mouse.

When they had completed this action, subjects were instructed to return their hand to its initial resting position beside the mouse. Subjects were reminded to never rest their hand directly on the mouse between responses. While the mouse button was not connected to the Macintosh Plus computer, the subjects' physical response of moving their hand from its resting position to the mouse was recorded as a response. A response was recorded as correct when the subject's physical approximation to the mouse button coincided with an accurate *nth* interval response, as dictated by the task difficulty level. At the simple task difficulty level, subjects were asked to press the mouse button every time a blue color slide appeared. At the moderately difficult task level, subjects were instructed to press the mouse/response button every *third* time a blue color slide was projected. Finally, at the difficult task level, subjects were instructed to press the response button every *fifth* time a blue color slide appeared.

In addition to randomly placing subjects into either a natural or computer-generated speech group, task difficulty level presentation was randomized. As was previously mentioned, subjects in each speech stimuli presentation group were randomly assigned to begin at different task difficulty levels. Thus, one-third of the subjects in both speech stimuli presentation conditions received instructions for and performed a simple task, followed by a moderately difficult task and a difficult task. A second third of the subjects began by receiving instructions for and performing a moderately difficult task, followed by a difficult task and a simple task. The remaining third of the subjects in both speech stimuli presentation conditions were given instructions for and asked to perform a difficult task, followed by a simple task and a moderately difficult task. This stimuli presentation method served to counter-balance any effects that may have been attributable to differential (asymmetrical) transfer between conditions.

After subjects were exposed to the instructions and completed the task at all three difficulty levels, the experiment concluded. The researcher removed the subject's headphone set and handed the subject a debriefing form that explained the experiment in greater detail and stated the researcher's hypotheses. The debriefing form also included a telephone number at which the researcher could be contacted if a subject desired further information about the study. Before a subject left the experimental area, the researcher again made sure that all questions were answered.

Results

Subjects.

Sixty subjects met the criteria for participation in the study. Exactly half were selected to receive computer-generated speech training. The remaining thirty subjects received natural speech training. Thirty-six of the subjects (60.0%) werer female; 24 (40.0%) were male. While seven different ethnic groups were represented, the majority of subjects were either Caucasian (23 subjects - 38.3%), Hispanic (19 subjects - 31.7%), Black/Afro-American (6 subjects - 10.0%), or Asian (6 subjects - 10.0%). Thirty-seven of the subjects (61.7%) were full-time students (enrolled in 12 or more units), with the remaining 23 students (38.3%) indicating part-time (less than 12 units) enrollment status. The mean age of the participants was 25.4, ranging from a minimum of 17 years of age to a maximum of 53 years of age. With nine occurences, the mode was nineteen.

The first step of analysis involved the computation of preliminary descriptive statistics. The means, standard deviations and confidence intervals for percent correct word identification were computed for both the natural and computer-generated speech groups at all three task difficulty levels. These results are displayed in Table 1 and graphed in Figure 2. In addition to descriptive statistics, normal probability plots were also computed. At all three task difficulty levels, it appeared that a linear relationship existed, indicating that the values obtained at each task difficulty level were normally distributed.

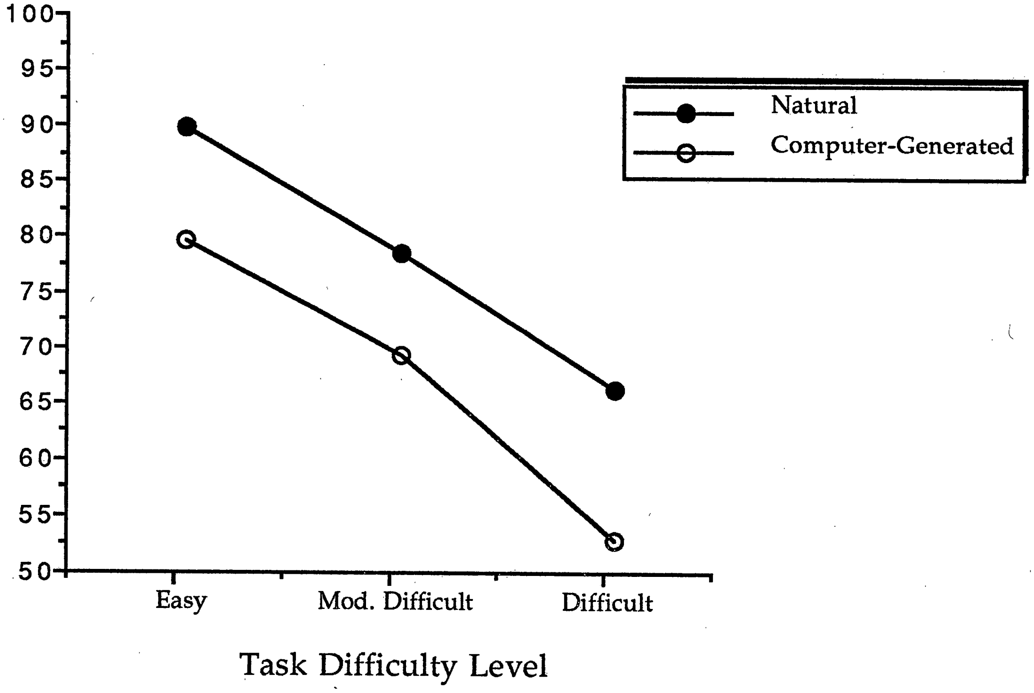
Since the dependent variable (percent correct word identification) was measured for each subject under three different conditions (task difficulty level), data was analyzed by means of a repeated measures MANOVA. In conducting all analyses, the statistical package SPSS-x, (release 3.1, for the VAX/VMS operating system) was implemented. Percent correct word

Table 1

Phonetically-balanced words correctly identified at each task difficulty level

Speech Group	Task Difficulty Level Easy	Moderately Difficult	Difficult
Natural Speech (N=30):			
M	88.9%	77.6%	65.1%
SD	14.2	16.3	21.8
Computer-Generated Speech (N=30):			
M	78.7%	68.2%	51.8%
SD	16.7	14.3	16.4

Figure 2. Percent of phonetically-balanced words correctly identified at each task difficulty level by speech group.



identification was loaded as the dependent variable, type of speech as the independent variable, and task difficulty as the concomitant (covariate) variable. Task difficulty was considered as a covariate to adjust for between-subject effects. Controlling for task difficulty allowed the researcher to ascertain whether observed differences between natural and computer-generated speech groups on percent correct word identification were truly attributable to the type of speech training a subject received.

Variables were transformed so that linear combinations of their differences, not the differences between the variables themselves, could be analyzed. The first analysis conducted examined the orthonormalized contrast that corresponded to between-subject effects. In examining the results of this analysis, significant findings were observed in regard to task difficulty ($F = 9.79$, $df = 1{,}57$, $p < .003$) and type of speech ($F = 5.94$, $df = 1{,}57$, $p < .018$). These findings suggest that, while variability due to task difficulty is significant, differences attributable to type of speech are significant even after differences between groups due to task difficulty are controlled.

Further analyses examined the orthnormalized contrast that corresponded to the percent correct word identification within-subject effect. While the significance of Mauchley's test of sphericity ($p < .000$) indicated a violation of assumptions of sphericity, adjustments to the numerator and denominator degrees of freedom were made by multiplying these degrees of freedom by lower-bound epsilon. Even after these adjustments were made, an analysis of type of speech by percent correct word identification revealed no statistically significant interaction effect. However, in further analyses a statistically significant main effect (F = 16.51, df = 1,27, p < .000) and within-subject effect (F = 33.94, df = 1,57.5, p < .000) was observed for percent correct word identification.

While the aforementioned analyses examined main, between- and within-subject effects for statistical significance, more specific analyses were necessary to properly address the researcher's four stated hypotheses. Student's t-tests and MANOVA analyses were conducted to test these four hypotheses. Percent correct word identification was the dependent variable and type of speech the independent variable in both t-test and MANOVA analyses; task difficulty was added as the covariate in all MANOVA analyses.

The researcher's first hypothesis examined percent correct word identification between natural and computer-generated speech groups at the simple task difficulty level. The researcher believed that both groups would correctly identify words at a similar performance rate. MANOVA analyses indicated that, at the simple task difficulty level, the sums of squares due to the regression was p < .177, indicating that variability attributable to the covariate (percent correct color identification) was not statistically significant. Further analyses based upon adjustments for the covariate revealed statistically significant within-subjects results (F = 9.02, df = 1,57, p < .004). Overshadowing these within-subjects results, however, was the statistically significant interaction noted between type of speech training received and percent correct word identification (*F* = 6.38, *df* = 1,57, *p* < .014). T-test results support these interaction findings (*t* = 2.55, *df* = 1,56.44, *p* < .013). The mean correct word identification of subjects in the natural speech group (88.9%) was statistically significantly higher than the mean correct word identification of subjects in the computer-generated speech group (78.7%). Thus, contrary to the researcher's first stated hypothesis, statistically significant performance level differences were noted between subjects in the natural speech group and subjects in the computer-generated speech group at the simple task difficulty level.

The researcher's second hypothesis assumed that: a) at the moderately difficult task level, subjects in both speech groups would suffer performance degradation, with subjects in the computer-generated speech group suffering significantly greater performance degradation; and b) the performance levels between these two groups would not differ significantly. MANOVA analyses revealed that the observed sums of squares due to the regression was statistically significant (*F* = 5.38, *df* = 1,57, *p* < .024), indicating that some of the variability at the moderately difficult task level was attributable to the covariate. After controlling for the statistical significance of the covariate, within-subject statistical significance was recorded (*F* = 71.72, *df* = 1,57, *p* < .000). This finding somewhat supports the first portion of the second hypothesis, with statistically significantly greater performance degradation noted for subjects in the computer-generated speech group as task difficulty increased (simple to moderate). However, contrary to expected findings, statistically significant performance degradation from simple to moderately difficult task level was also observed among the natural speech group. T-test results also failed to support the second portion of the second hypothesis. The 77.6% correct word identification noted among natural speech group subjects was statistically significantly higher than the 68.2% correct word identification recorded for subjects in the computer-generated speech group ($t = 2.36$, $df = 1{,}57.05$, $p < .022$).

The researcher's third stated hypothesis predicted statistically significant performance degradation for both speech groups from the moderately difficult to difficult task level. MANOVA analyses again revealed statistically significant variability due to the covariate ($F = 7.45$, $df = 1, 57$, $p < .008$). After controlling for this variability, within-subject statistically significant results were still observed ($F = 307.11$, $df = 1{,}57$, $p < .000$). These findings support the third hypothesis.

While statistically significant performance degradation from the moderately difficult to difficult task level was expected for both groups, degradation was expected to be greater among subjects in the computer-generated speech group, setting up conditions for the fourth stated hypothesis: that subjects in the computer-generated speech group would experience statistically significantly lower performance rates at the difficult task level than subjects in the natural speech group. To test this assumption, MANOVA and t-test analyses were again conducted. As at the simple task difficulty level, MANOVA analyses revealed a statistically significant interaction between type of speech and percent correct word identification ($F = 6.16$, $df = 1,57$, $p < .016$). T-test results support this finding ($t = 2.67$, $df = 1,53.80$, $p < .010$), with the 65.1% correct word identification recorded among natural speech subjects statistically significantly higher than the 51.8% recorded for subject in the computer-generated speech group.

Discussion

The statistically significant findings presented in the results section failed to substantiate all of the stated hypotheses. In reviewing the hypotheses, the first stated hypothesis predicted similar percent correct word identification between the natural speech group and the computer-generated speech group at the simple task difficulty level. Contrary to this prediction, statistically significant findings were observed at the simple task difficulty level between speech groups. Considering that the covariate did not add a statistically significant amount of variability, this finding suggests that observed differences are likely due to type of speech. While all available precautions were taken to create computer-generated speech that replicated natural speech as closely as possible, the researcher believes that the quality of the computer-generated speech implemented in the study accounted for the unexpected statistically significant difference observed. The use of a higher quality computer-generated speech software systems (e.g., DECTalk, Prose, etc.) may provide a truer reflection of the comprehensibility of computer-generated speech at simple task levels.

At the moderately difficult task level, differences between natural speech and computer-generated speech groups were also found to be statistically significant, contrary to the researcher's stated hypothesis. While it was predicted that performance degradation among subjects in the computer-generated speech group would be significantly greater than the degradation observed among subjects in the natural speech group, statistically significant levels of degradation were observed among both groups. The researcher believes that some of the degradation in percent correct word identification in the natural speech group is due to the secondary task included to alter task difficulty level. Statistical analyses somewhat support this contention; significant findings were observed indicating that some of the variability that occurred at the moderately difficult task level was attributable to the covariate. While the performance degradation experienced among the natural speech group on the primary task was not as severe as that experienced by subjects in the computer-generated speech group, this finding suggests the need for more stringent control of natural speech in future research.

At the difficult task level, statistically significant performance differences were observed between moderately difficult and difficult task levels for both groups, supporting the third hypothesis. The fourth hypothesis was also supported as statistically significant performance levels were observed between subjects in the natural speech group and subjects in the computer-generated speech group.

An overall examination of the findings reveals mixed support for the researcher's stated hypotheses. Contrary to expected findings, statistically significant findings were observed between natural and computer-generated speech groups at the simple task difficulty level. Both groups exhibited statistically significant performance degradation from the simple to the moderately difficult task level. While only the computer-generated speech group was expected to experience performance degradation, degradation among the natural speech group is largely attributable to the actual difficulty of the concomitant variable at the moderately difficult task level. Observed performance degradation among subjects in the computer-generated speech group might have been even greater if the quality of the computer-generated speech had supported the first hypothesis and performance at the simple task difficulty level been higher. The remaining portion of the study followed the researcher's stated expectations. Both speech groups suffered statistically significant performance degradation from the moderately difficult to the difficult task level. Statistically significant differences were also observed between the computer-generated speech and natural speech groups at the difficult task level, supporting cited research and hypotheses that subjects trained via natural speech will outperform subjects trained with computer-generated speech, especially as the task increases in difficulty.

Appendix A

Basic Phoneme Chart

AE	short "a" as in "last"	EH	short "e" as in "best"
IH	short "i" as in "fit"	AA	short "o" as in "cot"
AH	short "u" as in "up"		
EY	long "a" as in "ace"	IY	long "e" as in "beet"
AY	long "i" as in "ice"	OW	long "o" as in "dose"
UW	long "u" as in "lute"		
OY	diphthong in "noise"	AW	diphthong in "loud"
ER	"further" or "further"	CH	"chin"
TH	"thin"	SH	"shin"
DH	"then"	ZH	"z" as in "pleasure"
NG	"sing"	WH	"which"
P	"pin"	T	"tin"
B	"bin"	K	"kin"
D	"din"	J	"gin"
G	"given"	F	"fin"
S	"sin"	V	"vim"
Z	"zen"	L	"light"
M	"might"	N	"night"
H	"hit"	R	"rate"
W	"wait"	Y	"yet"
UH	"u" sound in book		
AX	schwa sound in "against"		
AO	intermediate "o'" as in "caught"		

Appendix B

Decibel levels (dB) for various sounds

DECIBELS	Environmental Noise	Specific Noise Source	DECIBELS
140			140
		50 hp siren (100 ft)	
130			130
		Jet takeoff (200 ft)	
120		Rock concert with amplifier (6 ft)	120
110	Casting shakeout area	Riveting machine*	110
		Cutoff Saw*	
100	Electric furnace area	Pneumatic peen hammer*	100
	Boiler room	Textile weaving plant*	
90	Printing press plant	Subway train (20 ft)	90
	Tabulating room	Pneumatic drill (50 ft)	
80			80
	Inside sports car (50 mph)	Freight train (100 ft)	
70		Vacuum cleaner (10 ft)	70
	Near freeway (auto traffic)		
60	Light store/accounting office	Speech (1 ft)	60
	Private business office		
50	Light traffic (100 ft)	Large transformer (200 ft)	50
	Average residence		
40	Min. levels, residential areas		40
	Studio (Speech)		
30		Soft whisper (5 ft)	30
	Studio for sound pictures		
20			20
10		Normal breathing	10
0			0

*Operator's position

Appendix C

Phonetically balanced words to be presented at the various task difficulty levels

Word	Familiarity	Concreteness
Group 1		
ASH	6.94	6.33
CAT	7.00	6.87
COUPON	6.87	6.53
FLOOR	6.90	6.80
LEECH	6.33	6.50
MILL	6.90	6.53
PADDLE	6.87	6.57
PIN	7.00	6.40
RIVER	7.00	6.47
SHIP	7.00	6.67
STAR	7.00	6.37
STREET	7.00	6.67
TOOTH	6.97	6.60
WALNUT	6.97	6.70
WHEAT	7.00	6.80
	$\bar{x}$=6.92	$\bar{x}$=6.59
Group 2		
BOW	6.97	6.53
CATTLE	6.94	6.43
FACE	7.00	6.63
FOOT	7.00	6.57
ICE	7.00	6.63
MOUSE	7.00	6.80
PEBBLE	6.80	6.37
PLANE	7.00	6.47
SATELLITE	6.80	6.40
SKY	7.00	6.47
STICK	6.80	6.57
TIE	6.94	6.37
TRUCK	7.00	6.77
WATER	7.00	6.73
WOLF	7.00	6.73
	$\bar{x}$=6.95	$\bar{x}$=6.56

Appendix C (continued)

Phonetically balanced words to be presented at the various task difficulty levels

Word	Familiarity	Concreteness
Group 3		
BOXER	6.90	6.57
COCK	6.84	6.33
FIELD	6.97	6.33
GRASS	6.97	6.77
HOUSE	6.97	6.77
OCEAN	6.97	6.63
PIE	6.97	6.40
POOL	6.97	6.67
SHEEP	7.00	6.67
SQUARE	7.00	6.53
STORM	7.00	6.47
TOMB	6.97	6.33
TRUNK	7.00	6.57
WEB	6.83	6.43
WORM	6.94	6.67
	$\bar{x}$=6.93	$\bar{x}$=6.54

References

Ainsworth, W.A. (1972). A Real-Time Speech Synthesis System. *IEEE Trans. Audio Electroacoust.,* **AU-20**, 397.

Ainsworth, W.A. (1974). Performance of A Speech Synthesis System. *International Journal of Man-Machine Studies,* **6**, 493-511.

Arnold, G.E. (1960). Alleviation of Alaryngeal Aphonia With the Modern Artificial Larynx: I. Evolution of Artificial Speech Aids and Their Value for Rehabilitation. *Logos,* **3**, 55-67.

Beasley, Daniel S., Bratt, Gene W., and Rintelmann, William F. (1980). Intelligibility of Time-Compressed Sentential Stimuli. *Journal of Speech and Hearing Research,* **23**, 722-731.

Beasley, D., Schwimmer, S., and Rintelmann, W. (1972b). Intelligibility of Time-Compressed CNC Monosyllables. *Journal of Speech and Hearing Research,* **15**, 340-350.

Benjafield, John, and Muckenheim, Ron (1989). Dates of Entry and Measures of Imagery, Concreteness, Goodness, and Familiarity for 1,046 Words sampled from the Oxford English Dictionary. *Behavior Research Methods,* **21(1)**, 31-52.

Bocca, E., and Calearo, C. (1963). Central Hearing Processes. In Jerger (Ed.), *Modern Developments in Audiology.* New York: Academic, 337-370.

Bornstein, Steven P., Randolph, Kenneth J., Maxon, Antonia, and Giolas, Thomas (1982). The Effects of Different Speech Stimuli on the Measurement of Speech Intelligibility Under Conditions of High Frequency Range and Frequency Response Irregularity. Paper presented at 103rd Meeting of The Accoustical Society of America.

Burrows, A.A. (1960). Acoustic Noise, An Informational Definition. *Human Factors,* **2(3)**, 163,168.

Calearo, C., and Lazzaroni, A. (1957). Speech Intelligibility in Relation to the Speed of the Message. *Laryngoscope,* **67,** 410-419.

Campbell, Richard A. (1974). Computer Audiometry. *Journal of Speech and Hearing Research,* **17,** 134-140.

Carhart, R. (1965). Problems in the Measurement of Speech Discrimination. *Archives of Otolaryngology,* **82,** 253-260.

Clark, John G., and Stemple, Joseph C. (1982). Assessment of Three Modes of Alaryngeal Speech With a Synthetic Sentence Identification (SSI) Task in Varying Message-To-Competition Ratios. *Journal of Speech and Hearing Research,* **25,** 333-338.

Cole, Ronald A. (1973). Perceiving Syllables and Remembering Phonemes. *Journal of Speech and Hearing Research,* **16,** 37-47.

Cole, R.A., and Jakimik, J. (1980). A Model Of Speech Perception. In R.A. Cole (Ed.), *Perception and Production of Fluent Speech..* Hillsdale, N.J.: Erlbaum.

Crouse, G. P. (1962). *An Experimental Study of Esophageal and Artificial Larynx Speech.* Unpublished thesis, Emory University, Atlanta.

Curry, E.T., and Snidecor, J.C. (1961). Physical Measurement and Pitch Perception in Esophageal Speech. *Laryngoscope,* **71,** 415-424.

de Haan, H.J. (1977). A Speech-Rate Intelligibility Treshold for Speeded and Time-Compressed Connected Speech. *Perception and Psychophysics,* **22,** 366-372.

de Haan, H.J., and Schjelderup, J.R. (1978). Treshold of Intelligibility/ Comprehensibility of Rapid Connected Speech: Method and Instrumentation. *Behavior Research Methods and Instrumentation,* **10,** 841-844.

de Quiros, J.B. (1964). Accelerated Speech Audiometry, an Examination of Test Results. *Translations of the Beltone Institute for Hearing Research,* **17**.

Di Carlo, L.M., Amster, W.W., and Herer, G.R. (1956). *Speech After Laryngectomy*. Syracuse: Syracuse University Press.

Di Carlo, L.M., and Taub, H. (1972). The Influence of Compression and Expansion on the Intelligibility of Speech by Young and Aged Aphasic (Demonstrated CVA) Individuals. *Journal of Communication Discoveries,* **5**, 299-306.

Doddington, G., and Schalk, T. (1981). Speech Recognition: Turning Theory to Practice. *IEEE Spectrum,* **18**, 26-32.

Dooling, D.J. (1974). Rhythm and Syntax in Sentence Perception. *Journal of Verbal Learning and Verbal Behavior,* **13**, 255-264.

Duffy, Joseph R., and Giolas, Thomas G. (1974). Sentence Intelligibility as a Function of Key Word Selection. *Journal of Speech and Hearing Research,* **17**, 631-637.

Egan, J.P. (1948). Articulation Testing Methods. *Laryngoscope,* **58**, 955-991.

Epstein, R., Giolas, T.G., and Owens, E. (1968). Familiarity and Intelligibility of Monosyllabic Word Lists. *Journal of Speech and Hearing Research,* **11**, 435-438.

Flanagan, J.L., Johnston, J.D., and Upton, J.W. (1982). Digital Voice Storage in A Microprocessor. *IEEE Tansactions on Communication,* **30**, 336-345.

Foulke, E. (1965). The Comprehension of Rapid Speech by the Blind--Part III. (Semi-Annual Progress Report, Cooperative Research Project No. 2430). Washington, D.C.: United States Department of Health, Education, and Welfare, Office of Education.

Foulke, E., and Sticht, T.G. (1969). Review of Research on the Intelligibility and Comprehension of Accelerated Speech. *Psychological Bulletin,* **77**, 50-62.

French, N.R., and Steinberg, J.C. (1947). Factors Governing the Intelligibility of Speech Sounds. *Journal of the Accoustical Society of America,* **19**, 90-119.

Gallant, John (1987). Low-Cost ICs Provide Flexibility for Applications Requiring Voice Output. *EDN,* 63-70.

Gardner, W. H., and Harris, H. E. (1961). Aids and Devices for Laryngectomees. *Archives of Otolaryngology,* **73**, 145-152.

Garvey, W.D. (1953). The Intelligibility of Speeded Speech. *Journal of Experimental Psychology,* **45**, 102-108.

Giolas, T.G., and Epstein, A. (1963). Comparative Intelligibility of Word Lists and Continuous Discourse. *Journal of Speech and Hearing Research,* **6**, 349-358.

Greenspan, Steven L., Nusbaum, Howard C., and Pisoni, David B. (1985). Perceptual Learning of Synthetic Words and Sentences. Paper presented at 110th Meeting of The Accoustical Society of America.

Guillemin, B.J., and Nguyen, D.T. (1984). Microprocessor-based Speech Processing Systems. *Journal of Speech and Hearing Research,* **27**, 311-317.

Hakkinen, Markku T., and Williges, Beverly H. (1984). Synthesized Warning Messages: Effects of an Alerting Cue in Single- and Multiple-Function Voice Synthesis Systems. *Human Factors,* **26(2)**, 184-195.

Hansen, John H., and Clements, Mark A. Objective Quality Measures Applied to Enhance Speech. Paper presented at 110th Meeting of The Accoustical Society of America.

Hawkins, J.S., Reising, J.M., Lizza, G.D., and Beachy, K.A. (1983). Is a Picture Worth a 1000 Words — Written or Spoken? Proceedings of the Human Factors Society 27th Annual Meeting, 970-972.

Hirsh, I.J., Reynolds, E., and Josepf, M. (1954). Intelligibility of Different Speech Materials. *Journal of the Accoustical Society of America.* **26**, 530-531.

Hofmann, Mark A., and Heimstra, Norman W. (1972). Tracking Performance With Visual, Auditory, or Electrocutaneous Displays. *Human Factors,* **14(2)**, 131-138.

House, A.S., Williams, C.E., Hecker, M.H.L., and Kryter, K.D. (1965). Articulation-Testing Methods: Consonantal Differentiation With a Closed-Response Set. *Journal of the Accoustical Society of America,* **37**, 158-166.

Hyman, M. (1955). An Experimental Study of Artificial Larynx and Esophageal Speech. *Journal of Speech and Hearing Disorders,* **20**, 291-299.

IEEE. (1969). IEEE Recommended Practice for Speech Quality Measurements. (IEEE No.297). New York: Author.

Kalikow, D.N., Stevens, K.N., and Elliott, L.L. (1977). Development of a Test of Speech Intelligibility in Noise Using Sentence Materials With Controlled Word Predictability. *Journal of the Accoustical Society of America,* **61**, 1337-1351.

Kent, R.D. (1973). The Imitation of Synthetic Vowels and Some Implications for Speech Memory. *Phonetica,* **28**, 1-25.

Kent, R.D. (1974). Auditory-Motor Formant Tracking: A Study of Speech Imitation. *Journal of Speech and Hearing Research,* **17**, 203-222.

Klapp, Stuart P., Kelly, Patricia A., and Netick, Allan. (1987). Hesitations in Continuous Tracking Induced by a Concurrent Discrete Task. *Human Factors,* **29(3)**, 327-337.

Knowles, W.B. (1963). Operator Loading Tasks. *Human Factors,* 155-161.

Koch, R. The AmBiChron. (1974). In S. Duker (Ed.) *Time-Compressed Speech: An Anthology and Bibliography.* Metuchen, N.J.: Scarecrow Press.

Konkle, D., Beasley, D., and Bess, F. (1977). Intelligibility of Time-Altered Speech in Relation to Chronological Aging. *Journal of Speech and Hearing Research,* **20**, 108-115.

Larkey, Leah S., and Danly, Martha. (1983). Fundamental Frequency and the Comprehension of Simple and Complex Sentences. Paper presented at 106th Meeting of the Accoustical Society of America.

Lehiste, I., and Peterson, G.E. (1959). Linguistic Considerations in the Study of Speech Intelligibility. *Journal of the Accoustical Society of America,* **31**, 280-286.

Liberman, A. (1970). The Grammars of Speech and Language. *Cognitive Psychology,* **1**, 301-323.

Loeb, Michel, and Binford, John R. (1964). Vigilance for Auditory Intensity Changes as a Function of Preliminary Feedback and Confidence Level. *Human Factors,* **6**, 445-458.

Logan, John S., Greene, Beth G., and Pisoni, David B. (1989). Segmental Intelligibility of Synthetic Speech Produced by Rule. *Journal of the Acoustical Society of America,* **86(2)**, 566-581.

Luce, Paul A., Feustel, Timothy C., and Pisoni, David B. (1983). Capacity Demands in Short-Term Memory for Synthetic and Natural Speech. *Human Factors,* **25(1)**, 17-32.

Lukatela, Georgije, Carello, Claudia, Kostic, Alexandar, and Turvey, M.T. (1988). Low-Constraint Facilitation in Lexical Decision with Single-Word Context. *American Journal of Psychology,* **101(1)**, 15-29.

Madell, J.R., and Goldstein, R. (1972). Relation Between Loudness and the Amplitude of the Early Components of the Averaged Electroencephalic Response. *Journal of Speech and Hearing Research,* **15**, 134-141.

Marics, Monica A., and Williges, Beverly H. (1988). The Intelligibility of Synthesized Speech in Data Inquiry Systems. *Human Factors,* **30(6)**, 719-732.

Martin, F.N., and Mussell, S.A. (1979). The Influence of Pauses in Competing Signal on Synthetic Sentence Identification Scores. *Journal of Speech and Hearing Disorders,* **44**, 282-292.

Massaro, D. (1972). The Role of Perceptual Images, Perceptual Units, and Processing Time in Auditory Perception. *Psychological Revue,* **79**, 124-145.

McCormick, E.J., and Sanders, M.S. (1987). *Human Factors in Engineering and Design* (6th Edition). New York: McGraw-Hill Book Company.

Miller, G.A., Heise, G.A., and Lichten, W. (1951). The Intelligibility of Speech as A Function of the Context of Test Material. *Journal of Experimental Psychology,* **41**, 329-335.

Miller, G.A., and Licklider, J.C.R. (1950). The Intelligibility of Interrupted Speech. *Journal of the Accoustical Society of America,* **22**, 167-173.

Neisser, U. (1967). *Cognitive Psychology.* New York: Appleton-Century-Crofts.

Nusbaum, Howard C., Greenspan, Steven L., and Pisoni, David B. (1986). Perceptual Attention in Monitoring Natural and Synthetic Speech. Paper presented at 110th Meeting of The Accoustical Society of America.

Nusbaum, H.C., and Pisoni, D.B. (1985). Constraints on the Perception of Synthetic Speech Generated by Rule. *Behavior Research Methods, Instruments, and Computers,* **17,** 235-242.

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

Owens, E., Benedict, M., and Schubert, E.D. (1972). Consonant Phonemic Errors Associated With Pure Tone Configurations and Certain Kinds of Hearing Impairment. *Journal of Speech and Hearing Research,* **15,** 308-322.

Owens, E., and Schubert, E.D. (1968). The Development of Consonant Items for Speech Discrimination Testing. *Journal of Speech and Hearing Research,* **11,** 656-667.

Pascoe, D.P. (1975). Frequency Responses of Hearing Aids and Their Effects on the Speech Perception of Hearing-Impaired Subjects. *Ann. Otol. Rhin. Laryngol. (Supplement 23),* **84,** 1-40.

Peterson, A.P.G., and Gross, E.E., Jr. (1972). *Handbook of Noise Measurement* (7th Edition). New Concord, Mass.: General Radio Company.

Pisoni, David B. (1981). Speeded Classification of Natural and Synthetic Speech in a Lexical Decision Task. *Journal of the Accoustical Society of America,* **70,** S98.

Pisoni, D.B. (1982). Perception of Speech: The Human Listener as A Cognitive Interface. *Speech Technology,* **1,** 10-23.

Pisoni, David B., and Koen, Esti. (1982). Some Comparison ofIntelligibility of Synthetic and Natural Speech at Different Speech-to-Noise Ratios. *Journal of the Accoustical Society of America,* **71,** UU1.

Rabiner, L.R. (1977). On the Use of Autocorrelation Analysis for Pitch Detection. *IEEE Transactions on Accoustics, Speech, and Signal Processing,* **25**, 24-33.

Schiavetti, Nicholas, Sitler, Ronald W., Metz, Dale E., and Houde, Robert A. (1984). Prediction of Contextual Speech Intelligibility From Isolated Word Intelligibility Measures. *Journal of Speech and Hearing Research,* **27**, 623-626.

Schon, T. (1970). The Effects on Speech Intelligibility of Time-Compression and Expansion on Normal-Hearing, Hard of Hearing, and Aged Males. *Journal of Audiotory Research,* **10**, 263-268.

Schroeter, Juergen, and Sondhi, M.M. A Hybrid Domain Articulatory Speech Synthesizer. Paper presented at 110th Meeting of The Accoustical Society of America.

Schwab, Eileen C., Nusbaum, Howard C., and Pisoni, David B. (1985). Some Effects of Training on the Perception of Synthetic Speech. *Human Factors,* **27(3)**, 395-408.

Sedory, S.E., Hamlet, S.L., and Connor, N.P. (1989). Comparisons of Perceptual and Acoustic Characteristics of Tracheoesophageal and Esophageal Excellent Esophageal Speech. *Journal of Speech and Hearing Disorders,* **54**, 209-214.

Sherwood, B.A. (1979, August). The Computer Speaks. *IEEE Spectrum,* 18-25.

Simpson, Carol A., and Marchionda-Frost, Kristine (1984). Synthesized Speech Rate and Pitch Effects on Intelligibility of Warning Messages for Pilots. *Human Factors,* **26(5)**, 509-517.

Simpson, Carol A., McCauley, Michael E., Roland, Ellen F., Ruth, John C., and Williges, Beverly H. (1985). System Design for Speech Recognition and Generation. *Human Factors,* **27(3)**, 115-141.

Skinner, M.W. (1980). Speech INtelligibility in Noise-Induced Hearing Loss: Effects of High Frequency Compensation. *Journal of the Accoustical Society of America,* **67**, 306-317.

Slowiaczek, Louisa, M., and Nusbaum, Howard C. (1985). Effects of Speech Rate and Pitch Contour on the Perception of Synthetic Speech. *Human Factors,* **27(6)**, 701-712.

Slowiaczek, Louisa M., and Pisoni, David B. (1982). Effects of Practice on Speeded Classification of Natural and Synthetic Speech. *Journal of the Accoustical Society of America,* **71**, S95.

Smith, Sidney L., and Goodwin, Nancy C. (1970). Computer-Generated Speech and Man-Computer Interaction. *Human Factors,* **12(2)**, 215-223.

Sorenson, J.M., and Cooper, W.E. (1980). Syntactic Coding of Fundamental Frequency in Speech Production. In R.A. Cole (Ed.), *Perception and Production of Fluent Speech.* Hillsdale, N.J.: Erlbaum.

Speaks, C. (1967). Intelligibility of Filtered Synthetic Sentences. *Journal of Speech and Hearing Research,* 10, 289-298.

Speaks, C., and Jerger, J. (1965). Methods for Measurement of Speech Identification. *Journal of Speech and Hearing Research,* **8**, 184-194.

Steeneken, H.J.M., and Houtgast, T. (1980). A Physical Method for Measuring Speech Transmission Quality. *Journal of the Accoustical Society of America,* **67**, 318-326.

Sticht, T., and Gray, B. (1969). The Intelligibility of Time Compressed Words as a Function of Age and Hearing Loss. *Journal of Speech and Hearing Research,* **12**, 443-448.

Thorndike, E., and Lorge, I. (1944). *The Teachers Word Book of 30,000 Words.* New York: Columbia University Teacher's College.

Toscher, Mark M., and Rupp, Ralph R. (1978). A Study of the Central Auditory Processes in Stutterers Using the Synthetic Sentence Identification (SSI) Test Battery. *Journal of Speech and Hearing Research,* **21**, 779-792.

Van Gieson, W.D. Jr., and Chapman, W.D. (1968). Machine-Generated Speech for Use With Computers. *Computers and Automation,* **17**, 31-34.

Wickelgren, W.A. (1965). Distinctive Features and Errors in STM for English Vowels. *Journal of the Acoustical Society of America,* **38**, 583-588.

Wickelgren, W.A. (1966). Distinctive Features and Errors in STM for English Consonants. *Journal of the Acoustical Society of America,* **39**, 388-398.

Wickelgren, W.A. (1969). Context-Sensitive Coding, Associative Memory, and Serial Order in (Speech) Behavior. *Psychological Revue,* **76**, 1-15.

Wickens, C.D. (1980). The Structure of Attentional Resources. In R. Nickerson (Ed.), *Attention and Performance,* (Vol. VIII), 239-257.

Wickens, C.D., Mountford, S.J., and Schreiner, W. (1981). Multiple Resources, Task-Hemispheric Intensity, and Individual Differences in Time-Sharing. *Human Factors,* **23**, 211-229.

Williamson, D., and Curry, D. (1984). Voice I/O Effectiveness Under Heavy Task Loading. *Journal of the American Voice I/O Society,* **1**, 12-23.

Wingfield, A. (1975). Accoustic Redundancy and the Perception of Time-Compressed Speech. *Journal of Speech and Hearing Research,* **18**, 96-104.

Wingfield, A., Buttet, J., and Sandoval, A.W. (1979). Intonation and Intelligibility of Time-Compressed Speech Supplementary Report: English Vs. French. *Journal of Speech and Hearing Research,* **22,** 708-716.

Wingfield, Arthur, Lombardi, Linda, and Sokol, Scott (1984). Prosadic Features and the Intelligibility of Accelerated Speech: Syntactic Versus Prosadic Segmentation. *Journal of Speech and Hearing Research,* **27,** 128-134.

Wolfe, Virginia I., and Ratusnik, David L. (1988). Acoustic and Perceptual Measurements of Roughness Influencing Judgments of Pitch. *Journal of Speech and Hearing Disorders,* **53,** 15-22.

Woodworth, R.S., and Schlosberg, H. (1954). *Experimental Psychology.* New York: Holt.

Yuchtman, Moshe, Nusbaum, Howard C., and Pisoni, David B. Consonant Confusion and Perceptual Spaces for Natural and Synthetic Speech. Paper presented at 110th Meeting of The Accoustical Society of America.

Yulsman, Tom. (1983). How Machines Mimic Speech. *Science Digest,* 96.