INFORMATION TO USERS

This was produced from a copy of a document sent to us for microfilming. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help you understand markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure you of complete continuity.

2. When an image on the film is obliterated with a round black mark it is an indication that the film inspector noticed either blurred copy because of movement during exposure, or duplicate copy. Unless we meant to delete copyrighted materials that should not have been filmed, you will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed the photographer has followed a definite method in "sectioning" the material. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For any illustrations that cannot be reproduced satisfactorily by xerography, photographic prints can be purchased at additional cost and tipped into your xerographic copy. Requests can be made to our Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases we have filmed the best available copy.

University
Microfilms
International

300 N. ZEEB ROAD, ANN ARBOR, MI 48106
18 BEDFORD ROW, LONDON WC1R 4EJ, ENGLAND

ELSON, ILA JEAN

SENSORY DYNAMICS OF SPEAKER RECOGNITION

The University of Oklahoma PH.D. 1980

THE UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

SENSORY DYNAMICS OF SPEAKER RECOGNITION

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

DOCTOR OF PHILOSOPHY

BY

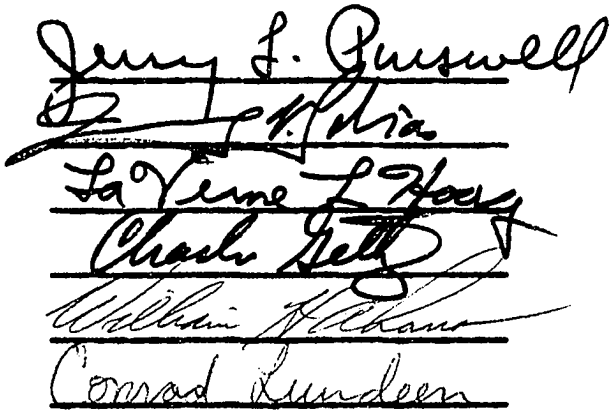
ILA J. ELSON

Norman, Oklahoma

1980

SENSORY DYNAMICS OF SPEAKER RECOGNITION

APPROVED BY



DISSERTATION COMMITTEE

ACKNOWLEDGMENTS

I gratefully acknowledge several persons and institutions for assistance in this study: W. Collins and J. Tobias of the Federal Aviation Administration Civil Aeromedical Institute, who allowed me to use the facility and equipment; W. Morris, who helped with the instrumentation; the Graduate College of the University of Oklahoma, which supplied some of the funds; and the dissertation committee, who read and offered constructive criticism on this manuscript.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

SENSORY DYNAMICS OF SPEAKER RECOGNITION

CHAPTER I

INTRODUCTION

Man-computer symbiosis is an expected development in cooperative interaction between men and electronic computers. Licklider (1960) first described this symbiotic partnership. He noted several prerequisites for the achievement of effective, man-machine systems, one of which was development in input and output equipment or, as it is seen from the human operator's point of view, displays and controls. Research now emerging in the laboratory promises transformation in input and output equipment that can extend man's capabilities and increase his productivity. This research is man-machine communication by voice.

Speaker recognition and speech recognition systems communicate by voice from the user to the machine. For speaker recognition, the task of the system is either to verify a speaker (i.e., a yes-no decision as to whether

the speaker is who he claims to be), or to identify the speaker from some known ensemble. The basic task of a speech recognition system is either to recognize the entire spoken utterance exactly (i.e., a phonetic or orthographic speech-to-text typewriter type of system), or else to understand the spoken utterance (i.e., to respond in some correct manner to what was spoken). The concept of understanding rather than recognizing the utterance is of most importance for systems which deal with fairly large vocabulary, continuous speech input; whereas the concept of exact recognition is of most importance for limited vocabulary, small speaker population, isolated word systems. The ideal system of the future will operate with both speaker and speech recognition functions.

With these system distinctions in mind, the intent of the present dissertation is to provide an empirical survey of the problems and progress of speaker recognition and to examine the effects of speaking effort and noise upon speaker recognition performance for human listening.

People speak more loudly in a noisy environment, or when momentarily deafened, and more softly in a quiet room or when sidetone is artificially increased. In 1911 the French otorhinolaryngologist Lombard described qualitatively this relation between the dynamics of listening and speaking. What he observed has come to be known as the Lombard reflex. Lombard's discovery and his subsequent findings were important

because of their contribution to several areas of scientific inquiry, one of which was the analyses of speech communication in noise. The present study was undertaken in order to examine the importance of three speaking efforts for the discrimination of speakers in ambient or broad-band noise. It is felt this human performance data will contribute theoretical information concerning how humans recognize voices and will indicate the performance of machine verification of talkers in noise.

The purpose of this dissertation is to investigate:

1. Whether a constant speech-to-noise ratio will result in similar speaker recognition scores for normal-normal, shout-shout, and whisper-whisper speaking effort comparisons,

2. Whether there is an optimum speaking effort comparison for speaker recognition with high noise levels, and

3. Whether a constant speech-to-noise ratio will result in similar recognition scores for normal-shout, whisper-normal, and whisper-shout speaking effort comparisons.

The balance of this work is divided into four chapters. In the following two chapters speaker recognition research by listening and machine analysis is reviewed. The fourth chapter describes the experimental methods for data collection, while the final chapter presents the results and discussion of this research. Recommendations are suggested for machine performance.

CHAPTER II

SPEAKER RECOGNITION SYSTEM: MAN

Speaker recognition is broadly defined as "any decision-making process that uses the speaker-dependent features of the speech signal" (Hecker, 1971, p.2). There are two basic recognition tasks: identification and discrimination. In the identification task, an attempt is made to identify the speaker of a particular sample of speech. An example of identification, given by Brown (1979, p.739), "would be the situation where you are in a room with a number of other people. Having turned your back, you hear somebody speak. The voice will therefore correspond to one of the people in the room." The discrimination task, classified by some researchers as a differentiation or same-different task, always involves two speech samples. The listener must decide as to whether the speech samples were produced by the same talker or by different talkers. Brown (1979, p.739) suggests an example of discrimination

is a situation where you telephone an office and are

> connected initially to the switchboard operator. Having asked to be put through to a certain extension, you then wait. The next voice to be heard may be either a different one (an unknown voice replying on the required extension), or the same one (the switchboard operator, to tell you that she cannot get a reply on that extension).

A practical application of both recognition tasks is called speaker verification (also known as authentication or validation). A speaker is said to be verified if his claimed identity as an individual or as a member of a group is confirmed by a recognition task. Verification tasks require that strong expectations be had as to the identity of the voice. One common example of a verification task occurs in the courtroom, where a witness is required to recognize the voice of one specific suspect (Tosi, 1975).

The remainder of this chapter consists of three sections. The first section discusses the nature of speaker variability. The second section reviews the five operational elements of a speaker recognition experiment. The final section analyzes the perceptual bases of speaker recognition.

Sources of Speaker Variability

It is well known that the pronunciation of a given word or phrase tends to vary from speaker to speaker. Acoustical analyses of utterances by several speakers typically reveal many dissimilarities. This effect is called interspeaker (between speakers) variability. It can be

attributed in part to organic differences in the structure of the vocal mechanism and in part to learned differences in the use of the vocal mechanism during speech production (Garvin and Ladefoged, 1963).

Not so well known is the fact that the same speaker rarely utters a given word twice in exactly the same way, even when the utterances are produced in succession. This is called intraspeaker (within speaker) variability. The success of any method of speaker recognition depends on the degree to which the sampled interspeaker variability is greater than the sampled intraspeaker variability.

This section consists of two major topics. The first topic outlines the acoustic correlates of speech sounds. The second topic illustrates some experimental studies which provide direct and indirect evidence of interspeaker and intraspeaker variability.

Acoustic Correlates of Speech Sounds

The acoustic manifestations of basic speech units are presented in the following paragraphs. No attempt is made to provide an exhaustive inventory of these correlations. The objective is instead to highlight the primary sources of speaker variability of interest for speaker recognition research.

The English language employs an inventory of about

forty to forty-five distinctively different speech sounds or phonemes, which may be associated with the acoustical characteristics of speech in many ways. Two of the most important descriptive categories are manner of articulation (e.g., vowel, sibilant, and nasal) and laryngeal action (e.g., voiceless, whispered, and voiced). Each attempt at uttering the same phoneme may result in a slightly different acoustical signal. Such acoustic variations in speech sounds are called allophones.

According to Stevens and Klatt (1974) the principal physiological manipulations that lead to distinctively different acoustic outputs from the larynx are: (1) adjustment of the stiffness or slackness of the vocal folds; (2) manipulation of the static opening to the glottis; and (3) control of external variables such as pressure across the glottis. During normal and shouted speech production the glottal opening is closed, and the associated sound wave type emitted from the glottal source consists of regular quasi-periodic air pulses. In contrast, whispered speech production is characterized by a narrowed glottis and constricted airway with turbulent air movement through the glottis. The vocal folds do not vibrate but are spread apart during whispered speech. It is the driving force of the transglottal air pressure which causes the opening and closing patterns of the vocal folds. This action regulates glottal area changes and associated waveform spectral

characteristics of speech sounds and speaking efforts.

Changes in the intensity of sound (perceived as changes in loudness) are achieved by decreasing the duty cycle of the glottal area pattern. The duty cycle is the percentage of time the glottis is open during each vibratory cycle. As the duty cycle decreases, the steepness of the slopes on resulting glottal volume velocity waves increases (Flanagan, 1958). The increase in steepness results in an increase in the amplitudes of all of the spectral energies in the glottal spectrum. For whispers, the random noise through the glottis permits a spectrum of approximately equal and lower amplitudes for all of the spectral frequencies.

The amplitudes of the various harmonic energies from the glottal source waveform are modified further by the resonant properties of the vocal tract (or vocal-tract transfer function). The peaks that are exhibited in the vocal-tract transfer function are called formants. They correspond to the natural frequencies associated with the shape of the vocal tract. Consequently the ranges over which a speaker's formants can vary depend to a great extent on the size of his head. Because the ranges cannot be altered at will, they can convey only speaker information.

Vowel sounds are produced with a relatively open vocal tract, and they require vibration of the vocal folds as a sound source. The first three formants are sufficient acoustic cues for listeners to identify the vowels of English.

The formant frequency relations that specify the vowels of English are inherently relational rather than absolute since different sized supralaryngeal vocal tracts and vowel-consonant combinations will produce different absolute formant frequencies. Both the linguistic and the socio-linguistic information conveyed by vowels depends largely on the relative positions of the formants. For example, when one considers that a speaker has vowel sounds which are typical of a Scottish speaker (i.e., when one interprets the socio-linguistic information conveyed by vowels), one probably does so by interpreting the relative formant structure of the vowels. On the other hand, the personal or speaker information conveyed by vowels depends partly on the absolute values of the formant frequencies (Ladefoged and Broadbent, 1957). Additional personal information is, of course, present in the relative positions of some of a speaker's vowels, insofar as these are idiosyncratic features of his speech and not aspects which identify him as belonging to a particular group.

Consonant sounds are produced with the vocal tract partially or completely occluded (Flanagan, 1972). The location of the sound source during consonant production may be laryngeal, supralaryngeal, or both. Among those consonants which depend upon vocal tract dynamics for their creation are the stop consonants. To produce these sounds a complete closure is formed at some point in the vocal tract. First

formant "cutback" is an important acoustic cue for delayed phonation onset which differentiates stop consonant sounds like /pa/ from /ba/. This first formant frequency transition rises from the low value that follows from the effective lengthening of the vocal tract by the stop occlusion.

The acoustical characteristics of nasals are basically similar to those of vowels. The primary acoustical speech parameters are the formant frequencies, amplitudes, and bandwidths. The most distinctive property, however, is the presence of antiresonances in the nasals due to the side-branching of the nasal passage from the vocal tract. In the production of an /m/, for example, it shows up as a relatively broad spectrum minimum near 1200 Hz (Fant, 1960). The spectra of nasal consonants (Glenn and Kleiner, 1968) and coarticulated nasal sounds (Su, Li, and Fu, 1974) seem to provide strong cues for the machine matching of speakers. In contrast, a nasal disguise markedly interferes with speaker identification for human listeners (Reich and Duke, 1979).

Many of the acoustic cues that differentiate sounds and speakers are very subtle (Shoup and Pfeifer, 1976). Sources of variability exist with the degree of vocal tract and glottal (opening between the vocal folds) constriction and with the place of articulation for vowels, consonants, and nasals.

Evidence of Inter- and Intraspeaker Variability

Experiments concerned with the recognition of speech by machine provide clear evidence of the presence of inter- and intraspeaker variability. These machines operate by detecting the acoustical correlates of the distinctive features of speech (Stevens, House, and Paul, 1966; Hemdal, 1967). Parker (1977, p.1052) defines acoustic cues as

> properties of the speech signal which relate the acoustic stream to phonological segments. An acoustic cue is justified to the extent that its presence or absence systematically causes a speaker of a given language to perceive or fail to perceive a given phonological segment in that language. Acoustic cues are determinable by observation of the physical stimuli and by observation of the effect that perturbation of that stimuli has on phonological identifications made by speakers.

Distinctive features, on the other hand, "are abstract dimensions in terms of which speakers are thought to organize phonological segments once they have been recognized, and it is in terms of these features that phonological rules are assumed to operate." Some examples of distinctive features are: voicing, duration, and place of articulation. The relation between distinctive features and acoustic cues is that of one to many.

Several factors appear to influence interspeaker variability. In a given experiment, the participation of speakers with certain speech pathologies (e.g., dysphonia), neuropathologies (e.g., cerebral palsy), psychopathologies (e.g., depression), or unique linguistic backgrounds may be

expected to increase interspeaker variability because of the greater organic and learned differences. Similarly, interspeaker variability tends to be reduced for groups of speakers (e.g., identical twins) whose organic and learned differences are small (Alpert, Kurtzberg, Pilot, and Friedhoff, 1963; Kersta, 1965).

The properties and measurement of interspeaker and intraspeaker variability among correlation matrices derived from spectra of continuous speech have been described by Li and Hughes (1974) and Li, Hughes, and House (1969). These studies report two important facts related to talker variability: (1) the long-term average correlation coefficients of speech spectra exhibit intertalker similarities and intratalker differences; and (2) the estimates of these coefficients converge to a set of stable values after about thirty seconds of speech material. These studies suggest data from statistical studies of speech spectra contain information about the physical characteristics of talkers which becomes more and more prominent as effects of context are averaged out. Since formant frequencies are relative measures of vocal-tract size (e.g., a shorter vocal-tract length results in higher resonant frequencies), it is likely, then, that these statistical measures reflect anatomical differences among talkers.

Atkinson (1976) attempted to quantify and compare inter- and intraspeaker variability in fundamental voice

frequency (F_0). For this acoustic speech parameter there is a virtual one-to-one correspondence to the related physiological parameter, vibration of the vocal folds. The fundamental voice frequency associated with a speech sound is independent of the formant frequencies of that sound. The formant frequencies are determined by the configuration of the vocal tract and remain practically constant with a steady-state articulation regardless of a high, low, or changing fundamental frequency. The results of this study show a great deal of variability in F_0 between different speakers and essentially as much variability within a single speaker as there was between several speakers of the same sex. Atkinson suggests this may imply that even if the speech stream can be normalized to a particular speaker, such normalization filters out very little of the variability that can obscure the prosodic features (in other words, the features superimposed upon all of speech like duration, F_0, and average speech power).

Linear prediction analysis enabled Sambur (1975) to evaluate and isolate effective features for the task of speaker identification. A novel probability of error criterion was used to determine the relative merits of the features. The experimental data base was collected over a 3½-year period and afforded the opportunity to investigate the variation over time of the measurements. The measurements that were found to be the most important were

the value of the second resonance (around 1000 Hz) in /n/, the value of the third or fourth resonance (1700-2000 Hz) in /m/, the value of the second, third and fourth formant frequencies in vowels, and the average fundamental frequency of the speakers.

The formant structure of three dipththongs (which generally consist of the formant frequency pattern going from that of one pure vowel to a second vowel or semivowel, as /ai/ in the word bite), four tense vowels (which refer to the muscle tension and physiological force behind the point of closure), and three retroflex sounds (which are defined articulatorily as a maneuver in which the tip of the tongue is curled back and placed posterior to the alveolar ridge of the hard palate, as in /t/) was examined in detail by Goldstein (1976) for possible speaker-identifying features. She found the most effective features in identifying speakers to be somewhat different compared to Sambur's result. It would seem useful to test the more effective features of these two studies using speakers who had not been involved in the original feature evaluations.

Evidence of intraspeaker variability was reported by Hargreaves and Starkweather (1963). They found that more incorrect identifications are made when the utterances represented by test and reference patterns are recorded on different days than when they are recorded on the same days. This observation constitutes evidence of day-to-day variations

in the speech of individual speakers. This finding is corroborated by Li, Dammann, and Chapman (1966).

Several factors have been shown to influence intra-speaker variability. Various diseases of the chest, larynx, and central nervous system, aging, and psychological stress are a few (Mysak, 1959; Hecker, Stevens, von Bismarck, and Williams, 1968).

Endres, Bamback, and Flösser (1971) presented some interesting evidence of intraspeaker variability for formant structure. In this paper an attempt was made to answer two questions: (1) does the formant structure of phonemes uttered by a certain speaker change over a long interval of time, and (2) can the formant structure be changed by disguise, or is it even possible to imitate the formant structure of another speaker? Spectrograms of utterances produced by seven speakers and recorded over periods of up to twenty-nine years showed that the frequency position of formants and pitch of voiced sounds shift to lower frequencies with increasing age of test persons. Speech spectrograms of texts spoken in a normal and a disguised voice revealed strong variations in formant structure. Imitators succeeded in varying the formant structure and fundamental frequency of their voices, but they were not able to adapt these parameters to match or even be similar to those of imitated persons.

This concludes the discussion of sources of speaker

variability. It is clear further study is needed to isolate the features that define and quantify inter- and intraspeaker variability.

Sources of Procedural Variability

The speaker recognition experiment can be conceived as five operational elements. Bricker and Pruzansky (1976) identify these elements to be: speaker ensemble, speech material, transmission system parameters, listeners, and tasks. This section shall discuss these operational elements and use them as a framework for organizing and summarizing the results of diverse experiments.

Speaker Ensemble

The term speaker ensemble refers to "the set of voices that the listener is exposed to during an experiment" (Bricker and Pruzansky, 1976, p.299). Sources of procedural variability exist in homogeneity, size, and characteristics of the population to which the experimenter wishes to generalize.

Level of performance depends on the homogeneity of the speaker group. Homogeneity refers to the perceptual similarity of the voices heard in a given test. In evaluating communication systems with five quartets of speakers, Stuntz (1963) found that some quartets consistently produced

lower scores than other quartets. Williams (1964) divided twelve adult male speakers into two ensembles of six by random assignment and collected recognition data on each ensemble by identical procedures. The final test scores were 62% correct for one ensemble and 50% correct for the other. Stuntz and Williams concluded speaker identification depends not only on the individual characteristics of each of the speakers but also on the characteristics of the other speakers with whom he is being compared. Carbonell, Grignetti, Stevens, Williams, and Woods (1965), and Clarke and Becker (1969) have attempted to identify these individual characteristics and use them to predict confusions among speakers. But this has not met with much success. The discovery of additional characteristics that do predict confusions is obviously a major goal of speaker recognition research.

There is little information indicating how large an ensemble must be to represent adequately the subpopulation of interest. Williams (1971) concluded an ensemble of five or six speakers appeared optimal for testing speaker identification. Pollack, Pickett, and Sumby (1954) used ensembles as large as sixteen without reaching the limit of information transmitted, suggesting that more speakers could have been identified. Hence, it is not yet clear what ensemble size is appropriate.

The most salient characteristics of ensembles are sex, age, and accent. There is experimental evidence

supporting almost perfect performance when listeners are required to recognize talkers cross the adult-child or the male-female barriers (Ptacek and Sander, 1966; Bennett and Weinberg, 1979). Both Ryan and Burk (1972) and Shipp and Hollien (1969) found that listeners could make age discriminations which were even finer than the young-old distinction. Identification of sex alone remains robust even when talkers whisper (Schwartz and Rine, 1968), utter only voiceless fricatives (Ingemann, 1968; Schwartz, 1968), use a monotone artificial larynx (Coleman, 1971) or esophageal speech (Weinberg and Bennett, 1971a), or are sampled from among only five- and six-year olds (Weinberg and Bennett, 1971b).

According to Lehiste and Meltzer (1973) male speakers are identified more easily than women and children, who can be confused with each other. These researchers also show that the confusions are much greater for machine synthesis of speech as compared to human production of speech. They consider formant structure to be a more important cue in speaker identification than fundamental frequency. For example, vowels produced with male formants, but female fundamental frequency, were assigned to a male speaker in 80.8% of instances, while vowels synthesized with female formants, but with male fundamental frequency, were assigned to a male speaker in only 18.6% of the cases.

Controlling Speech Gestures

Experimenters have used three techniques for controlling speech gestures of the speaker ensemble: (1) requiring speakers to vary their sound source in some way; (2) asking some speakers to mimic others; and (3) selecting the text to be spoken.

Source Manipulation

McGehee (1937) studied the importance of pitch by having talkers alter their normal pitch in some manner. Results indicate that when the target speaker disguised his voice by changing the pitch, identification was reduced by about 13%. Listener's identification performance for Polish vowels, spoken at normal, higher than normal, and lower than normal pitch was not always best for the normal pitch condition (Dukiewicz, 1970). However, Matsumoto, Hiki, Sone, and Nimura (1973) found that a listener's speaker discrimination performance was much better when the pitches of two samples were the same than when they differed by 40 Hz.

Manifestations of task-induced stress in the acoustic speech signal was studied by Hecker et al. (1968). For each of ten subjects, numerous verbal responses were obtained while the subject was under stress and while he was relaxed. Listeners could identify the stressful responses of some subjects with better than 90% accuracy and of others only at

chance level. Spectrograms of the test phrases showed that task-induced stress can produce a number of characteristic changes in the acoustic signal, most attributable to modifications in the amplitude, frequency, and detailed waveform of the glottal pulses.

Reich and Duke (1979) investigated the effects of selected vocal disguises upon speaker identification by listening for comparison with previous research using a spectrographic speaker-identification method. Speakers produced sentences in six different ways. They included simulating a hoarse voice quality, a hypernasal voice quality, an extremely slow rate of speaking, and a person "70 to 80 years of age." In addition to recording a natural voice utterance, each speaker was asked to disguise his voice in a manner which he felt would most effectively conceal his identity. In general the results indicate certain vocal disguises (e.g., slow-rate and nasal) do interfere with speaker identification accuracy for humans, and that some disguises (e.g., nasal voice quality) are not equally effective for spectrograph and computer methods. These researchers suggest this variability is due to listener uncertainty as to which acoustic cues are necessary for accurate speaker identification.

Several investigators have studied the importance of voicing by comparing recognition results from whispered speech with results from normal speech. Williams (1964)

found that recognition for normally spoken sentences was much better than for whispered speech. Pollack et al. (1954) studied the effects of whispered speech as a function of duration of speech sample. They concluded that a whispered speech sample had to be over three times the duration of normal speech for equivalent levels of speaker identification.

An artificial larynx with a fixed frequency of 85 Hz as the sound source was used by Coleman (1973) for eliminating possible glottal source variations. He reported speaker pairs were identified as being composed of the same or different speakers with better than 90% accuracy when the normal interspeaker differences in the glottal source characteristics were equalized. The errors for his twenty-eight listeners were concentrated in relatively few speaker pairs. Male speakers as a group retained a significantly greater degree of identifiability for these experimental conditions than did females. Coleman concluded that there was a lot of speaker identity information left when source variation was eliminated, and that female speakers might be more successful in disguising their voices in this way than males.

Mimicry

There are practical and theoretical interests for studying the effects of mimicry on speaker recognition. On the practical level, any automatic speaker recognition system

that uses the voice as a means of verifying the identity of an individual will have to contend with the problem of impostors. On a more basic level, results of studies using mimics may reveal some information about the relative importance of organic versus learned factors in listener's recognition of speakers.

Carbonell, Grignetti, Stevens, and Woods (1965) examined the effects of mimicking on the recognition of unfamiliar speakers. The mimicking was done by two of the investigators, who also served as critical listeners in comparing their various versions of impostor utterances with the prototypes. Fewer mimicked utterances were falsely accepted than other foreign utterances, suggesting that impostors are not particularly successful in deceiving listeners. This conclusion can only be regarded as tentative though. It could have been that the investigators were not particularly good mimickers.

Rosenberg (1973) conducted two speaker verification studies in order to compare listener performance with performance of an automatic speaker verification system. In the first experiment, where customers (people who wanted to be verified) were paired only with casual impostors (people speaking in their natural voices), both the false acceptance rate and the false rejection rate were about 4%.

The second experiment involved professional mimics. When a customer was compared with an impostor using his

natural voice, the false acceptance rate was about 4%, as it had been in the first experiment. When a customer was compared with a mimic impostor, the mean false acceptance rate was about 20%.

Selection of Text

Variations in the anatomy of the vocal mechanism of different talkers may be reflected in isolated utterances of different speech sounds. Some experimenters have considered the importance of anatomical features and their corresponding speech signal parameters by comparing talker recognition performance for different isolated phonemes.

Russian phonemes were studied by Ramishvili (1966). He found identification performance to be better than chance for all phonemes; in particular, identification was best for vowels, with the exception of the high back vowel /u/, and voiced consonant identification was better than unvoiced. The range of mean identification scores was 90% for /e/ to 30% for /k/. Dukiewicz (1970) also concluded that recognition was poorest for the back vowel /u/, which has a low second formant.

Recognition for words containing a front stressed vowel was found to be slightly better than for words with back stressed vowels (Stevens, Williams, Carbonell, and Woods, 1968). It was argued that better recognition for

words with front vowels might be due to the higher second formant, and that more speaker identification information is contained in the higher frequencies. Bricker and Pruzansky (1966) reported identification accuracy improved directly with the number of phonemes in the sample even when duration was controlled. There was slightly better familiar speaker naming performance for /a/ than for the high front vowel /i/ excerpted from the stressed vowel of a disyllable.

In general, the information conveyed by vowels is more important for speaker recognition than the information conveyed by other speech sounds.

Speech Material

The second essential operational element and major source of procedural variability of any speaker recognition experiment is that one or more voice samples be presented on every trial.

The effect of the duration of the sample on performance was studied by Pollack et al. (1954) in one of the earliest experiments on speaker recognition. The task was the identification of familiar voices with the sample excerpted from connected speech. Performance was found to improve rapidly up to approximately one second, and more slowly thereafter. Later studies (Bricker and Pruzansky, 1966; Murry and Cort, 1971) established that performance

depends on the size of the sample of the talker's repertoire rather than on the duration per se.

In addition to duration and size of the sample of the talker's repertoire, mode of presentation has been studied. Williamson (1961) compared the fixed-sequence and free-comparison modes of presentation for test items involving only two speech samples. In the fixed-sequence mode, the speech samples to be compared by the listener are arranged in a fixed temporal order that is determined entirely by the experimenter. In contrast, the free-comparison mode allows the listener to control the presentation of speech samples, with each test item consisting of a limited time interval. Williamson found the fixed-sequence mode to produce considerably higher scores than the free comparison mode. However, it cannot be concluded that the fixed-sequence mode is generally superior to the free-comparison mode. Short-term memory effects make the fixed-sequence mode appear less attractive when the number of speech samples per test item is large. Such memory effects are not encountered with the free-comparison mode.

Doehring and Ross (1972) assessed voice recognition by a matching to sample procedure in thirty right-handed adults with normal hearing. In this experiment the subject was required to indicate which of three voices speaking a nonsense syllable matched the speaker of a sample vowel. There was a significant difference as a function of the

temporal position of the matching voice, with recognition being most accurate when the matching voice was first and least accurate when it was third. This finding is most likely indicative of a short-term memory process associated with the fixed-sequence mode of presentation.

Transmission System

The transmission system has posed no procedural problem in itself. However, sources of procedural variability exist in controlling the signal.

Controlling the Acoustic Signal

Several experimenters have been concerned with manipulating the acoustic signal by varying parameters of the transmission system. They have attempted to determine what features of the speech signal carry information relevant to the identity of the speaker. The speech signal has been altered by: (1) selective filtering; (2) playing the speech backward; and (3) analysis and synthesis techniques.

Filtering

Of considerable interest has been the question of what portions of the broad-band frequency spectrum contribute most to speaker identifiability. The effects of low-pass,

high-pass, and band-pass filtering for isolated vowels, words, and sentences have been examined as means for answering this question.

Results of filtering studies seem to support these conclusions: (1) removal of spectral energy below 500 Hz and above 3000 Hz has no significant influence upon speaker recognition accuracy (Pollack, Pickett, and Sumby, 1954; Peters, 1954), (2) low-pass filtering frequencies below 500 Hz greatly reduces speaker identifiability, yet still allows listeners to assess personality factors related to hypertension (Skalbeck, 1955; Starkweather, 1956), (3) speaker identification scores are best when the octave band containing the fundamental frequency is emphasized rather than other octaves (Peters, 1956), and (4) frequency ranges in which a vowel's spectral energy are concentrated carry more speaker information than do ranges of low energy (Compton, 1963; Dukiewicz, 1970).

Backward Speech

Several studies have employed backward-played speech as a form of distortion. Playing speech backward does not affect the frequency spectrum of the speech at any one point in time, but it disturbs the temporal patterns of speech, such as directions of formant transitions and patterns of fundamental frequency. Different temporal patterns of speech

are considered to reflect learned differences of speakers, while differences in frequency spectrum are considered to reflect organic differences.

The ability of listeners to identify familiar speakers is appreciably impaired when speech is played backward. Clarke, Becker, and Nixon (1966) played sentences backward both with and without filtering. On the average, performance was slightly better for 250 Hz high-pass backward speech than for unfiltered backward speech. Low passing the backward speech at 500 or 1000 Hz showed a slight decrement in performance. Performance of Williams' (1964) subjects was better for backward sentences than for whispered sentences played forward, but poorer than for undistorted sentences. Bricker and Pruzansky (1966) compared results for speech played backward with recognition for undistorted speech. They used speech materials of varying length and composition. Reduction in recognition of the backward material was greatest for monosyllables and least for sentences. There was an approximate 10% reduction in recognition for vowel excerpts played backward. This finding suggests that there is some speaker information in a 100 msec vowel excerpt.

It appears that temporal cues are important for speaker recognition. However, Hecker (1971, p.42) points out that "it is not clear whether listeners use temporal clues per se or whether their judgments depend on a perceptually realistic speech signal."

Analysis and Synthesis Techniques

With the development of speech processing devices such as vocoders and sophisticated computer techniques, the speech signal can be subjected to detailed acoustical analyses in order to explore its speaker-dependent parameters. Several studies have attempted to discover the relative contribution of the vocal-tract function (also referred to as the articulatory characteristics) and the glottal-source characteristics to speaker identification. However, a definitive answer for the tract-source issue is not available.

An early attempt to manipulate formant frequencies by use of a vocoder is reported by Shearme and Holmes (1959). These investigators used a channel vocoder to produce a unique form of spectral distortion that was designed to obscure a primary speaker-dependent effect of articulation, namely the relative spacing of the formant frequencies. Several seconds of speech were processed through the vocoder that shifted the spectral envelope in the frequency range of the first formant upward by 100 Hz, and the second and third formant regions by 300 Hz. Laryngeal frequency characteristics were eliminated by using a fixed voicing frequency. Treated and untreated speech samples were presented to listeners in a discrimination test. When only one voice of the pair was treated by frequency shifting, the false acceptance rate for different-speaker pairs was only about

7%, but the false rejection rate for same-speaker pairs was about 90%. When both voices were treated, the false acceptances went up to almost 50%; no same-speaker pairs with both samples treated were presented. The nature of the experimental design makes it difficult to determine how much the difference in listeners' responses for the treated and untreated conditions was due to the shift of frequencies and how much to shifts in criterion. Nonetheless, this study presents evidence that the characteristics of the spectral envelope of speech play a part in recognition of a speaker's voice, and that the important characteristic is connected with formant positions.

Miller (1964) conducted a more direct investigation of the relative contribution of articulatory and glottal-source characteristics employing a computerized analysis-synthesis technique called inverse filtering. The computer synthesized speech signals which might be encountered if it were physiologically possible to interchange either vocal tracts or the larynges of two speakers. In other words, the vocal-tract transfer function and the glottal-source waveform were separated. The speech was recombined in several intriguing ways. Examples include one case in which the tract information for two speakers, who spoke the nonsense word /hod/ at about the same pitch and duration, were interchanged. Listeners judged the same-tract samples to sound more like the original utterances than did the same-source

samples. Tract data for six speakers was combined with two artificial glottal pulse shapes to discover the importance of the glottal source. Listeners reported that, although the effects of the glottal wave were noticeable, the voices still sounded like those of their tract donors. Two other resynthesis procedures also suggested that the synthesized voice sounded like the speaker whose tract was represented. However, Miller (1964), Hecker (1971), and Bricker and Pruzansky (1976) point out that these results were influenced by factors inherent to the technique of inverse filtering.

Matsumoto, Hiki, Sone, and Nimura (1973) also used inverse filtering to estimate glottal waveform and formant frequency patterns among voice samples with the same fundamental pitch frequency. It was concluded from their results that the magnitude of the contribution of the vocal tract characteristics (the deviation of the formant frequencies) to the perceptual difference of personal quality varied widely according to the kind of vowel, and that this was not the case with the glottal source characteristics (the slope of the glottal source spectrum and the fluctuation of the fundamental pitch). Only in the case of the Japanese vowel /a/ was the contribution of the vocal tract characteristics larger than that of the glottal source characteristics.

Evaluating the Transmission System

During the design, development, and testing of speech

handling or processing systems, there is a need for evaluation and for optimization criteria. Speech quality procedures can include many factors: intelligibility (main evaluation criterion in the past), preference, loudness, timbre and rhythmic character, systematic amplitude or time distortion, clarity and naturalness, and speaker recognizability. Only some of these factors can be measured objectively (Hecker and Guttman, 1967; Rothauser, Urbanek, and Pachl, 1968; Nakatani and Dukes, 1973). Among the tests which have been used to evaluate systems with respect to speaker recognizability are the modified speaker-naming test (Stevens, Hecker, and Kryter, 1962; Stuntz, 1963) and the voice-attribute rating test (Voiers, Cohen, and Mickunas, 1965). Some of these studies provide opportunity for examining the relationship between intelligibility and speaker recognizability.

Hecker and Williams (1965) were concerned with the inter-relation among results of paired-comparison preference tests, speaker-identification tests, phonetically balanced (PB) word-intelligibility tests, and nonsense-syllable articulation tests. Five simulated communication systems, representing different types and degrees of distortion, were used to process the recorded speech materials. Two systems were found to be equivalent in speech quality, and two systems were found to be equivalent in intelligibility. But four systems were equivalent in speaker identifiability. Only the most severe form of speech signal degradation

(peak-clipping followed by bandpass filtering) produced a statistically significant reduction in the scores obtained with the modified speaker-naming test. Clarke, Becker, and Nixon (1966) corroborated this finding.

The effects of three speech transmission systems on verification accuracy by human listeners were demonstrated by McGonegal, Rabiner, and McDermott (1978). This effect is potentially an important one for digital communications problems over telephone lines where the transmission system could vary from one which gives a high quality coded representation of the signal to a low-bit-rate vocoder. This study showed the false alarm rate (i.e., a customer is rejected) to be significantly higher when the test and reference utterances were transmitted by different systems than when transmitted by the same system. The miss rate (i.e., an impostor is accepted) was not significantly different for similar comparisons except for one of the conditions. The overall conclusion of these researchers was that speaker verification by human listeners cannot be performed as accurately over mixed speech transmission systems as over the same transmission system. One may also surmise that when analysis-synthesis systems are evaluated, speaker recognition scores can be more sensitive indicators of system performance than intelligibility scores.

The use of voice-attribute ratings represents another approach to the problem of evaluating communication

systems. Voiers, Cohen, and Mickunas (1965) explored the feasibility of measuring how well a given communication system preserves the perceptual parameters involved in speaker recognition. Performance standards were provided by ratings of unprocessed speech samples on ten semantic-differential scales. Analyses of these ratings permitted the identification of five perceptual factors, each of which was represented by two scales: pitch-magnitude, loudness-roughness, animation-rate, clarity-beauty, and normality. Various vocoder systems were evaluated in terms of the degree to which each of these factors was transmitted. The results showed that some factors did not appear to be sufficiently affected by the vocoders to warrant their inclusion in similar evaluation programs.

McDermott, Scagliola, and Goodman (1978) conducted an experiment to study the perceptual characteristics of speech processed by adaptive differential PCM (pulse code modulation). They created eighteen three-bit and four-bit coders spanning a wide range of quantizer adaptation parameters. Subjects judged differences between coders and rated the quality of each coder individually. The difference data revealed three important perceptual characteristics: overall clarity, signal versus background degradation, and rough versus smooth impairment. These characteristics were strongly correlated with coder design parameters and objective performance measures.

Listeners

Sources of procedural variability for the listener element in a speaker recognition study are size and training. There may be other variables of importance though because systematic investigations of listener differences are missing.

Williams (1964, 1971) suggests that a group size of twelve is large enough for achieving stable average performance. This recommendation was based upon data received from three groups of twelve listeners each who performed identifications equally well using three types of material.

Training can affect differences between listener groups. For example, Clarke and Becker (1966) report that their naive listeners achieved 58% correct recognition in a four-alternative forced-choice task, whereas five speech science students who had worked extensively with the ensemble of twenty speakers scored 67%. The interlistener error variability ranged from about 8% to 26%.

Tasks

The final operational element and source of procedural variability for speaker recognition researchers deals with the task chosen by the experimenter. According to Brown (1979) listener tasks may be grouped into two basic classes: designation and evaluation. As each is discussed, associated performance measurements will also be noted.

Designation

Designation tasks always involve a test stimulus--the voice sample to be classified--and one or more response categories defined by other voice-sample stimuli. There are two types of designation task: short-term memory and long-term memory. Short-term memory tasks "involve the stimulus voice sample being presented after a delay of at most a few minutes from the presentation of the reference voice sample" (Brown, 1979, p.732). In long-term memory tasks, "the stimulus voice sample is presented after a longer delay. In this case, the reference voice pattern will be stored in long-term memory either through rehearsal or because the reference voice pattern was already in long-term memory before the experimental session" (Brown, 1979, p.732).

Short-term memory tasks may involve simultaneous or sequential presentation. Examples of sequential short-term memory tasks are Coleman (1973), Doehring and Ross (1972), and Shearme and Holmes (1959). There are few, if any examples of simultaneous short-term memory tasks in the literature.

When only one comparison stimulus is presented, the listener's task is simply to decide whether the test stimulus was or was not uttered by the same talker. This basic structure is common to procedures called same-different testing, discrimination, and verification (Rosenberg, 1973; McGonegal, Rabiner, and McDermott, 1978). Since this task

allows a no-match response, the question of the listener's criterion for matching must be considered. Clarke and Becker (1969) and Matusmoto et al. (1973) required subjects to estimate their confidence in each same-different judgment and then used the confidence ratings to construct a relative operating characteristic (ROC) curve. A criterion-free measure of performance was derived from the empirical relative operating characteristics. Another measure called the similarity index can be derived for each pair of speakers in a same-different experiment by finding the mean confidence rating with which two different talkers were judged to be the same. Matsumoto et al. (1973) used similarity indices as input to a multidimensional scaling procedure in order to construct perceptual spaces for their speaker ensembles.

Long-term memory tasks may be subdivided into familiar and familiarized reference voice patterns. In tasks involving familiar reference voice patterns, these patterns have been acquired through social or business contact. Examples of experiments using the familiar reference voice pattern are Bricker and Pruzansky (1966), Compton (1963), Dukiewicz (1970), and Pollack et al. (1954). Familiarized reference voice patterns include patterns which have been presented to the listener at a much earlier stage in the experimental session, so that they have been stored in long-term memory through rehearsal. Williams (1964) and Stevens et al. (1968) are examples of experiments that involved

familiarization of reference voices.

The performance measure most commonly used with familiar and familiarized voice patterns is P(c), the percentage of responses correct. However, Pollack et al. (1954) preferred to estimate the amount of information transmitted, designated T, from the square stimulus-response contingency matrices. They were able to show that the relative information transmitted, which is defined as T/T max, was independent of response set size, at least over a limited range. Other investigators (Bricker and Pruzansky, 1966) have reported the entire stimulus-response (or confusion) matrix.

Evaluation

The term evaluation is applied to tasks that require the listener to judge the value of the stimulus--voice on some attribute, dimension, or characteristic. Voice samples may be presented in two manners: rating and comparison. Rating refers to evaluation tasks involving a single voice sample at a time and one or more attribute scales. Accordingly, comparison refers to evaluations of two or more samples (e.g., pair-wise similarity judgments). Presently the literature contains no examples of comparative evaluation.

Several investigators have used the rating task to study listener perceptions of a wide variety of charac-

teristics, from the objectively verifiable through the purely subjective. Among the former are sex (e.g., Schwartz, 1968) and age (e.g., Shipp and Hollien, 1969). Intermediate between objective and subjective characteristics are psychophysical scales. Singh and Murry (1978) and Clarke and Becker (1969) reported the performance of their listeners expressed in terms of the degree of relation between the ratings and the objective measures thought to be their chief physical correlates. In addition, a speech signal processing system was studied via psychophysical scales by McDermott, Scagliola, and Goodman (1978). A subjective procedure, the semantic differential technique, has been used by Voiers (1964, 1965), Holmgren (1967), and Clarke and Becker (1969). Listeners are required here to rate voice samples on a number of bipolar adjective scales such as beautiful-ugly or sharp-dull. The procedure produces data that can be factor-analyzed in various ways, some of which can be mapped onto a perceptual space.

Perceptual Bases of Speaker Recognition

The first of several inquiries into the perceptual bases of speaker recognition occurred in 1937 and 1944 by McGehee. These early studies attempted to determine why some voices could be identified more readily than others and why certain voices tended to be confused by listeners.

Underlying most studies of this kind is the assumption that a listener makes use of only a small number of perceptual parameters in discriminating between voices and in identifying familiar speakers. Because the listener is not conscious of this, he cannot be asked directly about the nature of these perceptual parameters. In order to explore the parameters, it is necessary to use indirect judgments and relatively complex analytical procedures. If the hypothetical parameters can be adequately defined and measured, they can provide a unique and compact description of a particular voice.

Voiers (1964, 1965) and Holmgren (1963, 1967) asked listeners to rate voices on several semantic-differential scales. In addition, several physical measures in the speech of the same speakers were obtained. These physical measures included the mean and variance of the amplitude of voiced speech sounds, the amplitude of voiceless speech sounds, and the fundamental frequency. Each investigator found four factors sufficient to account for most of the variability of the speaker effect. Holmgren gave these factors the names intensity, quality, pitch, and rate. Only a few of the scales were found to be correlated with expected physical measures. Even though difficulties of interpretation were encountered in these studies according to Bricker and Pruzansky (1976), these studies nevertheless revealed some evidence for the perceptual significance of glottal fundamental frequency, or

pitch, as well as a perceptual factor related to speech intensity or effort.

Clarke and Becker (1969) used various correlation techniques to study the relations between physical measures and three kinds of data taken from human listeners: ratings of selected attributes, ratings on semantic differential scales, and performance on a matching task. Their principal interest was in determining how well matching performance could be predicted from perceptual data or physical measurements. Their basic physical measures were mean glottal period, period variability, long-term power spectrum, and utterance duration. Six psychophysical scales were chosen to characterize speaker individuality. These were pitch, pitch variability, sibilant intensity, click-like elements, breathiness, and rate. Only two of these, however, were reported to have substantial correlations with physical measures. These were pitch with mean period and rate with duration. This last finding differs from the results of Voiers (1965) and Holmgren (1967); neither of whom found a strong relation between the semantic differential factor related to rate and a physical measure of duration. Clarke and Becker (1969) report both physical measurements and semantic differential data predict matching task performance to a statistically significant, but practically unimpressive, degree.

A study of the relation between acoustical parameters derived from analysis-by-synthesis and same-different

performance has been reported by Matsumoto et al. (1973). Their method involved constructing a psychological auditory space by applying multidimensional scaling to similarity indices derived from the matching task. An interpretable rotation was achieved by orienting the axes along directions taken by certain physical parameters of the speech samples. Three experiments using this technique were conducted to assess the relative contribution of tract and source parameters to the individuality of voices. The glottal source measures were glottal fundamental mean frequency, variance of the same, and slope of the glottal spectrum. Vocal tract parameters were the first three formant center frequencies. They interpret their findings to be in support of several conclusions: (1) mean glottal fundamental frequency makes a greater contribution to perceived voice individuality than any other parameter; (2) the other source parameters (spectral slope and pitch period variability) make a contribution, regardless of the vowel spoken and independent of tract characteristics, and (3) formant frequencies make a contribution varying in magnitude depending on the vowel spoken. However, Bricker and Pruzansky (1976) suggest that these results are not conclusive for several reasons. For example, only the five Japanese vowels, uttered in isolation at a pitch dictated by the experimenter were used. Also, the conclusion about mean fundamental, while consistent with that of other experimenters, is based on an experiment in

which the parameter was manipulated deliberately over a range of 40 Hz. Still, this study does suggest a promising technique for speaker recognition research.

Singh and Murry (1978) report a study that investigated and acoustically defined some of the perceptual parameters used to distinguish among normal male and female voices. For ten male and ten female speakers, eight acoustical measures were obtained, and psychophysical ratings of four commonly used descriptive terms were made by ten speech pathologists and ten women listeners. The acoustical measures for each speaker were: fundamental frequency (F_0), formant frequencies one (F_1) and two (F_2), the ratio of F_2 to F_1, the difference between F_2 and F_1, the total duration of the vocalic portion of the speech sample, and the total number of pitch shifts during the speech sample. The psychophysical method of equal appearing intervals was used to obtain judgments of similarity between normal male and female voices for pitch, nasality, hoarseness, and breathiness. Then a multidimensional analysis technique, INDSCAL, was used to analyze the acoustical measurements and ratings. The results of both the INDSCAL analysis and correlational analyses suggested that the critical perceptual attribute of voice is related to the differences between the parameters (both acoustic and perceptual) of male and female voices. In other words, the most important factor in these listeners' strategies for discriminating voices was the concept of male

versus female. Singh and Murry suggested two possible implications for this finding. First, the perceptual strategies of the judges reflected cultural stereotypes of the male voice versus the female voice. And second, different voice production strategies were used by male and female speakers. For example, the number of pitch shifts significantly predicted breathiness for females but not for males. As the subjects shift pitch direction, complete vocal fold approximation is less probable. According to Singh and Murry, this may be a learned behavior to generate the breathy "sexy" female voice. Also, F_0 was used by male speakers to produce the phenomenon of hoarseness, an aspect of voice more culturally acceptable for males than for females.

One other attempt has been made to relate dimensions of voice perception to acoustic parameters of voice production. In this study a four-dimensional INDSCAL analysis of similarity ratings was employed to derive psychological dimensions of talker similarity (Walden, Montgomery, Gibeily, Prosek, and Schwartz, 1978). Correlations between the thirteen acoustic measurements and the four INDSCAL dimensions revealed that fundamental frequency and word duration were moderately correlated with two of the psychological dimensions. The other two dimensions were not convincingly correlated with any of the acoustic measurements, but were best described as representing voice quality and talker age. The emergence of word duration and mean fundamental frequency as dimensions

was consistent with the findings of previous investigations of voice perception and speaker recognition (Voiers, 1964; Holmgren, 1967; Clarke and Becker, 1969; Wolf, 1972; and Matsumoto et al., 1973). However, the finding that talker age is an influential dimension was unexpected and previously unidentified. These authors suggested it was probable that the listeners were making their similarity judgments along some other dimension of voice perception which was highly correlated with age.

Haggard and Summerfield (1979) present evidence suggesting that listeners extract and store parameters characterizing the style of a speaker rather than matching a raw sound image. This evidence is based on data that demonstrates improvement in performance can result from increasing the length of either the claimant utterance or the stored sample even when the other cannot be increased. This observation, together with the results of the previous two perceptual studies, indicates the human perception process to differ from that of the machine. In other words, such perceptual dimensions as sex, talker age, and sentence duration may indicate the importance of visual information for talker recognition. Speech research may indeed reach a point where a small set of highly valid descriptive parameters will enable the superior storage and processing power of digital computers to return a performance better than the analogous human performance.

Qualitative contrasts between machine and human performance may illuminate the major discrepancies in the type of process each system uses for talker recognition. The next chapter outlines the progress for machine verification of speakers.

CHAPTER III

SPEAKER RECOGNITION SYSTEM: MACHINE

A typical speaker recognition by machine (SRM) study involves having speakers read some selected speech materials and then transmitting the speech wave or speech spectrum to an analog-to-digital converter which transforms the speech to numbers usable by a computer. According to Bricker and Pruzansky (1976), the major effort of SRM research has been aimed at developing computer procedures for (1) defining and extracting a set of acoustic parameters that might carry speaker information; (2) reducing these parameters to a set of features used in the decision process of recognizing speakers; and (3) developing a decision rule that is tailored to the set of features and the task being studied.

Machine recognition tasks can be divided into two types: speaker verification and speaker identification. For both tasks the machine has stored reference features of many speakers. In the speaker verification task, a cooperative speaker saying a test utterance claims to be Person X.

The machine's decision is either to accept or reject that speaker as Person X. In a speaker identification task, the speaker of the test sample does not claim an identity. The machine must decide the identity of the speaker from the ensemble of possible speakers. The verification task is judged not only more tractable but also of more practical interest. It consequently has received more emphasis in terms of a system design (Atal, 1976; Flanagan, 1976; Martin, 1976; and Rosenberg, 1976).

Machine experiments have identified several acoustic parameters that carry speaker information, such as the formant frequencies and their bandwidths, the fundamental voice frequency, the amplitude of the fundamental frequency, and vocal tract shapes (Wakita, 1976). This information can be useful for testing models of the human perceptual and decision processes. The remainder of this chapter will briefly discuss research problems and major contributions of this field. First, some of the data collection considerations prior to digitalization are presented. Then computer procedures are described. Finally, the process of speaker recognition by listening and machine is compared.

Data Collection

Three stages leading to the recognition response can be delineated. They are the speaker ensemble, speech material, and transmission system.

Speaker Ensemble

Ideally, the data base for speaker recognition studies should include a large number of speakers. Collection of speech data from a large number of speakers poses many practical problems. Consequently, most studies involve small populations of ten to thirty speakers. But a few studies have tested speech data from one hundred or more speakers. These are the Texas Instruments entry control system which has made over 150,000 verifications over a period of a year on a population of 180 users, and the Bell Labs telephone system which made over 4500 verifications over a five-month period on a population of approximately 100 users (Rosenberg, 1976). Other large test populations have been reported by Bricker, Gnanadesikan, Mathews, Pruzansky, Tukey, Wachter, and Warner (1971), Das and Mohn (1971), and Hair and Rekieta (1972). All show promising results with high recognition rates.

The composition and characteristics of the speaker population are other important considerations. Most evaluations have included only male talkers of similar age and regional dialect. In the few studies that contain both males and females, identification was based on spectral data with no pitch information, and confusions did cross sex boundaries (Bricker et al., 1971; Pruzansky, 1963; Pruzansky and Mathews, 1964). Relevant temporary and chronic speech irregularities

have not been considered in recognition experiments to date.

Some studies have examined how well a system can resist the effects of determined mimics. Lummis and Rosenberg (1971) and Doddington (1974) employed professional mimics to provide imitations of their speakers. Both studies found mimic acceptance significantly greater than acceptance of causal impostors. In contrast, Hair and Rekieta (1972) observed some increase in similarity for individual features in their mimic verification scheme, but the attempt at acceptance when all features were combined was unsuccessful. According to Rosenberg (1976) the evaluation of a system's resistance to mimics depends on the definition of a skilled mimic. Those who depend on caricature for their imitations would be poor candidates for systems which make strongly physiologically correlated measurements rather than measurements correlated with behavior of learned characteristics. In considering a related mimicry problem, systems whose measurements are physiologically oriented may not be able to discriminate among close family members, especially identical twins. More study of mimicry and identical twins is needed.

Speech Material

In addition to the desirability of testing a large population of speakers there are other considerations to be examined with regard to the experimental data base.

Researchers have been concerned about the amount of time over which the reference samples are collected, the time elapsed between the last reference sample and the test sample, and the minimum number of samples needed to form a representative reference pattern. For this discussion a reference pattern is considered "a set of averaged or weighted features that are determined from several reference samples of a given speaker" (Bricker and Pruzansky, 1976, p.315).

Many studies show that utterances spoken during one recording session tend to be similar and do not represent variations of speech patterns that may be expected over a period of time (Das and Mohn, 1971; Furui, Itakura, and Saito, 1972; Hargreaves and Starkweather, 1963; Luck, 1969; and Sambur, 1975). The problem of long-term variation has been examined in some detail in the investigations of Furui et al. (1972, 1973, 1974, 1975). These researchers were able to examine the effects of intervals of up to 18 months between samples. The best identification and verification rates were obtained when the reference data was calculated from four samples taken at intervals of three months. The data base used by Sambur also extended over a long period--up to 3½-years. The only feature reported by Sambur to be strongly affected by this interval was average pitch.

Fewer utterances (generally five or six) have been used in reference patterns for speaker identification experiments than for speaker verification studies, one of which

used as many as 125 utterances collected over eight days (Luck, 1969). However, very high verification has been achieved using only ten reference utterances collected over a two month period (Doddington, 1970; Lummis, 1973).

Most SRM experiments have required that the text of the reference and test utterances be the same so that corresponding speech events can be compared. Speech sounds, whole words, phrases, and sentences have been used. Recently there has been an increasing interest in computer-based techniques for text-independent speaker recognition. The term "text-independent" has been used in several different contexts. For example, Atal (1974) has used the term in the sense of choosing independent randomized test frames from a single sentence to use against the remaining frames as a reference set. Sambur (1976) has used the term in an experiment in which the sentences in the test set were different from those in the reference set, even though each speaker read precisely the same list of sentences. Although useful insight has been gained by these approaches, they are linguistically constrained. Markel and Davis (1979) report findings suggesting that if the pertinent parameters are averaged over sufficiently long intervals of time, such as 30 s or more, the features obtained are essentially free of linguistic constraint, showing speaker recognition performance comparable with some text-dependent speaker recognition experiments. A very large data base consisting of over thirty-six hours of unconstrained extemporaneous

speech from seven speakers recorded over a period of more than three months was analyzed to determine these important results. In many practical situations, a text-independent speaker identification system could overcome problems which may arise if the speaker is uncooperative, and obviously there is a great interest for communication over channels which have no linguistic constraints.

Furui et al. (1972) were concerned about the length of speech required to keep the phonetic sequence from influencing the long-term average spectrum. They concluded that the sample length necessary for determining average spectrum patterns is of the order of ten seconds. Atal (1972) described a procedure that resulted in high identification accuracy for speech of two seconds in duration, even though the texts of the test and reference were different.

Transmission System

The recording environment and the conditions governing the transmission of the speech signal to the processor are factors which may determine the ultimate success of a speaker-recognition system. Two problems of interest are multi-talker environments and background noise.

Parsons and Weiss (1975) and Parsons (1976) report progress towards enhancing intelligibility of speech in environments where the interference is the speech of another

talker. Specifically, the Fourier transform of the speech signal is dissected into components belonging to each talker. The sound of the desired voice is then reconstructed from its components as identified by the dissection. It is not clear yet, however, whether this separation can be done automatically by machine.

Preliminary data indicates that the addition of white noise distorts the acoustical manifestations of speaker and speech identity. Doddington and Hydrick (1975) and Sambur and Jayant (1976) advise signal-to-noise ratios in excess of 20 dB and 17.5 dB, respectively, for acceptable speaker and speech recognition performance. Speech material (sentences) and speech parameter measurement differed for these two studies. Neely and Reddy (1971) investigated speech recognition in the presence of noise. Three types of noise--teletype idling, teletype typing, and machine room (fans and air conditioners)--were recorded using an omnidirectional microphone. Each kind of noise was mixed with utterances at two different signal-to-noise ratios: 15 and 25. Recognition results indicate very poor scores (below 43%) for all three types of noise at 15 dB. For the 25 dB condition, accuracy improved somewhat. That is, recognition accuracy was 67% for idling noise, 43% for typing noise, and 54% for machine room noise.

A stand-alone noise suppression algorithm has been presented for reducing the spectral effects of acoustically

added noise in speech (Boll, 1979). This method suppresses stationary noise from speech by subtracting the spectral noise bias calculated during nonspeech activity. It appears to be a promising technique for speaker recognition systems.

In an earlier paper (McGonegal, Rabiner, and McDermott, 1978), the effects of several transmission systems on speaker verification by human listeners were reported. It was shown that the transmission system played a significant role in the speaker verification process. A later study by McGonegal, Rosenberg, and Rabiner (1979) shows the effects of the transmission system on an existing automatic speaker verification system in which the measured features were pitch and gain as a function of time for a specified utterance. In this experiment, there were ten male and ten female customers and forty male and forty female impostors. Fifty utterances were recorded using a conventional telephone connection over a period of two months. All utterances were post-processed by an ADPCM coding system and LPC vocoding system. When the reference and test utterances were subjected to different transmission systems, no significant difference in the verification accuracy of this automatic system was found. This result may verify that pitch and gain are robust features for use in a speaker verification system.

Computer Procedures

The machine analysis of speaker recognition may be considered consisting of three steps: (1) defining and extracting acoustic parameters, (2) data reduction and feature evaluation, and (3) choosing decision rules.

Defining and Extracting Acoustic Parameters

Acoustic parameters that have been considered in SRM studies have been divided by Bricker and Pruzansky (1976) into three groups: (1) time-varying measures made on prescribed text material; (2) parameters measured from specific speech events; and (3) measures of long-time average spectra.

The simplest measures of the first type that have been investigated are time-by-frequency matrices of spectral energies for words and phrases (Li, Dammann, and Chapman, 1966; Ramishvili, 1966; Pruzansky and Mathews, 1964; and Pruzansky, 1963). These data matrices, which may be regarded as digital spectrograms, at first represent different utterances by the same speaker and are later combined to form a single reference matrix for each speaker. The rows of the matrix correspond to the frequency bands of a spectrum analyzer. The columns correspond to the temporal locations of the sampled spectra, while each matrix cell describes a measured amplitude level. The degree of similarity between

a test matrix and the reference matrix determines the speaker recognition decision. Several investigators, arguing that spectral measures are sensitive to degradation of the transmission system, have studied recognition performance using parameters such as pitch, formant, and intensity contours of sentences (Atal, 1972b; Doddington, 1970; Lummis, 1973; Rosenberg and Sambur, 1975).

Even though they have the same text and are spoken by the same speaker, speech events in two utterances are seldom synchronized in time. Differences in speaking rates can cause this effect. In comparing time-varying parameters, it is necessary that these be derived from similar speech events in the reference and test utterances. Approximate time synchronization can be achieved by aligning the beginning and the end of the two utterances. More accurate synchronization can be achieved by a promising time normalization procedure first developed by Doddington (1970). This procedure time registers the test sentence with the reference sentence by nonlinear warping of the time axis of the test utterance. Dynamic programming has also been found useful for performing nonlinear time-warping (White and Neely, 1976; Sakoe and Chiba, 1978; Tappert and Das, 1978).

An alternative procedure has been to identify a number of "landmarks" in the utterances and to align the utterances by linear stretching or compression of the time scales between the landmarks (Das and Mohn, 1971). Several

studies have attempted to avoid the alignment problem by selecting acoustic parameters measured at specific speech events in an utterance (Glenn and Kleiner, 1968; Luck, 1969; Wolf, 1972; Hair and Rekieta, 1972; Su, Li, and Fu, 1974; Sambur, 1975). This procedure does require some prior segmentation and recognition of the linguistic component of the speech signal. Das and Mohn (1971) attempted to segment automatically; however, the system was unable to make a verification decision for about 10% of the utterances.

Some researchers have studied the effectiveness of measures that do not require temporal alignment or segmentation. These have included various measures of the long-time average spectra of prescribed words and sentences and of extemporaneous speech (Bricker et al., 1971; Li and Hughes, 1972; Kosiel, 1973; Zalewski, Majewski, and Hollien, 1975; Furui, Itakura, and Saito, 1975; and Doherty, 1976). Hollien and Majewski (1977) carried out two experiments in which long-term spectra were extracted from controlled speech samples. Three speaker conditions (normal speech, stressed speech, and disguised speech) were investigated. Results demonstrate high levels of correct speaker identification for normal speech, slightly reduced scores for speech under stress, and markedly reduced correct identifications for disguised speech. It may be that measures which do not require temporal alignment will work only for normal speech but require too much speech to be practical.

Recently developed speech analysis and synthesis techniques, such as linear prediction (Atal and Hanauer, 1971; Atal, 1975; Makhoul, 1975; Steiglitz, 1977; and Atal and Schroeder, 1978), allow convenient automatic extraction of a large number of acoustic parameters. This particular representation of the speech waveform appears to be quite promising for further evaluation of speaker-characterizing features.

Data Reduction and Feature Evaluation

SRM experiments have used dimensionality reduction techniques for two reasons according to Bricker and Pruzansky (1976). The first is simply to make the recognition system computationally more efficient by reducing the number of measures representing a single speech sample. The second, and more intrinsically interesting reason, is to identify a set of acoustic attributes that carry sufficient speaker information to allow recognition of a speaker. An excellent review of the statistical methods used in automatic recognition of speech and speakers is presented by Jelinek (1976).

The original set of acoustic parameters from a particular utterance can be thought of as a set of coordinates in multidimensional space. Mohn (1971) distinguished two ways of reducing the dimensionality of this space: subsetting and transformation. Subsetting involves selecting a smaller

number of measures from the original set without changing them. Transformation involves selecting a new set of coordinates in the space such that the new coordinates are a linear combination of the original ones.

Several different subsetting procedures have been used in SRM experiments: analysis of variance (Das and Mohn, 1971; Pruzansky and Mathews, 1964; Wolf, 1972), probability of error criterion (Rosenberg and Sambur, 1975; Sambur, 1975), and dynamic programming (Chang, 1973; Cheung and Eisenstein, 1978). Most procedures allow a ranking of the original set of measures according to their ability to discriminate speakers. Wolf (1972), using analysis of variance, found that some useful parameters were various measures of fundamental frequency, glottal source spectrum shape, and features of vowel and nasal consonant spectra. His speech samples were all collected on a single day. Sambur (1975), using the probability-of-error criterion he developed, was able to evaluate parameters, taking into consideration intersession variability. Speech samples were collected over a 3½-year period. His results showed that average fundamental frequency of a speaker remains relatively stable in a particular recording session, but shifts significantly from one session to another. Some features that had high intersession stability and low intraspeaker variability were certain parameters measured during nasals, and the second, third, and fourth formants of selected vowels.

Cheung and Eisenstein (1978) applied dynamic programming to the selection of feature subsets in text-independent speaker identification. The resulting subset of features showed a lower average identification error in comparison to that of the strategy employed by Sambur.

However, when dealing with a large number of speakers, say a thousand or more, the methods described earlier need some modification. One method is to subdivide the speaker set into a number of small groups in a meaningful way. The problem of grouping the speakers is commonly labeled as "clustering," and there are a number of algorithms for accomplishing it (Kashyap, 1976; Rabiner and Wilpon, 1979).

Discriminant analysis, a transformation method, also has been used in SRM experiments as a dimensionality reduction technique (Bricker et al., 1971; Mohn, 1971; Smith, 1962). It has been shown to produce efficient dimensionality reduction; however, little attempt has been made to interpret the discriminant coordinates (Bricker and Pruzansky, 1976). In one study, the first coordinate effectively separated male and female speakers, but further coordinates, although useful statistically, could not be interpreted in terms of meaningful characteristics of speakers (Bricker et al., 1971).

Decision Rules

Decision techniques are all based on the computation

of a distance which quantifies the degree of dissimilarity between the feature vectors associated with pairs of utterances. For identification tasks, the distance is computed between the test sample and each of the reference patterns. The speaker associated with the smallest distance is chosen as the speaker of the test utterance. The most common distance metric is the Euclidian distance $d = (\Sigma(x-y)^2)^{\frac{1}{2}}$; various other distance metrics have been investigated (Bricker et al., 1971; Atal, 1972, 1976; Furui et al., 1972; Itakura, 1975; Gupta, Bryan, and Gowdy, 1978).

For speaker verification, only the distance between the test sample and the reference pattern of the claimed speaker need be computed. If the distance is less than some threshold value, the speaker is accepted; otherwise, he is rejected. The selection of threshold is part of the design of the verification scheme, and in an experimental system, actually is determined after distances are computed. There are two kinds of error possible in verification: a customer can be falsely rejected, or an impostor can be falsely accepted. The equal error threshold, the most commonly used threshold determination, is the distance at which the two types of errors are equal. Additionally, however, the machine's decision can be made with an error "mix" (i.e., misses versus false alarms) that is tailored to the gravity of the requested transaction.

An important extension of decision techniques is the

use of sequential strategies. In a sequential strategy, each identification or verification trial is transformed into a series of trials. At each trial there is an option to defer decision to the next trial if the distance associated with the trial is too close to the predetermined decision threshold. The procedure terminates on the trial in which a decision is made. In practice, a limit is placed on the number of trials, with a special decision category if the limit is exceeded. A sequential procedure is used if trial-to-trial measurements are not strongly correlated. Sequential strategies are especially useful in on-line speaker verification implementations (Doddington, 1974).

Lummis (1973) computed five different distance measures for each of five different features (pitch, gain, and three formant frequencies, all as functions of time) and for all possible combinations of features. In this manner, he was able to evaluate the effectiveness of his features for speaker verification. He obtained very low error rates when using only pitch and gain contours of a sentence. Including formant information did not improve his results except when the impostor set included professional mimics. It appeared that, overall, formants were more difficult to mimic than pitch and gain. This may be explained by noting that formant values are constrained, to a much greater degree than are gain and pitch, by the words being pronounced. However, false acceptance rates showed that gain and pitch

functions of some speakers were more difficult to imitate than their formants (Lummis and Rosenberg, 1971). In an extension of this study, Rosenberg and Sambur (1975) tailored their decision rules (choice of distance measures and set of features) to each customer in their experimental system. They were able to reduce the false acceptance rate for professional mimics from over 20% to an acceptable 4%.

Comparison With Speaker Recognition by Man

Researchers conducting human speaker recognition studies can only infer from the listener's responses the nature of the human perceptual and decision processes, while in machine recognition experiments the features of the speech signal and the decision rule are specified precisely. An example of this specification is an automatic speech recognition approach described by Zwicker, Terhardt, and Paulus (1979). This system provides effective preprocessing of speech by implementing the basic functional laws which control the essential auditory sensations of loudness, pitch, roughness, timbre, and subjective duration. By comparing results achieved by machines and human beings, one may gain some insight into the less understood listener process. This comparison focuses on one question: How accurate are the automatic methods compared to the human listeners in quiet and noisy environments?

Clarke and Becker (1969) obtained machine recognition

scores which could be compared with the achievement of their listeners on an aural four-choice identification test and an aural discrimination test. The average listener scores on these two tests were 63-67% and 90%, respectively. For the discrimination test, the score was the optimal point on the median ROC curve. Machine results indicate that the long-term spectrum was superior (63%) to any other parameter or set of parameters. The machine scores were based on the identical speech samples that were heard in the aural tests.

In the comparison involving the identification task, it is noted that only the machine score obtained for the entire long-term spectrum closely resembles the listener score; the other machine scores are considerably lower. Also, in the comparison involving the discrimination task, only the machine score obtained for the long-term spectrum (83%) approaches the listener score. It appears from these comparisons the human listeners are generally able to extract more speaker-dependent information from the speech signal than is contained in relatively simple physical measures, including fundamental frequency. Some physical measures, however, appear to contain much information that is relevant to speaker recognition. For example, Atal (1974) achieved 98% correct identification for .5 s of speech based only on spectral envelope measures. This study employed ten speakers with utterances collected over a month.

As pointed out by Clarke and Becker (1969) and

Haggard and Summerfield (1979), human listeners do not necessarily use the same parameters that have been found advantageous for machine recognition. Hecker (1971) suggests that this may be due to several factors: (1) a given listener may modify his strategy as a test proceeds, especially if he is given some indication of his immediate performance; (2) auditory memory from test item to test item may differ; and (3) prior experience in differentiating among speakers with perceptually similar voices may favorably influence listener scores. This may be visual and auditory information.

Rosenberg (1973) compared the speaker verification performance of listeners with machine verification results based on the same speech material. The listeners were asked to respond to whether a pair of test and reference utterances were spoken by the same or different speakers. The speech utterance was a 2-s long all-voiced sentence "We were away a year ago." Rosenberg found that the false acceptance error rate for machines was significantly lower than the average performance of listeners for both casual and mimic impostors. However, the performance of the best listeners compared favorably with, or even exceeded, the performance of the automatic system. Bricker and Pruzansky (1976) note that no comparisons of individual test items were reported by Rosenberg, so one does not know if listener and machine errors were made on the same samples. This information might contribute to our knowledge about the perceptual data used by listeners.

Lummis (1973) also reports interesting findings comparing man and machine recognition scores for identical twin impostors. The same twin utterances were almost always accepted by listeners; whereas the machine analysis, using only pitch and gain as a function of time, correctly rejected all of the utterances. It seems that organic cues were more salient to the listeners than the timing and intensity cues, which differed substantially between the identical twins as evidenced by the machine results.

These evaluations were carried out in sound booths with high quality recordings. Eventually, however, one must consider whether these conditions represent a fair approximation to conditions that are expected in a practical application. One speaker recognition system has been studied in a noisy environment (Doddington and Hydrick, 1975). Another experiment has examined the effects of noise on speaker recognition scores for human listeners (Clarke, Becker, and Nixon, 1966). Both man and machine recognition performance was seriously affected by the background noise.

This concludes the empirical survey of the problems and progress of speaker recognition by man and machine. The remainder of this paper reports the research conducted to examine the effects of speaking effort and noise upon speaker recognition performance for human listeners.

CHAPTER IV

EXPERIMENTAL METHODS

This chapter reports the experimental data collection methods. It consists of a description of subjects, speech material, recordings, test tapes, and listening procedures.

Subjects

Williams (1971) concludes an ensemble of five or six speakers to be optimal for testing speaker recognition. Accordingly, six males, ranging in age from 30 to 55 years, were selected for speaking the speech material in this study. Henceforth, these subjects are referred to as talkers. The criteria for talker selection were: (1) none were to have either a pronounced regional dialect or speech abnormalities, (2) all were to record the speech material placing emphasis on identical words, and (3) all were to register equivalent duration for the speech utterance. These criteria were chosen to minimize some of the prosodic cues responsible for simple talker discrimination. Three independent evaluations,

performed by the researcher and two laboratory personnel, selected the six talkers from a pool of twelve according to the above criteria.

Eleven paid volunteers, four males and seven females, ranging in age from 19 to 30 years, were recruited to be the listeners. It has been suggested one may regard results based on groups of such a size as typical (Williams, 1964). The criteria for selection of the eleven listeners were that they have hearing within normal limits (i.e., pure-tone air-conduction thresholds in both ears no greater than 10 dB hearing level (ANSI, 1969) at octave frequencies 250 through 8000 Hz as indicated by a calibrated diagnostic audiometer, Grason-Stadler model 1701) and that they be unfamiliar with the talkers' voices.

Speech Material

The speech material consisted of one sentence: "My name is Miller; cash this bond, please." This utterance was chosen for two reasons: (1) it has been used in previous studies because it includes a wide variety of speech sounds or phonemes (Wolf, 1972; Sambur, 1975), and (2) it has face validity for the talker discrimination task.

Recordings

Each talker produced the chosen sentence at three

different speaking efforts: whisper, normal, and shout. The system employed in recording the speech is shown schematically in Figure 4.1. The anechoic chamber in which the sentences were recorded contained a microphone, a sound level meter used as an amplifier, a Daven Volume Units (VU) meter, and a loudspeaker. The Daven VU-meter was used by the talker during speech to monitor level, while the loudspeaker provided communication for the researcher from the control room into the anechoic chamber. The speech signal was passed through an amplifier to an Ampex tape recorder model 440B located outside the chamber. The operating speed was 3.25 ips and the recordings were made on Scotch 138 magnetic tape.

Prior to recording the speech, the reading on the Daven VU-meter was calibrated to match the reading on the Ampex VU-meter. A precise calibration was accomplished with an audio oscillator, Hewlett Packard model 201CR, which provided the steady sound required for visual confirmation of the correct adjustment of the meters (i.e., the same reading on both meters). This calibration procedure allowed the researcher to monitor the level at which each talker spoke the sentence.

Next, the settings on the Hewlett Packard 350D attenuator, which would determine the amount of effort at which all talkers would speak the whisper, normal, and shout sentences, were identified. The respective attenuator settings were 45, 39, and 21 dB; these correspond to peak

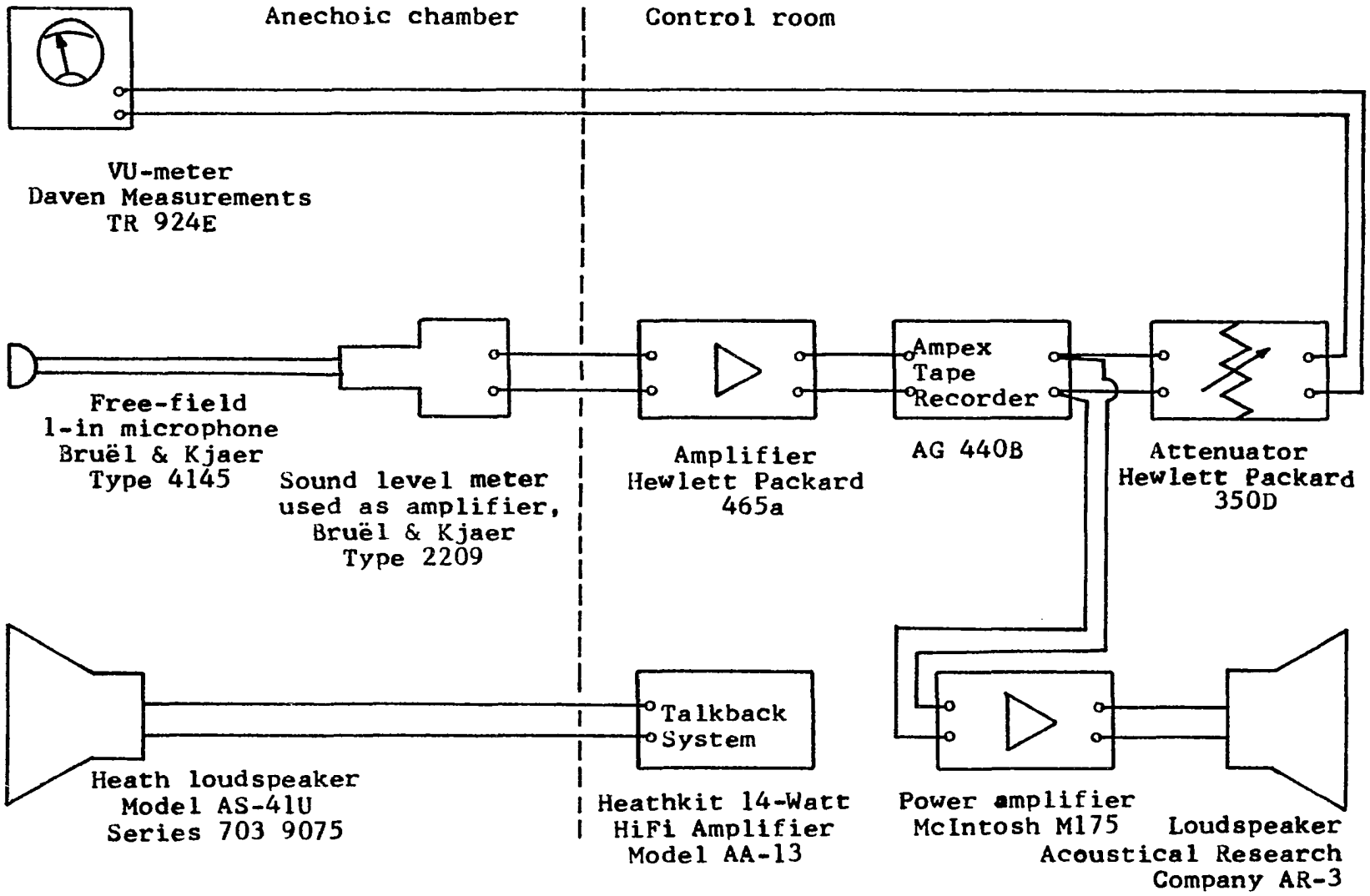


Fig. 4.1. Block diagram of experimental setup for recording speech.

speech levels of 59, 65, and 83 dB SPL measured 1 m from the AR-3 loudspeaker with a Bruël and Kjaer Type 2209 sound level meter. These SPL's are the most common quoted figures for the loudness of the three speaking efforts and are generally measured at a distance of 1 m from the lips (Waltzman and Levitt, 1978). Other speaking efforts not used in this study, such as raised and very loud, correspond to 6-dB increments in voice level between the normal effort value of 65 dB SPL and the shout effort value of 83 dB SPL. This measurement process was achieved by playing a 1000 Hz tone at the peak speech level for each speaking effort.

Each talker produced all of his material in a single session seated in the anechoic chamber. His head was placed in a head positioner with his lips approximately .5 m from the microphone. He was instructed to speak naturally, using the same speed for every sentence and stressing the same words of the sentence. These were Miller and bond. For each speaking effort, the talker was told to monitor his voice level on the Daven VU-meter and to speak the entire sentence at the level necessary to peak the needle at the zero designation.

Upon completion of a practice session for each speaking effort, the researcher recorded five speech samples each of whispers, normal efforts, and shouts, in that order, for each talker. The practice and recording sessions for each speaking effort lasted approximately fifteen minutes.

During those sessions the attenuator was adjusted to its predetermined setting for each speaking effort, thereby forcing the talker to use the appropriate amount of effort. The success of each talker for speaking the sentence at the level necessary to peak the VU-meter needle at the zero dB designation was monitored by the researcher. These calibration and recording procedures enabled the researcher to produce a tape with constant peak level for every speaking effort and talker.

The instructions to the talkers were as follows:

> This is a speaker verification experiment. A verification task consists of distinguishing a known cooperative talker from the set of all other talkers. In order to do this, the talker in question identifies himself and gives a natural voice utterance. Then the computer must decide the question, "Is this speaker so-and-so?"
>
> Suppose you go to the bank one morning to find the money withdrawal slip-check system replaced by a voice-check system. The system compares your utterance of a selected key phrase with previous recorded samples. Similarity of your on-the-spot utterance with the previous recorded samples governs your acceptance or rejection to withdraw money.
>
> In this experiment, it is important for you to maintain consistency between recordings of the key phrase. I suggest that you try merely to speak naturally, using the same speed for every sentence and stressing the same words which I will identify. We will record your voice at three different speaking efforts. They are whisper, normal, and shout. You will need to repeat the key phrase several times so that you can become familiar with the speaking effort requested and the utterance.
>
> The key phrase is as follows: My name is Miller; cash this bond, please. Always stress the same words. These words are Miller and bond. For each speaking effort, you must watch this VU-meter and speak the entire sentence at the level necessary to peak the needle at the zero designation. For example, the whisper speaking effort will be spoken like this (researcher performs correctly for the talker). Do you understand? Let's begin practice.

Test Tapes

One sentence for each speaking effort (whisper, normal and shout) from each of the six talkers' repertoire of utterances was selected, according to the previously stated criteria, for constructing the test tapes. This decision minimized intratalker variability of the speech material.

The eighteen sentences, three for each talker, were then prepared to form six test tapes. This was done by editing the chosen utterances from the original speech recording tape and splicing them onto separate tape loops. Then these loops were used with Ampex tape recorders PR10 and AG 440B to re-record the speech material onto the test tapes. Each test tape was composed of 120 pairs of voices. The test tapes are distinguished by the speaking efforts chosen to construct the pairs. These speaking effort comparisons were: (1) normal-normal (N-N), (2) whisper-whisper (W-W), (3) shout-shout (S-S), (4) normal-shout (N-S), (5) whisper-normal (W-N), and (6) whisper-shout (W-S). There was a 1-s interval within pairs and a 5-s interval between pairs. Each sentence was about 3 s in duration, measured by a stopwatch.

Of the 120 pairs constructed for each test tape, fifty percent or sixty pairs presented a comparison of voices which belonged to the same talker. This meant there were ten pairs for each of the six talkers in which each talker's voice was compared with itself. For the three test tapes in which different speaking efforts were compared (whisper-

shout, whisper-normal, and normal-shout) five of the ten pairs exchanged speaking effort order. For example, the test tape that presented the normal-shout pairs contained five comparisons in the shout-normal order. This procedure effectively counterbalances a possible order effect.

The remaining sixty pairs presented a comparison of voices which belonged to different talkers. This meant there were ten pairs for each of the six talkers in which that talker's voice was compared with one of the other five voices. For the three test tapes in which different speaking efforts were compared, five of the ten pairs exchanged speaking effort order.

Prior to recording the speaking effort comparisons on the test tapes, a 1000 Hz calibration tone was recorded on each test tape for one minute. The level of this tone was recorded at the level that corresponded to the peak speech reading on the Ampex VU-meter. All recordings were made on Ampex 631 magnetic tape, and each test tape was approximately 24 minutes long.

Listening Procedure

Eleven listeners participated in thirty-one same-different discrimination tasks for speaker recognition. This task minimized short-term memory factors (the listener simply compared two utterances and decided whether or not they were spoken by the same talker) and matched the machine verification

process. Each task consisted of 120 discriminations. The a priori probability of the same talker's voice occurring was 0.5. The subjects listened to the six speaking effort comparisons presented in ambient or broad-band noise in three groups. There were four listeners in two of the groups and three listeners in the third group. The experiment lasted four consecutive days for each group of listeners.

The test tapes were produced from an Ampex tape recorder model AG 440B through a power amplifier and loudspeaker. These tapes were presented with broad-band, thermal noise produced by a Grason-Stadler noise generator model E10588A. Figure 4.2 illustrates the equipment used for producing the speech and noise.

The listeners were seated either 4.7 or 5.3 m from the loudspeaker, which was positioned on the floor. They were instructed to exchange seats on every test tape presentation in order to counterbalance for sound level variability due to the unequal distance of the seats from the loudspeaker. The listeners were given fifteen minute rest periods after every test tape with a two minute rest after every forty pairs for each test tape. Test tape conditions and noise levels were presented in a randomized order with the three groups of listeners receiving different orders.

The level of the calibration tone on the test tapes was read on a Bruël and Kjaer Type 2209 sound level meter placed approximately 5 m from the loudspeaker, with the

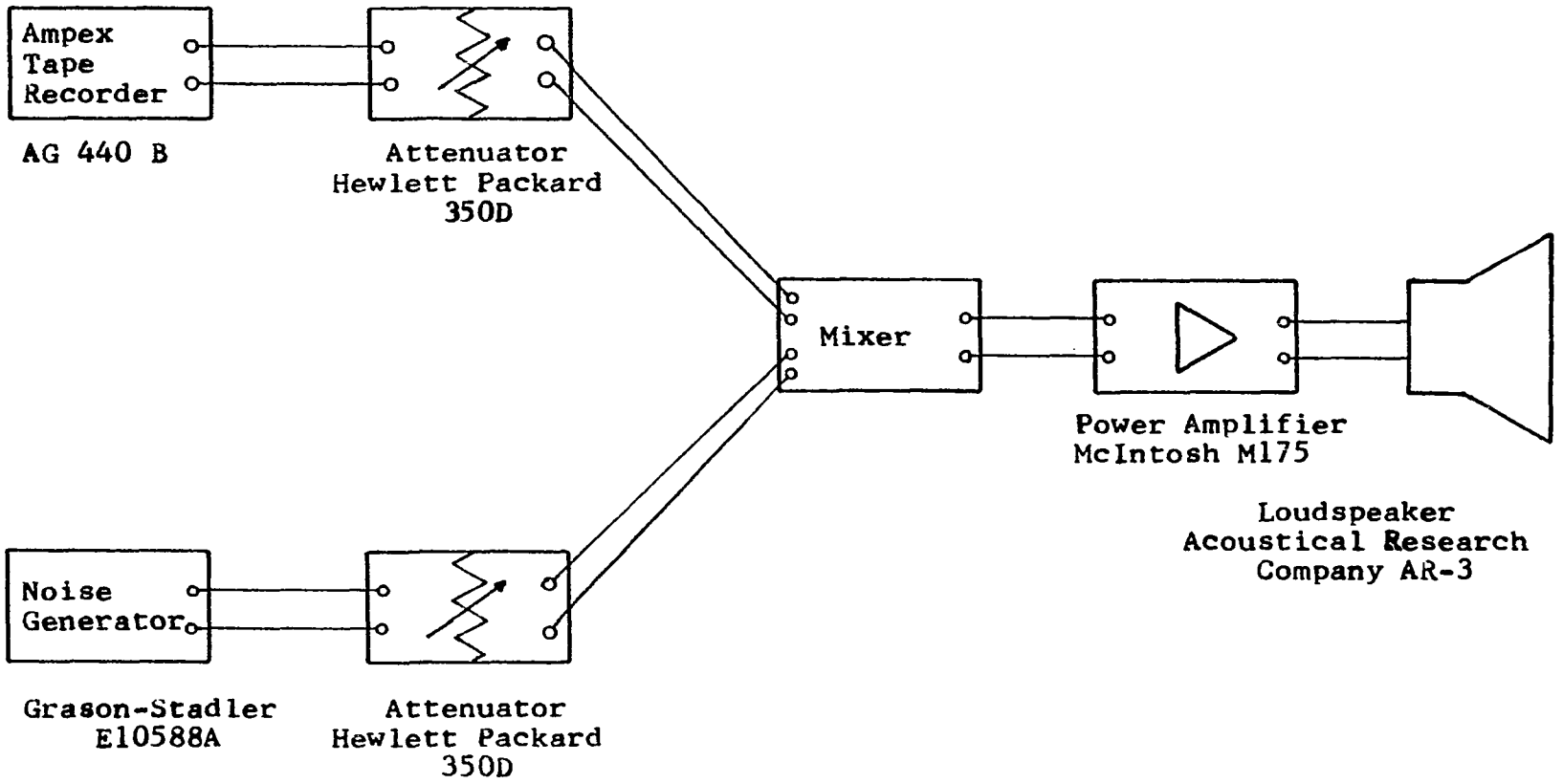


Fig. 4.2. Block diagram of experimental setup for mixing speech and broad-band noise.

Bruël and Kjaer microphone Type 4145 facing the loudspeaker. The peak speech level equaled the level of the 1000 Hz calibration tone. Therefore when the 1000 Hz calibration tone was adjusted with a Hewlett Packard attenuator to a comfortable listening level (Richards, 1975), this also determined the peak speech level. All test tapes were presented at 62 dB SPL. The listeners agreed this was a comfortable listening level.

Broad-band, white noise was chosen as the masking noise in this study. It is a signal containing energy at all frequencies in the speech spectrum at approximately equal intensities. Figure 4.3 shows that the acoustic spectra for the broad-band noises in this study are limited, however, by the frequency response of the loudspeaker. The octave band analysis also shows that there is a small amount of amplitude distortion present in the broad-band noise. This distortion may be due to limitations in the range of the loudspeaker or to ambient noise contributing to the measurement. It should have a negligible influence on speaker recognition.

All six test tapes were presented at the ambient noise level of 59.1 dB SPL and at three broad-band noise levels of 62.3, 65.7, and 69.8 dB SPL. In addition to those noise levels, the N-N and S-S comparison tapes were presented at 71.7, 72.6, and 74 dB SPL, and the W-W comparison tape was presented only at 72.6 dB SPL. These noise levels were included for the purpose of determining the shape of the

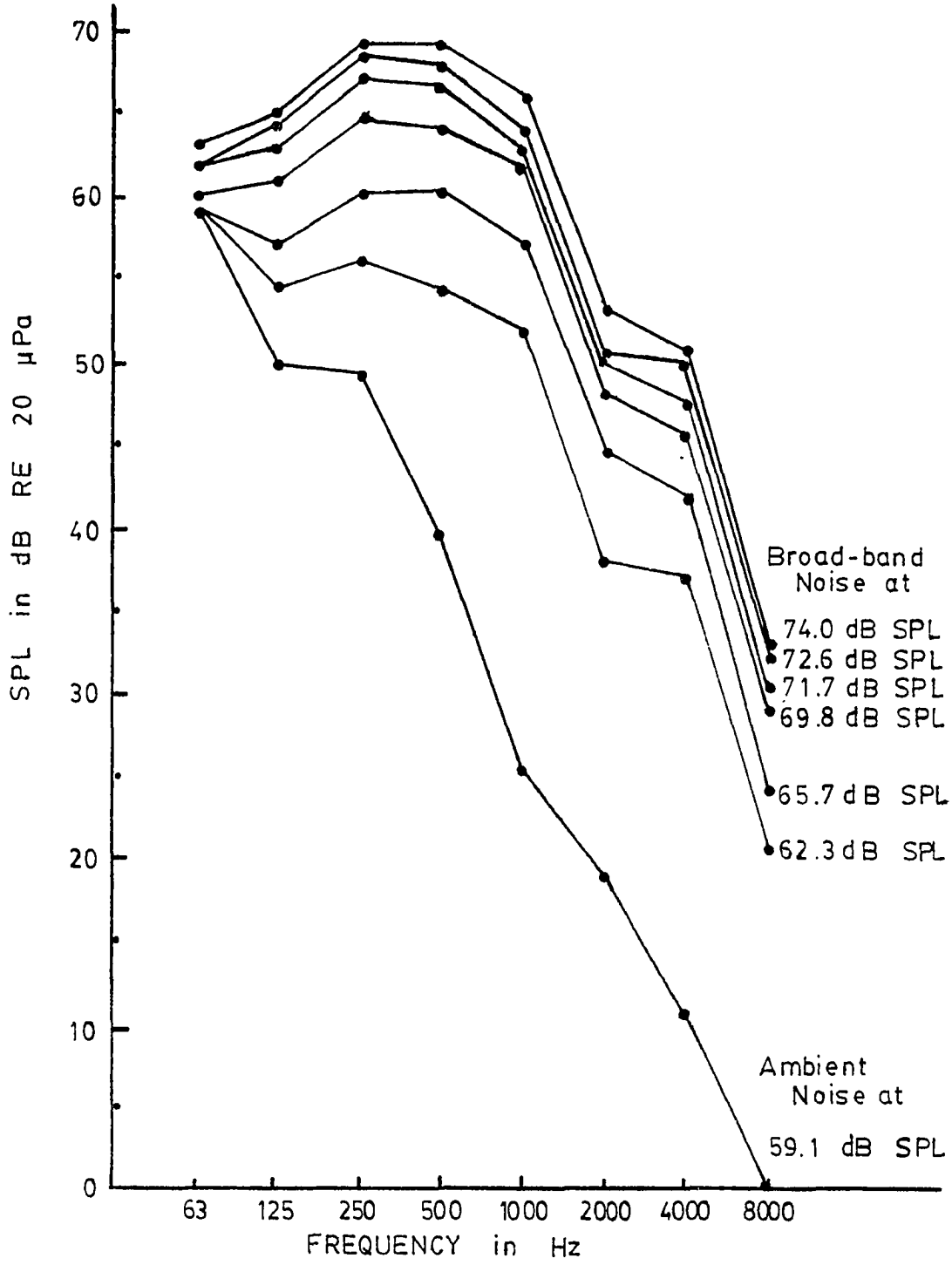


Fig. 4.3. Octave band analysis and calculated overall level for the ambient or broad-band noise.

psychometric function between the 90% and 50% correct performance levels. For this study 90% and 50% correspond to the training criterion and chance performance levels, respectively. The specific broad-band noise levels were identified in a pilot study.

Noise levels were read on a Bruël and Kjaer Type 2209 sound level meter placed approximately 5 m from the loudspeaker, with the Bruël and Kjaer microphone Type 4145 facing the loudspeaker. The level of the noise was adjusted with a Hewlett Packard attenuator 350D. The tone and noise levels were measured before each test tape presentation.

The SPL's stated in this study were calculated from octave band data. The frequency range used was 63 through 8000 Hz. This range includes the frequencies associated with the fundamentals of voiced sounds and the high consonant area of sibilants. The SPL's were calculated because the broad-band sound level meter measurements were felt to be confounded by very low frequency noise.

On the first day of the experiment the listeners were repeatedly presented a familiarization tape of the normal speaking effort for each talker, using the test tape utterance. The intent was to train the listeners to discriminate the voices and thereby minimize individual differences in ability that may exist for voice recognition. The listeners were instructed to associate each talker on the familiarization tape with a number. Specifically, the

first talker's speaking efforts on the tape were identified as number one, the second talker's speaking efforts on the tape were identified as number two, and so on. The listeners were then presented a test tape and asked to identify the voices by number. During this time correct identifications were given to the listeners immediately following each utterance. Upon completion of this thirty minute session of voice identification practice, the listeners were tested for voice discrimination accuracy for the normal speaking effort. That is, they were presented with pairs of voices and asked to compare the two voices and decide whether or not they were spoken by the same talker. This testing continued until every listener correctly discriminated 90% of the voices for four consecutive tests of twenty trials. Testing lasted approximately thirty minutes. Throughout this training period for the normal speaking effort, each test presented a different randomized order of voices.

The whisper and shout speaking effort training periods proceeded in a similar manner. Total training time for all three speaking efforts was approximately two hours and forty-five minutes. The listeners were not informed as to which whisper, shout, and normal speaking effort belonged to each talker in order to match a machine verification process. The computer system stores acoustic information belonging to a normal speaking effort only. So when the machine is accessed in a noisy environment or with a

different speaking effort or both, the machine and listeners in this study face a similar problem.

The listeners were read the following instructions prior to the training sessions:

> The experiment in which you are participating is concerned with the discrimination of talkers or voices. You are now going to hear six talkers repeating one sentence using the normal speaking effort. Each talker is to be associated with a particular number from one to six. Specifically, the first talker's speaking efforts on the tape are identified as number one, the second talker's speaking efforts are identified as number two, and so on. Your task is to learn the correct talker-number associations and, thereafter, write down a talker's number each time you hear his voice.
>
> In the beginning you will have to guess. However, following each of your responses (each time you write down a number), you will be told the correct talker number. If your response was correct, write down the talker number alongside the number you have just written; if your response was incorrect, circle the number you have written and write down the correct talker number.
>
> Do not concern yourself with what the talkers are saying; your task is only to learn to identify the different talkers.
>
> Following this familiarization session in which you will always be informed as to the correct talker, there will be a test session during which time you will not receive this information. During this test session you will be presented pairs of voices and your task will be to simply compare the two voices and decide whether or not they are spoken by the same talker. You will be tested for voice discrimination accuracy for this speaking effort until every listener correctly discriminates 90% of the voices for four consecutive tests of twenty trials. Upon completion of this criterion you will be prepared to begin the actual experiment.
>
> Do you have any questions?

At the beginning of each day of testing, the training phase was repeated. The listeners were again required to meet the criterion of four consecutive tests of twenty trials for each speaking effort with 90% discrimination accuracy. This lasted about an hour.

Following training the listeners were read these instructions:

> This is a speaker verification experiment. This verification task consists of distinguishing pairs of voices. Your task will be simply to compare two voices and decide whether or not they are spoken by the same talker. You will also need to indicate your degree of confidence in the correctness of your response. The answer sheets given to you show you report one of three answers--very sure response is correct, quite likely to be correct but not certain, or a best guess but not very confident. Try to use all three categories but only where appropriate. To sum up, you make two marks on your answer sheet for every pair of voices presented. You answer "yes" or "no" and you indicate the degree of confidence that you have in your response being correct. We are now ready to begin.
>
> Are there any questions?

The same-different discrimination task was analyzed according to signal detection methods (Clarke, Becker, and Nixon, 1966; Matsumoto et al., 1973; Elman, 1979; Haggard and Summerfield, 1979). For a same-different discrimination task, a signal detection parameter indicates the magnitude of differences between stimuli; that is, the distance between distributions of perceived stimulus differences when identical stimuli are presented, and the distance between distributions of perceived stimulus differences when different stimuli are presented. The rating scale procedure allowed the listener six response categories, each corresponding to a different criterion. The listener's responses were sorted according to the confidence ratings they received, and within each rating category the correct "same" responses were separated from the incorrect "same" responses. Hit and false alarm rates were then calculated for each rating scale category

from the raw data. The response bias measure chosen for this study (McNicol, 1972, p. 126) was calculated from these hit and false alarm rates. Hit rates, Prob(yes|same-talker pair), for the speaker recognition task are defined as the proportion of "yes" responses given that the same talker's voice has been presented. False alarm rates, Prob(yes|different-talker pair), are the proportion of "yes" responses given that a different talker's voice has been presented. In the rating scale task one capitalizes on the fact that a listener can hold several criteria of different degrees of strictness simultaneously, thus allowing the determination of several pairs of hit and false alarm rates within the one block of trials. For this experiment one block of trials was equivalent to 120 discriminations. The signal detection parameter, P(C), was used as the measure of sensitivity of discrimination. P(C) is the proportion of correct same-talker and different-talker judgments for each listener.

CHAPTER V

RESULTS AND CONCLUSIONS

This final chapter presents the results of statistical analyses and the discussion of their implications.

Results

For the normal-normal (N-N), shout-shout (S-S), and whisper-whisper (W-W) speaking effort comparisons, the results demonstrate similar speaker recognition scores for the normal and shout efforts in noise, with quite dissimilar scores for the whisper effort. For the normal-shout (N-S), whisper-normal (W-N), and whisper-shout (W-S) speaking effort comparisons, the results revealed poor recognition scores for all speaking effort comparisons at all noise levels. These findings are illustrated in Figure 5.1.

The percentage correct recognition scores were analyzed with a within-subjects analysis of variance with three factors: groups (3), speaking effort comparisons (6), and noise levels (4). The groups factor accounted for the

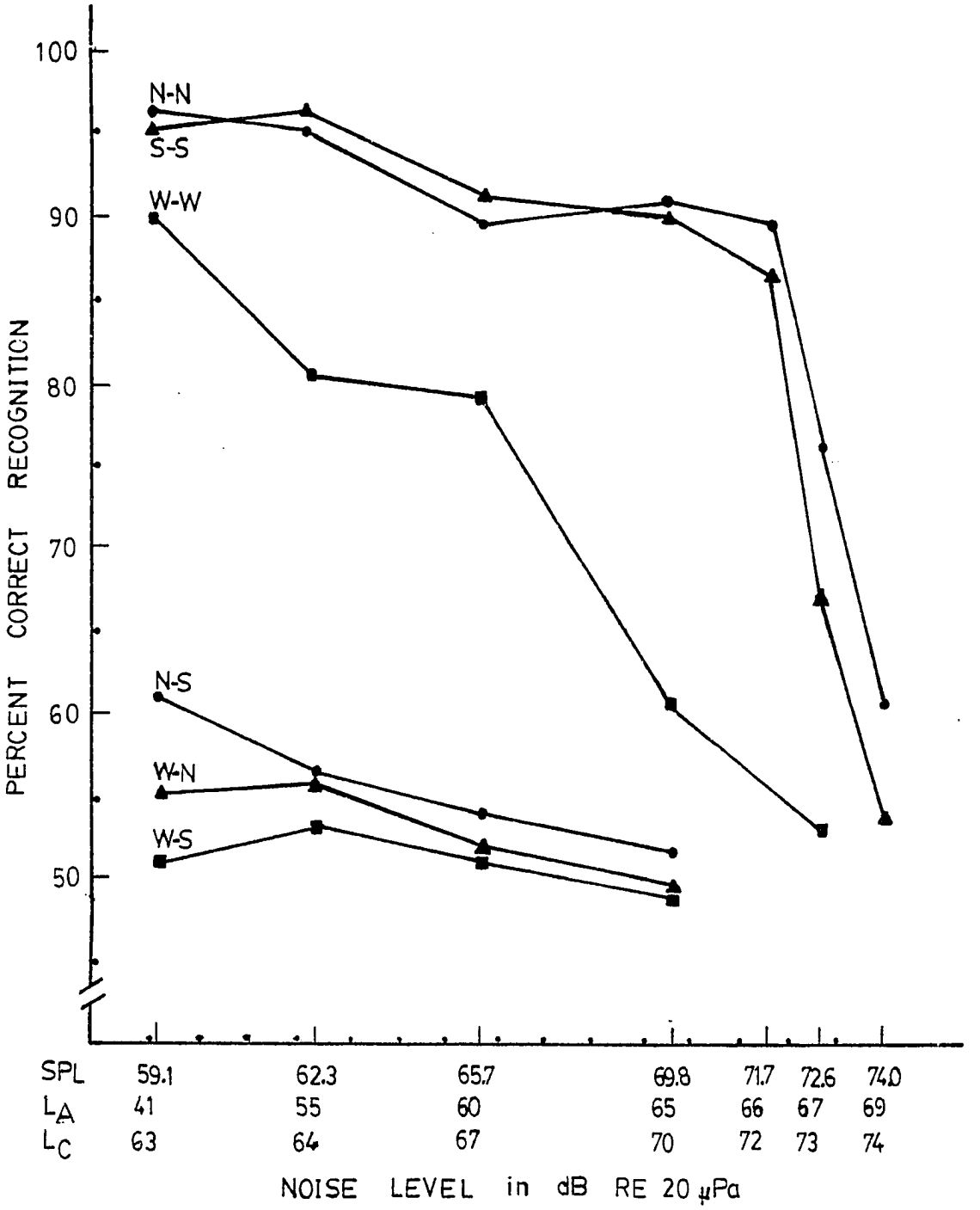


Fig. 5.1. Mean correct talker recognition scores as a function of noise level and speaking effort comparison (N = normal, S = shout, and W = whisper). Noise levels are reported in SPL (sound pressure level), L_A (A-weighted sound level), and L_C (C-weighted sound level). Speech level is held constant at 62 dB SPL.

variance attributable to the sequential effects in the series of test tapes and the environmental conditions present for each group of listeners. The six speaking effort comparisons were N-N, S-S, W-W, N-S, W-N, and W-S. The four noise levels were the ambient noise level of 59.1 dB SPL, and three broad-band noise levels of 62.3, 65.7, and 69.8 dB SPL. Table 1 summarizes the results of this mixed model, split-plot design. Speaking effort (E), noise (N), and their interaction (EN) were significant. The assumption of homogeneity of variance for the four sets of error terms in this design was tested with F max tests and found tenable. Tests of equality and symmetry of the variance-covariance matrices indicated assumptions were met and so the within-subjects F ratios were considered to be distributed as the F distribution. Percentage correct scores for each listener are presented in the Appendix.

Tests of simple main effects revealed significant differences between speaking effort comparisons at all four noise levels (59.1, 62.3, 65.7, and 69.8 dB SPL), computed values of $F(5,40) = 199.95$, 174.90, 191.90, and 179.22, $p<.01$, respectively. In addition, there were significant differences between noise levels for the N-N, W-W, N-S, and W-N speaking effort comparisons, computed values of $F(3,36) = 4.65$, 65.98, 7.84, and 12.43, $p<.01$, in order.

A posteriori comparisons among means were made by Scheffe's ratio and are given in Table 2. This ratio indicated

TABLE 1

ANALYSIS OF VARIANCE SUMMARY

Source	SS	DF	MS	F	Prob
Between Listeners					
Groups (G)	10.16	2	5.08	.03	.971
List w groups	1421.17	8	177.64		
Within Listeners					
Effort (E)	81728.69	5	16345.73	415.66*	.000
EG	393.24	10	39.32	1.56	.155
E x list w groups	1009.26	40	25.23		
Noise (N)	3187.24	3	1062.41	24.73*	.000
NG	257.71	6	42.95	1.69	.166
N x list w groups	610.18	24	25.42		
EN	3000.06	15	200.00	10.40*	.000
ENG	576.39	30	19.21	1.33	.144
EN x list w groups	1738.37	120	14.48		

*$p < .001$.

TABLE 2

RESULTS OF COMPARISONS AMONG MEANS BY SCHEFFE'S RATIO

Comparison (E at N1, N2, N3, or N4)	F
$(\overline{E1N1} + \overline{E2N1} + \overline{E3N1})/3 - (\overline{E4N1} + \overline{E5N1} + \overline{E6N1})/3$	981.41**
$(\overline{E1N2} + \overline{E2N2} + \overline{E3N2})/3 - (\overline{E4N2} + \overline{E5N2} + \overline{E6N2})/3$	798.65**
$(\overline{E1N3} + \overline{E2N3} + \overline{E3N3})/3 - (\overline{E4N3} + \overline{E5N3} + \overline{E6N3})/3$	775.69**
$(\overline{E1N4} + \overline{E2N4} + \overline{E3N4})/3 - (\overline{E4N4} + \overline{E5N4} + \overline{E6N4})/3$	614.64**
$(\overline{E1N1} + \overline{E2N1})/2 - \overline{E3N1}$	11.00
$(\overline{E1N2} + \overline{E2N2})/2 - \overline{E3N2}$	71.69**
$(\overline{E1N3} + \overline{E2N3})/2 - \overline{E3N3}$	42.61**
$(\overline{E1N4} + \overline{E2N4})/2 - \overline{E3N4}$	262.15**
$\overline{E3N1} - (\overline{E4N1} + \overline{E5N1} + \overline{E6N1})/3$	420.42**
$\overline{E3N2} - (\overline{E4N2} + \overline{E5N2} + \overline{E6N2})/3$	214.07**
$\overline{E3N3} - (\overline{E4N3} + \overline{E5N3} + \overline{E6N3})/3$	246.87**
$\overline{E3N4} - (\overline{E4N4} + \overline{E5N4} + \overline{E6N4})/3$	43.52**
$\overline{E4N1} - (\overline{E5N1} + \overline{E6N1})/2$	14.57*

*p<.05, $F'(5,40)(k-1) = 12.25$.
**p<.01, $F'(5,40)(k-1) = 17.55$.

Comparison (N at E1, E2, E3, E4, E5, or E6)	F
$\overline{E1N1} - \overline{E1N3}$	10.98*
$\overline{E3N1} - (\overline{E3N2} + \overline{E3N3} + \overline{E3N4})/3$	105.30**
$\overline{E3N2} - \overline{E3N4}$	86.71**
$\overline{E3N3} - \overline{E3N4}$	76.16**
$\overline{E4N1} - \overline{E4N4}$	16.80**
$\overline{E5N2} - \overline{E5N4}$	11.57*

NOTE: E1-E6 = N-N, S-S, W-W, N-S, W-N, and W-S speaking effort comparisons, respectively. N1-N4 = 59.1, 62.3, 65.7, and 69.8 dB SPL, respectively.

*p<.05, $F'(3,36)(k-1) = 8.61$.
**p<.01, $F'(3,36)(k-1) = 13.20$.

the N-N, S-S, and W-W means to be significantly different from the N-S, W-N, and W-S means at all noise levels (59.1, 62.3, 65.7, and 69.8 dB SPL), computed values of $F(5,40)$ = 981.41, 798.65, 775.69, and 614.64, $p<.01$, respectively. The W-W comparison mean was significantly different from the N-S, W-N, and W-S means at all noise levels (59.1, 62.3, 65.7, and 69.8 dB SPL), computed values of $F(5,40)$ = 420.42, 214.07, 246.87, and 43.52, $p<.01$, in order. The N-N and S-S effort comparison means did not reach significance when paired with the W-W effort at the ambient noise level of 59.1 dB SPL, computed value of $F(5,40)$ = 11.00, $p>.05$. However, for the other noise levels (62.3, 65.7, and 69.8 dB SPL) a significant difference between these means was noted, computed values of $F(5,40)$ = 71.69, 42.61, and 262.15, $p<.01$. The N-S mean was found to differ significantly from the W-N and W-S means at the ambient noise level, computed value of $F(5,40)$ = 14.57, $p<.05$.

The simple main effect between noise levels for an effort comparison indicated the W-W effort to be the most affected. Scheffe's ratio demonstrated significant differences between noise levels 59.1 and 62.3, 65.7, 69.8 dB SPL, computed value of $F(3,36)$ = 105.30, $p<.01$; 62.3 and 69.8 dB SPL, computed value of $F(3,36)$ = 86.71, $p<.01$; and 65.7 and 69.8 dB SPL, computed value of $F(3,36)$ = 76.16, $p<.01$, for the W-W comparison. Other significant differences between means were noted for the N-N comparison at 59.1 and 65.7 dB SPL

(computed value of $F(3,36) = 10.98$, $p<.05$), for the N-S comparison at 59.1 and 69.8 dB SPL (computed value of $F(3,36) = 16.8$, $p<.01$), and for the W-N comparison at 62.3 and 69.8 dB SPL (computed value of $F(3,36) = 11.57$, $p<.05$).

An estimate of the strength-of-association in this analysis of variance was obtained through the use of the correlation ratio. This measure represents the proportion of variance in the dependent variable that may be accounted for by the independent variable. The correlation ratio indicated speaking effort to represent 87% of the variance in this experiment. Noise and the effort-noise interaction accounted for 3% each.

The results of a linear regression analysis for each psychometric function are presented in Table 3. This analysis was performed in order to obtain slope estimates that could be compared with other psychometric functions. The independent variable was noise level (SPL), and the dependent variable was talker recognition or percentage correct. For the N-N and S-S psychometric function, the regression line was fitted to four points in the range of 50% to 90%. The W-W function was fitted to five points in the range of 50% to 90%. The remaining functions were fitted to four points in the range of 50% and 60%.

The best talker recognition scores for each speaking effort comparison and noise level are illustrated in Figure 5.2. The trends in Figure 5.1 appear to be present for the

TABLE 3

LINEAR REGRESSION RESULTS

Speaking Effort Comp.	Estimated Regr. Coeff.	Estimated Std. Deviation	Std. Regr. Coeff.*
Normal-Normal	b_0 = 603.402	60.809	0
	b_1 = -7.266	0.844	-0.798
Shout-Shout	b_0 = 721.540	59.514	0
	b_1 = -8.979	0.826	-0.858
Whisper-Whisper	b_0 = 245.020	12.847	0
	b_1 = -2.607	0.194	-0.878
Normal-Shout	b_0 = 112.774	13.999	0
	b_1 = -0.875	0.217	-0.527
Whisper-Normal	b_0 = 91.479	9.368	0
	b_1 = -0.580	0.145	-0.523
Whisper-Shout	b_0 = 60.790	9.073	0
	b_1 = -0.134	0.141	-0.145

*Regression line was forced through the point of origin.

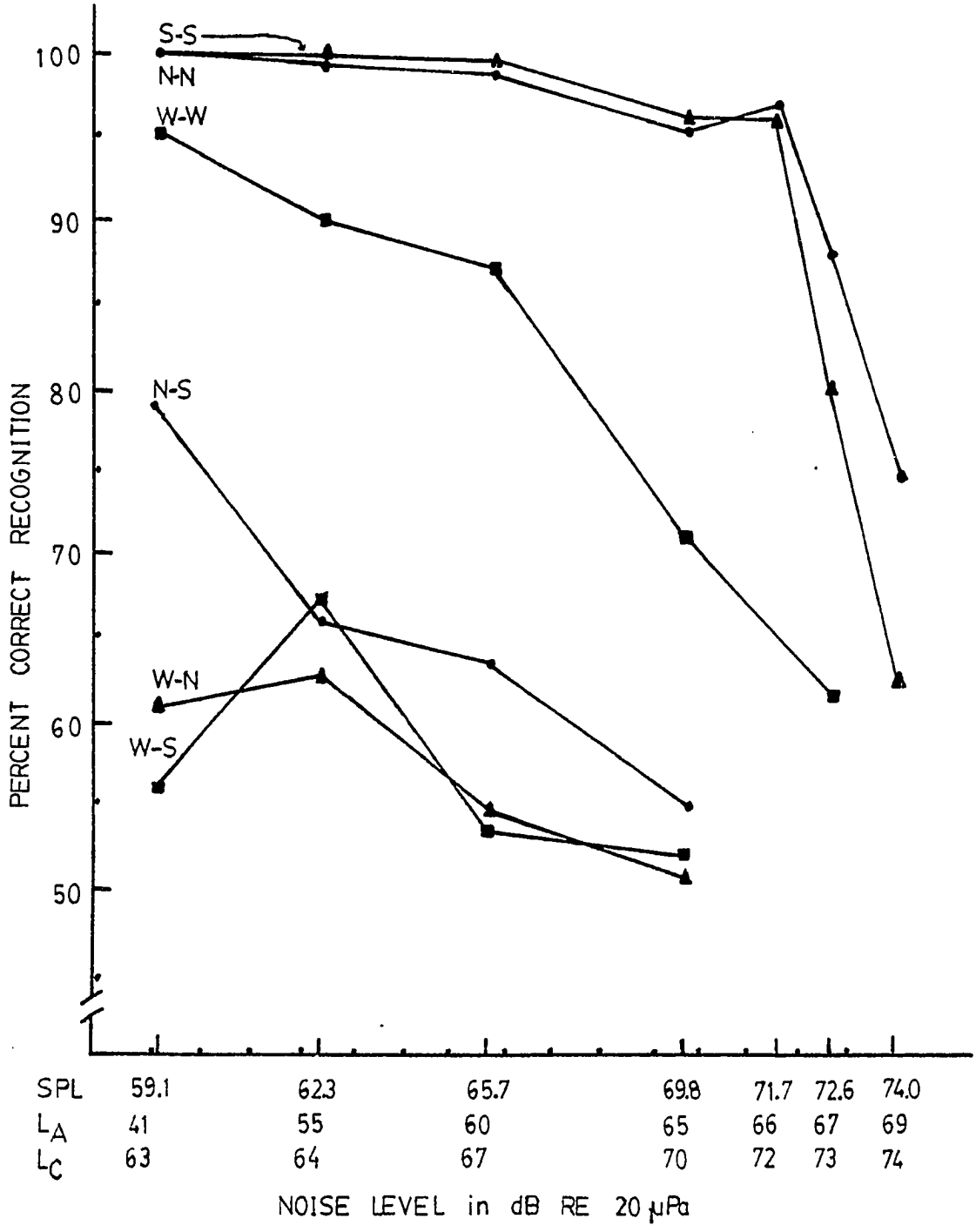


Fig. 5.2. Best individual human performances for talker recognition as a function of noise level and speaking effort comparison (N = normal, S = shout, and W = whisper effort). Noise levels are reported in SPL (sound pressure level), L_A (A-weighted sound level), and L_C (C-weighted sound level). Speech level is held constant at 62 dB SPL.

best performance data. Generally, the best performance scores by human listeners for speaker recognition are more accurate than recognition scores by machine verification. The performance data in Figure 5.2 indicate that a machine process will have difficulty verifying talkers who change speaking effort.

Figure 5.3 presents the percentage correct talker recognition scores for each talker as a function of noise level and speaking effort comparison. This illustration depicts the similar effect of noise upon all voices for the three speaking efforts. These particular noise levels were selected because they approximate 90% and 75% correct recognition. In addition, the percentage correct scores for each talker are developed from the same-talker pairs only. They reflect the noise effect on recognition.

Talker confusion matrices for the N-S, W-N, and W-S speaking effort comparisons at two noise levels (59.1 and 62.3 dB SPL) appear in Tables 4, 5, and 6. These matrices were constructed from the eleven listeners' combined discrimination data in order to examine the variability of the confusions among talkers and speaking effort comparisons, and to note the emergence of any special confusion patterns. The proportions in these tables show diffuse confusions among talkers and speaking effort comparisons. The highest and lowest proportions of talker confusions generally varied among speaking effort comparisons and noise levels. The

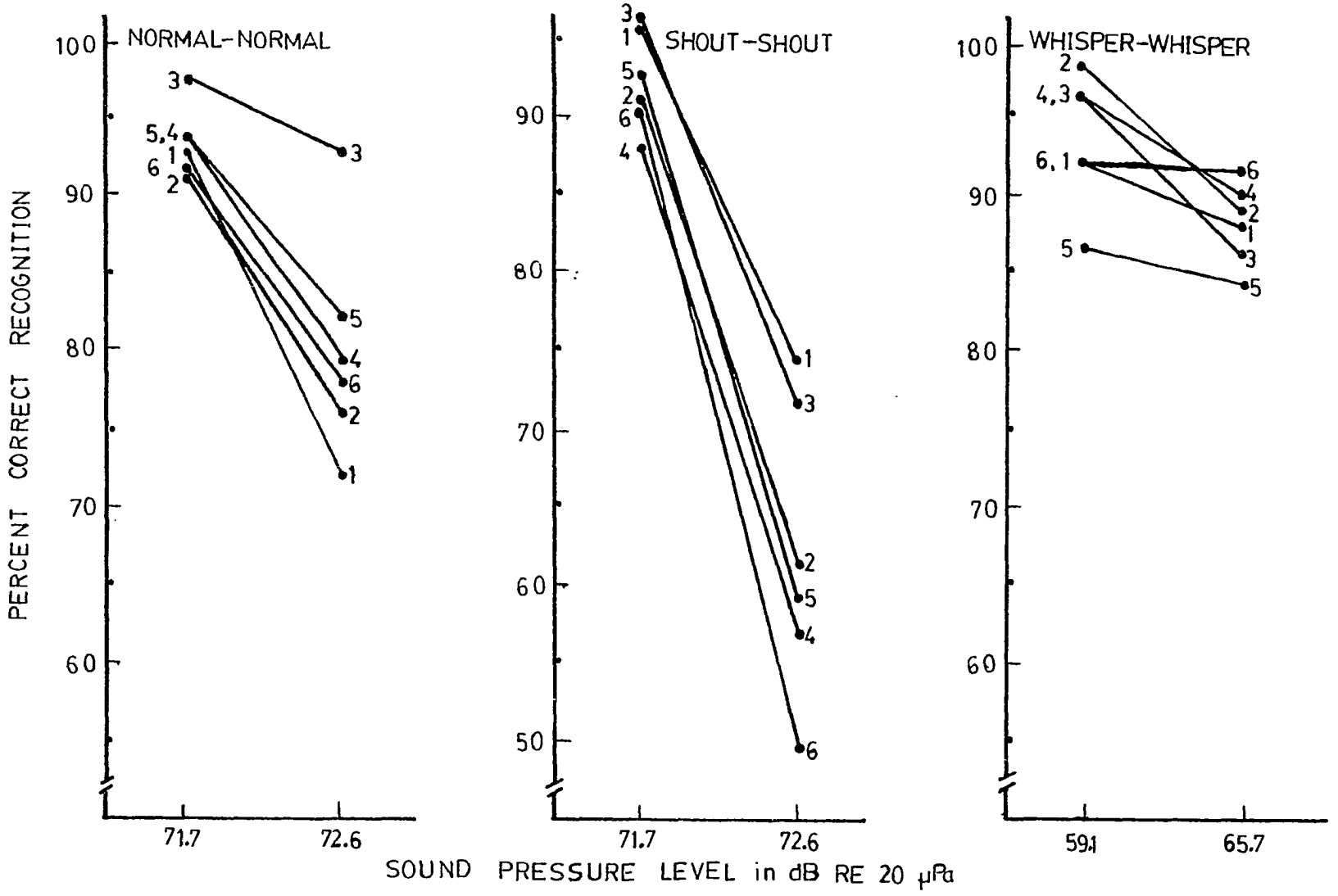


Fig. 5.3. Percent correct talker recognition scores for each talker (#1-#6) as a function of noise level and speaking effort comparison.

TABLE 4

TALKER CONFUSION MATRICES FOR THE NORMAL-SHOUT SPEAKING EFFORT COMPARISON AT TWO NOISE LEVELS (dB SPL)

Proportion of Errors for Eleven Listeners at 59.1 dB SPL and 62.3 dB SPL

		Talker								Talker					
		1	2	3	4	5	6		1	2	3	4	5	6
	1	.45	.18	.34	.31	.18	.27	1	.49	.25	.38	.40	.15	.31
	2		.51	.22	.34	.40	.36	2		.51	.29	.22	.29	.25
	3			.38	.40	.20	.27	3			.50	.27	.22	.40
Talker	4				.57	.25	.27	4				.61	.20	.27
	5					.60	.15	5					.69	.18
	6						.43	6						.54

NOTE: There are six talkers, identified by the numbers 1 through 6 on the horizontal and vertical borders of the confusion matrix. Total number of possible talker confusions for the matrix is 21. The listener's task was to indicate whether the two voices he heard, as identified on the horizontal and vertical axes of each matrix, were the same or different talkers. Each listener's response sheet was checked for incorrect discriminations for each of the 21 talker confusions. Then the total number of errors for the eleven listeners was computed and converted to a proportion for each of the 21 talker confusions.

Each talker-number combination on the matrix diagonal (e.g., 1-1, 2-2, 3-3, 4-4, 5-5, or 6-6) is composed of 110 same-talker pairs (i.e., one talker's voice was heard twice) or 110 possible discrimination errors. Each talker in these pairs was heard in the normal and shout speaking effort. There were 55 same-talker pairs presented in the normal-shout speaking effort order and 55 same-talker pairs presented in the shout-normal speaking effort order. Total number of possible discrimination errors on the matrix diagonal for eleven listeners is 660.

Each talker-number combination off the matrix diagonal (e.g., 1-3, 2-5, 4-6, etc.) is composed of 44 different-talker pairs (i.e., two talkers' voices were heard) or 44 possible discrimination errors. There were 22 different-talker pairs presented in the normal-shout speaking effort order and 22 different-talker pairs presented in the shout-normal speaking effort order. Total number of possible discrimination errors off the matrix diagonal for eleven listeners is 660.

TABLE 5

TALKER CONFUSION MATRICES FOR THE WHISPER-NORMAL SPEAKING EFFORT COMPARISON AT TWO NOISE LEVELS

Proportion of Errors for Eleven Listeners at 59.1 dB SPL and 62.3 dB SPL

		Talker					
		1	2	3	4	5	6
Talker	1	.61	.38	.34	.36	.31	.45
	2		.50	.20	.22	.40	.34
	3			.67	.27	.54	.52
	4				.47	.25	.36
	5					.49	.47
	6						.38

		Talker					
		1	2	3	4	5	6
	1	.57	.50	.34	.47	.22	.45
	2		.33	.34	.50	.43	.38
	3			.60	.36	.18	.40
	4				.46	.31	.36
	5					.39	.36
	6						.46

NOTE: There are six talkers, identified by the numbers 1 through 6 on the horizontal and vertical borders of the confusion matrix. Total number of possible talker confusions for the matrix is 21. The listener's task was to indicate whether the two voices he heard, as identified on the horizontal and vertical axes of each matrix, were the same or different talkers. Each listener's response sheet was checked for incorrect discriminations for each of the 21 talker confusions. Then the total number of errors for the eleven listeners was computed and converted to a proportion for each of the 21 talker confusions.

Each talker-number combination on the matrix diagonal (e.g., 1-1, 2-2, 3-3, 4-4, 5-5, or 6-6) is composed of 110 same-talker pairs (i.e., one talker's voice was heard twice) or 110 possible discrimination errors. Each talker in these pairs was heard in the normal and whisper speaking effort. There were 55 same-talker pairs presented in the whisper-normal speaking effort order and 55 same-talker pairs presented in the normal-whisper speaking effort order. Total number of possible discrimination errors on the matrix diagonal for eleven listeners is 660.

Each talker-number combination off the matrix diagonal (e.g., 1-3, 2-5, 4-6, etc.) is composed of 44 different-talker pairs (i.e., two talkers' voices were heard) or 44 possible discrimination errors. There were 22 different-talker pairs presented in the whisper-normal speaking effort order and 22 different-talker pairs presented in the normal-whisper speaking effort order. Total number of possible discrimination errors off the matrix diagonal for eleven listeners is 660.

TABLE 6

TALKER CONFUSION MATRICES FOR THE WHISPER-SHOUT SPEAKING EFFORT COMPARISON AT TWO NOISE LEVELS

Proportion of Errors for Eleven Listeners at 59.1 dB SPL and 62.3 dB SPL

		Talker								Talker					
		1	2	3	4	5	6		1	2	3	4	5	6
Talker	1	.60	.43	.38	.31	.47	.34	1	.58	.36	.34	.36	.47	.34
	2		.42	.45	.29	.31	.27	2		.37	.52	.31	.38	.25
	3			.70	.50	.38	.43	3			.67	.45	.52	.50
	4				.57	.25	.36	4				.47	.29	.40
	5					.61	.38	5					.50	.38
	6						.59	6						.50

NOTE: There are six talkers, identified by the numbers 1 through 6 on the horizontal and vertical borders of the confusion matrix. Total number of possible talker confusions for the matrix is 21. The listener's task was to indicate whether the two voices he heard, as identified on the horizontal and vertical axes of each matrix, were the same or different talkers. Each listener's response sheet was checked for incorrect discriminations for each of the 21 talker confusions. Then the total number of errors for the eleven listeners was computed and converted to a proportion for each of the 21 talker confusions.

Each talker-number combination on the matrix diagonal (e.g., 1-1, 2-2, 3-3, 4-4, 5-5, or 6-6) is composed of 110 same-talker pairs (i.e., one talker's voice was heard twice) or 110 possible discrimination errors. Each talker in these pairs was heard in the whisper and shout speaking effort. There were 55 same-talker pairs presented in the whisper-shout speaking effort order and 55 same-talker pairs presented in the shout-whisper speaking effort order. Total number of possible discrimination errors on the matrix diagonal for eleven listeners is 660.

Each talker-number combination off the matrix diagonal (e.g., 1-3, 2-5, 4-6, etc.) is composed of 44 different-talker pairs (i.e., two talkers' voices were heard) or 44 possible discrimination errors. There were 22 different-talker pairs presented in the whisper-shout speaking effort order and 22 different-talker pairs presented in the shout-whisper speaking effort order. Total number of possible discrimination errors off the matrix diagonal for eleven listeners is 660.

results of a *t* test for correlated measures between noise levels for each speaking effort comparison demonstrated that the twenty-one talker confusion scores did not significantly differ between noise levels for each speaking effort comparison. The obtained *t* values for the N-S, W-N, and W-S comparisons were .4790, .1965, and .0669, respectively. These values were less than the critical value of $t(20) = 2.086$, $p = .05$. In addition, the twenty-one talker confusion scores did not significantly change among speaking effort comparisons at 59.1 and 62.3 dB SPL. The results of a one-way within-subjects analysis of variance indicated no significant speaking effort comparison effect for 59.1 dB SPL and 62.3 dB SPL, computed values of $F(2,40) = 2.734$ and 3.150, in order, $p > .05$. The primary pattern to emerge was the evidence for increased confusions for same-talker relative to different-talker pairs for all the speaking effort comparisons at most noise levels. This pattern contrasts with the W-W speaking effort comparison results. For that condition errors increased for different-talker relative to same-talker pairs at 59.1 dB SPL and 65.7 dB SPL. Table 7 shows this contrast and gives the results of *t* tests for correlated means. This table also shows that there were no significant differences between the number of same-talker and different-talker errors for the normal-normal and shout-shout speaking effort comparisons. It is speculated that there were decreased errors for different-talker relative to same-talker pairs for the N-S, W-N, and W-S speaking efforts because of greater dissimilarity

TABLE 7

MEAN NUMBER OF SAME-TALKER VERSUS DIFFERENT-TALKER ERRORS

Speaking Effort Comparison	Noise (dB SPL)	Mean Number of Errors: Same Talker	Mean Number of Errors: Different Talker	t_{obt}
Normal-Normal	71.7	4.00	6.50	1.03
	72.6	11.81	12.54	1.02
Shout-Shout	71.7	4.50	6.90	1.24
	72.6	22.54	17.72	1.12
Whisper-Whisper	59.1	3.72	8.45	2.51*
	65.7	7.09	17.63	3.41**
Normal-Shout	59.1	30.00	16.50	2.75*
	62.3	33.40	16.27	3.61**
Whisper-Normal	59.1	31.20	21.60	3.17**
	62.3	28.09	22.50	1.43
Whisper-Shout	59.1	35.10	22.60	2.90*
	62.3	31.50	23.70	3.31**

NOTE: Maximum number of pairs (or errors) is 1320 for each SPL and effort comparison (660 same-talker and 660 different-talker).

*$p<.05$, $t(10) = 2.228$.
**$p<.01$, $t(10) = 3.169$.

of the comparisons (in other words, an easier discrimination task). Conversely, it was similarity of utterances that led to decreased errors for same-talker relative to different-talker pairs for the W-W speaking effort. And finally, there was no statistically significant difference between speaking effort presentation orders within same-talker pairs for the N-S, W-N, and W-S speaking effort comparisons. Table 8 gives the total number of errors associated with each presentation order and talker at two noise levels. None of the *t* tests for correlated means, presented in this table, was significant.

Discussion

The objectives of this research were to answer three questions: (1) Will a constant speech-to-noise ratio result in similar recognition scores for all levels of speaking efforts; (2) Is there an optimum speaking effort for speaker recognition with high noise levels; and (3) Will a constant speech-to-noise ratio result in similar recognition scores for normal-shout, whisper-normal, and whisper-shout speaking effort comparisons? The results indicate the answers to these questions to be: (1) A constant speech-to-noise ratio does not result in similar recognition scores for the N-N S-S, and W-W speaking effort comparisons. Indeed, whispering talkers are difficult to recognize in noise compared to talkers using the normal or shout effort. (2) The normal

TABLE 8

TOTAL NUMBER OF ERRORS AS A FUNCTION OF SPEAKING EFFORT PRESENTATION ORDER AND TALKER AT TWO NOISE LEVELS

Talker	N-S	S-N	W-N	N-W	W-S	S-W
			59.1 dB SPL			
1	24	26	30	38	33	33
2	30	27	27	29	21	26
3	22	20	40	34	38	40
4	26	37	25	27	33	30
5	36	30	21	33	36	32
6	20	28	22	20	34	31
Mean	26	28	27	30	32	32
$t$obtained	.613		1.00		.349	
			62.3 dB SPL			
1	26	28	33	30	30	34
2	30	27	16	21	22	19
3	25	30	37	30	30	44
4	30	38	30	21	30	22
5	37	39	27	16	26	29
6	32	28	27	24	29	26
Mean	30	31	28	23	27	29
$t$obtained	.889		2.00		.430	

NOTE: Maximum number of pairs (or errors) is 110 for each talker and speaking effort comparison (N-S, W-N, or W-S) at one noise level (55 pairs for each presentation order). The 110 pairs are same-talker comparisons only.

*p<.05, $t(5)$ = 2.571.
**p<.01, $t(5)$ = 4.032.

speaking effort is the most favorable condition for speaker recognition with high broad-band noise levels. And (3) A constant speech-to-noise ratio results in similar recognition scores for the N-S, W-N, and W-S speaking effort comparisons. Equally important, these poor recognition scores for all noise levels are in sharp contrast to the results for the N-N, S-S, and W-W speaking effort comparisons.

Past speaker recognition research has not investigated the effect of speaking effort in noise for human or machine analysis. Although one study has reported the effect of noise upon speaker recognition for man (Clarke, Becker, and Nixon, 1966) and one study has reported the effect of noise upon speaker recognition for machine (Doddington and Hydrick, 1975), a direct comparison of the present results with these studies is impossible because speech level, type of noise, and speech material were either not specified or differed from this experiment. Nevertheless, there are other studies which clarify the present findings.

The main conclusion to emerge in this study was that listeners cannot recognize talkers who change their speaking effort. The work of Glave and Reitveld (1975), Thomas (1969), and Meyer-Eppler (1957) indicate changing spectral information to be the special clue for the perception of speaking efforts. In the case of shouting, Glave and Rietveld examined a spectral analysis of vowels spoken with and without effort. They discovered the spectra of vowels

spoken with effort have much more energy in the higher frequency regions and show a decrease in energy at the lower end of the frequency scale. Other researchers have noted the increase in energy for the middle- and high-frequency components of the speech spectrum; however, these reports indicate that low frequency components increase in energy also (Licklider, Hawley, and Walkling, 1955; Pearsons, Bennett, and Fidell, 1976). Relatively speaking, low frequency changes are not as great as the middle- and high-frequency changes. In order to determine if the dependence of loudness upon physiological effort could be explained in terms of the spectral structure of the signal, Glave and Rietveld compared empirical effort and noneffort loudness data with calculated effort and noneffort loudness of the same stimuli. Intensity, fundamental frequency, and duration were held constant in their experiments. Loudness was computed according to Zwicker's model of loudness calculation, which is based upon the functional correspondence between masking and loudness, and which is substantiated by data on the critical frequency bandwidths of the ear (Kryter, 1970). With respect to the relative level, listener loudness and calculated loudness agreed in effort-dependent loudness differences. The comparative loudness data suggested that effort-dependent speech loudness changes were due to pure spectral changes. Spectral changes are also present for whispered speech. Meyer-Eppler (1957) has shown that changes

of pitch in normal speech are replaced in whispered speech by shifts of some formant regions accompanied by added noise between the higher formants. A spectrographic analysis of whispered vowels and words revealed these results. In addition to this study, Thomas (1969) presents evidence of spectral change for whispered vowels. He notes the approximately equal amplitudes for formants one and two for all vowel samples. These studies suggest the main conclusion of this experiment can be explained in terms of changing spectral information. That is, if the listeners attended to spectral clues for making their talker recognition judgments, then one would predict poor recognition performance for those comparisons (N-S, W-N, and W-S) in which the talker changed his speaking effort. By the same reasoning, good recognition scores would be expected for those comparisons in which the talker did not change his speaking effort.

Although this explanation provides a plausible account for the present results, it is necessary to consider two alternative explanations. First, is the main result due to a listener response bias? The term response bias (or response criterion) refers to listeners' tendencies to favor one response over the other, independent of stimulus discriminability. According to this hypothesis, a listener would either respond with many different-talker differentiations (which represents a criterion that attempts to avoid errors) or respond with many same-talker judgments. One

would expect, then, to find poor recognition scores for the N-S, W-N, and W-S comparisons for all noise levels with either bias operating. The effect of response bias for these comparisons at the noise level of 62.3 dB SPL was assessed according to a method suggested by McNicol (1972, p. 126). The aim of this method was to find the point on the listener's rating scale at which he was equally disposed to same-talker and different-talker responses. This non-parametric analysis of response bias indicated there was no preference for same-talker or different-talker judgments. In other words, listeners averaged either a 3.7 or 3.5 beta measure for the six-point rating scale for each of the speaking effort comparisons. The data is given in Table 9. As a result it can be concluded that a response bias is not responsible for the poor recognition for the N-S, W-N, and W-S speaking effort comparisons.

A more viable alternative explanation for the main results is that the listeners attended to pitch differences among the speaking efforts for talker recognition judgments. Fundamental voice frequency information was not held constant in this study. Ladefoged and McKinney (1963) and Hirano, Ohala, and Vennard (1969) present evidence that the fundamental frequency increases with vowel level. Atkinson (1976) suggests that there is as much intraspeaker variability for fundamental voice frequency as interspeaker. If perceptually different fundamental frequencies are present in the shouts

TABLE 9

BETA SCORES FOR EACH LISTENER AT 62.3 dB SPL

Listener	Speaking Effort Comparison		
	N-S	W-N	W-S
1	4.27	3.26	3.29
2	4.63	4.20	4.61
3	3.64	3.88	4.00
4	3.25	3.30	3.81
5	1.91	1.87	2.56
6	3.62	4.19	3.82
7	3.31	4.05	3.11
8	5.40	5.14	5.22
9	4.38	3.40	2.42
10	3.51	3.21	3.08
11	3.64	2.77	2.59
Mean	3.77	3.57	3.50
Std. Dev.	.89	.86	.88

NOTE: Beta is defined as the rating scale category at which the Prob(same-talker response|same-talker pair) + Prob(same-talker response|different-talker pair) = 1. Small beta scores indicate a preference for same-talker responses and large beta scores a preference for different-talker responses.

and whispers as compared to the normal speaking effort, and the listeners utilized relative pitch information for talker differentiations, then it is likely there would be poor talker recognition for the N-S, W-N, and W-S speaking effort comparisons. Matsumoto et al. (1973) demonstrate that the relative contribution of the "mean fundamental pitch frequency" to the perception of the personal quality of voice is greater than that of formant frequencies and glottal source spectrum slope. Future studies should be directed towards illuminating the relative importance of spectral bandwidth (or other kinds of spectral information) and pitch changes to the listener for discriminating the dynamic nature of speaker recognition.

A second and less surprising finding of the present research was that talker recognition in noise is poorest for the whisper speaking effort comparison. Two previous studies (Pollack, Pickett, and Sumby, 1954; Williams, 1971) have shown poor talker identification for whispered speech. According to Pollack, Pickett, and Sumby, whispering talkers are difficult to identify because their voices sound alike in average pitch. While this is true, there is an additional explanation for the poorer talker recognition of the whisper-whisper speaking effort comparison in noise compared to the N-N and S-S speaking efforts for this study. This explanation focuses upon the consonant-vowel intensity ratio within the syllable associated with each speaking effort and its

effect on the intelligibility of speech sounds in noise. For the normal conversational effort the level of the average consonant is substantially lower than that of the average vowel, approximately -15 dB. This disparity is smaller for whispered speech syllables (-7 dB) and approximately the same for shouted speech syllables (-14 dB) (Fairbanks and Miron, 1957). It also interacts with sex of the speaker. In other words, the consonant range for women was about 4 dB smaller than that of the men. The intelligibility of words heard in noise and spoken with different amounts of effort was measured by Pickett in 1956. The results show less than 5% deterioration in intelligibility over the range from a moderately low voice to a very loud voice. Beyond these points intelligibility decreases abruptly. In an important analysis of listener errors he determined the intelligibility of different sound-producing movements associated with the extremes of speaking effort. This analysis showed shouting reduces the intelligibility of the initial and final parts of the syllable while whispering degrades the intelligibility of all parts of the syllable. It is likely, then, that the consonant-vowel intensity ratio and associated spectral characteristics of whispers contributed to poorer recognition scores.

The real significance of the findings of this work lie with the implications for machine recognition of speakers. Lombard noted in 1911 that a speaker with normal hearing

unconsciously changes his voice level when the ambient noise level changes. During communication, the speaker makes these matches to compensate for changes in signal-to-noise ratio. He matches his voice level inversely to apparent changes in signal level, and directly to changes in the noise level. The effort to compensate for these changes in the signal-to-noise ratio, or to match directly changes in the noise level, typically falls about halfway short (in decibel units). This is probably because a speaker considers that he has doubled his own vocal level in half as many decibels as it takes to double the loudness of the signal or the noise (Lane, Tranel, and Sisson, 1970; Lane and Tranel, 1971). The Lombard, or voice reflex, effect results in speech with characteristics different from those of speech that is produced with the normal speaking effort. The present research results for the normal-shout speaking effort comparison show that this change of characteristics, while an effective way to combat noise interference and improve speech intelligibility, destroys speaker recognition for humans. The obvious question suggested by the rather surprising listener performance is, "Will a machine experience the same degree of difficulty verifying a talker who raises his speaking level in a noisy environment?" Indeed, is not the speaking effort change as serious a problem as noise for effective computer-based speaker recognition systems? Several directions for further work are evident. These are identification of the

speech parameters which accurately verify talkers who change speaking effort, determination of the talking and noise level ranges at which various machine verification systems perform effectively, and differentiation of individual and sex variations associated with speaking effort changes in noise. Finally, studies should pursue whether a whisper speaking effort or speaker recognition task or both would make an efficient and sensitive speech evaluation test for high quality speech systems. According to Webster (1978), no intelligibility test, of and by itself, is difficult enough to be an efficient test for evaluating highly intelligible systems such as vocoders or speech compression systems. Even 1000 nonsense syllables have an intelligibility of greater than 90%.

In summary, the main result to emerge in this study was that listeners cannot recognize talkers who change their speaking effort. A second and less surprising finding was that talker recognition in noise is poorest for the whisper speaking effort comparison. Explanations for these results suggest the unique spectral information and consonant-vowel intensity ratio within the syllable associated with each speaking effort to be the special cues for not discriminating the dynamic nature of speaker recognition. The present results also imply the Lombard reflex to be a serious problem for effective computer-based speaker recognition systems. Future research must identify those speech parameters which accurately verify talkers who change speaking effort.

LITERATURE CITED

Alpert, M., Kurtzberg, R.L., Pilot, M., and Friedhoff, A.J. (1963). "Comparison of the spectra of the voices of twins." J. Acoust. Soc. Am. 35, 1877(A).

Anon. (1969). "American National Standards Institute Specifications for Audiometers," ANSI S3.6-1969. American National Standards Institute, New York.

Atal, B.S. (1972). Text-independent Speaker Recognition. Paper presented at the Eighty-Third Meeting of the Acoustical Society of America, Buffalo, New York, April 18-21.

Atal, B.S. (1972). "Automatic speaker recognition based on pitch contours." J. Acoust. Soc. Am. 52, 1687-1697.

Atal, B.S. (1974). "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification." J. Acoust. Soc. Am. 55, 1304-1312.

Atal, B.S. (1975). "Linear prediction of speech--recent advances with applications to speech analysis," in Speech Recognition, edited by D.R. Reddy (Academic, New York).

Atal, B.S. (1976). "Automatic recognition of speakers from their voices." Proc. IEEE 64, 460-473.

Atal, B.S., and Hanauer, S.L. (1971). "Speech analysis and synthesis by linear prediction of the speech wave." J. Acoust. Soc. Am. 50, 637-655.

Atal, B.S., and Schroeder, M.R. (1978). "Linear prediction analysis of speech based on a pole-zero representation." J. Acoust. Soc. Am. 64, 1310-1318.

Atkinson, J.E. (1976). "Inter- and intraspeaker variability in fundamental voice frequency." J. Acoust. Soc. Am. 60, 440-445.

Bennett, S., and Weinberg, B. (1979). "Acoustic correlates of perceived sexual identity in preadolescent children's voices." J. Acoust. Soc. Am. 66, 989-1000.

Boll, S.F. (1979). "Suppression of acoustic noise in speech using spectral subtraction." IEEE Trans. Acoust. Speech and Signal Processing 27, 113-120.

Bricker, P.D., Gnanadesikan, R., Mathews, M.V., Pruzansky, S., Tukey, P.A., Wachter, K.W., and Warner, J.L. (1971). "Statistical techniques for talker identification." The Bell System Technical Journal 50, 1427-1454.

Bricker, P.D., and Pruzansky, S. (1966). "Effects of stimulus content and duration on talker identification." J. Acoust. Soc. Am. 40, 1441-1449.

Bricker, P.D., and Pruzansky, S. (1976). "Speaker recognition," in Contemporary Issues in Experimental Phonetics, edited by N.J. Lass (Charles C. Thomas, Springfield, Illinois).

Brown, R. (1979). "Memory and decision in speaker recognition." Int. J. Man-Machine Studies 11, 729-742.

Carbonell, J.R., Grignetti, M.C., Stevens, K.N., Williams, C.E., and Woods, B. (1965). Speaker Authentication Techniques (BBN Rep. 1296) (Bolt Beranek & Newman, Cambridge, MA).

Chang, C.Y. (1973). "Dynamic programming as applied to feature subset selection in a pattern recognition system." IEEE Trans. Syst. Man and Cybern. 3, 166-171.

Cheung, R.S., and Eisenstein, B.A. (1978). "Feature selection via dynamic programming for text-independent speaker identification." IEEE Trans. Acoust. Speech and Signal Processing 26, 397-409.

Clarke, R.R., and Becker, R.W. (1969). "Comparison of techniques for discriminating among talkers." J. Speech Hear. Res. 12, 747-761.

Clarke, R.R., Becker, R.W., and Nixon, J.C. (1966). Characteristics that Determine Speaker Recognition (ESD-TR-66-636) (Electronic Systems Division, Air Force Systems Command, Hanscom Field).

Coleman, R.O. (1971). "Male and female voice quality and its relationship to vowel formant frequencies." J. Speech Hear. Res. 14, 565-577.

Coleman, R.O. (1973). "Speaker identification in the absence of intersubject differences in glottal source characteristics." J. Acoust. Soc. Am. 53, 1741-1743.

Compton, A.J. (1963). "Effects of filtering and vocal duration upon the identification of speakers, aurally." J. Acoust. Soc. Am. 35, 1748-1752.

Das, K.K., and Mohn, W.S. (1971). "A scheme for speech processing in automatic speaker verification." IEEE Trans. Audio Electroacoust. 19, 32-43.

Doddington, G.R. (1970). A Method of Speaker Verification. Paper presented at the Eightieth Meeting of the Acoustical Society of America, Houston, Texas, November 3-8.

Doddington, G.R. (1974). Speaker Verification (RADC-TR-74-179) (Rome Air Development Center, Griffiss AFB, NY) (NTIS No. AD 785 135).

Doddington, G.R., and Hydrick, B.M. (1975). Speaker Verification II (RADC-TR-75-274) (Rome Air Development Center, Griffiss AFB, NY) (NTIS No. AD-A018 901).

Doehring, D.G., and Ross, R.W. (1972). "Voice recognition by matching to sample." J. Psycholinguistic Res. 1, 233-242.

Doherty, E.T. (1976). "An evaluation of selected acoustic parameters for use in speaker identification." J. Phonetics 4, 321-326.

Dukiewicz, L. (1970). "Frequency-band dependence of speaker identification," in Speech Analysis and Synthesis (Vol. II), edited by W. Jassem (Institute of Fundamental Technical Research, Polish Academy of Sciences, Warsaw, Poland).

Elman, J.L. (1979). "Perceptual origins of the phoneme boundary effect and selective adaptation to speech: A signal detection theory analysis." J. Acoust. Soc. Am. 65, 190-207.

Endres, W., Bamback, W., and Flosser, G. (1971). "Voice spectrograms as a function of age, voice disguise, and voice imitation." J. Acoust. Soc. Am. 49, 1842-1847.

Fairbanks, G., and Miron, M.S. (1957). "Effects of vocal effort upon the consonant-to-vowel ratio within the syllable." J. Acoust. Soc. Am. 29, 621-626.

Fant, G. (1960). Acoustic Theory of Speech Production (Mouton, The Hague).

Flanagan, J.L. (1958). "Some properties of the glottal sound source." J. Speech Hear. Res. 1, 99-116.

Flanagan, J.L. (1972). Speech Analysis, Synthesis, and Perception (Springer-Verlag, NY) 2nd ed.

Flanagan, J.L. (1976). "Computers that talk and listen: Man-machine communication by voice." Proc. IEEE 64, 405-415.

Furui, S., Itakura, F., and Saito, S. (1972). "Talker recognition by long-time averaged speech spectrum." Electronics and Communications in Japan 55-A, 54-61.

Furui, S., Itakura, F., and Saito, S. (1975). "Personal information in the long-time averaged speech spectrum." Rev. Elec. Commun. Lab 23, 1133-1141.

Garvin, P.L., and Ladefoged, P. (1963). "Speaker identification and message identification in speech recognition." Phonetica 9, 193-199.

Glave, R.D., and Rietveld, A.C.M. (1975). "Is the effort dependence of speech loudness explicable on the basis of acoustical cues?" J. Acoust. Soc. Am. 58, 875-879.

Glenn, J.W., and Kleiner, N. (1968). "Speaker identification based on nasal phonation." J. Acoust. Soc. Am. 43, 368-372.

Goldstein, U.G. (1976). "Speaker-identifying features based on formant tracks." J. Acoust. Soc. Am. 59, 176-182.

Gupta, V.N., Bryan, J.K., and Gowdy, J.N. (1978). "A speaker-independent speech-recognition system based on linear prediction." IEEE Trans. Acoust. Speech and Signal Processing 26, 27-33.

Haggard, M., and Summerfield, Q. (1979). "Perceptual and memory factors in simulated machine-aided speaker verification." Int. J. Man-Machine Studies 11, 717-728.

Hair, G.G., and Rekieta, T.W. (1972). Techniques for Objective Speaker Identification. Paper presented at the Eighty-Fourth meeting of the Acoustical Society of America, Miami Beach, Florida, November 28-December 1.

Hargreaves, W.A., and Starkweather, J.A. (1963). "Recognition of speaker identity." Language and Speech 6, 63-67.

Hecker, M.H.L. (1971). "Speaker recognition: An interpretive survey of the literature." ASHA Monographs 16.

Hecker, M.H.L., and Guttman, N. (1967). "Survey of methods for measuring speech quality." J. Audio Engr. Soc. 15, 400-403.

Hecker, M.H.L., and Williams, C.E. (1965). "On the interrelation among speech quality, intelligibility, and speaker identifiability." Proc. 5th Int. Cong. Acoust., Liege, 1-4.

Hecker, M.H.L., Stevens, K.N., von Bismarck, G., and Williams, C.E. (1968). "Manifestations of task-induced stress in the acoustic speech signal." J. Acoust. Soc. Am. 44, 993-1001.

Hemdal, J.F. (1967). "Some results from the normalization of speaker differences in a mechanical vowel recognizer." J. Acoust. Soc. Am. 41, 1594(A).

Hirano, M., Ohala, J., and Vennard, W. (1969). "The function of laryngeal muscles in regulating fundamental frequency and intensity of phonation." J. Speech Hear. Res. 12, 616-628.

Hollien, H., and Majewski, W. (1977). "Speaker identification by long-term spectra under normal and distorted speech conditions." J. Acoust. Soc. Am. 62, 975-979.

Holmgren, G.L. (1963). Speaker Recognition (AFCRL-63-119) (Air Force Cambridge Research Lab, Office of Aerospace Research, Bedford, MA).

Holmgren, G.L. (1967). "Physical and psychological correlates of speaker recognition." J. Speech Hear. Res. 10, 57-66.

Ingemann, F. (1968). "Identification of the speaker's sex from voiceless fricatives." J. Acoust. Soc. Am. 44, 1142-1144.

Itakura, F. (1975). "Minimum prediction residual principal applied to speech recognition." IEEE Trans. Acoust. Speech and Signal Processing 23, 67-72.

Jelinek, F. (1976). "Continuous speech recognition by statistical methods." Proc. IEEE 64, 532-556.

Kashyap, L.L. (1976). "Speaker recognition from an unknown utterance and speaker-speech interaction." IEEE Trans. Acoust. Speech and Signal Processing 24, 481-488.

Kersta, L.G. (1965). "Environmental influence on the speech of family members shown by spectrographic speech matching." J. Acoust. Soc. Am. 38, 935(A).

Kosiel, U. (1973). "Statistical analysis of speaker-dependent differences in long-term average spectrum of Polish vowels," in Speech Analysis and Synthesis (Vol. III), edited by W. Jassem (Institute of Fundamental Technical Research, Polish Academy of Sciences, Warsaw, Poland).

Kryter, K.D. (1970). The Effects of Noise on Man (Academic, New York).

Ladefoged, P., and Broadbent, D.E. (1957). "Information conveyed by vowels." J. Acoust. Soc. Am. 29, 98-104.

Ladefoged, P., and McKinney, N.P. (1963). "Loudness, sound pressure, and subglottal pressure in speech." J. Acoust. Soc. Am. 35, 454-460.

Lane, L., and Tranel, B. (1971). "The Lombard sign and the role of hearing in speech." J. Speech Hear. Res. 14, 677-709.

Lane, H., Tranel, B., and Sisson, C. (1970). "Regulation of voice communication by sensory dynamics." J. Acoust. Soc. Am. 47, 618-624.

Lehiste, I., and Meltzer, D. (1973). "Vowel and speaker identification in natural and synthetic speech." Language and Speech 16, 356-364.

Li, K.-P., Dammann, J.E., and Chapman, W.D. (1966). "Experimental studies in speaker verification, using an adaptive system." J. Acoust. Soc. Am. 40, 966-978.

Li, K.-P., and Hughes, G.W. (1972). Talker Differences as They Appear in Correlation Matrices from Spectra of Continuous Speech. Paper presented at the Eighty-Fourth Meeting of the Acoustical Society of America, Miami Beach, Florida, November 28-December 1.

Li, K.-P., and Hughes, G.W. (1974). "Talker differences as they appear in correlation matrices of continuous speech spectra." J. Acoust. Soc. Am. 55, 833-837.

Li, K.-P., Hughes, G.W., and House, A.S. (1969). "Correlation characteristics and dimensionality of speech spectra." J. Acoust. Soc. Am. 46, 1019-1025.

Licklider, J.C.R. (1960). "Man-computer symbiosis." IRE Trans. Hum. Factors Electron. 1, 4-11.

Licklider, J.C.R., Hawley, M.E., and Walkling, R.A. (1955). "Influence of variations of speech intensity and other factors upon the speech spectrum." J. Acoust. Soc. Am. 27, 207(A).

Lombard, E. (1911). "Le signe de l'elevation de la voix." Ann. Maladies Oreille, Larynx, Nez, Pharynx 37, 101-119.

Luck, J.E. (1969). "Automatic speaker verification using cepstral measurements." J. Acoust. Soc. Am. 46, 1026-1030.

Lummis, R.C. (1973). "Speaker verification by computer using speech intensity for temporal registration." IEEE Trans. Audio Electroacoust. 21, 165-173.

Lummis, R.C., and Rosenberg, A.E. (1971). Test of an Automatic Speaker Verification Method with Intensively Trained Professional Mimics. Paper presented at the Eighty-Second Meeting of the Acoustical Society of America, Denver, Colorado, October 19-22.

McDermott, B., Scagliola, C., and Goodman, D. (1978). "Perceptual and objective evaluation of speech processed by adaptive differential PCM." The Bell System Technical Journal 57, 1597-1617.

McGehee, F. (1937). "The reliability of the identification of the human voice." J. General Psychology 17, 249-271.

McGehee, F. (1944). "An experimental study in voice recognition." J. General Psychology 31, 53-65.

McGonegal, C.A., Rabiner, L.R., and McDermott, B.J. (1978). "Speaker verification by human listeners over several speech transmissions systems." The Bell System Technical Journal 57, 2887-2900.

McGonegal, C.A., Rosenberg, A.E., and Rabiner, L.R. (1979). "The effects of several transmission systems on an automatic speaker verification system." The Bell System Technical Journal 58, 2071-2087.

McNicol, D. (1972). A Primer of Signal Detection Theory (George Allen & Unwin, London).

Makhoul, J.I. (1975). "Linear prediction: A tutorial review." Proc. IEEE 63, 561-580.

Markel, J.D., and Davis, S.B. (1979). "Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base." IEEE Trans. Acoust. Speech and Signal Processing 27, 74-82.

Martin, T.B. (1976). "Practical applications of voice input to machines." Proc. IEEE 64, 487-501.

Matsumoto, H., Hiki, S., Sone, T., and Nimura, T. (1973). "Multidimensional representation of personal quality and its acoustical correlates." IEEE Trans. Audio Electroacoust. 21, 428-436.

Meyer-Eppler, W. (1957). "Realization of prosodic features in whispered speech." J. Acoust. Soc. Am. 29, 104-106.

Miller, J.E. (1964). Decapitation and Recapitation, a Study of Voice Quality. Paper presented at the Sixty-Eighth Meeting of the Acoustical Society of America, Austin, Texas, October 21-24.

Mohn, W.S. (1971). "Two statistical feature evaluation techniques applied to speaker identification." IEEE Trans. Computers 20, 979-987.

Murry, T., and Cort, S. (1971). "Aural identification of children's voices." J. Auditory Res. 11, 260-262.

Mysak, E.D. (1959). "Pitch and duration characteristics of older males." J. Speech Hear. Res. 2, 46-54.

Nakatani, L.H., and Dukes, K.D. (1973). "A sensitive test of speech communication quality." J. Acoust. Soc. Am. 53, 1083-1092.

Neeley, R.B., and Reddy, D.R. (1971). "Speech recognition in the presence of noise." Proc. 7th Int. Cong. Acoust., Budapest, 3, 177-180.

Parker, F. (1977). "Distinctive features and acoustic cues." J. Acoust. Soc. Am. 62, 1051-1054.

Parsons, T.W. (1976). "Separation of speech from interfering speech by means of harmonic selection." J. Acoust. Soc. Am. 60, 911-918.

Parsons, T.W., and Weiss, M. (1975). Enhancing Intelligibility of Speech in Noisy or Multi-talker Environments (RADC-TR-75-155) (Rome Air Development Center, Griffiss AFB, NY) (NTIS No. AD-A013 767).

Pearsons, K.S., Bennett, R.L., and Fidell, S. (1976). Speech Levels in Various Environments (BBN Rep. 3281) (Bolt Beranek & Newman, Cambridge, MA).

Peters, R.W. (1954). Studies in Extra Messages: Listener Identification of Speakers' Voices Under Conditions of Certain Restrictions Imposed Upon the Voice Signal (NM 001-064-01, Rep. 30) (U.S. Naval School of Aviation Medicine, Pensacola, Florida).

Peters, R.W. (1956). Studies in Extra Messages: The Effect of Various Modifications of the Voice Signal Upon the Ability of Listeners to Identify Speakers' Voices (NM 001-104-500, Rep. 61) (U.S. Naval School of Aviation Medicine, Pensacola, Florida).

Pickett, J.M. (1956). "Effects of vocal force on the intelligibility of speech sounds." J. Acoust. Soc. Am. 28, 902-905.

Pollack, I., Pickett, J.M., and Sumby, W.H. (1954). "On the identification of speakers by voice." J. Acoust. Soc. Am. 26, 403-406.

Pruzansky, S. (1963). "Pattern-matching procedure for automatic talker recognition." J. Acoust. Soc. Am. 35, 354-358.

Pruzansky, S., and Mathews, M.V. (1964). "Talker-recognition procedure based on analysis of variance." J. Acoust. Soc. Am. 36, 2041-2047.

Ptacek, P.H., and Sander, E.K. (1966). "Age recognition from voice." J. Speech Hear. Res. 9, 273-277.

Rabiner, L.R., and Wilpon, J.G. (1979). "Application of clustering techniques to speaker-trained isolated word recognition." The Bell System Technical Journal 58, 2217-2233.

Ramishvili, G.S. (1966). "Automatic voice recognition." Engineering Cybernetics 5, 84-90.

Reich, A.R., and Duke, J.E. (1979). "Effects of selected vocal disguises upon speaker identification by listening." J. Acoust. Soc. Am. 66, 1023-1029.

Richards, A.M. (1975). "Most comfortable loudness for pure tones and speech in the presence of masking noise." J. Speech Hear. Res. 18, 498-505.

Rosenberg, A.E. (1973). "Listener performance in speaker verification tasks." IEEE Trans. Audio Electroacoust. 21, 221-225.

Rosenberg, A.E. (1976). "Automatic speaker verification: A review." Proc. IEEE 64, 475-487.

Rosenberg, A.E., and Sambur, M. (1975). "New techniques for automatic speaker verification." IEEE Trans. Acoust. Speech and Signal Processing 23, 169-176.

Rothauser, E.H., Urbanek, G.E., and Pachl, W.P. (1968). "Isopreference method for speech evaluation." J. Acoust. Soc. Am. 44, 408-418.

Ryan, W.J., and Burk, K.W. (1972). Predictors of Age in the Male Voice. Paper presented at the Eighty-Fourth Meeting of the Acoustical Society of America, Miami Beach, Florida, November 28-December 1.

Sakoe, H., and Chiba, S. (1978). "Dynamic programming algorithm optimization for spoken word recognition." IEEE Trans. Acoust. Speech and Signal Processing 26, 43-49.

Sambur, M.R. (1975). "Selection of acoustic features for speaker identification." IEEE Trans. Acoust. Speech and Signal Processing 23, 176-182.

Sambur, M.R. (1976). "Speaker recognition using orthogonal linear prediction." IEEE Trans. Acoust. Speech and Signal Processing 24, 283-289.

Sambur, M.R., and Jayant, N.S. (1976). "LPC analysis-synthesis from speech inputs containing quantizing noise or additive white noise." IEEE Trans. Acoust. Speech and Signal Processing 24, 488-494.

Schwartz, M.R. (1968). "Identification of speaker sex from isolated, voiceless fricatives." J. Acoust. Soc. Am. 43, 1178-1179.

Schwartz, M.R., and Rine, H.E. (1968). "Identification of speaker sex from isolated, whispered vowels." J. Acoust. Soc. Am. 44, 1736-1737.

Shearme, J.N., and Holmes, J.N. (1959). "An experiment concerning the recognition of voices." Language and Speech 2, 123-131.

Shipp, T., and Hollien, H. (1969). "Perception of the aging male voice." J. Speech Hear. Res. 12, 703-710.

Shoup, J.E., and Pfeifer, L.L. (1976). "Acoustic characteristics of speech sounds," in Contemporary Issues in Experimental Phonetics, edited by N.J. Lass (Charles C. Thomas, Springfield, Illinois).

Singh, S., and Murry, T. (1978). "Multidimensional classification of normal voice qualities." J. Acoust. Soc. Am. 64, 81-87.

Skalbeck, G.A. (1955). An Experimental Study of Several Factors in Speaker Recognition (Unpublished master's thesis, Univ. of Washington).

Smith, J.E.K. (1962). Decision-theoretic Speaker Recognizer. Paper presented at the Sixty-Fourth Meeting of the Acoustical Society of America, Seattle, Washington, November 7-10.

Starkweather, J.A. (1956). "Content-free speech as a source of information about the speaker." J. Abnorm. Soc. Psychol. 52, 394-402.

Steiglitz, K. (1977). "On the simultaneous estimation of poles and zeros in speech analysis." IEEE Trans. Acoust. Speech and Signal Processing 25, 229-234.

Stevens, K.N., Hecker, M.H.L., and Kryter, K.D. (1962). An Evaluation of Speech Compression Systems (RADC-TR-62-171) (Rome Air Development Center, Griffiss AFB, NY).

Stevens, K.N., House, A.S., and Paul, A.P. (1966). "Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation." J. Acoust. Soc. Am. 40, 123-132.

Stevens, K.N., and Klatt, D.H. (1974). "Current models of sound sources for speech," in Ventilatory and Phonatory Control Systems, edited by G.A. Cavagna and E.M. Camporesi (Oxford University Press, London).

Stevens, K.N., Williams, C.E., Carbonell, J.R., and Woods, B. (1968). "Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material." J. Acoust. Soc. Am. 44, 1596-1607.

Stuntz, S.E. (1963). Speech Intelligibility and Talker Recognition Tests of Air Force Communication Systems (ESD-TR-63-224) (Electronic Systems Division, Air Force Systems Command, Hanscom Field).

Su, L.-S., Li, K.-P., and Fu, K.S. (1974). "Identification of speakers by use of nasal coarticulation." J. Acoust. Soc. Am. 56, 1876-1882.

Tappert, C.C., and Das, S.K. (1978). "Memory and time improvements in a dynamic programming algorithm for matching speech patterns." IEEE Trans. Acoust. Speech and Signal Processing 26, 583-586.

Thomas, I.B. (1969). "Perceived pitch of whispered vowels." J. Acoust. Soc. Am. 46, 468-470.

Tosi, O.I. (1975). "The problem of speaker identification and elimination," in Measurement Procedures in Speech, Hearing, and Language, edited by S. Singh (Univ. Park Press, Baltimore).

Voiers, W.D. (1964). "Perceptual bases of speaker identity." J. Acoust. Soc. Am. 36, 1065-1073.

Voiers, W.D. (1965). Performance Evaluation of Speech Processing Devices II: The Role of Individual Differences (AFCRL-66-24) (Air Force Cambridge Research Lab, Office of Aerospace Research, Bedford, MA).

Voiers, W.D., Cohen, M.F., and Mickunas, J. (1965). Evaluation of Speech Processing Devices I: Intelligibility, Quality, Speaker Recognizability (AFCRL-65-826) (Air Force Cambridge Research Lab, Office of Aerospace Research, Bedford, MA).

Wakita, H. (1976). "Instrumentation for the study of speech acoustics," in Contemporary Issues in Experimental Phonetics, edited by N.J. Lass (Charles C. Thomas, Springfield, Illinois).

Walden, B.E., Montgomery, A.A., Gibeilly, G.J., Prosek, R.A., and Schwartz, D.M. (1978). "Correlates of psychological dimensions in talker similarity." J. Speech Hear. Res. 21, 265-275.

Waltzman, S.B., and Levitt, H. (1978). "Speech interference level as a predictor of face-to-face communication in noise." J. Acoust. Soc. Am. 63, 581-590.

Webster, J.C. (1978). "Speech interference aspects of noise," in Noise and Audiology, edited by D. Lipscomb (Univ. Park Press, Baltimore).

Weinberg, B., and Bennett, S. (1971). "A study of talker sex recognition of esophageal voices." J. Speech Hear. Res. 14, 391-395.

Weinberg, B., and Bennett, S. (1971). "Speaker sex recognition of 5- and 6-year-old children's voices." J. Acoust. Soc. Am. 50, 1210-1213.

White, G., and Neeley, R. (1976). "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming." IEEE Trans. Acoust. Speech and Signal Processing 24, 183-188.

Williams, C.E. (1964). The Effects of Selected Factors on the Aural Identification of Speakers (ESD-TR-65-153) (Electronic Systems Division, Air Force Systems Command, Hanscom Field).

Williams, C.E. (1971). "Aural speaker recognition and speech communication system evaluation." Proc. 7th Int. Cong. Acoust., Budapest, 3, 89-92.

Williamson, J.A. (1961). An Investigation of Several Factors Which Affect the Ability to Identify Voices as Same or Different (Unpublished dissertation, Univ. Edinburgh).

Wolf, J.J. (1972). "Efficient acoustic parameters for speaker recognition." J. Acoust. Soc. Am. 51, 2044-2056.

Zalewski, J., Majewski, W., and Hollien, H. (1975). "Cross correlation of long-term speech spectra as a speaker identification technique." Acustica 34, 21-24.

Zwicker, E., Terhardt, E., and Paulus, E. (1979). "Automatic speech recognition using psychoacoustic models." J. Acoust. Soc. Am. 65, 487-498.

APPENDIX

TABLE 10

PERCENTAGE CORRECT SCORES FOR THE NORMAL-NORMAL SPEAKING EFFORT COMPARISON

Listener	dB SPL						
	59.1	62.3	65.7	69.8	71.7	72.6	74.0
1	99.1	93.3	97.5	85.8	86.6	85.0	75.8
2	99.1	93.3	90.8	92.5	97.5	88.3	70.0
3	85.0	99.1	98.3	90.8	95.0	74.1	69.1
4	99.1	89.1	73.3	86.6	90.0	76.6	70.8
5	90.8	90.0	84.1	93.3	79.1	70.0	44.1
6	99.1	94.1	87.5	94.1	90.0	73.3	48.3
7	100.0	97.5	90.8	93.3	89.1	72.5	58.3
8	98.3	98.3	91.6	95.0	96.6	80.8	47.5
9	95.8	90.8	90.0	90.0	91.6	78.3	62.5
10	96.6	99.1	90.0	93.3	90.8	85.8	61.6
11	99.1	99.1	93.3	90.0	89.1	71.6	62.5
Mean	96.5	94.8	89.7	91.3	90.4	77.8	60.9
Std. Dev.	4.6	3.8	6.7	3.0	5.0	6.3	10.5

TABLE 11

PERCENTAGE CORRECT SCORES FOR THE SHOUT-SHOUT SPEAKING EFFORT COMPARISON

Listener	dB SPL						
	59.1	62.3	65.7	69.8	71.7	72.6	74.0
1	86.8	96.6	87.5	89.1	77.5	63.3	59.1
2	100.0	100.0	93.3	95.8	96.6	80.8	58.3
3	97.5	100.0	96.6	95.0	93.3	70.8	63.3
4	95.8	97.5	85.8	94.1	83.3	61.6	50.0
5	95.8	91.6	87.5	90.8	73.3	58.3	50.0
6	96.6	92.5	90.8	87.5	89.1	58.3	50.0
7	100.0	96.6	96.6	87.5	89.1	65.8	45.8
8	97.5	97.5	96.6	95.8	94.1	60.8	56.6
9	95.8	93.3	84.1	85.8	89.1	69.1	54.1
10	99.1	99.1	100.0	85.0	94.1	76.6	55.8
11	85.0	98.3	97.5	88.3	79.1	73.3	55.8
Mean	95.4	96.6	92.3	90.4	87.1	67.1	54.4
Std. Dev.	4.9	2.9	5.4	4.0	7.7	7.5	5.0

TABLE 12

PERCENTAGE CORRECT SCORES FOR THE WHISPER-WHISPER SPEAKING EFFORT COMPARISON

Listener	dB SPL				
	59.1	62.3	65.7	69.8	71.7
1	78.3	68.3	72.5	60.0	61.6
2	93.3	90.0	85.8	71.6	55.8
3	91.6	86.6	80.8	60.8	61.6
4	95.0	61.6	72.5	58.3	46.6
5	85.8	78.3	80.8	61.6	62.5
6	94.1	80.0	85.0	65.8	53.3
7	89.1	84.1	78.3	64.1	57.5
8	93.3	86.6	87.5	71.6	45.8
9	91.6	90.8	75.0	54.1	50.0
10	92.5	87.5	73.3	54.1	51.6
11	85.0	71.6	80.8	56.6	54.0
Mean	89.9	80.4	79.3	61.6	54.5
Std. Dev.	5.0	9.6	5.4	6.1	5.8

TABLE 13

PERCENTAGE CORRECT SCORES FOR THE NORMAL-SHOUT SPEAKING EFFORT COMPARISON

Listener	dB SPL			
	59.1	62.3	65.7	69.8
1	58.3	50.0	51.6	52.5
2	79.1	65.8	64.1	55.0
3	74.1	63.3	60.0	50.0
4	65.0	55.8	48.3	50.8
5	50.0	54.1	53.3	51.6
6	55.0	57.5	52.5	52.5
7	55.0	57.5	52.5	53.3
8	65.8	56.6	55.0	54.1
9	62.5	60.0	58.3	52.5
10	59.1	66.6	54.1	51.6
11	52.5	50.0	50.0	50.0
Mean	61.4	57.9	54.5	52.1
Std. Dev.	9.0	5.6	4.6	1.5

TABLE 14

PERCENTAGE CORRECT SCORES FOR THE WHISPER-NORMAL SPEAKING EFFORT COMPARISON

Listener	dB SPL			
	59.1	62.3	65.7	69.8
1	60.8	55.8	50.8	49.1
2	59.1	60.0	54.1	50.0
3	57.5	55.8	54.1	50.8
4	51.6	50.0	52.5	49.1
5	48.3	48.3	51.6	50.0
6	59.1	63.3	55.0	49.1
7	56.6	62.5	55.0	50.0
8	55.8	63.3	52.5	51.6
9	61.6	63.3	53.3	51.6
10	55.0	56.6	53.3	52.5
11	48.3	53.3	52.5	51.6
Mean	55.7	57.4	53.1	50.4
Std. Dev.	4.6	5.4	1.3	1.1

TABLE 15

PERCENTAGE CORRECT SCORES FOR THE WHISPER-SHOUT SPEAKING EFFORT COMPARISON

Listener	dB SPL			
	59.1	62.3	65.7	69.8
1	50.8	48.3	51.6	49.1
2	56.6	46.6	50.0	50.0
3	52.5	58.3	53.3	51.6
4	50.0	53.3	52.5	49.1
5	48.3	54.1	52.5	50.8
6	46.6	52.5	54.1	50.0
7	52.5	46.6	53.3	50.8
8	56.6	55.0	54.1	53.3
9	55.8	57.5	50.8	50.0
10	51.6	67.5	54.1	52.5
11	46.6	53.3	51.6	50.0
Mean	51.6	53.9	52.5	50.6
Std. Dev.	3.6	5.9	1.4	1.3