SPEAKER RECOGNITION: AN INTERPRETIVE SURVEY OF THE LITERATURE

ASHA MONOGRAPHS
NUMBER 16 A PUBLICATION OF THE AMERICAN SPEECH AND HEARING ASSOCIATION

SPEAKER RECOGNITION
AN INTERPRETIVE SURVEY OF THE LITERATURE

The fact that the National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration of the U.S. Department of Justice, furnished the financial support of the activity described in this publication does not necessarily indicate concurrence of the Institute with the statements and conclusions contained therein.

Speaker Recognition

An Interpretive Survey of the Literature

MICHAEL H. L. HECKER
Stanford Research Institute
Menlo Park, California

ASHA Monographs Number 16
AMERICAN SPEECH AND HEARING ASSOCIATION
Washington, D.C.
January 1971

ASHA MONOGRAPHS NUMBER 16

SERIES EDITOR
Gerald M. Siegel, Ph.D.

BUSINESS MANAGER
Kenneth O. Johnson, Ph.D.

PUBLICATIONS DEPARTMENT
James A. Leadon, *Manager*
Margaret S. Tokunaga, *Assistant Manager*

PUBLICATIONS BOARD

Frederic L. Darley, Ph.D., *Chairman*
Robert C. Bilger, Ph.D.
Elizabeth Carrow, Ph.D.
David P. Goldstein, Ph.D.
Katherine S. Harris, Ph.D.
James Jerger, Ph.D.
Kenneth O. Johnson, Ph.D.
Gerald M. Siegel, Ph.D.
Robert T. Wertz, Ph.D.
Dean E. Williams, Ph.D.

AMERICAN SPEECH AND HEARING ASSOCIATION

OFFICERS

President
Jack L. Bangs, Ph.D.
Houston Speech and Hearing Center

President-Elect
Robert Goldstein, Ph.D.
University of Wisconsin

Vice President for Administration
Harold L. Luper, Ph.D.
University of Tennessee

Vice President for Education and Scientific Affairs
Rolland Van Hattum, Ph.D.
State University College at Buffalo

Vice President for Standards and Ethics
Kenneth L. Moll, Ph.D.
University of Iowa

Vice President for Clinical Affairs
Gerald G. Freeman, Ph.D.
Oakland Schools Speech and Hearing Clinic, Pontiac, Michigan

Vice President for Planning
Phillip A. Yantis, Ph.D.
University of Washington

Executive Secretary
Kenneth O. Johnson, Ph.D.

Contents

Preface

In 1968 the U.S. Department of Justice awarded a grant to the Department of State Police, State of Michigan, for research on methods for identifying a person by his voice. As a subcontractor under this grant, Stanford Research Institute was requested to prepare a monograph reviewing past and present research on all methods of speaker recognition. Reports of experimental studies in this field appear in professional journals, conference transactions, and governmental documents, some of which are not generally available. This monograph catalogs, describes, and evaluates the various techniques of speaker recognition that have been developed. In order for it to be useful to readers with different backgrounds, it was written in a tutorial style.

Wherever possible, the monograph uses the terminology found in the literature. Ambiguous or otherwise unsuitable terms occurring in the literature were replaced by new terms, and other new terms were created to underline important distinctions. Studies are described to the extent to which they demonstrate the principles and measurement of speaker recognition; more studies were examined than are described. Recommendations for future research are made throughout the monograph, usually in connection with comments on specific methodological limitations.

The author gratefully acknowledges the many helpful suggestions he has received from his colleagues during the preparation of this monograph. In particular, he thanks Dr. James M. Pickett of Gallaudet College, Dr. Arthur S. House of Purdue University, Dr. Dorothy A. Huntington of Stanford University, and Dr. Karl D. Kryter, Dr. Frank R. Clarke, Dr. James R. Young, Mr. Richard W. Becker, and Mr. Fausto Poza of Stanford Research Institute. Dr. Joan E. Miller of the Bell Telephone Laboratories kindly furnished Figures 8 and 9, and Dr. Victoria A. Fromkin of the University of California, Los Angeles, provided the photographs appearing in Figures 27 and 28. Figure 11 was prepared from data computed by Mr. John Ohala of the University of California, Los Angeles. The author assumes complete responsibility for the accuracy of the material presented.

This monograph is dedicated to Rose Marie.

M. H. L. H.

Menlo Park, California
April 1970

Abstract

Speaker recognition is defined as any decision-making process that uses the speaker-dependent features of the speech signal. The origin and nature of these features are discussed first. This background material is followed by detailed descriptions of three general methods of speaker recognition. Each method is illustrated by many experimental studies. Speaker recognition by listening appears to be the most accurate and reliable method at the present time. Speaker recognition by visual comparison of spectrograms is used in the field of criminology, but this method must be studied further to determine its validity. Speaker recognition by machine is limited by various design shortcomings. More research on this method is expected to improve recognition performance.

Chapter I

GENERAL INTRODUCTION

When a person speaks, he produces a complex acoustic signal containing various kinds of information. This signal serves primarily to convey a linguistic message; listeners who are familiar with the language can transcribe, or at least repeat, what the speaker said. Besides conveying a message, the speech signal reflects some of the anatomy and physiology of the speaker. For example, listeners can often determine the speaker's sex, his approximate age, his emotional state, and whether or not he is suffering from an illness (such as the common cold). Of particular interest is the ability of listeners to distinguish among the speech characteristics of different speakers. This ability provides the basis for one method of speaker recognition.

For the purposes of this monograph, the term speaker recognition is defined very broadly. It refers to any decision-making process that uses the speaker-dependent features of the speech signal. 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. The discrimination task always involves two speech samples. In this task, a decision is rendered as to whether the speech samples were produced by the same speaker or by different speakers. Both recognition tasks have a number of practical applications, one of which is called speaker authentication (or verification). A speaker is said to be authenticated if his claimed identity as an individual or as a member of a group is confirmed by a recognition task.

There are three general methods of speaker recognition. These are speaker recognition by listening, speaker recognition by visual comparison of spectrograms, and speaker recognition by machine. Each of these methods is described in considerable detail in a separate chapter of this monograph. Speaker recognition by listening is, of course, the method used in everyday life. It has been studied for a longer period of time, and appears to be more accurate and reliable than either of the other methods. A possible limitation of this method is that it is entirely subjective. No matter how accurate and reliable listeners may be, they are unable to explain the criteria underlying their decisions.

Speaker recognition by visual comparison of spectrograms is considered to be a more objective method. Spectrograms are visual displays of the speech signal; they exhibit graphic features which can be discussed in a fairly objective

manner. But these features are still interpreted subjectively in arriving at an overall decision. For this reason, there has been much interest in a third method, namely speaker recognition by machine. Although machine decisions are inherently objective, they are often less accurate than comparable human decisions. Current research efforts in speaker recognition by machine are specifically directed toward overcoming this limitation.

All methods of speaker recognition are based on the fact that a given word or phrase tends to be uttered differently by different speakers. There is much variability in the speech signal, and some of this variability is undoubtedly related to particular speaker differences. The nature of speaker variability is discussed in Chapter II. That chapter is included to provide the reader with an understanding of the principles of speaker recognition.

This monograph is an interpretive survey of the literature on speaker recognition. It consists of descriptions of various methodological factors and many experimental studies. Most of the experimental studies are critically evaluated with respect to their design, results, and conclusions. Tutorial material has been included in order to accommodate readers without an expert knowledge of speech science.

Chapter II

INTERSPEAKER AND INTRASPEAKER VARIABILITY

A. INTRODUCTION

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. Interspeaker variability in the speech signal 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). Organic differences may be determined by heredity, sex, and age, while learned differences may be related to geographical, social, and cultural factors.

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. In generating an utterance, a speaker strives to produce appropriate respiratory, laryngeal, and articulatory activity. However, he is unconcerned about the details of the resulting speech signal because many features of this signal are not critical to communication.

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. Both forms of variability are extremely difficult to quantify. Because speaker variability is a reflection of many differences in speech production, it cannot be meaningfully expressed in terms of a single measure. Its measurement requires an understanding of how specific differences in speech production are manifested in the speech signal, but such an understanding is not yet available.

This chapter consists of two major sections. The first section deals with the possible sources of speaker variability. The second concerns experimental studies which provide direct and indirect evidence of speaker variability. An appreciation of the relative magnitudes of interspeaker and intraspeaker variability may be derived from many of these studies.

B. SOURCES OF SPEAKER VARIABILITY

The English language employs an inventory of about 40 different speech

sounds,[1] each of which is produced in a particular manner. All speech sounds are normally uttered in connected speech, but some of them may also be uttered in isolation. The production of isolated speech sounds has been studied in considerable detail. For these speech sounds, a generally accepted acoustical theory of speech production describes the relationship between articulatory states and features of the corresponding speech signal (Fant, 1960; Flanagan, 1965). A brief summary of this theory will be presented here in order to discuss the likely sources of speaker variability.

1. Classification of Speech Sounds

The speech sounds are traditionally classified according to how they are produced (Wise, 1957). As indicated in Table 1, vowels are classified on the

TABLE 1. Phonetic classification of vowels.

Tongue Height	*Front-Back Location of Highest Part of Tongue*		
	Front	*Central*	*Back*
High	i		u
	ɪ		ʊ
	e		o
Mid	•	ɔ	•
	ɛ	ʌ	ɔ
	æ		ɒ
Low	a		ɑ

basis of tongue height and front-back location of the highest part of the tongue. Thus, [i] and [ɪ] are called high front vowels, [u] and [ʊ] are high back vowels, and [ɒ] and [ɑ] are low back vowels. The central vowel [ə] is of special interest because of its neutral articulatory position. Vowel production requires vibration of the vocal folds as a sound source; this activity is termed voicing.

The classification of consonants is shown in Table 2. Consonants are classified in terms of manner and place of production. Manner of production refers to how the vocal mechanism is manipulated in order to produce the sound. This manipulation characteristically either includes or excludes voicing. Furthermore, the manipulation typically involves the forming of a closure or a narrow constriction in the vocal tract. Stop consonants, for example, are produced by momentarily interrupting the air flow with a complete closure of the vocal tract. Fricative consonants, on the other hand, are produced by constricting a portion of the vocal tract sufficiently to generate turbulent noise.

[1]The term speech sound usually refers to a phoneme. Occasionally, it refers to features of the speech signal that give rise to a phoneme percept.

TABLE 2. Phonetic classification of consonants.

Manner of Production	*Place of Production*					
	Bilabial and Labiodental	*Dental*	*Alveolar*	*Palato-alveolar*	*Palatal*	*Velar*
Voiceless Stops	p		t			k
Voiced Stops	b		d			g
Voiceless Fricatives	f	θ	s	ʃ		
Voiced Fricatives	v	ð	z	ʒ		
Nasals (Voiced)	m		n			ŋ
Glides (Voiced)	w				j	
Retroflex (Voiced)				r		
Lateral (Voiced)			l			

Place of production describes the front-back location of the closure or constriction in the vocal tract. Bilabial and labiodental consonants, for example, are produced by closing or constricting the vocal tract either with both lips or with the lower lip and the upper teeth. Similarly, the production of alveolar consonants involves placing the tongue against the alveolar ridge behind the upper teeth. Some common English words and their phonetic transcriptions are listed in Table 3 to help the reader identify each of the speech sounds represented in Tables 1 and 2.

TABLE 3. List of English words and their phonetic transcriptions.

Word	*Transcription*
rake	/ r e k /
boost	/ b u s t /
one	/ w ʌ n /
those	/ ð o z /
yet	/ j ɛ t /
gang	/ g æ ŋ /
bath	/ b a θ /
thief	/ θ i f /
pull	/ p ʊ l /
vision	/ v ɪ ʒ ə n /
marshal	/ m ɑ r ʃ ə l /
watchdog	/ w ɒ t ʃ d ɔ g /

2. *Isolated Utterances of Speech Sounds*

Figure 1 shows a midsagittal section through the structures of the head and neck that form the vocal mechanism. In a strict sense, the vocal mechanism also

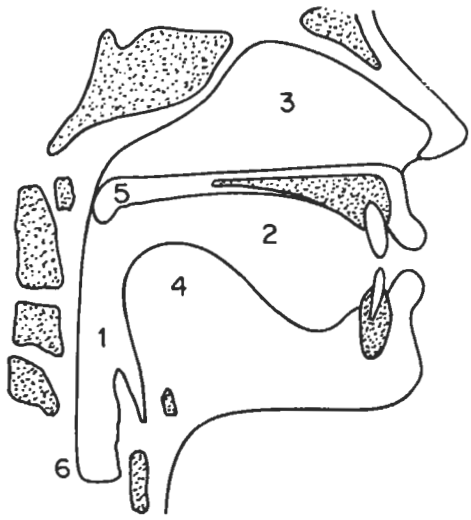

FIGURE 1. Diagrammatic midsagittal section of vocal mechanism during production of vowel [u]. Structures shown include (1) pharyngeal cavity, (2) oral cavity, (3) nasal cavities, (4) tongue mass, (5) soft palate, and (6) vocal folds.

includes the entire respiratory system, which provides the air stream that drives the vocal folds. The vocal tract consists of the pharyngeal cavity and the oral cavity. The illustrated configuration of the vocal tract is appropriate for the production of the vowel [u].

Vowel production may be described with reference to Figure 2. When the vocal folds are set in vibration, they create a source of acoustic energy. This source is called the glottal source; the glottis is the narrow opening between the vocal folds. The waveform of the glottal source consists of a sequence of pulses.[2] One pulse occurs during each vibratory cycle of the vocal folds, and the number of pulses generated per second determines the fundamental frequency of the voice. A typical adult male has a median fundamental frequency in the range 100–150 Hz. The detailed shape of the pulses is influenced by many anatomical and physiological factors, including the dimensions of the vocal folds, control of the laryngeal muscles, and intensity of voicing (vocal effort). These factors tend to differ among speakers. Even for a particular speaker, several factors can differ considerably from utterance to utterance. Thus, the glottal waveform can contribute to both interspeaker and intraspeaker variability.

The spectrum of the glottal source consists of many frequency components, or harmonics. The first harmonic is the fundamental frequency, and the higher harmonics occur at multiples of the fundamental frequency. There is a considerable reduction in the amplitude of each succeeding harmonic. The vocal tract modifies the glottal spectrum in accordance with its properties as an

[2]It is possible to derive the glottal waveform from a knowledge of the time-varying glottal area and the subglottic air pressure (Flanagan, 1958). The glottal area may be measured by means of high-speed motion pictures of the vibrating vocal folds. The subglottic pressure may be measured by a rubber balloon that is positioned in the esophagus.

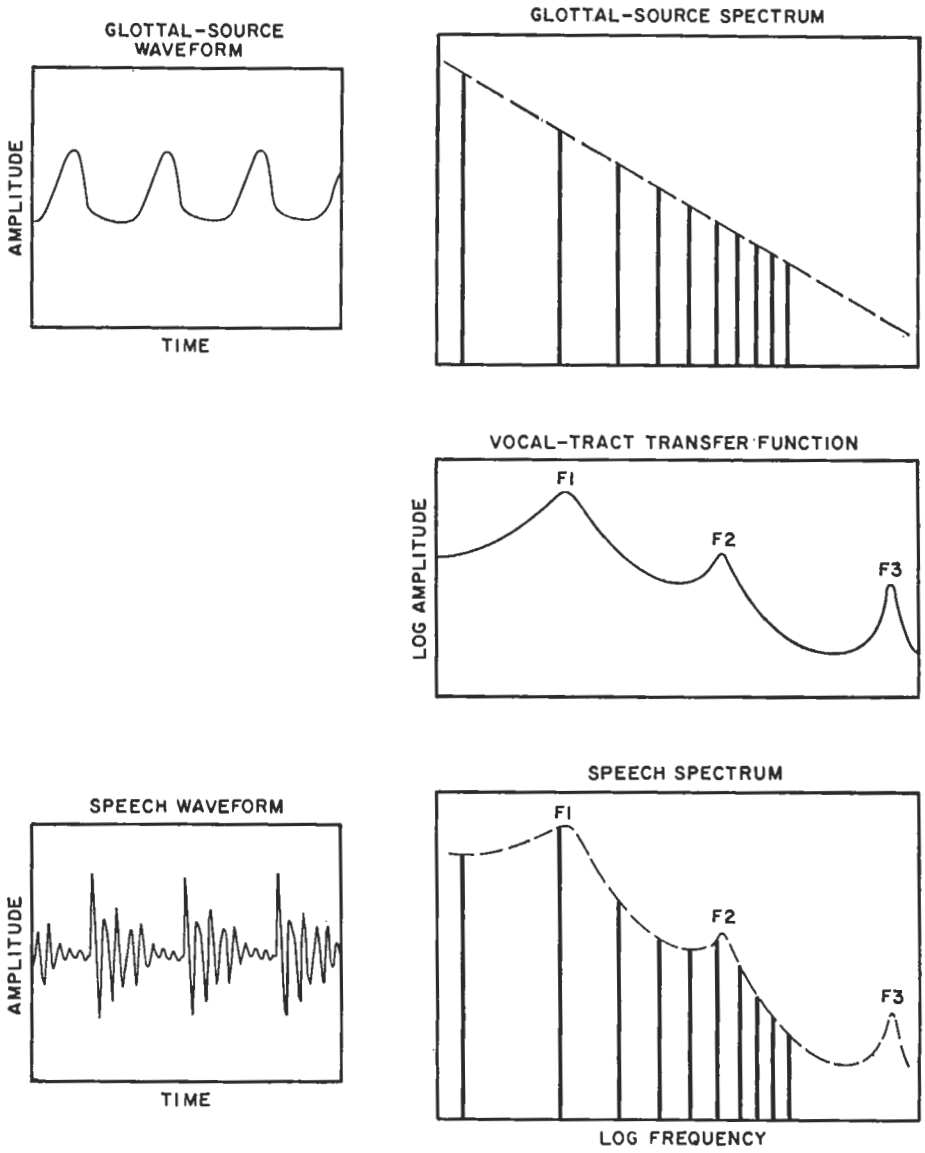

FIGURE 2. Acoustical description of vowel production.

acoustic resonator. In transferring acoustic energy from the glottis to the lips of the speaker, the vocal tract selectively emphasizes certain portions of the glottal spectrum. The resonant properties of the vocal tract are independent of the fundamental frequency. They may be described by what is called the vocal-tract transfer function.[3]

[3]The transfer function of an acoustic resonator is a mathematical description of the relationship between any input signal and the resulting output signal.

The vocal-tract transfer function exhibits several peaks that correspond to the natural frequencies of the vocal tract; these peaks are referred to as formants. Each formant is characterized by a center frequency, a relative amplitude, and a bandwidth. The formant frequencies and amplitudes depend greatly on the configuration of the vocal tract (Stevens and House, 1961). In this manner, the vocal-tract transfer function reflects the identity of a specific vowel. The formant bandwidths are almost independent of the vocal-tract configuration. They represent energy losses in the vocal tract, including radiation from the mouth, absorption by the walls of the vocal tract, and radiation through the glottis into the trachea and lungs.

Approximate formant frequencies for six vocal-tract configurations that are associated with specific vowels are given in Table 4. In general, the frequency

TABLE 4. Approximate formant frequencies for vocal-tract configurations associated with specific vowels. Data are typical for isolated utterances of vowels from adult male speakers.

Vowel	*Formant Frequency (Hz)*		
	F1	*F2*	*F3*
[ə]	500	1500	2500
[i]	250	2000	3000
[ɛ]	500	1750	2500
[a]	750	1250	2500
[ɔ]	500	750	2500
[u]	250	750	2000

of the first formant (F1) is low for high vowels (see Table 1) and high for low vowels, and the frequency of the second formant (F2) is low for back vowels and high for front vowels. Usually, the frequency of the third formant (F3) provides further differentiation among the vowels. The vocal-tract configurations represented by the tabulated data are used only for isolated utterances of the respective vowels. In connected speech, the vowels are produced with less extreme vocal-tract configurations. Thus, the formant-frequency values for a given vowel produced in connected speech are closer to the values for the neutral vowel [ə].

The vocal-tract transfer function also reflects individual differences in the dimensions of the vocal tract. For example, the exact formant frequencies for a given vowel uttered by a particular speaker are indicative of vocal-tract length. Low formant-frequency values signify a relatively long vocal tract, whereas high values signify a shorter vocal tract (Stevens and House, 1961). Similarly, the exact formant bandwidths would be expected to differ among speakers. Different vocal tracts presumably have different energy losses. However, formant bandwidths are difficult to measure (Dunn, 1961), and their contribution to speaker variability has not been examined.

The spectrum existing at some distance in front of the speaker is called the

speech spectrum. This spectrum includes the effects of the glottal spectrum and the vocal-tract transfer function.[4] The speech waveform consists of a sequence of damped oscillations, each of which is initiated at the instant of glottal closure. A computerized procedure has been developed for recovering the glottal waveform from the speech waveform (Mathews, Miller, and David, 1961). The speech signal is processed by a variable filter which is adjusted to exactly cancel the effect of the vocal-tract transfer function. This procedure, called inverse filtering, is carried out in synchronism with the fundamental frequency.

Voiceless fricative consonants are produced by forming a narrow constriction in the vocal tract and forcing air through this constriction at a high velocity. The front-back location of the constriction (referred to as place of production in Table 2) depends on the particular fricative. To produce the fricative [s], for example, the constriction is formed by placing the upper surface of the tip of the tongue against the alveolar ridge. This articulatory configuration is shown in Figure 3.

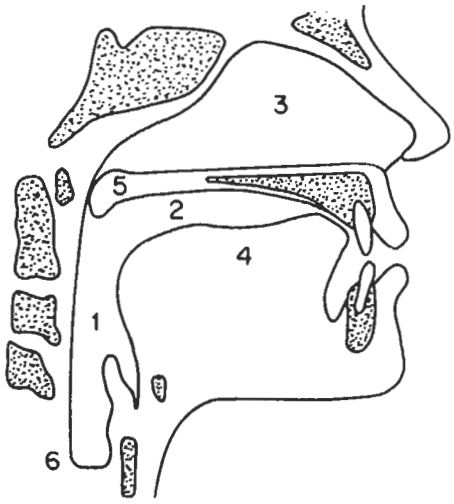

FIGURE 3. Diagrammatic midsagittal section of vocal mechanism during production of fricative consonant [s]. Structures are identified in Figure 1.

The air flow becomes turbulent in the vicinity of the constriction, and this creates an acoustic source of random noise. The noise spectrum, which extends uniformly over a wide frequency range, is modified by a vocal-tract transfer function that exhibits several peaks and a few abrupt depressions (Heinz and Stevens, 1961). As in the case of vowel production, the peaks correspond to the natural frequencies of the vocal tract.[5] The depressions are attributable to

[4]The speech spectrum also includes a high-frequency emphasis that is due to acoustic radiation from the mouth into the air.

[5]The vocal-tract configurations associated with voiceless fricatives represent higher energy losses than the vocal-tract configurations associated with vowels. There are losses at the

acoustical cancellation by the portion of the vocal tract behind the noise source. They reflect the location and the dimensions of the constriction in the vocal tract. Since different speakers undoubtedly form slightly different constrictions to produce the same fricative, the exact frequencies and bandwidths of these depressions are likely to contribute to interspeaker variability.

The production of voiced fricative consonants is similar to the production of voiceless fricative consonants, except that the glottal source is in use. Voiced and voiceless fricatives having a common place of production are produced with very similar vocal-tract configurations.

Nasal consonants are produced by coupling the nasal cavities to the vocal tract, closing the vocal tract in front of the point of coupling, and using the glottal source. To couple the nasal cavities to the vocal tract, the back of the soft palate is lowered. The location of the vocal-tract closure (place of production) depends on the particular consonant. For the consonant [m], the vocal tract is closed at the lips; for [n], the closure is formed by pressing the tongue against the alveolar ridge; and for [ŋ], the closure is formed by touching the back of the tongue to the lowered soft palate. A midsagittal section of the vocal mechanism during the production of the consonant [n] is shown in Figure 4.

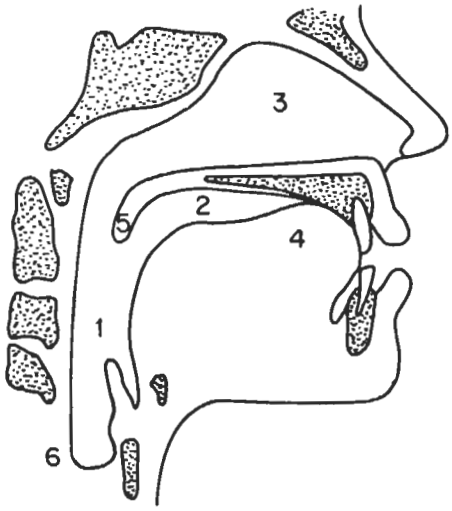

FIGURE 4. Diagrammatic midsagittal section of vocal mechanism during production of nasal consonant [n]. Structures are identified in Figure 1.

The glottal spectrum is modified by a transfer function that describes the acoustical path between the glottis and the nostrils of the speaker. This transfer function exhibits both peaks and abrupt depressions (Fujimura, 1962). The

constriction and also at the vocal folds, which are partially abducted. For this reason, the formant bandwidths tend to be greater in the case of voiceless fricatives.

peaks correspond to the natural frequencies of the vocal mechanism, including the nasal cavities.[6] Assuming that the dimensions of the various parts of the vocal mechanism differ from speaker to speaker, the frequencies and bandwidths of the peaks should be speaker dependent. The depressions are attributable to acoustical cancellation by the closed oral cavity; they reflect the location of the closure. Thus, the depressions are also expected to contribute to speaker variability.

3. Connected Speech

The preceding section has described how various speaker differences are likely to be reflected in isolated utterances of vowels, fricative consonants, and nasal consonants. Speaker differences are expected to have an even greater effect on the production of connected speech. Both organic and learned differences may influence the speech signal (Garvin and Ladefoged, 1963). While organic differences presumably affect each individual speech sound, learned differences are probably reflected in transitions from one speech sound to another, and in patterns of fundamental frequency, overall amplitude, and speaking rate.

As mentioned earlier, the vocal-tract configurations used to produce vowels in connected speech are more neutral than the vocal-tract configurations used for isolated utterances of vowels. Since the articulatory structures are continuously in motion, they do not have time to form extreme vocal-tract configurations. The formant-frequency values listed in Table 4 may be viewed as target values that are not fully reached when the vowels are produced in connected speech. These target values, and the degree to which they are reached in specific consonantal environments, are highly speaker dependent.

In connected speech, each speech sound is more or less affected by the preceding and following speech sounds because articulatory configurations cannot be changed instantaneously. For example, the duration and formant structure of a given vowel in a consonant-vowel-consonant utterance depend not only on the identity of the vowel but also on the particular consonants. Such context-dependent modifications of speech sounds are called coarticulation effects. Superimposed on the coarticulation effects is interspeaker and intraspeaker variability. The excursions and relative timing of the many articulatory adjustments required for the pronunciation of a given word vary among speakers and utterances.

During the production of connected speech, vowels occurring in the vicinity of nasal consonants may be nasalized. Coupling the nasal cavities to the vocal tract may affect the vowel spectrum in several respects. Since there is now an additional energy loss in the vocal tract, the formant bandwidths may be increased. The nasal cavities may also provide acoustical cancellation and thus

[6] Because of high energy losses in the nasal cavities, the peaks have relatively large bandwidths.

introduce abrupt depressions at particular frequencies. These effects of the nasalization of vowels are likely to be more pronounced for some speakers than for others.

In analyzing connected speech, it is often difficult to separate coarticulation effects from speaker variability. For example, the acoustical characteristics of stop consonants are determined both by the context in which the consonants occur and by the manner in which the speaker has learned to coordinate his respiratory, laryngeal, and articulatory activities.

The linguistic description of spoken English includes the so-called prosodic features of speech, intonation and stress. Both of these features can serve to make the semantic contents of an utterance more specific. A change in intonation may turn a statement into a question. Stress selectively emphasizes the syllables represented by an utterance, and this may lead to a unique interpretation. It is important to realize that intonation and stress are linguistic concepts and not acoustic entities. Each prosodic feature has many acoustical correlates, such as variations in fundamental frequency, overall amplitude, spectral balance, and speech-sound duration. The acoustical correlates of intonation and stress are not yet fully understood, but they appear to be highly speaker dependent.

C. EXPERIMENTAL EVIDENCE OF SPEAKER VARIABILITY

1. Evidence of Interspeaker Variability

Experiments concerned with the recognition of speech by machine provide clear evidence of the presence of interspeaker variability. Speech recognizers generally require different adjustments for different speakers in order to maintain an optimal level of performance. In the first such machines, these adjustments were made manually and largely by trial and error. The feasibility of incorporating automatic adjustment schemes into more sophisticated machines is currently being studied.

For example, the computer-simulated machine of Hemdal and Hughes (1967) has been modified to adapt automatically to the speech characteristics of different speakers (Hemdal, 1967). This machine operates by detecting the acoustical correlates of the so-called distinctive features of speech.[7] The digitized output of a conventional spectrum analyzer is processed by a distinctive-feature classification routine to select time segments that represent vowels. For each of these time segments, the frequencies of the first three formants (F1, F2, and F3), and the duration of the sound in which the time segment occurs, are measured. Table 5 and Figure 5 show how these measurements are then used to label each time segment as representing a particular vowel. The

[7] The concept of the distinctive features is that each speech sound (phoneme) may be uniquely specified by a small number of binary decisions about basic phonological states (Jakobson, Fant, and Halle, 1963).

TABLE 5. Acoustical correlates of distinctive features used for vowel recognition.

Distinctive Feature	*Acoustical Correlate*
Acute/Grave	High F2/Low F2
Compact/Diffuse	High F1/Low F1
Flat/Plain	F1 + F2 < Threshold/F1 + F2 > Threshold
Tense/Lax	Longer duration, farther from a neutral position in F1-F2 plane/ Shorter duration, closer to neutral position

distinctive-feature boundaries and the neutral position shown in Figure 5 are shifted for each new speaker according to the results of a preliminary analysis of several utterances.

As mentioned in Section B3, interspeaker variability may be superimposed upon coarticulation effects. Using a speech-analysis procedure developed by

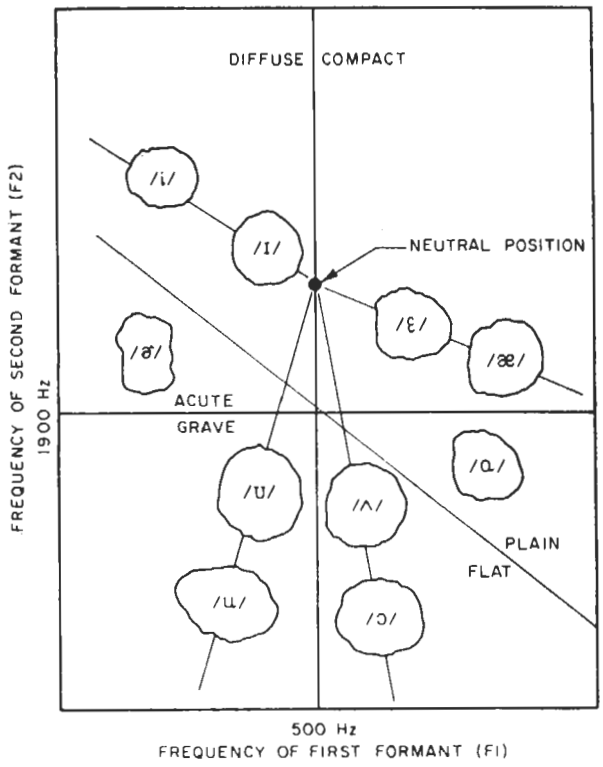

FIGURE 5. Formant-frequency plane showing distinctive-feature boundaries used for vowel recognition. (Reprinted, by permission, from Hemdal and Hughes, 1967.)

Bell, Fujisaki, Heinz, Stevens, and House (1961), Stevens, House, and Paul (1966) measured formant frequencies at many points in time throughout various consonant-vowel-consonant utterances. Three adult male speakers participated in this study; they differed with respect to physical height and therefore presumably also with respect to vocal-tract length. For each speaker, the contour of the second-formant frequency during the vocalic portion of the syllable /dɪd/ is shown in Figure 6. The arrows to the right of the figure indicate the

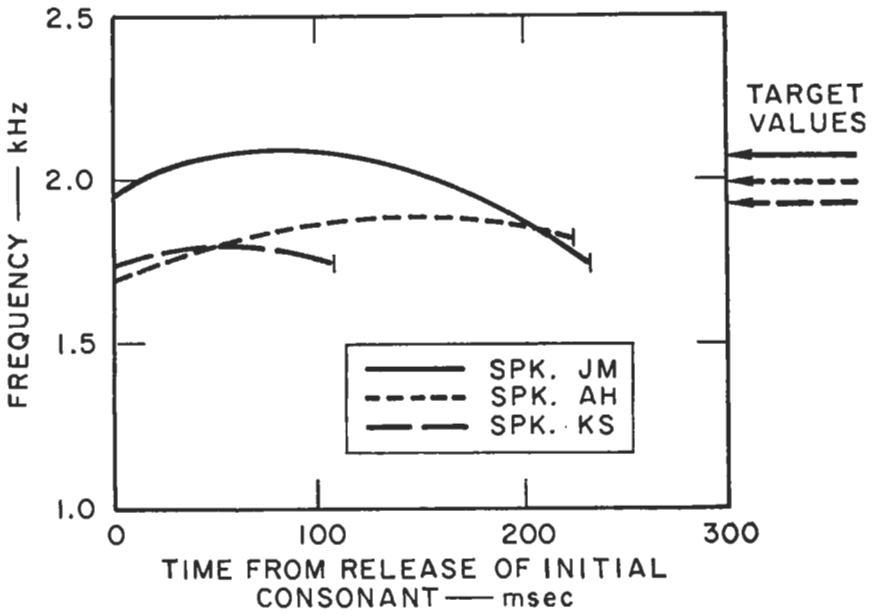

FIGURE 6. Second formant-frequency contours for vocalic portion of syllable /dɪd/ uttered by three speakers. Arrows indicate second-formant frequencies for isolated utterances of vowel [ɪ]. (Reprinted, by permission, from Stevens et al., 1966.)

respective target values of the second-formant frequency (i.e., the values for isolated utterances of the vowel [ɪ]). As was expected, the shortest speaker (J. M.) produced the highest target value, and the tallest speaker (K. S.) produced the lowest target value. There was considerable variability in the degree to which the target values were reached, in vowel duration, and in the symmetry of the contours. The variability at the consonant-vowel and vowel-consonant boundaries was generally found to be greater than the variability at the midpoint of the vowel.

There is some evidence that the interspeaker variability exhibited in isolated utterances of a vowel is basically independent of the particular vowel. A study by McGee (1965) employed spectrographic sections (amplitude-frequency plots) that represented the vowels [i] and [ɔ] as uttered by 19 male speakers.

Each section was divided into 34 frequency bands, and an energy level was estimated for each band. A 19 × 34 data matrix was formed for each vowel; also, by averaging the data matrix over speakers, a 1 × 34 vowel matrix was calculated. The vowel matrix was then subtracted from the data matrix to obtain a residual matrix which was assumed to describe the speaker characteristics associated with the vowel. A statistical analysis demonstrated that the two residual matrices were extremely similar. It was concluded, therefore, that the individual differences exhibited in the utterances of the vowel [i] were equivalent to the individual differences exhibited in the utterances of the vowel [ɔ].

Several factors appear to influence interspeaker variability. In a given experiment, the participation of speakers with certain speech pathologies (e.g., dysphonia or dysarthria), neuropathologies (e.g., cerebral palsy or Parkinsonism), psychopathologies (e.g., hysteria or depression), or unique linguistic backgrounds (e.g., early exposure to a regional dialect or foreign language) 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 whose organic and learned differences are small. Organic differences are presumably minimal in the case of identical twins; there is evidence that the long-term speech spectra of monozygotic twins are more similar than the spectra of either dizygotic twins or sex- and age-matched nontwin pairs (Alpert, Kurtzberg, Pilot, and Friedhoff, 1963).

The effects of both genetic and environmental factors may be studied by comparing speech samples from different members of a family. Kersta (1965a) conducted several experiments that employed speech samples from three members of each of eight families, the father, the mother, and a 5- or 6-year old son. Each family member produced four normal and four whispered utterances of the sentence *You and I were there.* A spectrogram[8] was prepared for each utterance, and spectrograms representing either normal or whispered utterances of a particular word were compared by trained observers who attempted to match sons with parents and husbands with wives. The results of this study, shown in Table 6, indicate that whispered speech provides a higher matching accuracy than normal speech. The similarity of spectrograms for sons and parents can be explained in terms of the obvious genetic factor and the family environment. While the similarity of spectrograms for husbands and wives can be attributed primarily to environmental factors, remote genetic factors may also be involved. A person of a given race, height, and age is likely to marry someone of the opposite sex with a similar description.

2. *Evidence of Intraspeaker Variability*

Experiments concerned with speaker recognition by machine provide evi-

[8]A spectrogram is a visual amplitude-frequency-time display of the speech signal. The preparation and use of spectrograms will be described in Chapter IV.

TABLE 6. Results of spectrogram-matching experiments, averaged over five words. Matching accuracy attributable to chance is 2.8%.

Experiment	*Matching Accuracy (%)*	
	Normal Speech	*Whispered Speech*
Sons Matched with Fathers	18.0	21.0
Sons Matched with Mothers	13.0	17.5
Sons Matched with Both Parents	15.0	28.0
Husbands Matched with Wives	—	23.0

dence of the presence of intraspeaker variability. These experiments will be described in considerable detail in Chapter V. Regardless of the techniques employed, machines can perform either an identification task or a discrimination task. In the identification task, an utterance from the speaker to be identified is converted into a test pattern which is compared with a number of reference patterns. Each reference pattern is constructed from several utterances by a particular speaker, and one reference pattern represents the speaker to be identified. For identification, the test pattern is associated with the most similar reference pattern. Hargreaves and Starkweather (1963) found that more incorrect identifications are made when the utterances represented by the 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.

In the discrimination task, the test pattern is compared with a single reference pattern, and a decision is made as to whether both patterns represent the same speaker. According to the results of a study by Li, Dammann, and Chapman (1966), performance can be optimized by constructing the reference pattern from utterances which are not all recorded at the same time. Such a reference pattern incorporates more of the intraspeaker variability that is sampled by the test pattern.

Several factors may influence intraspeaker variability. One such factor is psychological stress. In a study by Hecker, Stevens, von Bismarck, and Williams (1968), stress was induced in experimental subjects by having them carry out an arithmetic task under time pressure. Verbal responses involving test phrases were obtained under stress and control conditions, and acoustical analyses of these responses indicated that the speech of many subjects was modified by stress. Most of the changes were attributable to differences in the amplitude, frequency, and detailed shape of the glottal pulses. Other changes resulted from differences in articulation.

The speech signal also reflects aging. Mysak (1959) found that aging is often accompanied by a rise in median fundamental frequency, a greater variability in fundamental frequency, and a slight reduction in speaking rate. The rise in fundamental frequency was attributed to physiological changes in the

larynx and also to psychological factors precipitated by social and economic changes. In a similar study by Ptacek, Sander, Maloney, and Jackson (1966), advanced age was found to be associated with a reduced range of fundamental frequency, a lower maximum vowel intensity, and a less rapid articulation. It was speculated that the reduced range of fundamental frequency is caused by aging of the laryngeal cartilages and muscles.

Another factor which may influence intraspeaker variability is disease. Various diseases of the chest, the larynx, and the central nervous system are known to affect particular aspects of speech production. Many disease processes are characterized by alternating periods of remission and relapse, and the speech signal may be modified accordingly.

It may be possible to devise techniques for reducing intraspeaker variability within a given experiment or other restricted situation. Techniques that reduce intraspeaker variability more than they reduce interspeaker variability would be expected to improve the reliability of speaker recognition. Kurtzberg, Alpert, and Friedhoff (1963) suggested a technique that requires each speaker to match his fundamental frequency to a standard tone and the intensity of his voice to an intensity standard. Obviously, this technique is only applicable in situations where the speakers are highly cooperative.

3. *Further Evidence of Speaker Variability*

Both interspeaker and intraspeaker variability are displayed in the spectrograms in Figure 7. These spectrograms show how the properties of the speech signal change with time. Frequency is plotted vertically, time horizontally, and relative amplitude is indicated by the darkness of the mark. Various portions of the spectrograms exhibit distinct spectral features. A rapid sequence of pulses occurs whenever the vocal folds are vibrating. The dark energy bands appearing during these intervals of vocal-fold vibration are the formants. Intervals with randomly fluctuating energy above 4 kHz represent fricative consonants, and narrow (50–100 msec) energy gaps are related to the production of stop consonants. Since the spectrograms portray different utterances of the same phrase, each spectral feature of one utterance has a grossly similar counterpart in another utterance. The variability in corresponding spectral features seems to be somewhat greater between the two speakers (interspeaker variability) than between the two utterances by the same speaker (intraspeaker variability).

As mentioned earlier, differences in the operation of the glottal source may contribute to speaker variability. It is difficult to appraise the significance of this contribution by examining spectrograms because spectrograms do not display glottal-source characteristics directly. Miller and Mathews (1963) extracted the glottal waveform from the speech signal by means of inverse filtering. An improved version of the technique developed by Mathews, Miller, and David (1961) was used to analyze isolated utterances of vowels from six

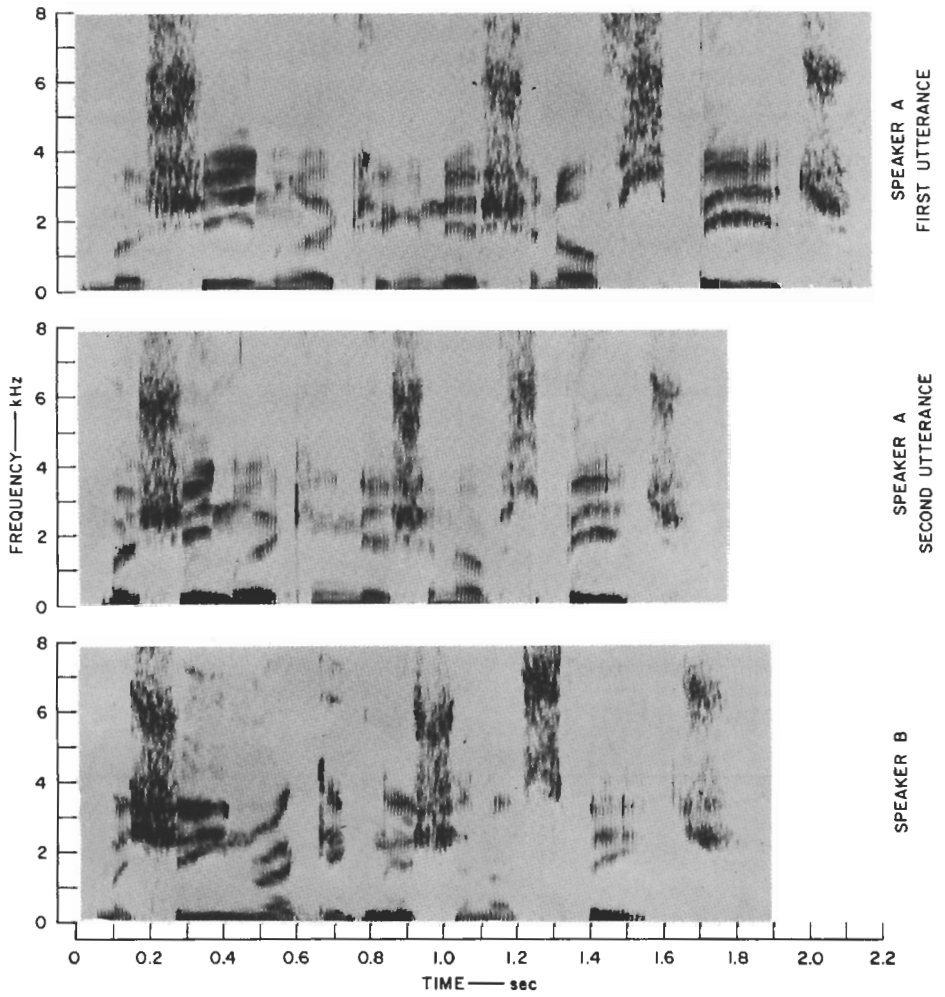

FIGURE 7. Sound spectrograms of three utterances of the phrase *Machine recognition of speech.* (Reprinted, by permission, from Young and Hecker, 1968.)

male speakers. Some typical results are illustrated in Figure 8. The left column shows data for four vowels uttered by the same speaker, and the right column shows data for a single vowel uttered by four speakers. Each utterance is represented by a portion of the speech waveform and by the corresponding portion of the glottal waveform.

The observation that the glottal waveform is similar for different vowels uttered by the same speaker suggests a relatively low intraspeaker variability. That the glottal waveform varies from speaker to speaker even though all speakers uttered the same vowel suggests a relatively high interspeaker variability. Thus, the glottal waveform would appear to be a valuable descriptor for differentiating among speakers. However, the glottal waveform also varies

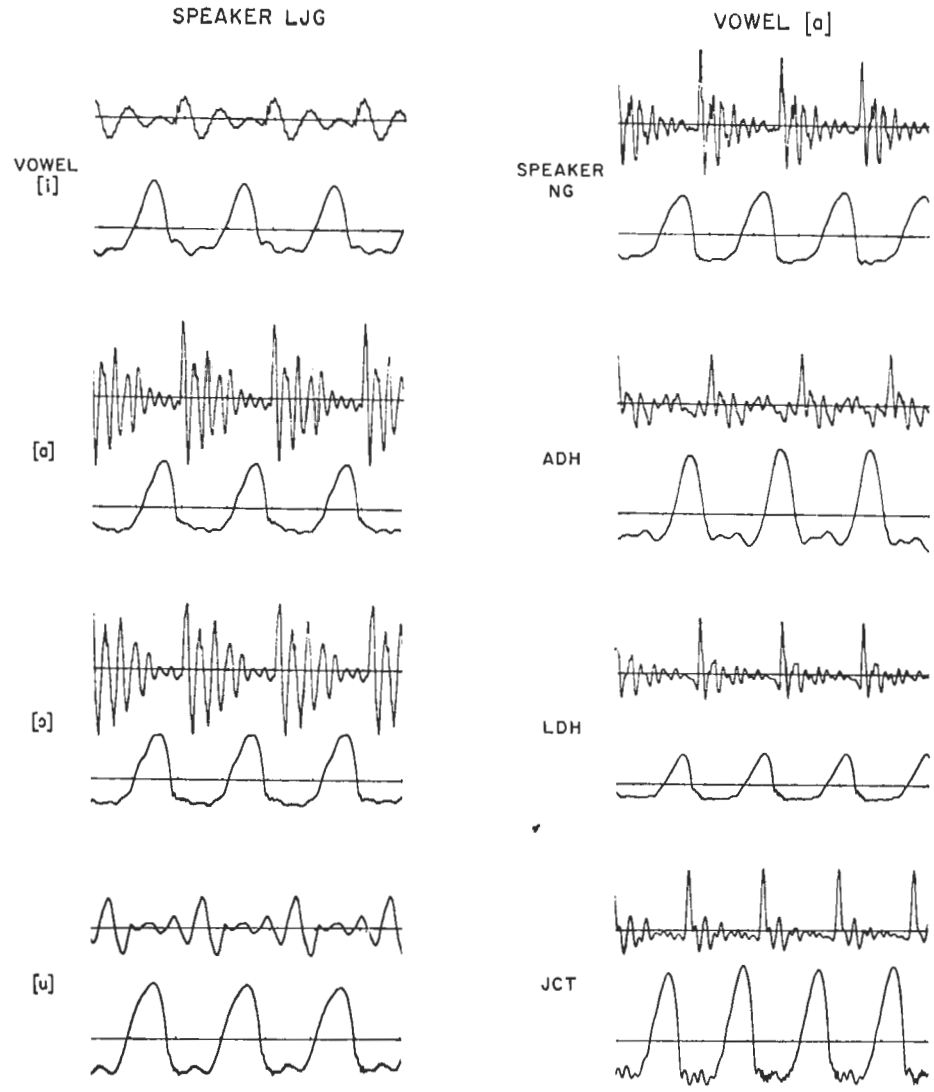

FIGURE 8. Waveforms of speech signal (upper traces) and glottal source (lower traces) for four vowels produced by Speaker LJG and for vowel [ɑ] produced by four speakers. (Reprinted, by permission, from Miller and Mathews, 1963.)

with the intensity of voicing (vocal effort) and with the fundamental frequency. These effects are illustrated in Figure 9. With a decrease in intensity or an increase in fundamental frequency, the interval during which the glottis is open tends to become a larger portion of the vibratory cycle. Therefore, in order to use the glottal waveform as a basis for speaker recognition, its amplitude and period must be taken into account.

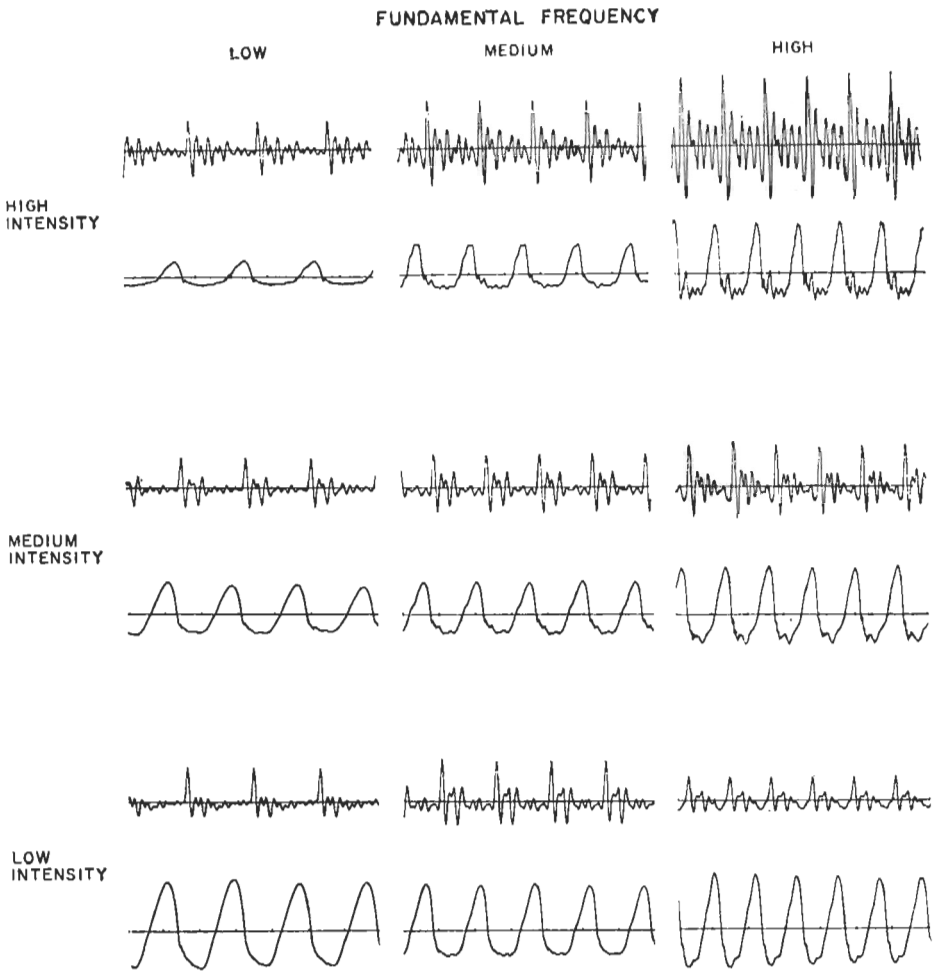

FIGURE 9. Waveforms of speech signal (upper traces) and glottal source (lower traces) for nine utterances of vowel [ɑ] produced by Speaker JCT at various levels of intensity and fundamental frequency. (Reprinted, by permission, from Miller and Mathews, 1963.)

Miller (1964) studied the relative importance of the vocal-tract transfer function and the glottal-source characteristics in speaker recognition by listening. Using inverse filtering, these two components of the speech signal were separated for short utterances by several speakers and then recombined so that the vocal tract of one speaker was effectively modulating the glottal source of another speaker (or an artificial source). Listeners consistently associated the hybrid utterances with the speaker whose vocal tract was represented. In other words, the vocal-tract transfer function appeared to carry more information about a speaker than the glottal-source characteristics. This study will be described in greater detail in Chapter III.

4. Studies of Physiological Activity

For purposes of speaker recognition, the speaker variability observed in the speech signal is of greater interest than the physiological differences that give rise to this variability. All methods of speaker recognition employ some form of analysis of the speech signal, and physiological measures are generally not available. However, studies of physiological activity are valuable in that they may lead to a better understanding of the nature of speaker variability. Such studies may provide further means for estimating the relative magnitudes of interspeaker and intraspeaker variability.

Hirano and Smith (1967) recorded the electrical activity in several articulatory muscles of four speakers. These investigators used bipolar, thin-wire electrodes that were injected directly into the muscle tissues. The placement of the electrodes for recording from the anterior portion of the genioglossus muscle and from the mylohyoid muscle is illustrated in Figure 10. Various consonant-

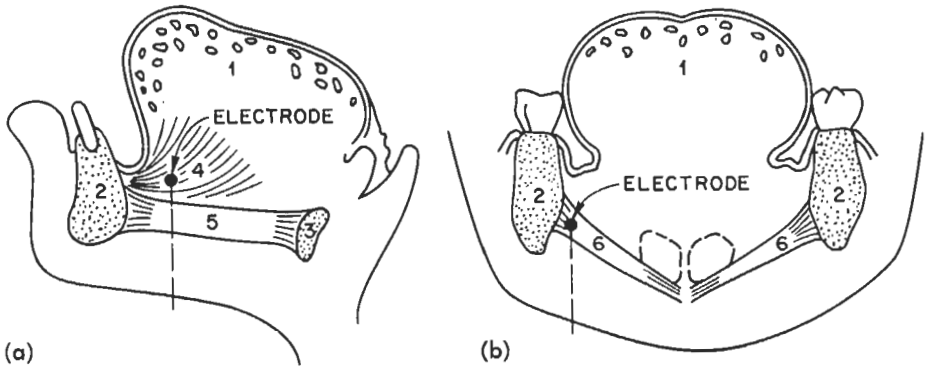

FIGURE 10. Diagrammatic midsagittal (a) and coronal (b) sections of floor of mouth, indicating positions of electrodes. Structures shown include (1) tongue mass, (2) mandible, (3) hyoid bone, (4) genioglossus muscle, (5) geniohyoid muscle, and (6) mylohyoid muscle. (Adapted, by permission, from Hirano and Smith, 1967.)

vowel-consonant syllables served as the speech material; these syllables were singly embedded in a suitable carrier phrase. Each item was uttered at least ten times in succession. A small computer was used to average the recorded action potentials over several utterances and to smooth and display the combined data.

Some typical results are presented in Figure 11. For each speaker, two traces are shown for a given muscle. These traces were obtained by averaging the action potentials over two different sets of five utterances. Thus, both interspeaker and intraspeaker differences may be observed in these physiological data. The traces for the genioglossus muscle appear to be characterized by one central prominence for Speaker GA, by two prominences for Speakers JO and GP, and by three prominences for Speaker TS. For a given speaker, how-

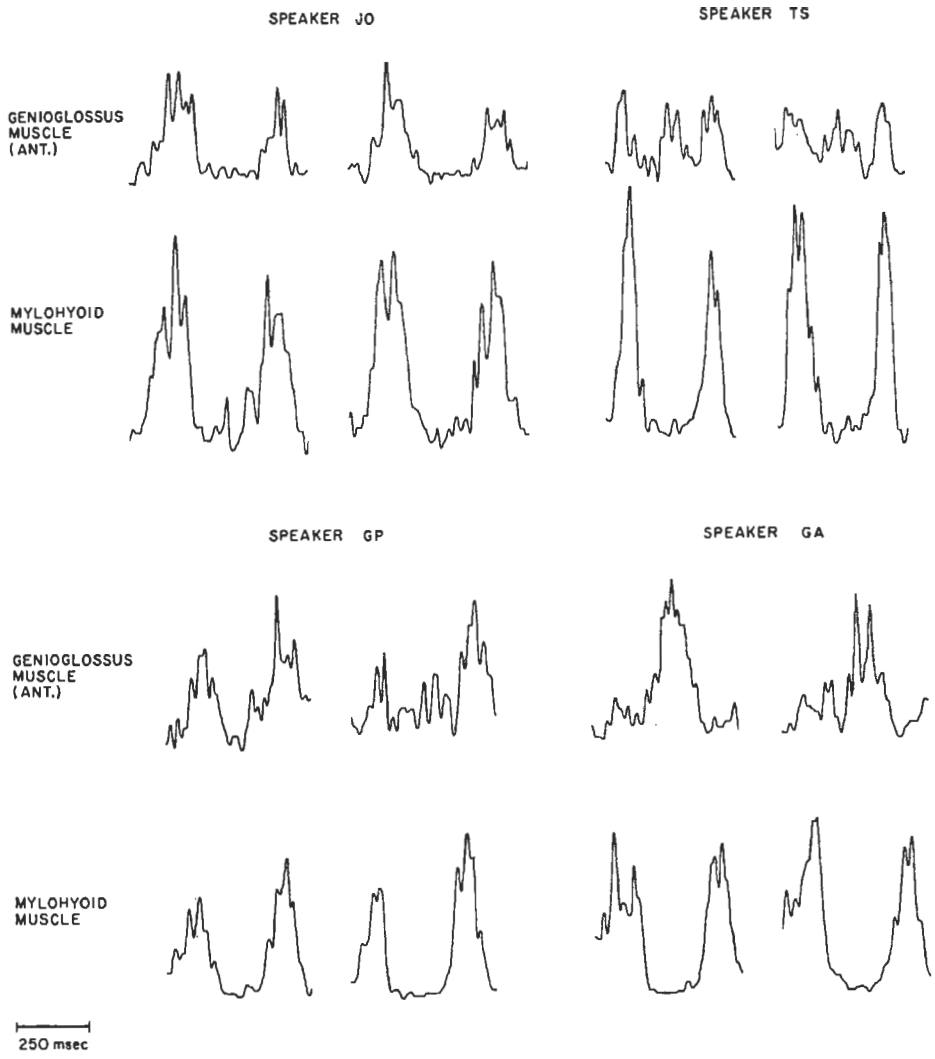

FIGURE 11. Electrical activity in genioglossus and mylohyoid muscles of four speakers during production of nonsense syllable /kɑk/. For each speaker, action potentials were averaged and smoothed over two sets of five utterances.

ever, the two traces are fairly similar. All traces for the mylohyoid muscle are characterized by two prominences, but again the interspeaker differences seem to be greater than the intraspeaker differences. To the extent that the electrical activity in these muscles is related to articulatory behavior, and to the extent that such behavior is, in turn, reflected in the speech signal, the interspeaker variability would be expected to exceed the intraspeaker variability. This prediction thus agrees with other estimates of the relative magnitudes of the two forms of speaker variability that are based on acoustic measures.

Chapter III

SPEAKER RECOGNITION BY LISTENING

A. INTRODUCTION

Research on speaker recognition by listening has several interrelated objectives. One objective is to determine the variables which affect listener performance and to develop test formats for controlling these variables. A second objective is an understanding of the perceptual bases of speaker recognition. It has been postulated that listeners use a small number of perceptual parameters in differentiating among voices; attempts have been made to define and measure these parameters. A third objective is knowledge of the acoustical manifestations of speaker identity. By modifying the speech signal in a selective manner and noting the effect on listener performance, it is often possible to determine what acoustical features carry information about the identity of the speaker. A fourth objective is to use speaker-recognition tests for evaluating communication systems. Each of these topics will be discussed in this chapter. Also, because of a long-standing concern about the reliability of the human listener, there will be a discussion of listener fallibility.

B. VARIABLES OF SPEAKER RECOGNITION

Several kinds of tests have been devised to study different aspects of speaker recognition by listening. All tests employ the same basic procedure: speakers drawn from a prescribed population are recorded reading selected speech material, the recordings are edited and presented to listeners, and the listeners carry out a recognition task. Each step in this procedure introduces many variables which can influence the resulting performance measure. It is the function of the test format to provide control over these variables so that they can be studied individually. The most important variables will now be described in detail.

1. Size and Homogeneity of Speaker Group

Recognition performance is found to be inversely related to the size of the speaker group. Using a test in which the listeners learn to associate a number with each of several unfamiliar speakers, Williams (1964) trained three groups

of listeners with groups of four, six, and eight speakers. After considerable training, the listeners working with the four- and six-speaker groups achieved above 60% correct identification, while the listeners working with the eight-speaker group barely achieved 40% correct identification. It was concluded that a group of five or six speakers may be optimal for this kind of test. Other tests have employed larger speaker groups.

The level of performance also 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. On the basis of this finding, the former quartets were considered to be more homogeneous. Within a particular speaker group, the incorrect identities assigned to each speaker are usually not equally distributed over all of the remaining speakers. Such confusions tend to involve certain speakers, and these speakers are presumed to form a homogeneous subset with the incorrectly identified speaker. Thus, homogeneity has been inferred to exist both between and within speaker groups.

Various perceptual and physical measures have been employed as indicators of the homogeneity of a speaker group. Although the correlation between these measures and recognition performance is relatively poor, the measures have been used to select speakers and to describe particular speaker groups. Carbonell, Grignetti, Stevens, Williams, and Woods (1965) based their selection of speakers on voice-attribute ratings which were obtained from expert listeners. The rated attributes included regional accent, articulatory precision, nasal resonance, and a number of attributes specified by semantic-differential scales.[1] The average ratings of the individual attributes were not found to be related to the observed confusions among the selected speakers. Silbiger (1966) proposed a similar method for selecting speakers and classifying voices.

Among the physical measures that have been employed as indicators of homogeneity are direct measures of the speech signal. Compton (1963) determined the fundamental frequency of isolated utterances of the vowel [i] by several speakers and found that speakers with similar ranges of fundamental frequency were often confused by the listeners. Indirect physical measures have also been used. Stevens, Williams, Carbonell, and Woods (1968) estimated the length of the vocal tract of each speaker from a profile photograph of the speaker. As mentioned in Chapter II, the length of the vocal tract affects the formant frequencies.

2. *Selection of Speech Material*

Performance is also influenced by the speech material. The most important

[1] A semantic-differential scale is a psychophysical scale that is defined at its endpoints by adjectives of opposite meaning, e.g., pleasant–unpleasant (Osgood, Suci, and Tannenbaum, 1957).

requirement of the speech material is that it must allow an adequate sampling of each speaker's phonetic repertoire. This requirement is usually expressed in terms of the duration of the speech samples. As shown in Figure 12, when

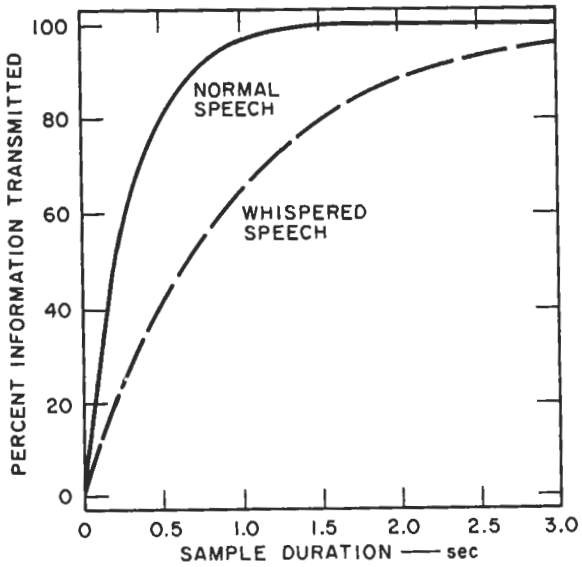

FIGURE 12. Transmission of speaker identity as a function of duration of speech sample. Data are shown for normal and whispered speech. (Reprinted, by permission, from Pollack et al., 1954.)

the duration of a sample of normal speech exceeds 1 sec, there is almost no further improvement in performance; a sample of whispered speech must have a duration of at least 3 sec before maximum performance is reached (Pollack, Pickett, and Sumby, 1954). This difference in duration illustrates the contribution of voicing to speaker recognition.

Bricker and Pruzansky (1966) conducted experiments employing sentences, disyllables, monosyllables, consonant-vowel excerpts, and vowel excerpts. Their results indicate that performance depends on the number of phonemes contained in the speech sample rather than on the duration of the sample per se. Williams (1964) obtained significantly higher scores for sentence tests than for tests consisting of monosyllables. Stevens, Williams, Carbonell, and Woods (1968) also obtained higher scores for phrases and disyllables than for monosyllables. Clarke, Becker, and Nixon (1966) used sentences containing from 3 to 11 syllables; performance was not significantly influenced by the number of syllables. Considering that the duration of a three-syllable utterance may easily approach 1 sec, this finding is not at variance with the data shown in Figure 12. Compton (1963) constructed a test from vowel segments of different duration and concluded that 500-750 msec is the optimal range of duration for this kind of speech sample.

3. *Size and Training of Listener Group*

The ability to recognize speakers varies from listener to listener. If only a few listeners participate in a given test, the average score will depend greatly on their particular abilities. As the number of listeners is increased, this dependence is reduced, and the test results become more useful for making comparisons between experiments using different listeners. Thus, the size of the listener group can have an appreciable effect on performance. Although this effect has not been studied specifically, most investigators have used at least ten listeners. An even larger listener group may be necessary in order to obtain generally meaningful results.

Williams (1964) studied the requirements for training listeners with a group of six unfamiliar male speakers. One experiment consisted of a series of training trials; each speaker read two sentences during a given trial. The listener responded to each sentence by operating one of six push buttons representing the six speakers. Following his response, a small light was flashed above the button that should have been pushed for correct speaker identification. Eight training trials were required for the average score for 16 listeners to reach above 50%. In another experiment, Williams compared two methods of training listeners. One was the push-button method described above. In the second method, each speaker was represented by a number from one to six, and the listener responded by writing down one of these numbers. The correct speaker number was announced after each response. No significant difference was found between these two methods of training. The write-down method has the advantage that the test can be administered to many listeners at the same time.

4. *Mode of Presentation of Speech Material*

Two modes have been used to present the speech material to the listeners: a fixed-sequence mode and a free-comparison mode. 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. Each test item typically consists of a sequence of several speech samples. When the sequence contains more than two speech samples, the first sample may be from the speaker to be identified, and the following samples may be response alternatives. In the free-comparison mode, the listener controls the presentation of speech samples. Each test item consists of a limited time interval during which several speech samples are continuously available to the listener. The listener is provided with a switch allowing him to hear the available samples, one after another, in any desired order. He can spend different amounts of time on the individual samples, and within the overall time limit he can usually switch back to particular samples before responding.

Williamson (1961) compared the fixed-sequence and free-comparison modes of presentation for test items involving only two speech samples. Listeners

were required to state whether the two samples were uttered by the same or by different speakers. The fixed-sequence mode produced considerably higher scores than the free-comparison mode. When the time interval of each test item in the free-comparison mode was increased from 5 sec to 10 sec, there was a slight improvement in performance. However, it cannot be concluded that the fixed-sequence mode is generally superior to the free-comparison mode. As will be explained later, 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.

5. *Task Assigned to Listeners*

A further variable affecting performance is the task assigned to the listener. The task determines what kind of judgment the listener must make; some judgments are apparently easier to make, or more reliable, than others. As shown in Table 7, listener tasks may be grouped into three basis classes. There are tasks

TABLE 7. Classification of listener tasks.

Description	*Speech Samples/Test Item*	*Example*
Tasks Involving Long-Term Memory	1	Listener identifies speaker who is personally known to him.
Tasks Involving Direct Comparisons of Speech Samples	2	Listener decides whether samples are similar enough to have been produced by same speaker.
	3 or More	Listener decides which reference sample is most similar to test sample.
Tasks Involving Voice-Attribute Ratings	1	Listener rates sample on many semantic-differential scales.

involving (1) long-term memory, (2) direct comparisons of speech samples, and (3) voice-attribute ratings.

Tasks involving long-term memory require that the listener has previously heard at least one of the speakers participating in the test. The listener may be assigned the task of deciding, for each test item, whether he has heard the speaker before. In most of these tasks, however, the listener is assumed to be familiar with the voices of all participating speakers. For example, if the listener is personally acquainted with the speakers, he can rely on his memory of their voices in identifying the speaker of a given test item. The ability of the listener to perform this task depends on the degree to which he is familiar with each speaker's voice. In another task, the listener does not know the speakers personally, but the test format offers him various opportunities to associate their names and voices.

Tasks involving direct comparisons of speech samples do not require that

the listener be familiar with the speakers' voices. First, consider tasks in which each test item consists of only two speech samples. In one such task, the listener rates the level of similarity of the samples. He usually responds with reference to a specific rating scale. In another task, the listener decides whether the two samples are similar enough to have been produced by the same speaker. This is called a discrimination task. Because of the binary nature of the listener's response, no rating scale is necessary. It is assumed that the listener makes use of an internal decision threshold; the observed level of similarity must reach this threshold in order for the listener to report that the samples were produced by the same speaker.

Different listeners tend to use widely different decision thresholds. While some listeners are willing to ascribe the two samples to a common speaker if the samples appear grossly similar, other listeners are willing to do so only if the samples appear almost identical. To overcome this problem, the listener is usually required to rate his confidence of the correctness of each decision. His response then consists of two parts, a same-different decision and a confidence rating. Such responses can be analyzed in a manner that eliminates the effect of different decision thresholds. The procedure will be described in Section C4.

Now consider tasks involving direct comparisons of three or more speech samples. If two of the samples represent a common speaker, the listener may be assigned the task of identifying these samples. In most tasks, however, the comparisons are made with respect to one particular sample, which is called the test sample, and the remaining samples serve as reference samples. Each reference sample usually represents a different speaker. The listener may be asked to decide which reference sample is most similar to the test sample. Assuming that the speaker of the test sample is represented by one of the reference samples, this task would be expected to result in the identification of the speaker of the test sample. In another task, the listener rates the levels of similarity between the test sample and each of the reference samples. An appropriate rating scale must be provided for this task.

Tasks involving voice-attribute ratings require a relatively lengthy speech sample. The listener rates this sample on a number of scales that are selected to measure certain perceptual attributes of the voice. In some tasks, the scales are related to specific aspects of speech production. The listener may rate a particular voice as having moderately low pitch, fairly precise articulation, and barely noticeable nasal resonance. Other tasks employ semantic-differential scales. A particular voice may be rated as being more rich than thin, more rumbling than whining, more hard than soft, more agitated than serene, and more rough than smooth. Up to 49 semantic-differential scales have been used to obtain such ratings. In both cases, the listener usually marks a response form that lists all scales.

C. TEST FORMATS

The variables described above are treated according to a particular test

format that is chosen primarily to suit the purpose of the experiment. The test format specifies the variable under study and provides as much control as possible over the remaining variables. Besides taking into account the speaker group, the speech material, the listener group, the mode of presentation, and the listeners' task, the test format is also concerned with the method by which the listeners' responses are analyzed. To some extent, the choice of the test format is influenced by practical considerations. These include the amount of work involved in preparing the desired recordings, presenting them to listeners, and analyzing the listeners' responses.

Several test formats that have been used repeatedly in various studies will be described and evaluated. All of these tests are of the forced-choice type. This means that the listener must respond categorically to each test item, no matter how uncertain he may feel about making a decision. The tests are labeled according to the listeners' task.

1. Speaker-Naming Test

In the speaker-naming test, the speech samples to be identified are presented sequentially to listeners who are already familiar with the voices of the participating speakers. Usually, each listener knows each speaker on a personal basis. The listener responds to each speech sample by writing the name of the speaker he believes has produced the sample. The advantage of this test is that it does not involve direct comparisons. Because there are no reference samples, the test is easy to construct and can be administered in a short period of time. A limitation is that all listeners may not be equally familiar with all voices. This could introduce bias into the responses of some listeners and thereby lead to large individual differences. From a practical standpoint, it is often difficult to find many listeners who are very familiar with the voices of several common speakers. Another limitation is the high accuracy of long-term memory for familiar voices; the listeners' responses may be only slightly affected by changes in the experimental variable.

2. Modified Speaker-Naming Test

The modified speaker-naming test allows the listeners to become familiar with the voices of previously unknown speakers. Before the test proper begins, each participating speaker identifies himself, usually by a number, a letter, or a pseudonym, and reads some training material. The actual test consists of a sequence of speech samples in which the speakers are heard at random. Each speech sample is followed by a suitable pause during which the listener responds. Before the next speech sample is presented, the true speaker may identify himself; this procedure ensures that the training of the listeners continues as the test progresses.

The obvious advantage of this test is that it does not require listeners who are already familiar with the voices of several common speakers. Another advan-

tage is that the test is not overly demanding with respect to its construction and administration. The training material and test items can usually be read by the speakers in the desired order during a single recording session. A limitation is that only about six speakers can be used in this test; listeners find it difficult to identify more than six speakers, even under optimal experimental conditions (Williams, 1964). Also, the level of performance tends to be relatively insensitive to changes in some experimental variables (Hecker and Williams, 1965).

3. *Multiple-Choice Identification Test*

In this test, the listener makes direct comparisons between a test sample and several labeled reference samples. The speaker of the test sample, who is to be identified, is known to be represented by one of the reference samples. Thus, the listener has only to decide which reference sample is most similar to the test sample. The speech samples may be presented in a fixed sequence, or they may be continuously available to the listener, who switches among them and makes free comparisons.

With the fixed-sequence mode of presentation, the number of reference samples is usually restricted to four because the demands on short-term memory would otherwise be excessive (Clarke, Becker, and Nixon, 1966). Figure 13

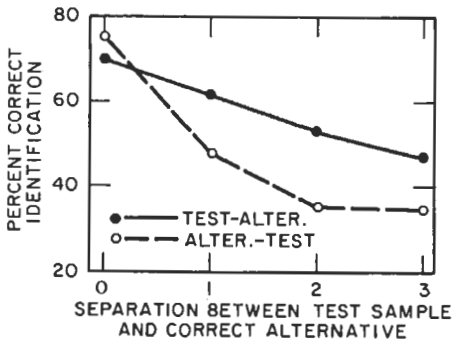

FIGURE 13. Effect on speaker identifiability of separation between test sample and correct alternative in four-choice identification test. Data are shown for both orders of presentation of test sample and alternatives. (Reprinted, by permission, from Clarke et al., 1966.)

illustrates the temporal biasing effect of short-term memory. The performance on a given test item depends on the separation between the test sample and the reference sample that is the correct response alternative. It is highest if these two samples follow each other directly and lowest if they are separated by the three remaining reference samples. This effect is especially pronounced when the test sample follows rather than precedes the reference samples. The effect probably increases as the number of reference samples is increased above four.

With the free-comparison mode of presentation, however, as many as eight reference samples have been employed (Stevens, Williams, Carbonell, and Woods, 1968). Here short-term memory plays a less important role, since the

listener can usually modify the temporal order of the speech samples several times before making his decision. The total number of speakers participating in the multiple-choice identification test may, of course, exceed the number of response alternatives offered on each test item.

The principal advantage of this test is that it is relatively easy to score, no matter which mode of presentation is used. With the fixed-sequence mode, if the number of reference samples is small, an additional advantage is that the construction and administration of the test are not overly time consuming. A different advantage may be realized with the free-comparison mode. Since the listener can return to particular speech samples and compare various aspects of the samples before arriving at a decision, he may perform better than he can with the fixed-sequence mode, where his decision must be based on a single, fleeting comparison between the test sample and each reference sample.

The limitation of this test, for the fixed-sequence mode of presentation, is the biasing effect of the short-term memory. This applies especially if the number of reference samples is large. If the number of reference samples is small, another limitation arises; different listeners may employ different decision strategies. In an ABX design,[2] for example, some listeners may compare X individually with A and B, while other listeners may disregard A and only compare X with B. This may lead to large individual differences. The limitation for the free-comparison mode of presentation is that the instrumentation requirements may be severe, particularly if many reference samples are used.

4. Discrimination Test

In the discrimination test, each test item consists of two speech samples. Either the fixed-sequence mode or the free-comparison mode of presentation may be employed. The listener decides whether the two samples were produced by the same or different speakers. As mentioned earlier, the listener is usually required to rate his confidence in the correctness of each same-different decision so that his decision threshold may be taken into account.

The listener's responses are analyzed in the following manner. Suppose three confidence ratings are used; the listener must report either that he is very sure his decision is correct, that he is fairly sure his decision is correct, or that his decision is only a best guess. This would allow the construction of a six-point similarity scale ranging from *same speaker, very sure* to *different speakers, very sure*. Responses to pairs of speech samples that are, in fact, by the same speaker and to pairs of speech samples that are, in fact, by different speakers are listed separately along this similarity scale.

For example, Table 8 lists possible responses of two listeners on a 200-item discrimination test employing three confidence ratings. There are as many test items in which the two samples are by the same speaker as there are test items in which the two samples are by different speakers. From an inspection

[2] In the ABX design, A and B are reference samples, and X is the test sample.

TABLE 8. Possible responses of two listeners on 200-item discrimination test employing three confidence ratings. Listeners use different criteria for deciding whether speech samples are by same speaker or by different speakers.

Same/Diff. Decision	*Confidence Rating*	*Similarity Scale*	*Listener X*		*Listener Y*	
			In Fact Same	*In Fact Diff.*	*In Fact Same*	*In Fact Diff.*
Same	+ + +	1	40	5	77	22
Same	+ +	2	25	8	13	20
Same	+	3	19	17	5	19
Diff.	+	4	9	20	2	12
Diff.	+ +	5	4	23	2	15
Diff.	+ + +	6	3	27	1	12
		Total Items	100	100	100	100

of the data it is evident that if the confidence ratings are not considered (i.e., if the six-point similarity scale is reduced to a simple same-different scale), Listener X would be 40 + 25 + 19 = 84% correct on samples by the same speaker and 20 + 23 + 27 = 70% correct on samples by different speakers. Listener Y, on the other hand, would be 95% correct on samples by the same speaker but only 39% correct on samples by different speakers. The two listeners are using different decision thresholds; Listener Y is much more inclined to decide *same* than Listener X.

The next step in the analysis of the listener's responses consists of a rearrangement of the data according to the six decision criteria inherent in the response format. As shown in Table 9, the first decision criterion (labeled A) includes only the first level of the six-point similarity scale. If a listener were to adopt this criterion, he would decide *same* only if he felt very sure that the speech samples are by the same speaker; if he had the slightest reservation he would decide *different*. The second decision criterion (labeled B) includes

TABLE 9. Probabilities (in percent) of listener deciding *same* when speech samples are, in fact, by same speaker and when speech samples are, in fact, by different speakers, for six decision criteria.

Decision Criterion	*Included Similarity Levels*	*Listener X*		*Listener Y*	
		In Fact Same	*In Fact Diff.*	*In Fact Same*	*In Fact Diff.*
A	1	40	5	77	22
B	1 2	65	13	90	42
C	1 2 3	84	30	95	61
D	1 2 3 4	93	50	97	73
E	1 2 3 4 5	97	73	99	88
F	1 2 3 4 5 6	100	100	100	100

two similarity levels. A listener adopting this criterion would decide *same* if he felt either very sure or fairly sure that the speech samples are by the same speaker. Each successive decision criterion includes one more similarity level. The last decision criterion (labeled F) thus includes all six similarity levels. Using this criterion, a listener would always decide *same*. The column entries of Table 9 are the cumulative sums of corresponding column entries of Table 8. Table 9 may be viewed as listing the probabilities of each listener deciding *same* when the speech samples are, in fact, by the same speaker and when the speech samples are, in fact, by different speakers.

The data of Table 9 are plotted in Figure 14. This figure shows that the

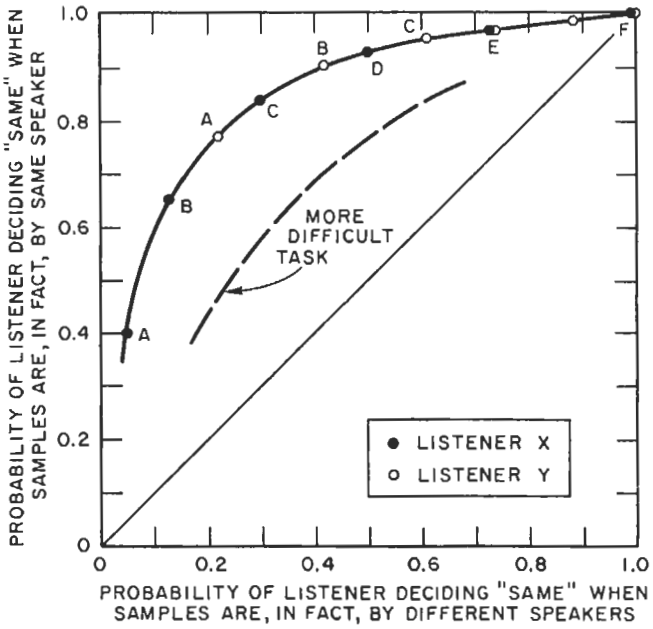

FIGURE 14. Receiver Operating Characteristic curves for two listeners in discrimination test using confidence ratings.

so-called Receiver Operating Characteristic (ROC) curves[3] for the two listeners coincide. Thus, although these listeners use different decision thresholds, they are equally capable of discriminating among speakers. The final measure of performance, namely the position and shape of the ROC curve, is not affected by differences in decision thresholds. If one listener had more difficulty with the discrimination task, his ROC curve would appear shifted toward the diagonal of Figure 14.

It is apparent that the discrimination test can be used to authenticate a

[3]For a detailed discussion on the theory and uses of ROC curves, see Egan, Schulman, and Greenberg (1959).

speaker who claims a particular identity. The speech sample submitted for authentication, which usually includes the name of the alleged speaker, is paired with a reference sample that has, in fact, been produced by the speaker whose identity is being claimed. If the listener decides that the two speech samples are similar enough to have been produced by the same speaker, he is effectively authenticating, or accepting, the challenging speaker. On the other hand, if the listener decides that the samples were produced by different speakers, he is rejecting the challenging speaker and designating him an imposter. The listener can commit two types of errors; he may falsely accept an imposter, or he may falsely reject a speaker who has the claimed identity. Depending on the specific application, one type of error can be more costly than the other, and the listener may adjust his decision threshold accordingly.

From a theoretical point of view, the discrimination test offers three advantages over the multiple-choice identification test. First, because each test item consists of only two speech samples, the test makes minimal demands on short-term memory. Also, the listener can employ only one decision strategy; he must appraise the level of similarity of the two samples presented to him. And finally, the effects of different listeners using different decision thresholds can be eliminated with the aid of confidence ratings. The practical advantages are that the test is relatively easy to construct and that it requires little time for administration. Many investigators consider this the test of choice. In comparison with many other tests, however, the discrimination test employing confidence ratings is more difficult to score. The results are usually expressed in the form of ROC curves.

5. *Identification-Discrimination Test*

As its name implies, this test has some of the features of the multiple-choice identification test and some of the features of the discrimination test. The identification-discrimination test has the same basic construction as the multiple-choice identification test; each test item consists of a test sample and several labeled reference samples. However, the test sample may have been produced by a speaker who is not represented by any of the reference samples. The listener is usually told this.

Assuming that the listener can find a reference sample resembling the test sample, he must now also consider the possibility that the speaker of the selected reference sample did not produce the test sample. The listener carries out two tasks in succession; first he decides which reference sample is most similar to the test sample, and then he decides whether the two samples are similar enough to have been produced by the same speaker. If the listener should find that none of the reference samples resemble the test sample, or that the selected reference sample and the test sample are not similar enough, he reports that he cannot identify the speaker of the test sample. Thus, in this test, there is one more response alternative than the number of reference samples.

The identification-discrimination test can be used to authenticate a speaker who requests access to certain information or facilities. Consider a situation in which several persons are authorized to have access, and in which a challenging speaker has not stated his name.[4] A speech sample from the challenging speaker is taken as the test sample. Each authorized person has previously recorded a suitable reference sample. The listener determines whether the test sample resembles any of the reference samples to a degree that is sufficient for the challenging speaker to be authenticated and thereby granted access. As a secondary function, the listener also determines the identity of the authenticated speaker. Occasionally, however, the listener may correctly authenticate a speaker on the basis of an incorrect identification.

This test has the advantage of being somewhat more realistic than the multiple-choice identification test. In most practical situations, especially when the number of available reference samples is small, there is always the possibility that a particular speaker cannot be identified. A limitation of the identification-discrimination test is that it is more difficult to score than either the multiple-choice identification test or the discrimination test. The listener can make three types of errors; he may either incorrectly identify a speaker who is represented by the reference samples, falsely reject such a speaker, or falsely accept a speaker who is not represented by the reference samples. Generally, these types of errors are not assigned equal weight in computing an overall measure of performance, so that they must be tallied separately. If confidence ratings are used, the analysis of the listeners' responses is even more complex.

6. Voice-Attribute Rating Test

This test is also concerned with speaker identification and discrimination, but it does not require the listener to perform these tasks directly. Instead, speech samples from different speakers are rated on a number of psychophysical scales. These scales are selected to measure certain perceptual attributes that are differentially shared by all voices. The speech samples are usually presented in a fixed sequence; the duration of each sample is sufficient for the listener to mark all scales. In an effort to determine the characteristics of the psychological space in which the speech samples were perceived by the listeners, the ratings are first subjected to an analysis of variance and then to a factor analysis. The separation of the speech samples in the psychological space provides information on the perceptual differences among the voices and hence on the feasibility of speaker identification and discrimination. The procedure involved in using this test will be described in more detail in Section D.

The voice-attribute rating test offers various opportunities to study the many factors underlying overall listener responses, and it is easier to construct and

[4] A challenging speaker is usually very cooperative; he can be asked to identify himself. Authentication can then be accomplished with the discrimination test, as described earlier.

administer than most other tests. However, the analysis of the ratings is complex and time consuming. Another limitation is the relatively poor reliability of the test; the results appear to depend greatly on the particular speakers, scales, and listeners employed.

D. PERCEPTUAL BASES OF SPEAKER RECOGNITION

In one of the earliest studies on speaker recognition (McGehee, 1944), an attempt was made to determine why some voices could be identified more readily than others, and why certain voices tended to be confused by listeners. Groups of listeners were required to rate the voices of several speakers in terms of their apparent unlikeness (i.e., uniqueness), agreeableness, pitch, and speaking rate. Although the results were difficult to interpret, this experiment may be regarded as the first of many inquiries into the perceptual bases of speaker recognition.

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 employ indirect judgments and relatively complex analytical procedures. If the hypothesized parameters can be adequately defined and measured, they may provide a unique and compact description of a particular voice. Perceptual parameters are usually explored by means of the voice-attribute rating test. Listeners rate different voices on a large number of psychophysical scales; semantic-differential scales are commonly used for this purpose. Statistical analyses of the ratings often permit a grouping of the scales suggesting the involvement of as few as four perceptual parameters.

The general procedure was outlined by Voiers (1964). In this study, 32 listeners rated each of 16 voices on 49 semantic-differential scales. An analysis of variance was performed to determine the major sources of variance in the ratings; they were the speakers, the listeners, and the speaker-listener interaction. For each of these three effects, a factor analysis was performed to determine how many orthogonal factors would suffice to account for most of the variability of the effect. Four factors were required to account for the variability of the speaker effect, each factor being represented by several of the 49 scales. On the basis of the labeling of the scales, the four factors were named *clarity*, *roughness*, *magnitude*, and *animation*. Similarly, six factors were required for the listener effect, and five factors for the interaction effect. The four factors required for the speaker effect were viewed as providing a coordinate system for a four-dimensional psychological space that contains all 16 voices. Thus, four numbers would specify the location of a particular voice in this space. Voices that are perceptually identical would presumably be described by the same set of numbers.

Unfortunately, the association of particular semantic-differential scales with

each of the factors required for a given effect is not unique; it depends on the placement of the coordinate system within the multidimensional space. A criterion for the orientation of the system axes must be selected, and this selection tends to be somewhat arbitrary. For the grouping of scales represented by the four factors named above, a normalized varimax criterion was selected. Although other criteria might have led to different groupings and hence to different factor names, the dimensionality of the observation space would have remained the same.

In a similar study by Holmgren (1963), 10 listeners rated each of 10 voices on only 12 semantic-differential scales. The smaller number of scales was still considered sufficient to provide for reliable differentiation among the speakers. Again, four factors were required to account for most of the variability of the speaker effect; these factors were given the names *intensity, quality, pitch,* and *rate.* Voiers (1965) tried various modifications of the voice-attribute rating test, including the use of rating forms featuring combined scales, filler items, and maximum connotative dissimilarity between adjacent items. None of these modifications resulted in the consistent emergence of more than four factors in the analysis of the variability of the speaker effect. Only in a study concerned with the use of this approach for evaluating communication systems were five factors required (Voiers, Cohen, and Mickunas, 1965). This study will be discussed in Section F.

There has been considerable interest in identifying acoustical correlates for each of the perceptual parameters that can be isolated by means of the voice-attribute rating test. Finding such correlates could promote a better understanding of the acoustical manifestations of speaker identity. In most studies, the search for acoustical correlates involves an intercorrelation between perceptual variables and physical measures, followed by a factor analysis.

Voiers (1965) performed a factor analysis of 23 variables, 16 of which were derived from listener ratings of unprocessed and processed speech samples. (Each of four perceptual factors was represented by data for unprocessed speech, vocoderized speech, lowpass-filtered speech, and highpass-filtered speech.) The remaining seven variables were physical measures related to average fundamental frequency, average speaking rate, average spectral characteristics, and speech level. For the case of the unprocessed speech, five factors emerged from the analysis; three of these were defined by perceptual variables and two primarily by physical measures. Each of the three perceptual factors was found to be more or less correlated with several physical measures. For the case of the processed speech, the physical correlates of some perceptual factors were altered depending on the type of processing involved.

In a related study by Holmgren (1967), two sets of data were obtained; these were average ratings of the voices of 10 speakers on 12 semantic-differential scales, and seven physical measures of the speech of the same speakers. The physical measures included the mean and variance of the amplitude of voiced speech sounds, the amplitude of voiceless speech sounds, and the fundamental frequency. A measure of duration was also taken. These two sets of

data were intercorrelated, and the resulting 19 × 19 intercorrelation matrix was subjected to a factor analysis. Five factors were required to account for over 90% of the total variance. Three of these factors were represented by both voice ratings and physical measures, and two were represented primarily by physical measures. Only a few of the scales were found to be correlated with expected physical measures (e.g., the scales *low-high* and *deep-shallow* were both highly correlated with the fundamental frequency). Most scales, however, were not so correlated (e.g., the scale *slow-fast* appeared to be relatively independent of the duration measure).

Both of these studies indicate that the dimensionality of the observation space is higher when perceptual variables and physical measures are combined than when only perceptual variables are considered. This would suggest that the two sets of data sample different portions of some common information on voice characteristics. A relatively low intercorrelation between the two sets of data would therefore be expected. In the ideal case, where both sets of data sample the same information, one set would be totally redundant and would not increase the dimensionality of the observation space if it were added to the other set. By definition, the two sets of data would be perfectly intercorrelated. The general finding that the perceptual variables cannot be clearly related to the physical measures tends to support this interpretation. These studies illustrate the difficulties encountered in searching for acoustical correlates of the perceptual parameters. The degree of success that can be achieved depends largely on the labeling of the perceptual factors (which, in turn, depends on the speakers, the listeners, and the choice of scales), and on the physical measures employed.

Clarke and Becker (1969) approached the problem somewhat differently. They used five graduate students of speech science, who first took a multiple-choice identification test. The students then attempted to define a set of psychophysical scales which would be relevant to the task of differentiating among voices. Six scales were specified: *pitch, pitch variability, rate, click-like elements, sibilant intensity,* and *breathiness.* These scales were used to rate the voices heard in the multiple-choice identification test, and decision rules were applied to the ratings so that the resulting scores could be compared with the mean score obtained on the multiple-choice identification test. Physical measures of fundamental frequency, long-term spectral energy, and duration (normalized for sentence length) were also obtained, and the same decision rules were applied to these data. It was found that the scores based on the ratings were much lower than the mean score achieved on the multiple-choice identification test; the scores based on the ratings were also lower than some of the scores based on physical measures. An analysis of the relations between the six scales and the physical measures revealed that the *pitch* and *rate* scales were highly correlated with the measures of fundamental frequency and duration, respectively. The remaining scales were not found to be correlated with any of the available physical measures.

From these observations it was concluded that the listener extracts more

information from speech samples than is contained in either voice-attribute ratings or relatively simple physical measures. Nevertheless, voice-attribute ratings and physical measures appear to contain much information that is relevant to speaker discrimination and identification. The investigators carefully point out that while their study shows what kinds of information can be used by listeners in differentiating among voices, it does not demonstrate that listeners do indeed make use of such information. Thus, the question about the perceptual bases of speaker recognition and their acoustical correlates remains largely unresolved.

E. ACOUSTICAL MANIFESTATIONS OF SPEAKER IDENTITY

Many experimental studies have attempted to determine what features of the speech signal carry information relevant to the identity of the speaker. Unfortunately, only a small number of these studies have been designed specifically to examine interspeaker and intraspeaker variability. Rarely has the speech signal been subjected to a detailed acoustical analysis in order to explore its speaker-dependent parameters. In most studies, the speech signal was degraded or distorted in a particular manner before it was presented to listeners. The amount by which test scores were thereby reduced was taken as an indication of the importance of those features of the speech signal that were distorted or eliminated.

Of considerable interest has been the question of what portions of the broadband frequency spectrum contribute most to speaker identifiability. Lowpass, highpass, and bandpass filtering of the speech signal have been employed as means for answering this question. The results of two studies suggest that the removal of spectral energy below 500 Hz or above 3 kHz does not have much effect on speaker-identification scores (Pollack, Pickett, and Sumby, 1954; Peters, 1954). But the results of a later study, which are shown in Figure 15, indicate that spectral energy outside of this frequency range may also contribute to speaker identifiability (Clarke, Becker, and Nixon, 1966). Lowpass filtering has sometimes been used to remove the message content of the speech signal. Passing only frequencies below 500 Hz greatly reduces speaker identifiability (Skalbeck, 1955) but apparently still allows listeners to assess personality factors related to hypertension (Starkweather, 1956). In one study, the speech signal was passed through single octave-band filters, and the highest speaker-identification scores were achieved using the filter covering the range 1.2–2.4 kHz (Peters, 1954). In a similar experiment, various octave bands of the broadband speech signal were individually emphasized; when the band containing the fundamental frequency was emphasized, the test scores were higher than the scores obtained without spectral emphasis (Peters, 1956).

The effect on speaker-identification scores of adding white noise to the speech signal has also been investigated. The results of one study (Clarke,

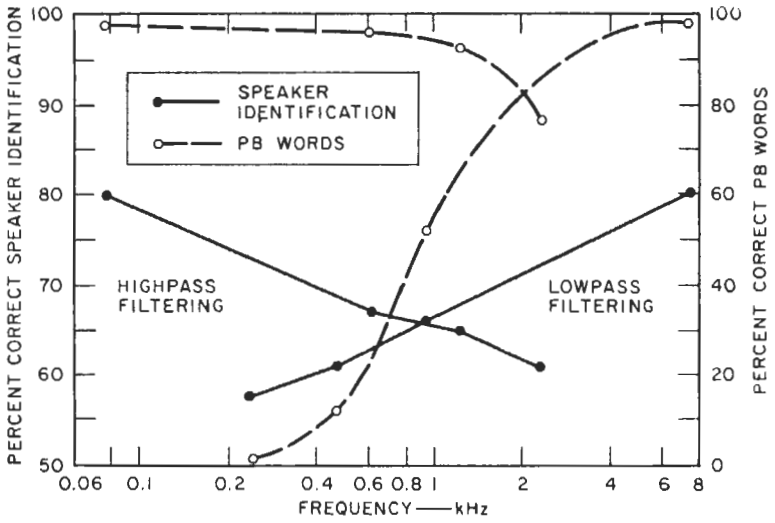

FIGURE 15. Effects of highpass and lowpass filtering on speaker identifiability and word intelligibility. (Reprinted, by permission, from Clarke et al., 1966.)

Becker, and Nixon, 1966) are shown in Figure 16. It is difficult to specify what parameters of the speech signal are disturbed most when random noise is added. Hence, not much is learned from experiments of this kind. Several

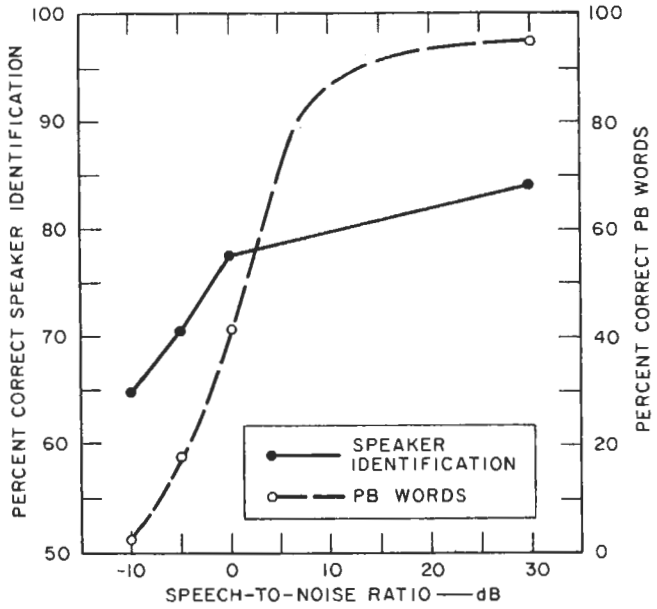

FIGURE 16. Effects of additive noise on speaker identifiability and word intelligibility. (Adapted, by permission, from Clarke et al., 1966.)

studies have employed backward-played speech as a form of distortion. The temporal reversal of the speech signal not only removes its message content, but the sequence of normal articulatory events is also disturbed. Consequently, the ability of listeners to identify familiar speakers is appreciably impaired (Skalbeck, 1955; Williams, 1964; Clarke, Becker, and Nixon, 1966; Bricker and Pruzansky, 1966). As indicated in Figure 17, this effect is noted for all types

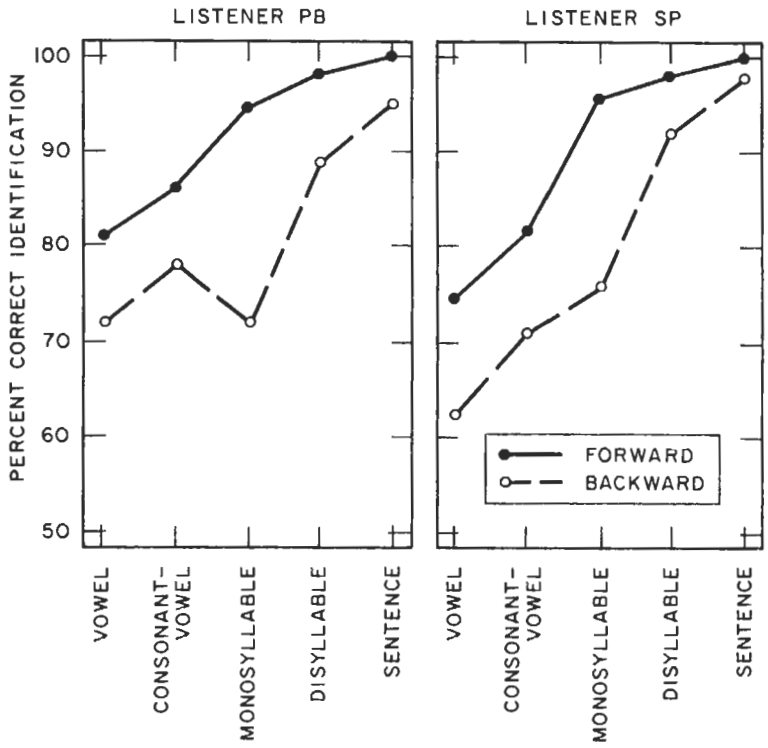

FIGURE 17. Speaker identifiability for various types of speech material presented in forward and backward order. Data are shown for two listeners. (Reprinted, by permission, from Bricker and Pruzansky, 1966.)

of speech material, including excerpted vowels, but is especially pronounced for monosyllables. Thus, temporal clues seem to be important for speaker identification. It is not clear whether listeners use temporal clues per se or whether their judgments depend on a perceptually realistic speech signal. The ability to identify speakers has been learned with natural speech over many years; this ability may not be readily transferable to a novel form of speech distortion that can exist only in the laboratory.

It will be recalled from Chapter II that both static and dynamic aspects of articulation may be expected to impart speaker-dependent characteristics on the speech signal. The contribution to speaker identifiability of only the static

aspects may be studied by restricting the speech material to sustained isolated vowels. Even when such vowels are artificially prolonged with tape loops, the speaker-identification scores are extremely low (Skalbeck, 1955). This demonstrates the importance of the dynamic aspects of articulation. There is some evidence that the relative identifiability of different speakers, and the particular confusions that arise, vary from vowel to vowel (Bricker and Pruzansky, 1966). Lowpass and highpass filtering of isolated vowel sounds have been used to investigate the relative contributions to speaker identifiability of various portions of the steady-state vowel spectrum. Removal of spectral energy below about 1 kHz or above about 4 kHz does not appreciably affect test scores (Compton, 1963; Ramishvili, 1966). Thus, the relevant spectral clues appear to be concentrated in the ranges of the second, third, and fourth formants.

Many studies describe efforts to separate the contributions to speaker identifiability of glottal-source characteristics and articulatory characteristics. The results of several experiments indicate that the involvement of the glottal source aids speaker identification. Whispered speech, for example, leads to significantly lower test scores than normal speech (see Figure 12); the scores for whispered speech are even lower than those for backward-played speech (Williams, 1964). A speaker-naming test employing isolated utterances of Russian phonemes produced higher scores for voiced speech sounds, especially for vowels, than for voiceless speech sounds (Ramishvili, 1966). Nevertheless, as would be expected on theoretical grounds, voiceless speech sounds do carry some information about the speaker. Isolated utterances of the voiceless fricatives [s] and [ʃ] allow listeners to identify at least the sex of the speaker (Schwartz, 1968; Ingemann, 1968).

Shearme and Holmes (1959) clearly demonstrated the important role played by articulatory characteristics. 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. Various portions of the spectrum were shifted upward in frequency by different amounts; the first formant was effectively raised by 100 Hz and the second and third formants by 300 Hz. Treated and untreated speech samples were presented to listeners in a discrimination test. When the paired speech samples were, in fact, produced by the same speaker but only one was treated, the listener believed that two speakers were involved. When the paired speech samples were, in fact, produced by different speakers, treatment of only one of the samples exaggerated the perceptual difference. In order to exclude possible clues arising from differences in intonation patterns, all speech samples were processed by the vocoder to remove the normal fluctuations in fundamental frequency. Although the formant-frequency translations used in this study are crude compared with the relative spacings of the formant frequencies which may distinguish different speakers, the study does suggest that articulatory characteristics are more important than glottal-source characteristics.

Miller (1964) conducted a more direct investigation of the relative contri-

butions of articulatory and glottal-source characteristics to speaker identifiability. In this study, a computer was used to synthesize speech signals which might be encountered if it were physiologically possible to interchange either the vocal tracts or the larynges of two speakers. By inverse filtering in synchronism with fundamental frequency, phonetically equivalent speech samples from several speakers were reduced to their two basic components: the vocal-tract transfer function and the glottal waveform (see Chapter II). These data were then recombined to produce various hybrid speech samples which could be compared with the natural speech samples.

The study consisted of four experiments. For the first experiment, two speakers were selected whose utterances of the word *hod* exhibited the same duration and the same fundamental frequency. When each hybrid sample was compared with the two natural samples, it sounded more like the speaker whose vocal tract was represented than like the speaker whose larynx was represented. Because the utterances of many speakers were incompatible with respect to duration and fundamental frequency, the second experiment employed several artificial glottal waveforms. While some of these waveforms were intended to be realistic, others were triangles, pulses, and sinusoids. These waveforms were individually combined with a vocal-tract transfer function that was derived from one speaker's utterance of the word *hod*. Listeners had the impression that all of the hybrid samples were produced by the same speaker, although there were obvious differences in speech quality. So far, the results indicated that, for a consonant-vowel-consonant syllable, the vocal-tract transfer function carries more information about the identity of a speaker than the glottal waveform.

The third experiment was designed to test whether this is also true for a sustained isolated vowel, where the vocal-tract transfer function is known to be relatively constant throughout an utterance. Six speakers produced isolated utterances of the vowel [ɑ], and the vocal-tract transfer functions extracted from these utterances were combined with two artificial but realistic glottal waveforms. The perceptual differences due to different vocal-tract transfer functions were found to be much greater than those due to different glottal waveforms. In the final experiment, each speech sample consisted of many repetitions of a single, 10-msec fundamental period. Two natural samples and two hybrid samples were constructed from representative vocal-tract transfer functions and realistic glottal waveforms. These speech samples were presented to listeners in a two-choice identification test; the reference samples were always the two natural samples, and the test sample was either a natural sample or a hybrid sample. The results showed that each hybrid sample tended to be matched with the natural sample having the same vocal-tract transfer function. Thus, this study provides further evidence that articulatory characteristics contribute more to speaker identifiability than glottal-source characteristics.[5]

[5]It is conceivable that the results of this study were influenced by factors inherent to the technique of inverse filtering in synchronism with fundamental frequency. For example, the

As mentioned in Section D, experimental efforts to identify acoustical correlates for each of the perceptual parameters isolated by the voice-attribute rating test have been largely unsuccessful. It is entirely possible, however, that further studies employing a greater number and variety of physical measures will provide insight into the relation between perceptual and acoustical variables. The desired acoustical correlates would allow a much more compact description of a given voice than is represented by the speech waveform. Numerous experiments have demonstrated that a complete physical description of speech is not necessary for performing speaker recognition by machine (i.e., without human intervention). But the relative importance of various descriptors of the speech signal has not been systematically examined. Further studies along these lines could contribute to a better understanding of the acoustical manifestations of speaker identity.

F. EVALUATION OF COMMUNICATION SYSTEMS

It is often necessary to evaluate the capability of a communication system to transmit speech of acceptable quality. The concept of speech quality is difficult to define; it probably involves such factors as message intelligibility, naturalness, speaker recognizability, and aesthetic appeal. Only some of these factors can be measured objectively (Hecker and Guttman, 1967). Message intelligibility can be estimated by transmitting lists of words over the system and determining the percentage of words correctly understood by a group of listeners. Several intelligibility tests have been developed for this purpose, such as the Harvard Phonetically Balanced (PB) Word Test (Egan, 1948). 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). For an extensive evaluation of speech-communication systems, it is common practice to use both an intelligibility test and a speaker-recognition test.

Studies in which both types of tests are used under the same conditions provide opportunities for examining the relationship between intelligibility and speaker recognizability. Hecker and Williams (1965) obtained scores for five speech-processing systems with a paired-comparison preference test of speech quality, two types of intelligibility tests, and a modified speaker-naming test. Two systems were found to be equivalent in speech quality, and two other 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 speak-

---

vocal-tract transfer function is assumed to be constant throughout each fundamental period. Whether this assumption could have biased some experiments cannot be conclusively determined.

er-naming test. In a related experiment, Clarke, Becker, and Nixon (1966) used the Harvard PB Word Test and a two-choice identification test to evaluate the effects of three types of speech-signal degradation: lowpass filtering, highpass filtering, and additive noise. The results, shown in Figures 15 and 16, indicate that speaker identifiability is less dramatically affected than intelligibility over a given range of degradation.

Both of these studies suggest that either speaker-recognition tests are inherently less sensitive than some intelligibility tests, or the acoustical manifestations of speaker identity are more resistant to signal processing than the acoustical manifestations of the semantic content. The first interpretation is regarded as the more likely. It should be pointed out that neither study tested analysis-synthesis systems such as vocoders. Vocoders are designed to accomplish a more economical transmission of the speech signal[6] with a minimal reduction in message intelligibility. In order to reach this objective, the speech signal is often processed in a manner which has a detrimental effect on speaker recognizability. The acoustical manifestations of speaker identity can be either totally destroyed or systematically transformed. In the latter case, listeners may or may not be able to learn a new set of clues for each speaker. Thus, when analysis-synthesis systems are evaluated, speaker-recognition scores can be more sensitive indicators of system performance than intelligibility scores.

Clarke, Becker, and Nixon (1966) found that the two-choice identification test produced a much greater interlistener variance than the Harvard PB Word Test. Apparently the individual listeners differed greatly in their ability to recognize speakers, but they were about equally able to understand what the speakers said. This finding is not surprising, considering how little time listeners have to become familiar with the voices heard in a speaker-recognition test; in contrast, listeners have spent many years learning to understand the spoken language. The large interlistener variance observed with the two-choice identification test can also be interpreted in another way. In the beginning of the test, different listeners may have selected different sets of clues for differentiating among the speakers. Some listeners may have inadvertently picked rather inefficient clues that they nevertheless continued to use throughout the test, whereas other listeners may have made better choices. This interpretation allows for the possibility that an individual listener may change his decision strategy as the test proceeds, with a corresponding change in his performance. Such temporal changes in the scores of an individual listener have not as yet been investigated.

The preceding discussion argues for a large number of listeners in evaluating communication systems with respect to speaker recognizability. It may also be beneficial to familiarize the listeners with the speakers' voices for a considerable period of time before data are collected, especially if unusual forms

[6]The processed speech signal may be transmitted over an analog link having less bandwidth, or over a digital link having a lower information rate, than would be required to transmit the unprocessed speech signal.

of speech processing are involved. Both of these precautionary measures may help to minimize the effects of individual differences in the final results.

The use of voice-attribute ratings represents an entirely different 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. This is primarily a practical consideration; theoretical questions concerning this technique have been considered in Section D.

G. LISTENER FALLIBILITY

The objective of most studies on speaker recognition by listening is, of course, to appraise the likelihood that a listener's judgment may be in error. In fact, one of the first studies of this kind was motivated by a legal question of listener fallibility that arose in the Lindbergh case of 1935 (McGehee, 1937). Lindbergh claimed that he recognized the voice of the defendant as the voice of his son's kidnapper heard almost three years earlier. Although his testimony was accepted by the court, the defense argued that such recognition was not entitled to much weight as evidence.

McGehee (1937) studied the reliability with which listeners can recognize unfamiliar voices. Groups of listeners participated in two experimental sessions that were separated in time from one day to five months. During the first session, they heard an unfamiliar speaker read a paragraph of text. During the second session, they heard the same paragraph read successively by five speakers, including the speaker from the first session. The ability of the listeners to recognize the speaker whom they heard before was investigated as a function of the time interval between the two sessions. The results, which are shown in Table 10, indicate that the reliability of recognition decreases rapidly as the time interval is extended beyond two weeks.

TABLE 10. Percent correct recognition of unfamiliar male speakers after various intervals of time. (Reprinted, by permission, from McGehee, 1937.)

Days			*Weeks*			*Months*		
1	*2*	*3*	*1*	*2*	*3*	*1*	*3*	*5*
83	83	81	81	69	51	57	35	13

The effect of increasing the number of unfamiliar speakers heard during the first session was also investigated. When one of two speakers heard during the first session spoke again during a second session two days later, 77% of the listeners recognized his voice. When five speakers participated in the first session, only 46% of the listeners could recognize one of their voices two days later. Vocal disguise was also found to be effective in lowering recognition scores. In this experiment, only one speaker was heard during the first session; he disguised his voice by changing its fundamental frequency. During the second session he used his normal voice. For a time interval of one day, correct recognition was reduced by 13%.

Stevens, Williams, Carbonell, and Woods (1968) investigated the ability of listeners to distinguish between familiar and unfamiliar voices. This study employed an identification-discrimination test. For each of 32 test items, a test sample and eight reference samples representing the same word were continuously available to the listener for free comparisons. In one experiment, four of the 32 test items involved utterances by unfamiliar speakers (i.e., speakers not represented by the reference samples); in another experiment, this number was increased to 16. No particular effort was made to select unfamiliar speakers whose utterances would be perceptually similar to any of the reference utterances. The results of these two experiments are shown in Table 11. It appears that in both cases the listeners were able to detect most of the unfamiliar speakers. Also, most of the familiar speakers were recognized as such (and correctly identified).

There are situations in which an imposter may try to deceive the listener by mimicking. Carbonell, Grignetti, Stevens, Williams, and Woods (1965) examined the effects of mimicking on the recognition of unfamiliar speakers. In the same identification-discrimination test described above, 16 of the 32 test items involved utterances by unfamiliar speakers; one-half of these utterances

TABLE 11. Percent correct recognition of familiar and unfamiliar male speakers by listening. Data are shown for two experimental conditions. (Reprinted, by permission, from Stevens et al., 1968.)

4 of 32 Test Items by Unfamiliar Speakers		
	Recognized As	
Speaker	*Familiar*	*Unfamiliar*
Familiar	88	12
Unfamiliar	6	94

16 of 32 Test Items by Unfamiliar Speakers		
	Recognized As	
Speaker	*Familiar*	*Unfamiliar*
Familiar	92	8
Unfamiliar	8	92

were carefully produced to sound as much as possible like certain reference utterances. The mimicking was done by two of the investigators, who also served as critical listeners in comparing their various versions of imposter utterances with the prototypes. Fewer mimicked utterances were falsely accepted than other foreign utterances, suggesting that imposters are not particularly successful in deceiving listeners.

In the absence of other data, however, this conclusion can only be regarded as tentative. The experiment employed isolated utterances of the bisyllabic words *sidewalk* and *dovetail.* Such utterances may be difficult to mimic because there is little opportunity to copy gross articulatory features; many of the perceptually distinguishing characteristics of these utterances may depend on properties of the vocal mechanism over which the speaker has no direct control. It is known that entertainers who impersonate famous people attempt to copy the prosodic features of their speech (patterns of intonation and stress). These features can be applied more readily to utterances of longer duration. A possible criticism of the experiment is that the investigators may not have been particularly good mimickers. Effective mimicking is a complex skill, if not an art.

It appears, then, that further studies must be undertaken in order to evaluate the reliability of speaker recognition by listening. Such studies are especially desirable in view of the many advantages of this method of speaker recognition over other methods to be described in the following chapters.

Chapter IV

SPEAKER RECOGNITION BY VISUAL COMPARISON OF SPECTROGRAMS

A. INTRODUCTION

This method of speaker recognition makes use of an instrument which converts the speech signal into a visual display. The instrument is called a sound spectrograph, and the display it provides is a sound spectrogram (or voiceprint). Spectrograms of different utterances of a given word or phrase are presented to a trained observer, who attempts to determine whether some utterances were produced by a common speaker. Because the method has obvious applications in criminology, many studies have been concerned with its reliability as a means of positive identification. The operation of the sound spectrograph will be described first in this chapter. The relevant variables of speaker recognition by visual comparison of spectrograms will be discussed next, and the commonly used test formats will be described. After this, the question of observer fallibility will be considered in some detail. Finally, this method will be compared with speaker recognition by listening.

B. SOUND SPECTROGRAPH

Various interests motivated the development of the sound spectrograph (Potter, Kopp, and Green, 1947). There was an interest in the detailed structure of the speech signal, which was insufficiently understood to settle certain fundamental arguments about speech production and perception. No instrument then available would simultaneously display the temporal and spectral properties of the speech signal. Another interest was to provide the deaf with a new form of visible speech.[1] It was believed that a deaf person could be taught to read a visible pattern which reflected the semantic contents of the speech signal. During the war, a need arose for an instrument which could be used to study speech-privacy systems. Several experimental sound spectrographs were built to meet these objectives.

The sound spectrograph consists of four basic parts: (1) a magnetic recording device, (2) a variable electronic filter, (3) a drum which is coupled to

[1] Alexander Melville Bell (1819–1905) developed a phonetic alphabet for teaching the deaf; he called this alphabet Visible Speech (see Wise, 1957).

the magnetic recording device and carries a sheet of special paper, and (4) an electric stylus which marks the paper as the drum rotates. The magnetic recording device is first used to record a short sample of speech; the duration of the speech sample corresponds to the time required for one revolution of the drum. The speech sample is then played back over and over again in order to analyze its spectral contents. For each revolution of the drum, the variable electronic filter passes only a certain band of frequencies, and the energy in this frequency band activates the electric stylus so that a straight line of varying darkness is produced across the paper. The darkness of the line at any point on the paper indicates how much energy is present in the speech signal at the specified time within the given frequency band. As the drum revolves, the pass-band of the variable electronic filter moves to increasingly higher frequencies, and the electric stylus moves parallel to the axis of the drum. Thus, a pattern of closely spaced lines is generated on the paper. This pattern, which is the spectrogram, has the dimensions of frequency, time, and amplitude (see Figure 7). The following paragraphs describe the operation of the sound spectrograph in greater detail.

Figure 18 shows a block diagram of an early commercial sound spectrograph.[2] In the record mode, a speech signal having a duration of less than 2.4 sec is recorded on the side of a rotating magnetic drum. The highest recorded frequency is about 8 kHz. During the reproduce mode, the magnetic drum revolves 3.33 times as fast as it does during the record mode, so that the highest reproduced frequency is $3.33 \times 8 = 26.6$ kHz. The reproduced signal is usually equalized to compensate for the falling spectrum of voiced speech sounds (see Chapter II), and then the signal is modulated with the output of a variable oscillator. Attached to the magnetic drum is a marking drum which holds a sheet of teledeltos paper. This paper, on which the spectrogram is to be recorded, blackens at points where a high-frequency current is passed through it. As the marking drum rotates, a stylus is passed across the teledeltos paper by means of a worm gear which is mechanically coupled to the drums. An additional coupling link controls the variable oscillator so that its frequency is linearly proportional to the position of the stylus.

The output of the modulator is applied to one of two bandpass filters. These filters have different bandwidths but are both centered on 35 kHz. When the stylus is in its initial position on the worm gear, the frequency of the variable oscillator is 35 kHz, so that there is an output from the bandpass filter only for a reproduced signal of zero frequency. Similarly, when the stylus is in its final position, the frequency of the variable oscillator is 61.6 kHz, and there is an output only for a reproduced frequency of $61.6 - 35.0 = 26.6$ kHz (highest reproduced frequency). Thus, the stylus first records the amplitude variations of the lowest frequency components of the speech signal. As the stylus moves gradually across the teledeltos paper, it records the amplitude variations of in-

[2] This sound spectrograph was manufactured by the Kay Electric Company, Pine Brook, New Jersey, and was marketed under the name Sona-Graph.

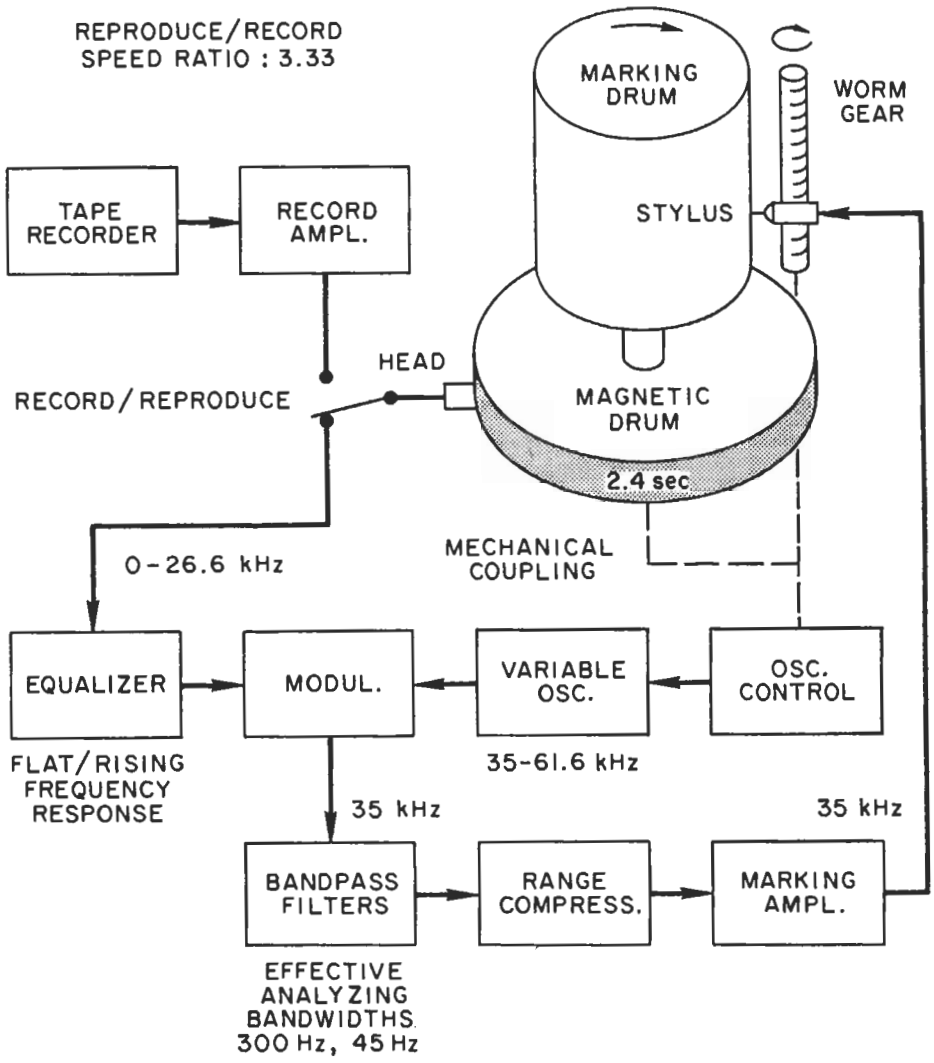

FIGURE 18. Block diagram of early commercial sound spectrograph.

creasingly higher frequency components, until the entire spectral range is finally covered. The output of the bandpass filter is led to the stylus via a range compressor which has the function of reducing the dynamic range of the signal from a nominal value of 40 dB to about 10 dB. The teledeltos paper cannot accommodate a wider dynamic range.

This sound spectrograph is calibrated so that one dimension of the resulting spectrogram has a linear 0–8 kHz frequency scale, the other dimension has a 0–2.4 sec time scale, and the darkness of the mark indicates relative amplitude. The two bandpass filters provide effective analyzing bandwidths of 300 Hz

and 45 Hz; the corresponding spectrograms are referred to as wide-band and narrow-band spectrograms. While both types of spectrograms are useful in speech research, only wide-band spectrograms are employed for speaker recognition. Because the wider bandpass filter can still resolve events that are separated by only about 5 msec, wide-band spectrograms show the individual vibratory cycles of the vocal folds as vertical striations. The formants are shown as curved horizontal bars. For this reason, a wide-band spectrogram is also called a bar spectrogram.

With the sound spectrograph just described, it takes about five minutes to prepare one spectrogram. Efforts to reduce this time, and to improve the quality of the spectrogram, have resulted in the high-speed sound spectrograph[3] (Presti, 1966). A block diagram of this instrument is shown in Figure 19. Instead of recording on a magnetic drum, a loop is formed from the magnetic tape containing the speech signal to be analyzed. If the tape was recorded at 7.5 ips, the loop will include up to 2.4 sec of speech signal. The loop is scanned by a rotating reproduce head at 12×7.5 ips. Assuming the highest frequency recorded at 7.5 ips to be 7 kHz, the highest reproduced frequency is $12 \times 7 = 84$ kHz. The reproduce head is mechanically coupled to the marking drum, and the reproduced signal is equalized and then modulated with the output of a variable oscillator.

The frequency of the variable oscillator depends on the position of the stylus on the worm gear. There is a choice between a linear and a logarithmic frequency scale; in both cases the oscillator covers the frequency range 126–210 kHz. This sound spectrograph uses three modulators. The output of the first modulator is applied to a 126-kHz bandpass filter. Because of its relatively high center frequency, this filter has very steep attenuation characteristics which improve the frequency resolution of the spectrogram. The function of the second modulator is to convert the frequency of the signal from 126 kHz to 30 kHz; the subsequent analyzing bandpass filters operate at 30 kHz. It is necessary to detect the output of the selected bandpass filter in order to accommodate an amplitude quantizer, described below. The third modulator gates a 12-kHz carrier according to the output of either the detector or the amplitude quantizer, and the modulated signal is used to mark the teledeltos paper.

The amplitude quantizer (Prestigiacomo, 1962) converts the detected (i.e., rectified and smoothed) signal into a sequence of narrow pulses. Whenever the amplitude of the input waveform passes through any one of eight levels that are separated by 6 dB, either in the direction of increasing amplitude or in the direction of decreasing amplitude, a pulse occurs at the output. Thus, the pulses are more closely spaced for rapid changes in amplitude than for slow changes in amplitude. As the stylus moves along the worm gear, and the effective center frequency of the analyzing bandpass filter is increased, the pulse

[3]The high-speed sound spectrograph is manufactured by the Voiceprint Laboratories, Somerville, New Jersey.

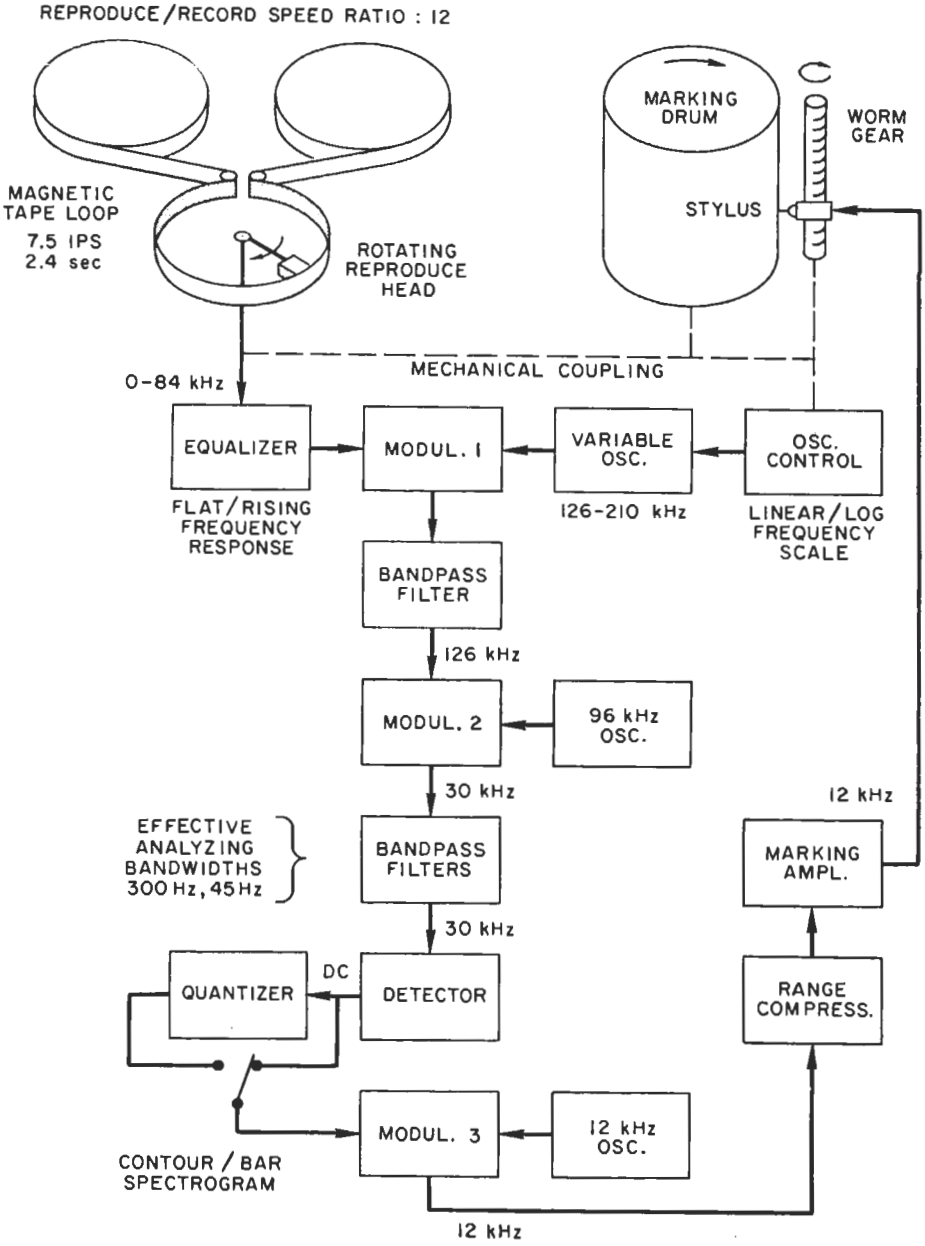

FIGURE 19. Block diagram of high-speed sound spectrograph. (Adapted, by permission, from Presti, 1966.)

marks on the teledeltos paper trace out a pattern of equal-amplitude contours. This pattern is called a contour spectrogram.

A comparison between wide-band bar and contour spectrograms of an

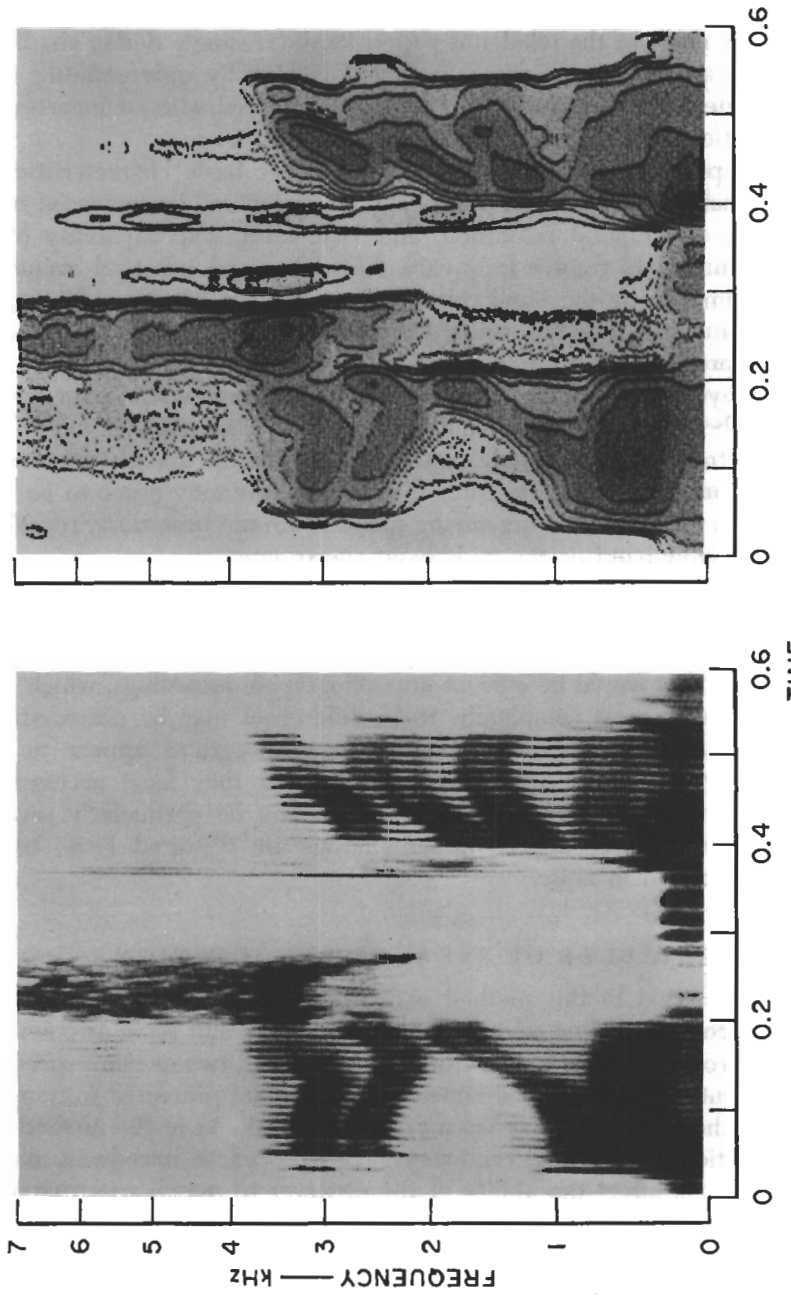

FIGURE 20. Wide-band bar (a) and contour (b) spectrograms of the identical utterance of the word *voiceprint*.

identical utterance is shown in Figure 20. It will be noted that the contour spectrogram has greater amplitude resolution but less temporal resolution[4] than the bar spectrogram. The contour spectrogram makes almost no demands on the dynamic range of the teledeltos paper; the increasingly darker shadings of higher-level amplitude contours, which aid in visually differentiating between amplitude maxima and minima, may be eliminated without incurring a loss of information.

The sound spectrograph has several limitations. A basic characteristic of all spectrum analyzers is that their frequency resolution can be increased only at the expense of temporal resolution, and vice versa. The capability of a particular instrument to resolve frequency differences and temporal events is determined primarily by the bandwidth of its analyzing bandpass filter. Although the sound spectrograph contains two bandpass filters, the choice of either filter represents a compromise. Some forms of speaker variability are probably displayed better in the narrow-band spectrogram, while other forms are displayed better in the wide-band spectrogram. Thus, one limitation of the sound spectrograph is that only certain features of the speech signal can be revealed at one time. Those features that might eventually prove to be the most useful ones for differentiating among speakers are not necessarily revealed in either the narrow-band or the wide-band spectrogram.

Because of the finite resolving power of the sound spectrograph, it is possible that spectrograms prepared from slightly different utterances of the same word cannot be told apart by human observers. While the differences among the utterances would be evident in oscillographic recordings, which describe the utterances most completely, these differences may be obscured by the sound spectrograph. Therefore, when two spectrograms appear to be identical in all respects, it cannot be concluded that they must necessarily represent the same speech signal. This limitation can be particularly severe in cases where the speech signals under analysis are distorted (e.g., band limited) or embedded in noise.

C. VARIABLES OF SPEAKER RECOGNITION

The procedure used in this method of speaker recognition is as follows: Speakers are recorded reading selected words or phrases that serve as cue material, bar spectrograms are prepared from the recordings, two or more spectrograms of different utterances of the same cue material are presented to trained observers, and the observers carry out a recognition task. As in the method of speaker recognition by listening, each step in this procedure introduces many variables which can affect the ability of the observer to match spectrograms that represent the same speaker. The most important variables will now be described in detail; it will be noted that some of them resemble the variables considered in Chapter III.

[4] The detected signal is subjected to additional lowpass filtering in the amplitude quantizer.

1. Size and Homogeneity of Speaker Group

Studies by Kersta (1962b), Young and Campbell (1967), and Stevens, Williams, Carbonell, and Woods (1968) have demonstrated that some speakers are considerably more difficult to identify by their spectrograms than other speakers. In one experiment (Stevens, Williams, Carbonell, and Woods, 1968), observers were required to match a spectrogram representing a speaker to be identified with one of eight reference spectrograms. Each reference spectrogram represented a different adult male speaker. The results of this experiment are shown in Figure 21. The solid bars indicate the percentage of error, averaged over various cue materials and six observers, for each of the eight reference speakers. (The cross-hatched bars pertain to an analogous experiment using speech samples and listeners; the results of the two experiments will be compared in Section F.) Thus, Speaker 7 was correctly identified much more often than Speaker 2. The results of other experiments (Kersta, 1962b; Anon., 1965) suggest that, on the average, female speakers are as easy to identify as male speakers. Because of the large variance in the identifiability of individual speakers, the speaker group should be as large as is practically possible.

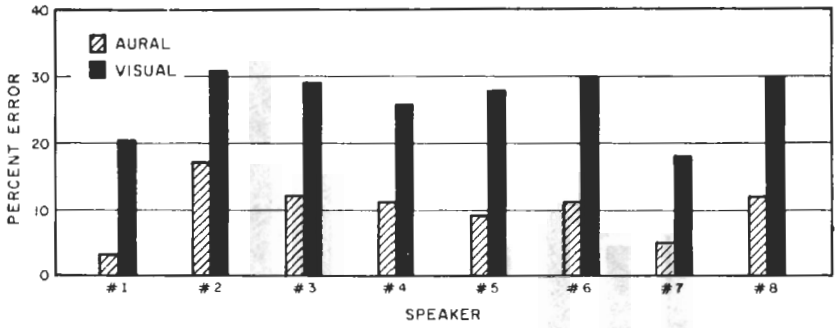

FIGURE 21. Percent error, averaged over cue materials and observers/listeners, for individual speakers in visual/aural eight-choice identification tests. (Reprinted, by permission, from Stevens et al., 1968.)

The speaker group should also be somewhat homogeneous. With respect to the method of speaker recognition by listening, the term homogeneity refers to the perceptual similarity among the voices heard in a particular test. Here, on the other hand, the term refers to the similarity in appearance among the spectrograms of the speakers participating in a test. Very little is known about the perceptual and physical correlates of this kind of speaker homogeneity. It is possible that speakers who have similar sounding voices do not produce similar appearing spectrograms. Thus, the criteria used to select a fairly homogeneous speaker group for a listening test may be inappropriate for a test involving comparisons of spectrograms.

2. *Selection of Cue Material*

Several studies have used as cue material the ten words most frequently occurring in telephone conversations: *I, you, it, me, on, the, is, and, a,* and *to* (Kersta, 1962a, 1962b; Anon., 1965; Young and Campbell, 1967). Stevens, Williams, Carbonell, and Woods (1968) used a number of monosyllabic and disyllabic words, a phrase, and a sentence. It has been repeatedly demonstrated that some cue materials are better vehicles for identification than others. This effect is illustrated in Figure 22. The solid bars indicate the percentage of

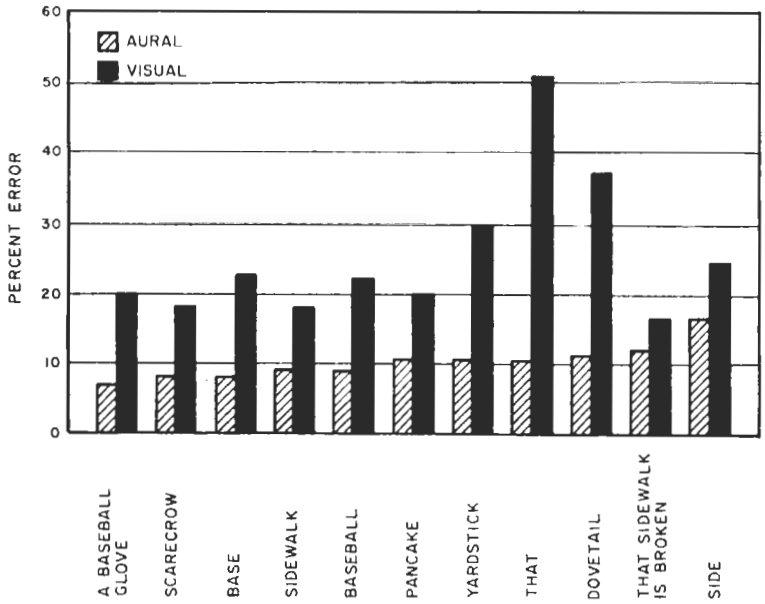

FIGURE 22. Percent error, averaged over speakers and observers/listeners, for cue materials in visual/aural eight-choice identification tests. (Reprinted, by permission, from Stevens et al., 1968.)

error, averaged over eight speakers and six observers, for each word, phrase, or sentence. Spectrograms of the monosyllabic word *that* were more often incorrectly matched than spectrograms of any other item. As shown in Figure 23, the observers' performance was proportional to the duration of the cue material. The highest level of performance was reached with the relatively long phrase and sentence, and the lowest level occurred with the short, monosyllabic words.

For disyllabic cue words, performance also appears to depend on the vowel receiving the primary phonetic stress (Stevens, Williams, Carbonell, and Woods, 1968). Figure 24 shows how words with stressed front vowels (see Table 1) make better vehicles for speaker identification than words with stressed back vowels. This finding may be explained in terms of the gross spec-

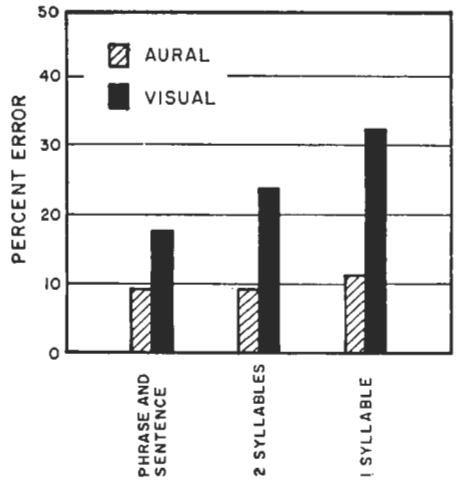

FIGURE 23. Percent error, averaged over speakers and observers/listeners, for cue materials of various durations in visual/aural eight-choice identification tests. (Reprinted, by permission, from Stevens et al., 1968.)

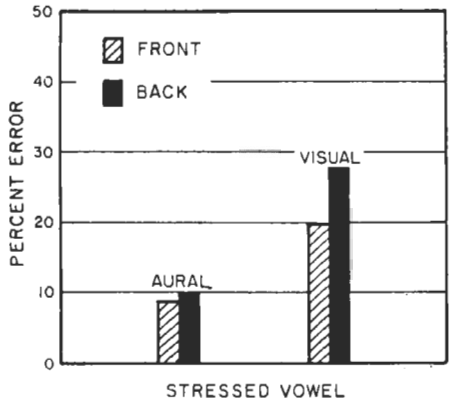

FIGURE 24. Percent error, averaged over speakers and observers/listeners, for stressed vowels of disyllabic cue words in visual/aural eight-choice identification tests. (Reprinted, by permission, from Stevens et al., 1968.)

tral features of front and back vowels. The frequency of the second formant determines the relative prominence of the high-frequency portion of the vowel spectrum (Stevens and House, 1961). Since the second-formant frequency is high for front vowels and low for back vowels (see Table 4), front vowels tend to have more high-frequency energy than back vowels. Considering that the sound spectrograph has a fixed spectral equalization and operates within a restricted dynamic range, a back vowel may be only partially recorded on the spectrogram, and it may consequently convey less information to the observer.

There is some evidence that an observer's performance can be increased by allowing him to compare spectrograms of several cue words simultaneously (Anon., 1965). For each participating speaker, the spectrograms of the different words may be mounted on a single card. Although the observer still compares corresponding spectrograms (i.e., spectrograms of the same cue word), he may now be more able to generalize any noted differences, and perhaps to

detect an atypical utterance that might have otherwise led to an error. This situation is very similar to the one in which cue material of relatively long duration is used. The observer's decision is based on a more extensive sample of each speaker's spectrographic characteristics.

3. *Context of Cue Material*

Another variable affecting performance is the phonetic context in which the cue material appears. Kersta (1962b) found that observers had more difficulty sorting spectrograms of cue words according to speaker when the words were uttered in context rather than in isolation. The effects of context were also studied by Young and Campbell (1967), who conducted two experiments employing five male speakers. Each speaker recorded four isolated utterances of each of the cue words *me, you,* and *it,* and four sentences, each of which contained the cue words *you* and *it.* Spectrograms were prepared from all recorded material. Ten observers were shown the 20 spectrograms of the cue word *me* ordered according to speaker. Several features of the spectrograms that were thought to be useful for speaker recognition were described and discussed.

In the first of the two experiments, the observers were given five spectrograms of the cue word *you* uttered in isolation. Each spectrogram served as a model for one of the five speakers. The observers were then asked to identify each of the remaining 15 spectrograms of the cue word *you* uttered in isolation by comparing it against the models. The same procedure was followed for the 20 spectrograms of the cue word *it* uttered in isolation, and the results obtained for the two cue words were combined.

In the second experiment, the observers worked with excerpts from the 20 spectrograms of the sentences. For each sentence uttered by a given speaker, excerpted spectrograms of the cue words *you* and *it* were mounted on a single card. The observers were given five such cards to serve as speaker models, and then they were asked to identify each of the remaining 15 cards, as in the first experiment. Table 12 shows the results of these two spectrogram-matching

TABLE 12. Results of two spectrogram-matching experiments employing cue words *you* and *it* uttered in isolation and in context. (Reprinted, by permission, from Young and Campbell, 1967.)

Speaker	*Correct Identification (%)*	
	Words in Isolation	*Words in Context*
RC	97.0	26.6
MY	91.1	40.0
BB	86.3	33.3
JR	71.4	50.0
RK	46.4	36.6
Average	78.4	37.3

experiments. The average performance of the observers was considerably lower when the cue words were uttered in the context of a sentence than when they were uttered in isolation. Also, the relative identifiability of the individual speakers appears to depend on whether the cue words were uttered in isolation or in context.

Young and Campbell mention two factors which might account for the observed difference in overall performance. A word uttered in context tends to have a shorter duration than the same word uttered in isolation, and since performance is proportional to the duration of the cue material (see Figure 23), the second experiment would be expected to produce lower scores. The other factor concerns the phonetic environment of the cue word. It will be recalled from Chapter II that the acoustical properties of a sequence of speech sounds depend somewhat on the identity of the preceding and following speech sounds (coarticulation effects). Since the second experiment did not provide a uniform phonetic environment for the cue words, it is reasonable to suspect that coarticulation effects may have obscured the spectral features used by the observers to differentiate among the speakers.

The effects of context on performance may be even greater than the data of Table 12 suggest. In the first experiment, the spectrograms of the cue words *you* and *it* were presented separately, but in the second experiment they were presented together. If the excerpted spectrograms of the two cue words had also been presented separately in the second experiment, the observer would have had less information per test item and less opportunity to detect an occasional atypical utterance. In addition, the likelihood that two speaker models appear indistinguishable would have been increased. Any of these factors might have reduced the observer's performance even further.

4. Characteristics of Transmission Link

It is not always possible to prepare the spectrograms from high-quality speech signals. In some situations, the speech signals may be available only at the receiver of a communication system (e.g., the telephone), and the characteristics of the transmission link may introduce various kinds of signal distortion which could influence performance. In other situations, the speakers may have uttered the cue material in a noisy environment, so that the speech signals are degraded by noise. While signal distortion and noise have been regarded as important variables, their effects on performance have not been studied systematically. It has been suggested that the primary limitation imposed by the telephone system, namely bandpass filtering, is not detrimental to this method of speaker recognition (Anon., 1965).

The greatest effect of a given form of signal degradation would be expected to occur in experiments in which only some of the spectrograms are prepared from degraded speech signals. Consider an experiment involving several reference spectrograms representing different speakers, and a number of test spectrograms to be identified. If one of the reference spectrograms and some of

the test spectrograms are prepared from degraded speech signals, but the remaining spectrograms are prepared from high-quality speech signals, the observer may not be able to distinguish between the spectral features attributable to individual speakers and those attributable to the signal degradation. Thus, he may be inclined to match spectrograms on the basis of the presence or absence of signal degradation, and his performance may be reduced accordingly.

5. *Type of Visual Display*

Kersta (1962a, 1962b) studied the ability of observers to sort spectrograms into groups that represent different speakers. In one series of experiments, the observers were given spectrograms of four utterances of a particular cue word by either 5, 9, or 12 male speakers (i.e., a matrix of 4×5, 4×9, or 4×12 spectrograms), and they were instructed to arrange the spectrograms into as many piles as there were speakers. Both bar and contour spectrograms were used in separate experiments. The results, shown in Figure 25, suggest that the

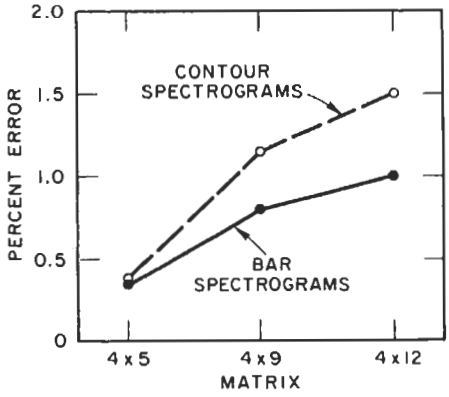

FIGURE 25. Percent error, averaged over cue words and observers, for matrices of four utterances by each of five, nine, or twelve speakers in spectrogram-matching experiments. Data are shown for two types of spectrograms. (Reprinted, by permission, from Kersta, 1962b.)

bar spectrogram is a better visual display for speaker recognition than the contour spectrogram. This finding may be related to the fact that the contour spectrogram does not provide as much temporal resolution as the bar spectrogram (see Section B).

Some investigators believe that the low-frequency region of the bar spectrogram (below about 3 kHz) is more important for speaker recognition than the high-frequency region. The lower formants of the speech signal are known to make extreme frequency excursions from one speech sound to the next, whereas the higher formants are relatively stable. It has been argued that the traces of the lower formants in the spectrogram therefore encompass most of the speaker variability. In order to allow a more detailed examination of the low-frequency region of the spectrogram without sacrificing the overall frequency range, some investigators use the logarithmic frequency scale of the sound spectrograph almost exclusively (Anon., 1965).

Conventional spectrograms are not the only useful visual displays of the speech signal. Pickett (1968) summarized the research on various other visual displays that have been developed to aid the deaf; some of these displays may also be applicable to the problem of speaker recognition. No study has been undertaken with the specific objective of finding a more optimal visual display for the present purpose than the bar spectrogram. Furthermore, the possible advantages of combining an appropriate auditory signal with the visual display have not been sufficiently explored.

A simple experiment by Ungeheuer (1965) will serve as an example of how speaker recognition may be accomplished with a special type of visual display. In this experiment, the speech signal was first amplitude normalized and then processed by two parallel channels, each of which consisted of a bandpass filter, a rectifier, and a 1.5-sec integrator. The two filters covered the frequency ranges 63–710 Hz and 1.4–2.8 kHz. The outputs of the integrators were applied to the horizontal and vertical inputs of an oscilloscope. For the duration of a speech sample by a particular speaker, the oscilloscope screen was photographed on a single sheet of film. The resulting display, which was called intensity club because of its general shape, was found to differ slightly among speakers.

6. *Number of Reference Spectrograms*

In experiments using reference spectrograms, the performance of an observer can usually be increased by providing him with more than one reference spectrogram for each participating speaker. The use of several reference spectrograms for each speaker allows the observer to consider intraspeaker variability as well as interspeaker variability. He is now in a better position to determine whether a feature of a test spectrogram is unique to the speaker represented by the test spectrogram or is occasionally also found in spectrograms representing other speakers.

Instead of presenting all reference spectrograms simultaneously, it is possible to present one set at a time (i.e., one spectrogram for each speaker), and to have the observer repeat his task with each new set. When the reference spectrograms are used in this manner, the observer is presumably unable to estimate the intraspeaker variability. The expected advantage of having all reference spectrograms available at the same time has not been formally demonstrated.

7. *Size and Training of Observer Group*

The ability to match test spectrograms to the proper reference spectrograms varies considerably from observer to observer. Figure 26 shows the results of the previously described experiment by Stevens, Williams, Carbonell, and Woods (1968), averaged over the speakers and the cue materials. The solid bars indicate the percentage of error for each of the six observers; it is apparent that Observers 4 and 5 reached a higher level of performance than any other observer.

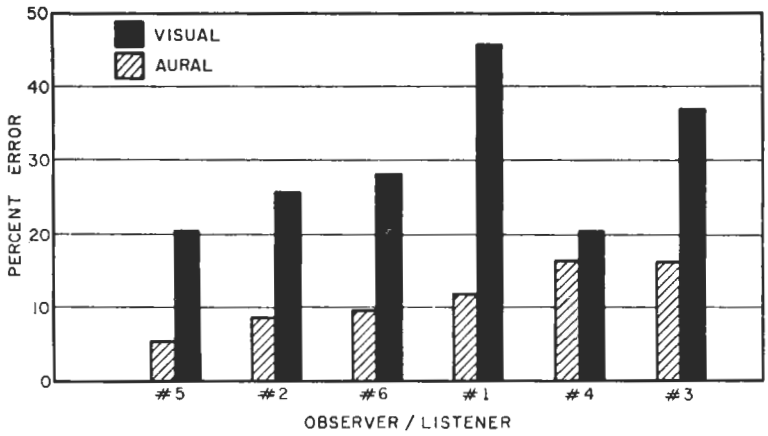

FIGURE 26. Percent error, averaged over speakers and cue materials, for individual observers/listeners in visual/aural eight-choice identification tests. (Reprinted, by permission, from Stevens et al., 1968.)

Young and Campbell (1967) obtained similar results. There are no reported studies explaining why some observers are more successful than others. Different observers may pay attention to different spectral features, weigh certain features differently, and apply different rules to arrive at their decisions.

From a practical point of view, the performance of an observer group can be optimized by selecting highly capable observers on the basis of preliminary tests. Another procedure for increasing overall performance involves combining the various decisions made by different observers for a particular test spectrogram to obtain a group decision. Stevens, Williams, Carbonell, and Woods (1968) found that a simple majority vote reduced the error score for the cue word *sidewalk* from 18% to 3%, and for the cue word *dovetail* from 36% to 22%.

Because of the individual differences in performance, the observer group should be as large as possible. Both male and female observers may be used. In several experiments (Kersta, 1962a, 1962b), the observers were female high-school students who worked in teams of two. This pairing of observers was found to increase the overall performance, probably on account of an exchange of information about decision strategies.

One of the most important variables of this method of speaker recognition is the training of the observer group. Unlike the training of listeners, which is undoubtedly facilitated by each listener's prior experience with different voices, the training of observers must deal with the problem of interpreting a novel visual display. The requirements for an optimal training program cannot, as yet, be specified. It is not known, for example, to what extent the spectrogram should be explained to the observers. While many investigators explain how various aspects of the speech signal are portrayed in the spectrogram, other investigators prefer that the observers regard the spectrogram as an arbitrary

display of acoustic energy. There is also no agreement on which spectral features are most useful for speaker recognition; different investigators emphasize different features in training their observers.

The duration of the training program determines how much skill and confidence each observer can acquire. Kersta (1962a, 1962b) trained one group of listeners for five days, which is not considered an overly long training program. Presently, Kersta is offering an intensive two-week course in the interpretation of spectrograms (voiceprints).[5]

8. *Task Assigned to Observers*

The task assigned to the observer may also influence performance. Observer tasks usually involve direct comparisons of spectrograms.[6] These tasks may be grouped into two classes on the basis of the number of spectrograms used per test item. One class includes all tasks in which the observer compares only two spectrograms for each test item, and the other class includes all tasks in which he compares three or more spectrograms.

When the observer is given only two spectrograms, he may be asked to rate their level of similarity. Or he may be assigned the task of deciding whether the spectrograms are similar enough to represent the same speaker. In the latter case, it is assumed that the observer makes use of an internal decision threshold, which is different for different observers. The decision threshold can be taken into account by requiring the observer to rate his confidence in the correctness of his decision. The observer's dual responses on a large number of test items are often analyzed to produce a ROC curve, as described in Chapter III.

In most of the tasks for which the observer is given three or more spectrograms, a test spectrogram is compared with several reference spectrograms. Each reference spectrogram usually represents a different speaker. The speaker who is represented by the test spectrogram may or may not be also represented by one of the reference spectrograms. If he is represented among the reference spectrograms, he may be identified by having the observer decide which reference spectrogram is most similar to the test spectrogram. In another task, the observer may be asked to rate the levels of similarity between the test spectrogram and each of the reference spectrograms. There are some tasks, however, in which the observer compares all spectrograms jointly. For example, the observer may be required to sort the spectrograms into a given number of groups representing different speakers.

---

[5] For information on this course, write to L. G. Kersta, Voiceprint Laboratories, Somerville, New Jersey.

[6] It is conceivable that a very experienced observer can examine a single spectrogram and identify the represented speaker by relying on his long-term memory of the spectral features exhibited by particular speakers. Also, an observer may be able to rate a single spectrogram on several scales that inquire about the perceptual dimensions of the visual pattern. Such tasks, however, have not been used to date.

D. TEST FORMATS

The variables described above are controlled by a particular test format. Three of the most commonly used test formats will be outlined and discussed in this section. There are obvious analogies between these test formats and certain formats discussed in Chapter III. Because of the nature of this method of speaker recognition, one test format does not have significant advantages over another. For example, all spectrograms constituting a given test item are typically presented simultaneously, so that there are no short-term memory effects. Similarly, all tests are about equally demanding with respect to their preparation. Instead of describing minor advantages and limitations, this section will concentrate on the various applications of the tests.

1. Multiple-Choice Identification Test

In the multiple-choice identification test, the observer is given one or more labeled reference spectrograms for each participating speaker. These reference spectrograms remain with the observer until he has responded to all test items. Each test item consists of a test spectrogram representing a speaker to be identified. The speaker represented by the test spectrogram is known to be also represented by one of the reference spectrograms. Thus, the observer has only to decide which reference spectrogram is most similar to the test spectrogram in order to identify the speaker represented by the test spectrogram. With this test, it is possible to treat one cue word at a time, or to treat several cue words simultaneously. In the latter case, the spectrograms of the different cue words uttered by each speaker may be mounted on a single card.

This test is well suited for exploring the variables of speaker recognition described in Section C. Many experimental studies have used the test to determine how selected variables should be managed to optimize the overall performance. However, the test has very few practical applications. In most practical situations, there is no assurance that the speaker to be identified is represented by one of the available reference spectrograms.

2. Discrimination Test

The discrimination test does not employ reference spectrograms. Each test item consists of two spectrograms, and the observer is asked to decide whether these spectrograms represent the same or different speakers. In many cases, the observer is also required to rate his confidence in the correctness of his decision. Consecutive test items usually involve different cue words. Although this test has been used for experimental purposes, its most often cited application is in the field of criminology, where it serves as an investigative tool.

This application of the test employs the following procedure: Suppose an obscene telephone call has been recorded, and a recording by a person who is suspected of having made the call is also available. First, the two recordings

are carefully transcribed to facilitate the search for suitable cue material.[7] In the selection of cue material, if there are only a few words common to both recordings, it may be necessary to include phonetically identical portions of different words. Pairs of spectrograms are then prepared for all selected cue materials. The observer examines each pair of spectrograms and determines the degree of similarity of the spectral features. One of three possible decisions is finally rendered; either the two recordings are ascribed to the same speaker, they are ascribed to different speakers, or the results are considered inconclusive. The third decision is made if there is a scarcity of cue material, or if the speech signal is severely degraded.

When the two recordings are ascribed to the same speaker, a display of several pairs of spectrograms may be prepared for the purpose of justifying this decision to laymen. Such displays have been exhibited and accepted as evidence in courts of law (Borders, 1966; McDade, 1968). Figure 27 shows a

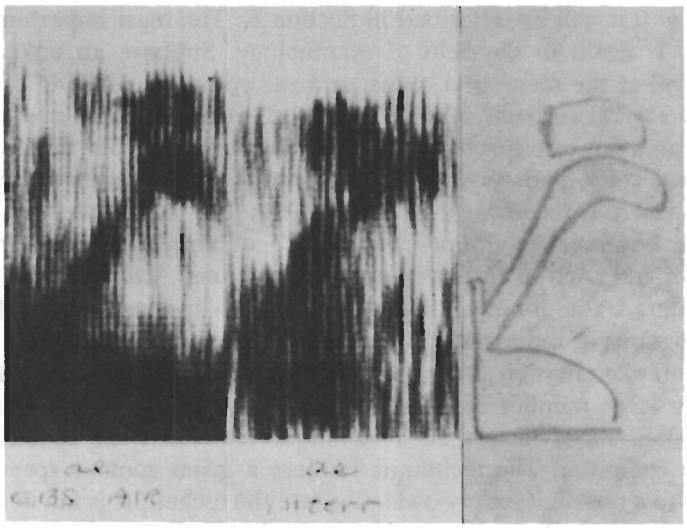

FIGURE 27. Display consisting of two spectrograms of the diphthong [aɪ] and a sketch of the gross pattern common to both spectrograms. (Reprinted, by permission, from Ladefoged and Vanderslice, 1967.)

display consisting of two spectrograms of the diphthong [aɪ] and a sketch of the gross pattern common to both spectrograms. The left-hand spectrogram was prepared from a televised interview of an unidentified youth who admitted to the crime of arson, and the right-hand spectrogram was prepared from a

[7] In listening to the two recordings, the investigator may use an informal version of the discrimination test described in Chapter III to estimate the likelihood that the recordings represent the same speaker. This estimate, which is independent of the results based on comparisons of spectrograms, may or may not be considered in arriving at the final decision.

recorded interrogation of Edward Lee King. On the basis of this and other evidence, King was convicted and sentenced to prison (McDade, 1968).

3. *Identification-Discrimination Test*

This test has some of the properties of the multiple-choice identification test and some of the properties of the discrimination test. Like the multiple-choice identification test, it uses reference spectrograms and test spectrograms, but the speaker represented by a given test spectrogram is not necessarily also represented by one of the reference spectrograms. The observer must therefore make two decisions; he must decide which reference spectrogram is most similar to the test spectrogram, and whether the selected reference spectrogram and the test spectrogram represent the same speaker. The latter decision is identical to the decision required in the discrimination test.

The identification-discrimination test has been used experimentally to determine the reliability with which speakers can be authenticated. This application of the test will be described in Section E. The most important practical application is again in the field of criminology. Suppose an unknown voice was recorded at the scene of a crime, and recordings were made of the interrogation of several suspects. Assuming that all of these recordings have a few words in common, test spectrograms may be prepared from the recording of the unknown voice, and corresponding sets of reference spectrograms may be prepared from the recordings of the suspects. For each test item (i.e., cue word), the observer determines the degree of similarity between the test spectrogram and each of the reference spectrograms. After all test items have been considered, the observer either identifies the speaker represented by the test spectrograms or reports that he is unable to do so.

It is possible to envision situations in which the observer would have to deal with a very large number of reference spectrograms. To avoid this problem, Kersta (1965b, 1966) developed a technique for classifying spectrograms by means of a computer. The technique reduces a given contour spectrogram of a cue word to a ten-digit binary code. Because the technique is inherently crude[8] and therefore relatively inaccurate, it cannot be used to replace the observer, but it may greatly simplify his task. In the foregoing example, one or two of the suspects might have been eliminated on the basis of their classification codes, so that the test would involve fewer reference spectrograms.

E. OBSERVER FALLIBILITY

The fallibility of the observer is a crucial issue because of the legal use of this method of speaker recognition (Borders, 1966; Ladefoged and Vanderslice, 1967; McDade, 1968; Bolt et al., 1970). Although a machine (the sound

[8] In order to minimize variability due to context, the technique is used only on those portions of cue words exhibiting the smallest spectral changes with time (Anon., 1965). Thus, the dynamic aspects of articulation, which are known to reflect important speaker differences, are purposely disregarded.

spectrograph) is used to prepare spectrograms, the interpretation of spectrograms is an art rather than a science. When this fact is pointed out to the members of a jury, they may be unable to evaluate the reliability of this means of identification. In the first trial in which spectrograms were allowed as evidence, the jury could not reach an agreement as to how much weight this evidence should be given (McDade, 1968). The previously mentioned conviction of Edward Lee King was reversed by a Court of Appeals because ". . . the voiceprint identification process has not reached a sufficient level of scientific certainty to be accepted as identification evidence in cases where the life or liberty of a defendant may be at stake . . ." (Kennedy, 1968).

The use of the term voiceprint, and the degree to which the analogy between voiceprints and fingerprints has been emphasized (Kersta, 1962a, 1962b; Anon., 1965; McDade, 1968), are rather unfortunate. There is an important but seldom considered difference between spectrograms and fingerprints. As was demonstrated in Chapter II, the intraspeaker variability of the speech signal can be substantial, and this variability is, of course, reflected in spectrograms representing a particular speaker. The variability exhibited by a particular person's fingerprints, on the other hand, is essentially zero (Ladefoged and Vanderslice, 1967; Bolt et al., 1970). Most of this variability is due to the fact that inadvertently left fingerprints are often incomplete or smeared. As a means of identification, fingerprints must be regarded as being considerably more foolproof than spectrograms (Anon., 1965).

Claims by Kersta and others of the reliability of this method of speaker recognition are based largely on the results of unpublished experiments. Thus, the scientific community cannot appraise the design of these experiments and the validity of the conclusions reached (Ladefoged and Vanderslice, 1967). The results of one series of published experiments (Kersta, 1962b) could not be duplicated by other investigators. Young and Campbell (1967), and also Stevens, Williams, Carbonell, and Woods (1968), obtained much higher error scores than those reported by Kersta (1962a, 1962b). Such disagreements make the publication of detailed descriptions of future experiments extremely desirable.

In the first experiments concerned with reliability, the observers were required to sort spectrograms into groups representing different speakers (Kersta, 1962a, 1962b). Later experiments used the multiple-choice identification test (Kersta, 1962c; Young and Campbell, 1967; Stevens, Williams, Carbonell, and Woods, 1968). There have been no reports of experiments using the discrimination test, which is commonly used in criminal proceedings. Ladefoged and Vanderslice (1967) argued that the reliability of the discrimination test cannot be predicted from the results of the published studies.

It has been claimed that performance is essentially unaffected by the loss of teeth, tonsils, or adenoids, the aging process, and attempts to disguise the voice, such as changing the fundamental frequency, whispering, mimicking another voice, or ventriloquism (Kersta, 1962c; Anon., 1965). However, in the absence of supporting experimental data, these claims cannot be considered

established facts. When the characteristics of the transmission link are unfavorable, so that the speech signal is degraded (see Section C4), many of the above-mentioned factors may be expected to reduce the reliability of this method.

According to Kersta (1962b) and others (Anon., 1965), the probability that two speakers have similar enough vocal-tract dimensions and articulation patterns to produce indistinguishable spectrograms is extremely small. This belief, which appears to underlie many experiments, has not been formally translated into a hypothesis that can be tested with a finite population of speakers. There is evidence that two arbitrarily selected speakers can occasionally produce very similar spectrograms (Ladefoged and Vanderslice, 1967). This situation is illustrated in Figure 28 for the cue word *you*. Findings of this kind

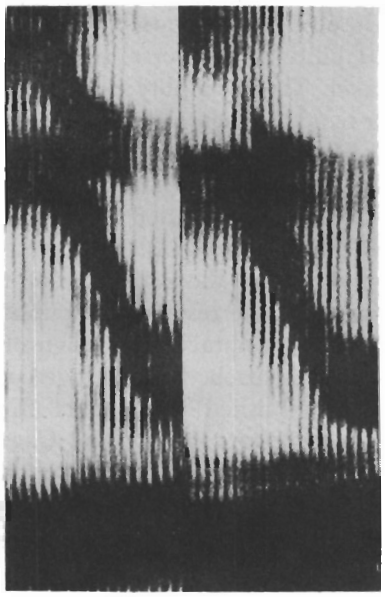

FIGURE 28. Similar spectrograms of the word *you*, uttered by two arbitrarily selected speakers. (Reprinted, by permission, from Ladefoged and Vanderslice, 1967.)

suggest that the range of one speaker's pronunciations of a given cue word (intraspeaker variability) may partially overlap the range of another speaker's pronunciations of the same cue word, and argue for the use of a large number of cue words in making an identification. There is also evidence of considerable similarity among spectrograms representing different members of a family (Kersta, 1965a), suggesting further sources of observer fallibility.

Stevens, Williams, Carbonell, and Woods (1968) examined the ability of observers to distinguish between familiar and unfamiliar speakers in a 32-item

identification-discrimination test. The observer was given eight reference spectrograms representing eight familiar speakers. There were two experimental conditions; either 4 or 16 of the 32 test spectrograms represented unfamiliar speakers who were not represented by the reference spectrograms. The results of this study are shown in Table 13. Most of the familiar speakers were recog-

TABLE 13. Percent correct recognition of familiar and unfamiliar male speakers by visual comparison of spectrograms. Data are shown for two experimental conditions. (Reprinted, by permission, from Stevens et al., 1968.)

4 of 32 Test Items by Unfamiliar Speakers		
	Recognized As	
Speaker	*Familiar*	*Unfamiliar*
Familiar	80	20
Unfamiliar	31	69
16 of 32 Test Items by Unfamiliar Speakers		
	Recognized As	
Speaker	*Familiar*	*Unfamiliar*
Familiar	90	10
Unfamiliar	47	53

nized as such and correctly identified. Many of the unfamiliar speakers, however were erroneously recognized as familiar speakers, especially when they appeared as often as the familiar speakers.

In view of the use of this method of speaker recognition in courts of law, the fallibility of the observer must be studied further (Bolt et al., 1970). Detailed descriptions of future experiments should be published or otherwise made available to the scientific community. Claims should be clearly differentiated from proven facts, and statements regarding the analogy between this method and fingerprinting should include appropriate qualifications. Although this method has obvious potential in various investigative and forensic applications, its reliability as a means of positive identification has not yet been sufficiently evaluated to allow its use at the level of confidence attributed to fingerprinting.

F. COMPARISON WITH SPEAKER RECOGNITION BY LISTENING

Certain properties of the speech signal, such as the formant structure in a vocalic interval and the spectral distribution of fricative noise, are more discernible in a spectrogram than in an auditory presentation (Potter, Kopp, and Green, 1947; Anon., 1965). This does not mean, however, that speaker recognition by visual comparison of spectrograms is inherently more accurate than

speaker recognition by listening. In fact, there is evidence indicating that the reverse is true.

Stevens, Williams, Carbonell, and Woods (1968) compared the two methods of speaker recognition with a single speaker group and a single observer/listener group under closely matched experimental conditions. Recorded utterances of various cue materials by eight selected male speakers were used to construct both visual and aural eight-choice identification tests. In order to match the format of the visual tests, the aural tests used the free-comparison mode of presentation. Six experimental subjects served as observers and listeners in these tests. Each subject was required to identify the speaker represented by a given test item and to rate the level of confidence with which he made this decision. He was allowed as much time as he needed to reach a decision, but a record was kept of the time expended.

It was found that the visual tests produced considerably higher error scores than the aural tests. Also, the confidence ratings were lower and the average time taken per test item was greater for the visual tests. These results suggest that speaker recognition by listening is the more accurate method, possibly because the observer's task is more difficult than the listener's task.

The errors made on both types of tests in identifying the individual speakers are compared in Figure 21. It is apparent that all speakers could be identified better with the aural tests than with the visual tests. The two speakers who were most often correctly identified in the aural tests (Speakers 1 and 7) were also most often correctly identified in the visual tests. Similarly, the speaker who was least identifiable in the aural tests (Speaker 2) was also least identifiable in the visual tests. Thus, there is evidence that the relative identifiability of different speakers is largely independent of the method employed.

Figure 22 compares the errors made on both types of tests in identifying the speakers by means of particular cue material. The cue materials are arranged according to the level of performance achieved on the aural tests; the highest level of performance (lowest error score) was obtained for the phrase *A baseball glove*. There is no indication that the data for the two types of tests are correlated. Cue materials that are good vehicles for aural identification are not necessarily good vehicles for visual identification, and vice versa.

As shown in Figure 23, the duration of the cue material was found to be an important variable in the visual tests, but not in the aural tests. For the visual tests, there is a progressive reduction in error scores as the duration of the cue material is increased. For the aural tests, however, there is a reduction only until the disyllabic cue words are reached. The average duration of these words was slightly less than 1 sec. Thus, the results of the aural tests are in agreement with the findings of Pollack, Pickett, and Sumby (1954), which are presented in Figure 12.

Figure 24 shows that disyllabic cue words with stressed front vowels are better vehicles for both visual and aural identification than disyllabic cue words with stressed back vowels. The front-vowel superiority appears to be less pronounced for the aural tests. This may be because the aural tests are not con-

fined to the relatively narrow dynamic range encompassed by the bar spectrogram. The inherently reduced high-frequency energy of back vowels may still be audible, but have insufficient amplitude to register in the spectrogram.

Figure 26 compares the errors made on both types of tests by the individual subjects (observers/listeners). The subjects are arranged according to the level of performance they achieved on the aural tests. Again, the data for the two types of tests do not appear to be correlated. Subjects who are especially capable listeners are not necessarily especially capable observers, and vice versa.

Stevens, Williams, Carbonell, and Woods (1968) also used identification-discrimination tests in comparing the two methods of speaker recognition. These tests involved only the cue words *sidewalk* and *dovetail.* The eight speakers who participated in the identification tests were used as familiar speakers in the identification-discrimination tests; 16 additional speakers were used as unfamiliar speakers. Four subjects served as observers and listeners in these tests.

The results obtained with the aural tests and the visual tests are presented in Tables 11 and 13, respectively. There were considerably more false acceptances of unfamiliar speakers in the visual tests than in the aural tests. When only 4 of the 32 test items represented unfamiliar speakers, there were also more false rejections of familiar speakers in the visual tests. Thus, speaker recognition by listening was found to be the more accurate method.

As mentioned earlier, the training of observers is an important variable of speaker recognition by visual comparison of spectrograms. The subjects used by Stevens, Williams, Carbonell, and Woods (1968) received only minimal training. Before each subject started his first visual test, which was regarded as a training test, he was given a brief explanation of the spectrogram. The average error score for the first visual test was 28%. There was very little improvement in the scores obtained on subsequent visual tests, suggesting that the subjects did not learn much about the interpretation of spectrograms as the study progressed. Young and Campbell (1967) trained their observers by showing them examples of several potentially useful spectral features. These investigators obtained an average error score of 22%. It is possible that more elaborate training procedures might lead to significantly lower error scores, but whether future visual tests would provide lower error scores than future aural tests is debatable.

Chapter V

SPEAKER RECOGNITION BY MACHINE

A. INTRODUCTION

Two approaches have been used to study the feasibility of speaker recognition by machine. One approach is to have the machine generate and examine amplitude-frequency-time matrices of specific cue material. The other approach is to have the machine extract speaker-dependent parameters from the speech signal and subject them to a statistical analysis. Each approach has led to a number of recognition techniques which will be described in detail. Some comments will then be made about the fallibility of speaker recognition by machine. A comparison between this method and speaker recognition by listening will also be made.

B. TECHNIQUES USING SPECIFIC CUE MATERIAL

In the recognition techniques to be described in this section, an amplitude-frequency-time matrix of specific cue material is prepared for each speaker, and these matrices are compared by means of a decision rule. Various cue materials have been used, including phrases, words excerpted from context, and even single phonemes (speech sounds) excerpted from context. For a given comparison, all data matrices represent the same phrase, word, or phoneme. The general procedure underlying these techniques will be outlined before experimental studies are described.

1. General Procedure

The utterances of specific cue material are usually processed by a spectrum analyzer consisting of a bank of bandpass filters, rectifiers, and smoothing circuits. The outputs of the analyzer are periodically sampled and amplitude quantized for further processing by a computer. Each utterance is represented in the computer by a data matrix having the format shown in Figure 29. The rows of the matrix correspond to the frequency bands of the spectrum analyzer, the columns correspond to the temporal locations of the sampled spectra, and each matrix cell describes a measured amplitude level. Such a matrix may be regarded as a digital spectrogram.

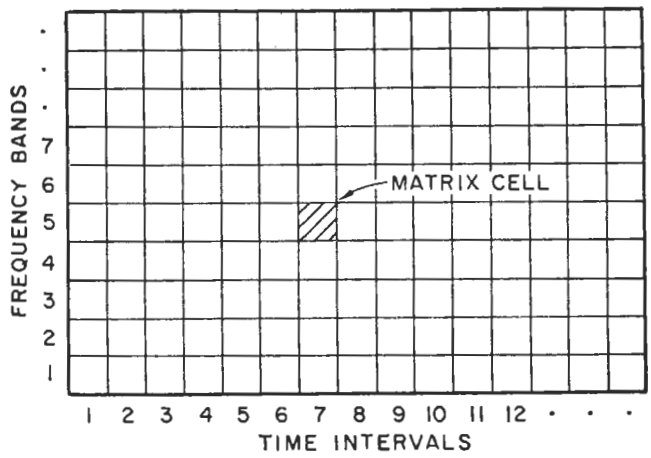

FIGURE 29. Format of data matrix used to represent an utterance of cue material. Each matrix cell gives the amplitude of energy present in a specific frequency band during a specific time interval.

The number of frequency bands used, the frequency range covered by these bands, the duration of each time interval, and the number of levels used to describe amplitude are variable. The number of time intervals used depends on the duration of each interval and on the duration of the utterance represented. Before any comparisons between matrices are attempted, it is usually necessary to normalize the matrices with respect to overall amplitude. This may be accomplished by uniformly increasing or decreasing all cell values in a given matrix until their sum is equal to a constant.

For each phrase, word, or phoneme used, several matrices representing different utterances by the same speaker are combined to form a single reference matrix for that speaker. A reference matrix is thus constructed for each speaker participating in a recognition experiment. The speaker to be recognized is represented by a test matrix. This test matrix usually represents a single utterance, but it may also be constructed from several utterances. Depending on the type of recognition to be performed, the test matrix is compared with all or only one of the reference matrices. The degree of similarity between the test matrix and each reference matrix is computed, and the results are used to arrive at a decision.

There are two basic recognition tasks, identification and discrimination. In the identification task, several reference matrices are used, and it is assumed that the speaker represented by the test matrix is also represented by one of the reference matrices. Thus, the reference matrix that is most similar to the test matrix is expected to identify the speaker represented by the test matrix. In the discrimination task, only one reference matrix is used, and the speaker represented by the test matrix may or may not be represented by this reference matrix. A decision threshold is selected to specify when the test and reference

matrices are similar enough to represent the same speaker. Recognition techniques using the discrimination task have been used to authenticate speakers who claim a particular identity. Two kinds of errors are encountered in such applications; a speaker who is, in fact, represented by the reference matrix may be falsely rejected, and an imposter may be falsely accepted. The decision threshold is often selected on the basis of an allowable percentage of false rejections.

Different decision rules may be applied in comparing two matrices. Because optimal decision rules cannot be practically implemented, several compromises have been made in the form of relatively simple measures of similarity. One of the most often used measures of similarity is an error measure, namely the sum of the squared differences between corresponding cell values of the two matrices considered. Thus,

$$\epsilon_{XA} = \sum^{N} (\delta_X - \delta_A)^2,$$

where N is the total number of cells in either Matrix X (test matrix) or Matrix A (reference matrix representing Speaker A), δ_X is the value of a particular cell in Matrix X, and δ_A is the value of the corresponding cell in Matrix A.

Some studies have used measures of the degree of cross correlation between two matrices. One coefficient of cross correlation is defined by

$$r_{XA} = \frac{\sum^{N} (\delta_X - m_X)(\delta_A - m_A)}{N \sigma_X \sigma_A},$$

where m_X is the mean of all cell values in Matrix X, m_A is the mean of all cell values in Matrix A, σ_X is the standard deviation of the cell values in Matrix X, and σ_A is the standard deviation of the cell values in Matrix A. This correlation coefficient is particularly attractive because it does not require that the two matrices be amplitude normalized.

Other decision rules may also be used. In a given comparison involving several reference matrices, the application of different decision rules may lead to different associations between the test and reference matrices. A numerical demonstration of the dependence of matrix associations on decision rules is presented in Table 14. Using the error measure, a greater similarity exists between Matrix X and Matrix A than between Matrix X and Matrix B, but the reverse is true when the respective matrices are cross-correlated.

Many recognition techniques are handicapped because of unavoidable inaccuracies in the temporal alignment of different utterances. In two utterances of the same word, especially if they were produced by different speakers, it is rare that corresponding spectral features are synchronized in time (see Figure 7). Thus, when such utterances are aligned with respect to one spectral feature,

TABLE 14. Demonstration of dependence of matrix associations on decision rules. Matrix X is associated with Matrix A when error measure is used and with Matrix B when correlation coefficient is used.

Cell Number	*Cell Value (δ)*			*Comparison*	
	A	B	X	XA	XB
1	21	30	21		
2	19	30	21		
3	20	20	20		
4	21	10	19		
5	19	10	19		
Σ	100	100	100		
N	5	5	5		
m	20	20	20		
σ	$2/\sqrt{5}$	$20/\sqrt{5}$	$2/\sqrt{5}$		
ϵ	*Error Measure*			8	324
r	*Correlation Coefficient*			0	1

they are usually not aligned with respect to other spectral features. This effect is attributable in part to differences in speaking rate and in part to involuntary articulatory perturbations. The problem of temporal alignment is encountered in combining several matrices that represent the same speaker, and in comparing test and reference matrices. Successive time intervals do not necessarily sample corresponding spectral features.

2. *Experimental Studies*

A summary description of six recognition techniques using specific cue material is presented in Table 15. For each experimental study, this table gives the cue material used, the configuration of the data matrix, the number of utterances included in the reference and test matrices, the recognition task, the decision rules, the number of speakers involved, and an overall measure of performance. These studies will be described more fully in the following paragraphs. Particular attention will be given to the manner in which the individual studies treat the problem of temporal alignment. Some studies inherently depend on an accurate temporal alignment but provide insufficient control over this variable; other studies either avoid the problem of temporal alignment or attempt to deal with it directly.

Pruzansky (1963) used three matrix configurations: (1) an amplitude-frequency-time configuration, as indicated in Table 15, (2) an amplitude-frequency configuration, obtained by averaging over all time intervals, and (3) an amplitude-time configuration, obtained by averaging over all frequency bands. The average recognition scores achieved with these three configurations were 89%, 89%, and 47%, respectively. The confusions which occurred among individual speakers for the amplitude-frequency-time configuration and for

TABLE 15. Summary description of six recognition techniques using specific cue material.

Study	Cue Material	Matrix Configuration				Utterances Incl.		Recogn. Task	Decision Rules	Speakers	Perform. %
		Frequency Bands	Range kHz	Interval msec	Amplitude bits	Ref. Matrix	Test Matrix				
Pruzansky (1963)	10 Words	17	0.2–7.0	10	10	3	1	Ident.	Cross-Correl.	10	89
Pruzansky and Mathews (1964)	10 Words	17	0.1–10.0	10	10	3	1	Ident.	Σd^2	10	93
Ramishvili (1965)	10 Words	7	0.2–10.0	50	2	10	1	Ident.	Σd^2	20	92
Li et al. (1966)	3 Phrases[a]	15	0.3–4.0	20	12	10+	1	Discr.	Adaptive	20	90
Glenn and Kleiner (1968)	Conson. [n]	25	1.0–3.5	–	6	10	10	Ident.	Cross-Correl.	30	93
Meeker (1967)	4 Vowels	19	0.2–8.0	40+	†	20	3	Discr.	Σd^2‡	11	95

[a] Used only first 500 msec of each utterance.
† Used relative frequencies of occurrence of three spectral slopes.
‡ Decision threshold: 1% false rejection.

the amplitude-frequency configuration are given in Tables 16 and 17. It is of interest to note that although the average scores were identical for these two configurations, both the distribution of errors and the relative identifiability of individual speakers differed considerably. For the same two configurations, Figure 30 shows the error scores obtained with each cue word. The speakers

TABLE 16. Confusions and recognition scores (in percent) for ten speakers when identification is based on amplitude-frequency-time matrices. (Reprinted, by permission, from Pruzansky, 1963.)

		Predicted Identity										
	Speaker	*MM*	*BM*	*PB*	*SP*	*JM*	*RG*	*CL*	*LK*	*LG*	*NG*	*Score %*
True Identity	MM	35	1	4	–	–	–	–	–	–	–	88
	BM	–	39	–	1	–	–	–	–	–	–	98
	PB	–	–	35	–	3	–	–	–	2	–	88
	SP	1	1	–	37	–	–	1	–	–	–	92
	JM	1	–	6	–	30	–	–	–	1	1	77
	RG	–	–	1	–	4	35	–	–	–	–	88
	CL	–	–	2	2	–	1	30	1	1	1	79
	LK	1	–	–	–	–	–	–	37	–	1	95
	LG	–	1	–	2	–	–	–	1	35	–	89
	NG	–	2	–	–	–	–	–	–	–	36	95
	Average Score											89

TABLE 17. Confusions and recognition scores (in percent) for ten speakers when identification is based on amplitude-frequency matrices. (Reprinted, by permission, from Pruzansky, 1963.)

		Predicted Identity										
	Speaker	*MM*	*BM*	*PB*	*SP*	*JM*	*RG*	*CL*	*LK*	*LG*	*NG*	*Score %*
True Identity	MM	37	–	–	–	–	–	–	–	2	1	93
	BM	1	37	–	–	–	–	1	1	–	–	93
	PB	–	–	30	–	3	3	–	–	3	1	75
	SP	–	1	–	38	–	–	–	–	1	–	95
	JM	–	–	3	–	33	–	–	1	2	–	85
	RG	–	–	–	–	4	36	–	–	–	–	90
	CL	–	–	–	–	–	–	33	2	1	1	87
	LK	–	1	–	–	–	–	–	34	2	2	87
	LG	–	–	1	–	–	–	–	–	36	2	92
	NG	–	–	1	–	2	–	1	–	–	34	89
	Average Score											89

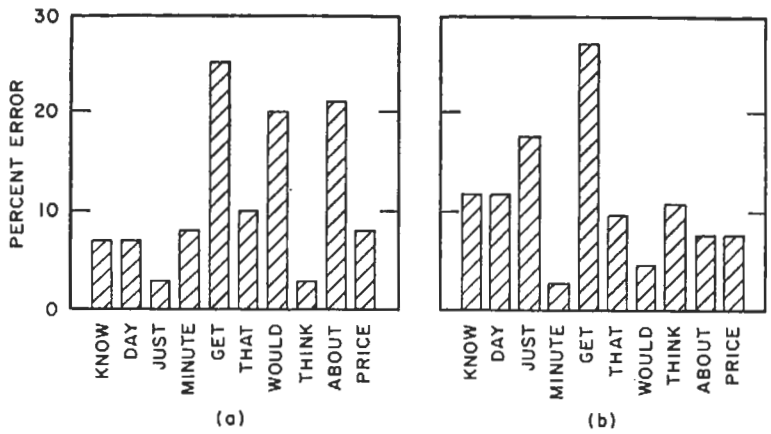

FIGURE 30. Percent error, averaged over speakers, for individual cue words. Data are shown for identification based on (a) amplitude-frequency-time matrices, and (b) amplitude-frequency matrices. (Reprinted, by permission, from Pruzansky, 1963.)

were correctly identified to a degree which depended not only on the particular cue word used, but generally also on the matrix configuration. Because the overall performance was not reduced by time averaging, it was concluded from this study that the long-term spectrum for specific cue material uttered by a particular speaker is distinctive.

Pruzansky and Mathews (1964) explored the feasibility of using only some of the matrix cells in computing recognition scores. The matrix configuration employed in the first part of this study is described in Table 15. For a given cue word, the 30 matrices (with all cell values included) from which the ten reference matrices were constructed were subjected to an analysis of variance to determine for each cell a ratio of interspeaker to intraspeaker variance (F ratio). It was hypothesized that cells having large F ratios would contribute more to correct identification than cells having small F ratios. Recognition scores were computed for various percentages of the total number of cells available for the given cue word. At first, only 1% of the cells having the largest F ratios were used; then more and more cells having progressively smaller F ratios were added, until all of the cells were finally used. The same process was then repeated for cells having small F ratios, using at first only 1% of the cells having the smallest F ratios and finally all of the available cells. The results obtained, averaged over ten cue words, are shown in Figure 31. It is evident that when the cells are used in order of decreasing F ratio (i.e., cells with highest F ratios are used first), there is little further improvement in performance after only 10% of the cells have been included.

In the second part of their study, Pruzansky and Mathews modified the earlier matrix configuration. F ratios were calculated for various matrices having fewer cells; the cell values were obtained by averaging the values of the former cells over several frequency bands, several 10-msec time intervals,

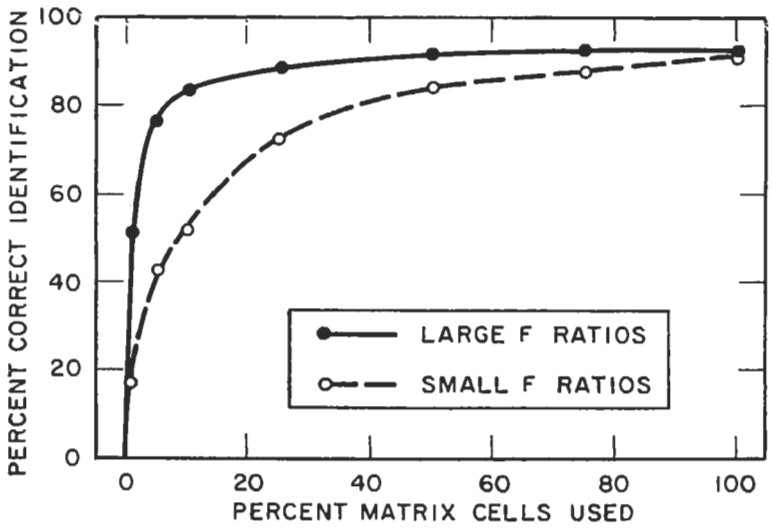

FIGURE 31. Percent correct identification as a function of the percentage of matrix cells used. Data are shown for matrix cells having large and small *F* ratios. (Reprinted, by permission, from Pruzansky and Mathews, 1964.)

or both. Recognition scores were then computed for these matrices, using a variable number of cells having large *F* ratios. The highest score was obtained for a matrix in which each cell encompassed one frequency band and all time intervals. This finding was interpreted as evidence that time segmentation is not essential for accurate speaker identification. It was also concluded that the information contained in a given frequency band is relatively independent of the contents of neighboring frequency bands, whereas the information contained in a given time interval is largely dependent on the contents of neighboring time intervals.

The study by Ramishvili (1965) differs from the two studies described above in that it employed a coarser matrix (see Table 15). Fewer and much wider frequency bands were used, the time intervals were much longer, and the amplitude was specified much less accurately. Nevertheless, approximately the same level of performance was achieved (92%). Perhaps the coarser frequency segmentation and amplitude specification were offset by the inclusion of a larger number of utterances in the reference matrices. Whatever the explanation for the high scores may be, this study casts some doubt on the earlier conclusion about the importance of high frequency resolution.

In the studies by Pruzansky (1963) and Pruzansky and Mathews (1964), and presumably also in the study by Ramishvili (1965), the matrices representing single utterances were aligned with respect to the time interval having the highest overall amplitude. This method of temporal alignment is considered extremely crude, and the resulting displacement of many spectral features may have affected the results obtained for short time intervals. In the study by

Pruzansky, for example, a score higher than 89% might have been obtained for the amplitude-frequency-time matrix if it had been possible to provide more accurate temporal alignment. The temporal resolution of spectral features is likely to benefit recognition only when the corresponding features of different utterances are properly aligned. A given degree of misalignment is expected to have less and less effect on performance as the time interval is increased; it would have no effect for the amplitude-frequency matrix. Poor temporal alignment also may have confounded the variables studied by Pruzansky and Mathews. The general conclusions reached by these investigators about the unimportance of time segmentation are therefore open to question.

Li, Dammann, and Chapman (1966) developed a recognition technique that used adaptive switching circuits to implement the decision rules.[1] This recognition system was trained to discriminate between utterances by a particular speaker and utterances by a group of imposters. Various forms of system programming were studied, including different decision rules and decision criteria. Many recordings of three cue phrases were used, and the number of utterances represented by the reference matrix was varied. Under all experimental conditions, the input data matrix consisted of 15 frequency bands and 25 time intervals (comprising a total duration of 500 msec). System performance reached 90% if the reference matrix represented at least ten utterances and if these utterances were originally recorded at different times, rather than in succession.

These investigators used a voice-operated switch to detect the beginning of each utterance entered into the recognition system. The switch triggered a digital timing sequence which effectively accomplished an initial temporal alignment. To ensure reliable operation of the switch, the cue phrases were restricted to begin with a stressed vowel. Even if a reasonably accurate initial alignment was achieved in this manner, it is likely that the subsequent displacement of spectral features in different utterances had a detrimental effect on performance.

The problem of temporal alignment may be avoided by using only one time interval which coincides with the occurrence of a particular phoneme. The recognition technique described by Glenn and Kleiner (1968) is based on a spectral analysis of the nasal consonant [n]. Thirty speakers recorded two different word lists containing numerous repetitions of this consonant, and for each speaker and word list ten occurrences of [n] were selected by visual inspection of spectrograms. Spectrographic sections (amplitude-frequency plots) of the selected consonants were manually quantized for computer processing. For each speaker, the spectral data representing one word list were averaged to form a 25-cell reference matrix, and the spectral data representing the other word list were averaged to form a 25-cell test matrix. When the speaker population was divided into three 10-speaker groups, an overall iden-

[1]The general theory and operating characteristics of adaptive switching circuits have been described by Widrow and Hoff (1960).

tification score of 97% was obtained; when all speakers were involved simultaneously, the overall score was 93%. It was concluded that the power spectrum generated during nasal phonation is highly speaker dependent.

Meeker (1967) viewed speech recognition (i.e., the identification of the phonemes occurring in connected speech) as a prerequisite for successful speaker recognition. A speech-recognition technique developed by Nelson, Herscher, Martin, Zadell, and Falter (1967) automatically selected the particular phoneme on which speaker recognition was to be based. When a vowel was selected, the subsequent analysis consisted of determining whether the average spectral slope in each of 19 frequency bands was positive, zero, or negative for the duration of the vowel. For 20 samples of a given vowel from a particular speaker, a 19 $\times$ 3 reference matrix was computed; this matrix listed for each of the 19 frequency bands the relative frequencies of occurrence of the three spectral slopes. Three further vowel samples from the same speaker were used to compute a similar 19 $\times$ 3 test matrix. Thus, the matrix configuration did not involve amplitude specifications. The results obtained when discrimination was based on single vowels, and when the error measures were averaged over four vowels, are shown in Tables 18 and 19, respectively. The relatively low scores obtained for some combinations of speakers and vowels (Table 18) can be attributed to inaccuracies in the automatic recognition of the vowels.

TABLE 18. Accuracy of rejection of imposters (in percent) for 11 speakers when discrimination is based on a single vowel. (Reprinted, by permission, from Meeker, 1967.)

Speaker	*Accuracy of Rejection (%)*			
	/i/	/ɛ/	/u/	/ʌ/
RC	92	33	47	85
HZ	80	85	70	43
AT	73	83	57	92
EG	82	53	88	93
JU	57	65	62	65
RT	82	52	53	70
GC	85	87	42	35
JF	40	83	22	90
JS	82	100	52	95
AS	83	37	2	37
PS	97	70	55	97

The level of performance which can be achieved using this approach depends heavily on the accuracy with which the desired phoneme can be located in each utterance. Automatic speech recognition is a particularly difficult problem which has both interested and discouraged investigators throughout the history of speech research. Although the underlying requirements are fairly well understood from a theoretical point of view (Young and Hecker, 1968), they cannot

TABLE 19. Accuracy of rejection of imposters (in percent) for 11 speakers when discrimination is based on four vowels. (Reprinted, by permission, from Meeker, 1967.)

Speaker	*Accuracy of Rejection (%)*
RC	100
HZ	100
AT	93
EG	100
JU	95
RT	93
GC	100
JF	97
JS	100
AS	71
PS	100

be readily incorporated into a practical recognition scheme. All presently available speech recognizers are prone to certain kinds of errors.

Carbonell, Grignetti, Stevens, Williams, and Woods (1965) proposed another method for avoiding the need of an overall temporal alignment. This method involves the automatic location of spectral landmarks in an utterance, but it does not depend on speech recognition per se. The basic concept is illustrated in Figure 32. For a given utterance of a particular cue word, in this case the word *baseball*, successive spectra are examined and certain obvious spectral landmarks are selected. In the example shown, the landmarks are: (1) an initial level increase in all frequency bands, corresponding to the release of the stop consonant [b] in *base*, (2) an initial level decrease in the two lowest frequency bands, corresponding to the cessation of voicing in *base*, and (3) a second level increase in all frequency bands, corresponding to the release of the stop consonant [b] in *ball*. The only requirement for a spectral landmark is that it can be reliably detected by relatively simple circuitry; no speaker recognition is performed with these spectra. The spectral landmarks are temporally independent in the sense that their relative positions in time may vary from utterance to utterance.

Associated with each spectral landmark are one or more spectra which are sampled and used for recognition purposes. These sampled spectra are located as close as possible to their respective landmarks in order to minimize the effects of temporal misalignment. The specific locations of the sampled spectra are selected on the basis of knowledge about which time intervals of the utterance are likely to provide maximum differentiation among speakers. In general, regions containing spectral transitions are more useful than regions of relative spectral stability (Stevens, House, and Paul, 1966). Preliminary results obtained with this method are promising; the data for the individual sampled spectra appear to be relatively independent, so that a high overall score may be obtained if these data are combined.

Schroeder (1968) suggested a method for actually accomplishing time normalization. In this method, pairs of parameters believed useful for speaker

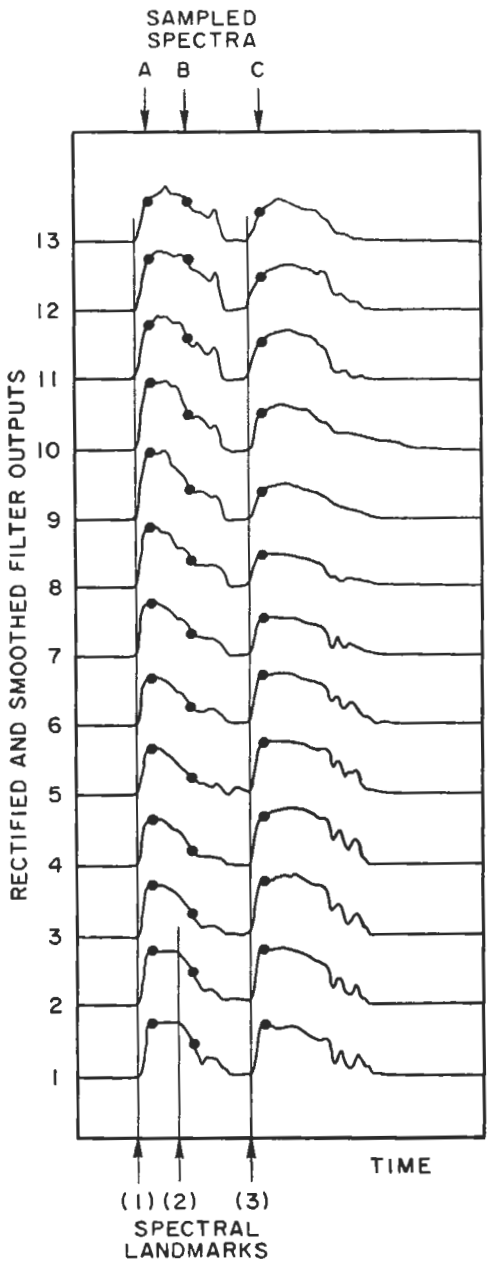

FIGURE 32. Typical output of spectrum analyzer for the cue word *baseball.* Shown are three temporally independent spectral landmarks that locate the spectra to be sampled. (Adapted, by permission, from Carbonell et al., 1965.)

recognition are extracted from the utterances to be compared. Each pair of parameters defines a plane in which each utterance is represented by the contour which is traced out as the parameter values change during the production of the utterance. Two utterances are compared by computing a measure of the similarity of their contours. If the two contours are sufficiently similar, as determined by a selected decision threshold, the represented utterances are ascribed to the same speaker. This method eliminates the effects of unknown timing differences between utterances, such as those introduced by variations in speaking rate. A demonstration of the similarity of two contours representing slow and fast utterances of the word *lion* by the same speaker is shown in Figure 33. In this example, the two parameters are formant frequencies.

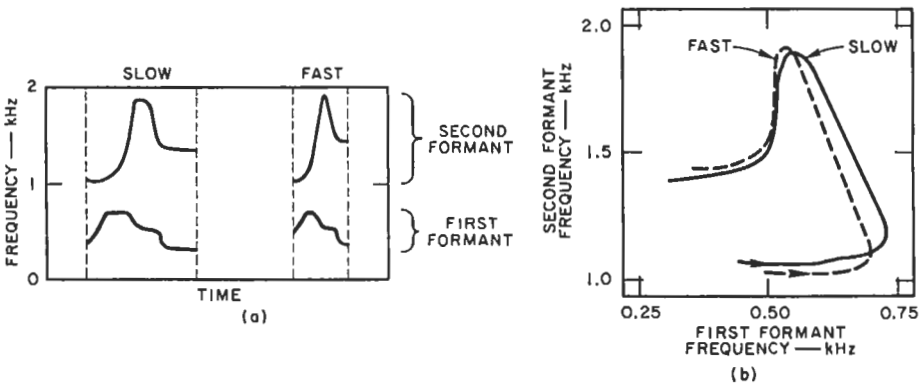

FIGURE 33. Demonstration of contour similarity despite differences in speaking rate. Shown are (a) formant-frequency traces from spectrograms of slow and fast utterances of the word *lion* by the same speaker, and (b) contours for the two utterances in the formant-frequency plane. (Adapted, by permission, from Schroeder, 1968.)

To date, there are no published studies in which this method of time normalization was used. The level of success that can be achieved with this method is expected to depend largely on the choice of parameters. Problems dealing with the selection and extraction of parameters appropriate for differentiating among speakers are considered in the following section.

C. TECHNIQUES USING STATISTICAL ANALYSES OF SPEECH PARAMETERS

The recognition techniques to be described in this section involve two distinct processes: (1) parameters thought to be useful for differentiating among speakers are extracted from the speech signal, and (2) decision rules are applied to combinations of parameter values that represent particular speech samples. It is conventional to regard the parameters as defining a multidimensional observation space in which the speech samples are located and in which the decision rules operate. For example, if only two parameters are extracted,

the observation space is a plane; a speech sample can be located in this plane by means of its two parameter values. In most cases, however, more than two parameters are extracted, and the observation space is more difficult to visualize. The choice of parameters influences the instrumentation requirements of the technique, the complexity of the decision rules, and the level of recognition that can be achieved.

1. Selection of Decision Rules

The decision rules are selected largely on the basis of the distribution in the observation space of speech samples representing different speakers. Three of the most frequently encountered distributions of speech samples are shown in Figure 34. For the sake of clarity, this figure involves only two parameters,

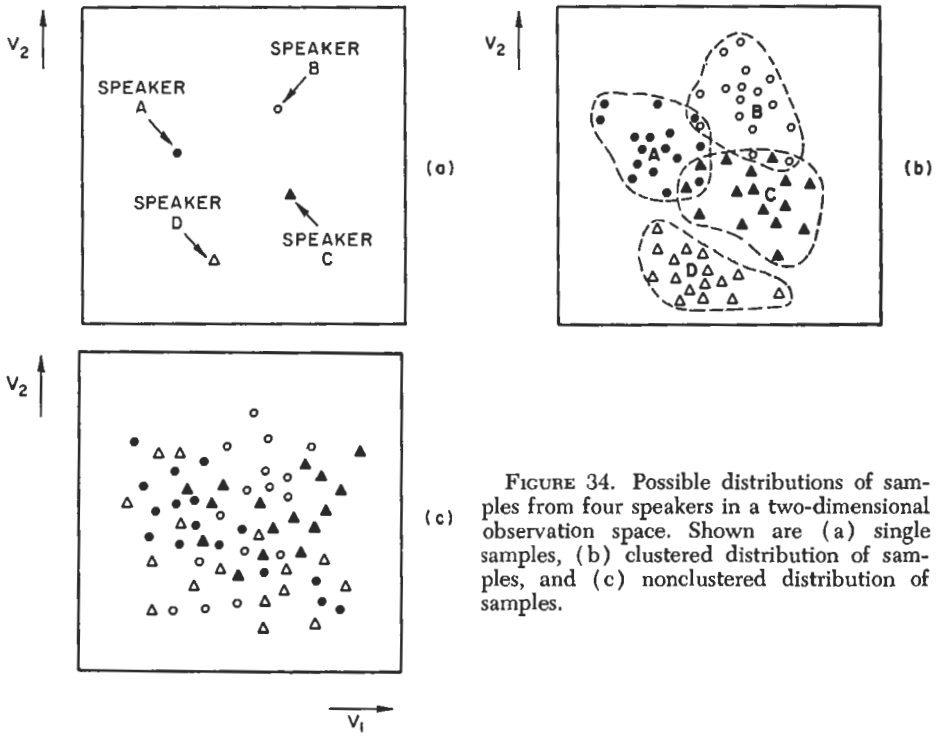

FIGURE 34. Possible distributions of samples from four speakers in a two-dimensional observation space. Shown are (a) single samples, (b) clustered distribution of samples, and (c) nonclustered distribution of samples.

labeled V_1 and V_2, and only four speakers. Part (a) illustrates the case where a single speech sample is available from each speaker. Although the particular four samples shown are well separated in the observation space, there is no assurance that another set of four samples would be similarly positioned and separated. The degree to which the selected parameters actually differentiate among the speakers is best evaluated by obtaining a large number of speech

samples from each speaker. Parts (b) and (c) of this figure illustrate two possible distributions of such samples; the samples from each speaker may be clustered together in a particular region of the observation space, or they may be scattered over the entire space. Even if the samples are clustered, there is usually some overlap of the regions associated with different speakers. In the illustration, the regions associated with Speakers A, B, and C are not as distinct as the region associated with Speaker D. When the samples are distributed in a manner that does not allow particular regions to be associated with particular speakers, it is still possible that the samples from each speaker are distributed in some characteristic pattern.

Decision rules applicable to the case where only one speech sample is available from each speaker usually use distance measures (Clarke, Becker, and Nixon, 1966). The assumption is made that as the differences between respective parameter values for two speech samples increase, the smaller is the probability that the two samples were produced by the same speaker. If it can be further assumed that the selected parameters are largely independent and that they are equally useful for differentiating among speakers, the simplest measure is the Euclidean distance between the two samples.[2] In view of these apparently restrictive assumptions, other distance measures have also been investigated, but none has proved significantly superior to Euclidean distance.

Two applications of decision rules using Euclidean distance measures are illustrated in Figure 35. In the first situation, in which an identification decision is to be rendered, a test sample from the speaker to be identified (Speaker X) is compared with four reference samples (from Speakers A, B, C, and D). The distances between the test sample and each of the reference samples are determined, and the speaker associated with the shortest distance (Speaker A) is assumed to have produced the test sample. In the second situation, in which a discrimination decision is to be rendered, the test sample is compared with only one reference sample (from Speaker A). A decision threshold is required in order to decide whether the two samples were produced by the same speaker or by different speakers. Various kinds of thresholds can be used; here the threshold is simply a circle centered on the reference sample. The radius of the circle indicates the maximum allowable distance at which a test sample can still be considered as having been produced by Speaker A. An alternative to using a fixed threshold is to vary the threshold progressively so as to provide several decision criteria for the preparation of a ROC curve (see Chapter III).

It is apparent that the two situations just described resemble the tasks performed by a human listener in the multiple-choice identification test and the discrimination test. Disregarding possible effects of memory from test item to

[2] In n-dimensional Euclidean space, which is an extension of ordinary three-dimensional space, the distance between any two points $x = (x_1 , \ldots , x_n)$ and $y = (y_1 , \ldots , y_n)$ is defined as

$$d = \left[\sum_{i=1}^{n} |x_i - y_i|^2 \right]^{1/2} .$$

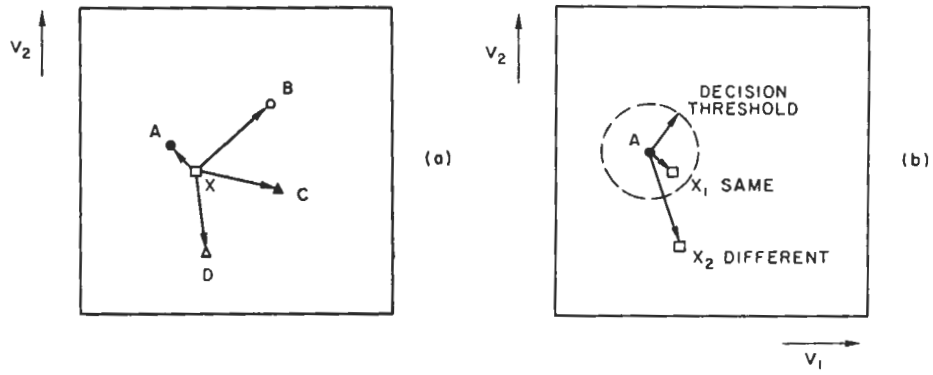

FIGURE 35. Decision rules using Euclidean distance measures between single samples. Shown are (a) distance measures between a test sample and four reference samples, suitable for an identification decision, and (b) a distance measure between two samples, suitable for a discrimination decision.

test item, the listener is required to identify a test sample by comparing it with only one reference sample from each speaker (including the speaker to be identified), or to discriminate between a single pair of speech samples. Of course, the listener does not necessarily extract the same parameters, or use the same decision rules, as a machine. Nevertheless, because of the gross similarity between these two types of tasks, recognition scores obtained using Euclidean distance measures may be compared with scores achieved by listeners on tests involving the same speech material. Such comparisons have been made, and will be discussed further in Section E.

Decision rules using Euclidean distance measures may also be applied when many speech samples are available from each speaker, provided that the samples are clustered. In such a case, each cluster may be represented by a single point, such as its centroid. All distances are measured between these points. This form of analysis, however, does not take advantage of all of the information inherent in the makeup of each cluster. More appropriate decision rules may be applied in the case of clustered distributions (Welch and Wimpress, 1961). These rules involve the use of boundaries to dissect the observation space into regions that can be associated with single speakers or small groups of speakers. When the decision boundaries are hyperplanes rather than nonlinear surfaces, the recognition technique is usually easier to instrument.

The application of decision rules using linear boundaries is illustrated in Figure 36. Since the observation space shown in Part (a) of this figure is a plane, the boundaries are straight lines. The first line, labeled l---l, separates the cluster representing Speaker A from the clusters representing Speakers C and D in a manner that provides approximately equal areas of overlap for each cluster. Line l---l passes through the middle of the cluster representing Speaker B and therefore does not separate this cluster from any of the other clusters. Considering now only data to the left of Line l---l, Line 2---2 separates

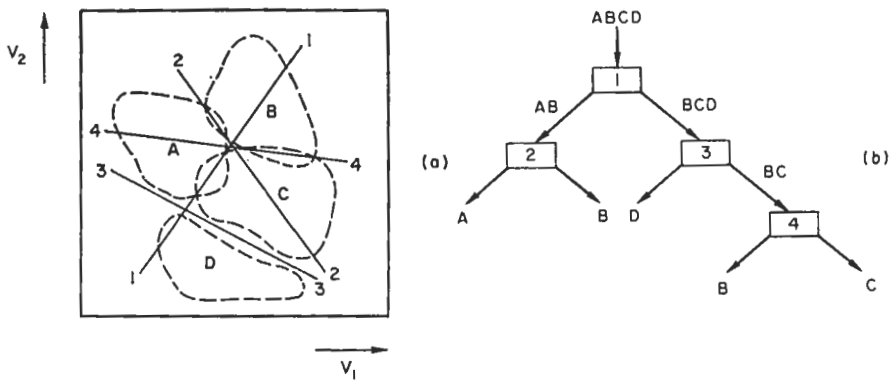

FIGURE 36. Decision rules using linear boundaries to dissect a clustered distribution of samples. Shown are (a) orientation of boundaries in observation space, and (b) decision tree for identifying test samples.

the two clusters representing Speakers A and B. Considering only data to the right of Line 1---1, Line 3---3 separates the cluster representing Speaker D from the clusters representing Speakers B and C. Finally, considering only data to the right of Line 1---1 and above Line 3---3, Line 4---4 separates the two clusters representing Speakers B and C. The corresponding decision tree for identifying a test sample is shown in Part (b) of this figure. When the test sample falls into any of the areas of overlap that were necessarily compromised when the lines were placed, the sample will be incorrectly identified. For example, a sample from Speaker A falling to the right of Line 1---1 would be incorrectly associated with Speaker C.

Decision rules using probabilistic measures (Sebestyen, 1962) are also applicable to a clustered distribution. These rules must always be applied in the case of a nonclustered distribution. Recognition techniques using these rules involve two modes of operation: a learning mode and a recognition mode. During the learning mode, as many speech samples as possible are obtained from each speaker. The distribution of samples for each speaker is then used to estimate a multivariate probability-density function which describes the most likely distribution of further data from the same speaker. Additional samples are more likely to have certain spatial locations than others, and this function provides a ranking of these locations according to their probability of being occupied. The probability-density functions thus computed for the participating speakers are stored. During the recognition mode, either a single test sample or a distribution of test samples may be obtained from a speaker to be identified. It is often possible to associate a single test sample with one speaker for whom the probability of occupancy of the location specified by the test sample is highest. An example of this procedure is shown in Figure 37, where the test sample would be associated with Speaker B. In some instances, however, the probabilities for several speakers may be identical, perhaps as a consequence of insufficient data acquisition during the learning mode. A

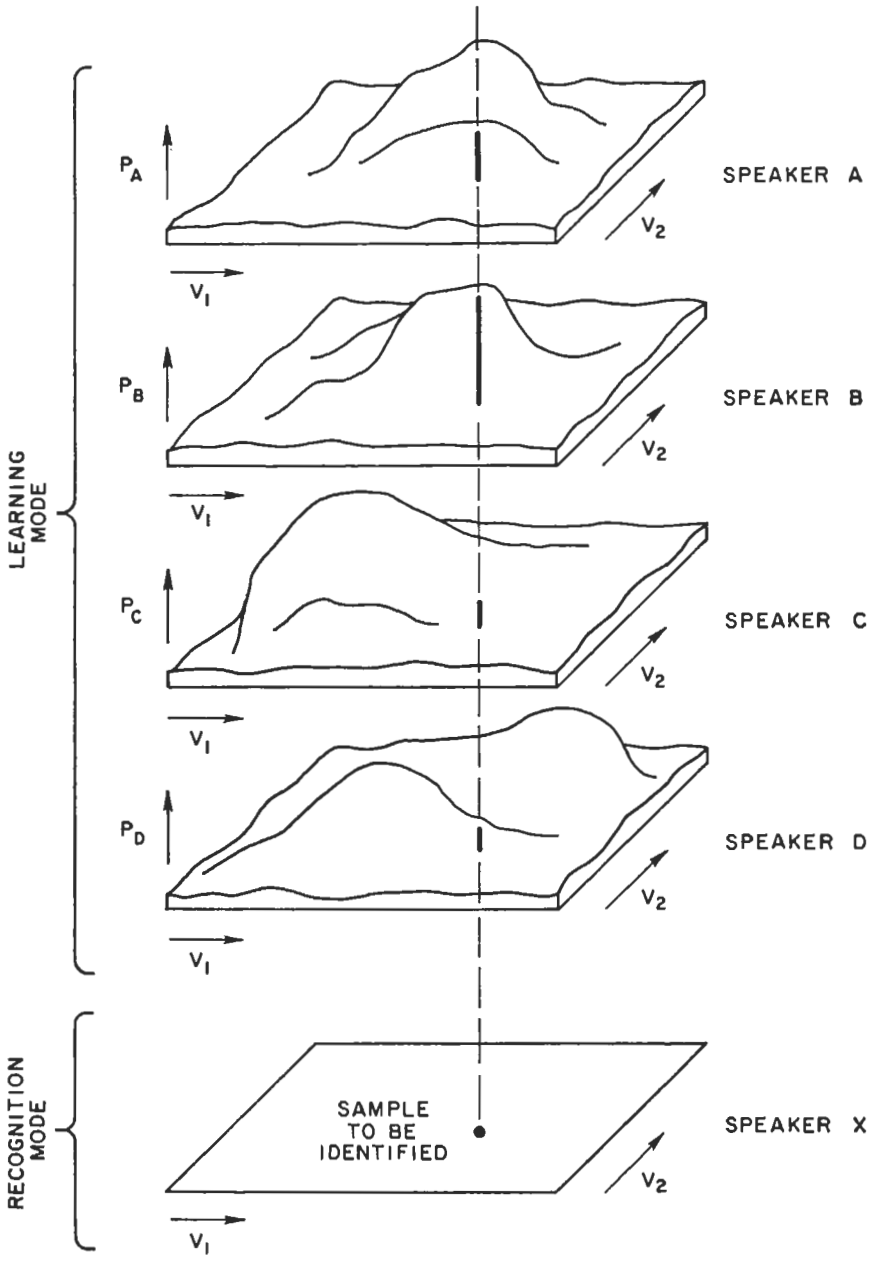

FIGURE 37. Decision rules using probabilistic measures for treating a nonclustered distribution of samples. Shown are estimated multivariate probability-density functions for four speakers, and specific measures for identifying a single test sample.

single test sample may also be identified incorrectly. More accurate identification is possible when a distribution of test samples is available; the test dis-

tribution is then associated with that speaker for whom the conditional probability of its occurrence is highest.

For a given set of parameters, several factors influence the level of recognition that can be achieved with decision rules using probabilistic measures. These include the number of speech samples processed during the learning mode, the number of speech samples processed during the recognition mode, the similarity of the speech materials used during the learning and recognition modes, bandwidth limitations imposed upon the speech signal during either or both modes, noise added to the speech signal during either or both modes, and the size and homogeneity of the speaker ensemble. Under optimal experimental conditions, when the speech is not degraded or distorted, a minimum of about two minutes of connected speech must be sampled during the learning mode. When the speech signal is degraded, the parameters may be extracted less accurately, and considerably more time may be required. Many of these factors also affect performance when the other decision rules are used.

2. *Selection of Speech Parameters*

Questions regarding the most appropriate speech parameters have generally not been resolved as well as have questions regarding optimal decision rules. Various kinds of parameters have been examined, using both waveform analyses and spectral analyses of the speech signal. Some studies have considered only parameters which are relatively easy to extract, such as the outputs of a bank of bandpass filters. Other studies have used more sophisticated parameters, such as formant frequencies, even at the expense of manual (rather than automatic) parameter extraction. The sophisticated parameters, however, do not necessarily produce higher recognition scores. It must be remembered that the formant frequencies are continually changing and that the recognition techniques discussed here do not take into account specific cue material. If formant frequencies were selected as parameters, recognition would be performed on the basis of comparing arbitrarily sampled formant-frequency values rather than average values representing specific vowels. Much rudimentary information about the distribution of formant frequencies is also contained in a series of spectral profiles obtained with a filter bank.

Clarke and Becker (1969) investigated the relative merits of various parameters and sets of parameters, including mean fundamental frequency, several measures of variability of fundamental frequency, the long-term spectrum, portions of the long-term spectrum, and overall duration. The speech samples consisted of short sentences, and the decision rules used Euclidean distance measures taken between single samples. Both identification decisions and discrimination decisions were obtained; the discrimination decisions were recorded in the form of ROC curves. The results, shown in Tables 20 and 21, indicate that the long-term spectrum was superior to any other parameter or set of parameters. This demonstrates that there are long-term spectral differences

TABLE 20. Recognition scores (in percent) obtained for various parameters and sets of parameters extracted from speech samples that were used in an aural four-choice identification test. On the aural test, listeners scored 63-67% correct.

Parameter or Set of Parameters	*Score (%)*
Mean Fundamental Frequency	42
Variability of Fundamental Frequency	32–35
Long-Term Spectrum (24 One-Third Octave Band Filters, 50–10,000 Hz)	63
Lower Portion of Long-Term Spectrum (12 One-Third Octave Band Filters, 50–630 Hz)	54
Upper Portion of Long-Term Spectrum (12 One-Third Octave Band Filters, 800–10,000 Hz)	52
Individual One-Third Octave Bands	28–42
Overall Duration (Normalized for Sentence Length)	32

among speakers.[3] It is interesting to note that the lower and upper portions of the long-term spectrum apparently contribute equally to speaker recognition. In fact, the individual filter bands were not found to produce greatly different recognition scores.

Hargreaves and Starkweather (1963) used only one set of spectral parameters. They recorded extemporary speech from each of 12 female speakers during 18 experimental sessions. Six sessions were held on each of three days that were spaced one week apart. The speech was processed by a spectrum analyzer consisting of 18 one-third octave band filters (covering the frequency range 100–5000 Hz), 18 detectors, and 18 integrators. Two alternate banks of 18 storage capacitors were switched at two-second intervals between positions of data acquisition and data reduction. Thus, every two seconds an 18-channel

TABLE 21. Recognition scores (in percent) obtained for various parameters and sets of parameters extracted from speech samples that were used in an aural discrimination test. The scores are optimal points on ROC curves. On the aural test, listeners scored 90% correct (optimal point on the median ROC curve).

Parameter or Set of Parameters	*Score (%)*
Mean Fundamental Frequency	66
Distribution of Fundamental Frequency	67
Long-Term Spectrum (17 One-Third Octave Band Filters, 100–4000 Hz)	83
Individual One-Third Octave Bands	58-69

[3] Pruzansky (1963) reached the same conclusion.

spectral profile was obtained for amplitude quantization and computer processing. Sixty spectral profiles, or speech samples, from each speaker were used to define linear boundaries in the observation space and to construct a 12-speaker decision tree. These speech samples represented in each case the first experimental session on the first and second days. The speech samples representing the remaining 16 sessions constituted the test samples. As shown in Table 22,

TABLE 22. Confusions and recognition scores (in percent) for 12 speakers participating in 16 experimental sessions. (Reprinted, by permission, from Hargreaves and Starkweather, 1963.)

	Speaker	*Predicted Identity*												*Score %*
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	
True Identity	1	14	–	1	–	–	–	–	1	–	–	–	–	88
	2	–	16	–	–	–	–	–	–	–	–	–	–	100
	3	–	–	15	–	–	1	–	–	–	–	–	–	94
	4	–	–	–	16	–	–	–	–	–	–	–	–	100
	5	4	–	1	–	8	–	1	2	–	–	–	–	50
	6	–	–	–	–	–	16	–	–	–	–	–	–	100
	7	–	–	–	–	–	–	15	1	–	–	–	–	94
	8	–	–	2	–	–	–	–	14	–	–	–	–	88
	9	–	–	–	1	–	–	3	1	11	–	–	–	69
	10	–	–	–	–	–	–	–	–	–	16	–	–	100
	11	–	–	–	–	–	–	–	–	–	–	16	–	100
	12	–	–	–	–	–	–	–	–	–	–	1	15	94
	Average Score													90

the accuracy of identification varied considerably from speaker to speaker; the average score was 90%. Most of the errors involved speech samples from the third day, which was the only day not represented by the decision tree. This observation is interpreted as evidence of day-to-day variations in the long-term spectrum of a given speaker.

A recognition technique reported by Smith (1962) used further processing of the extracted parameters before decision rules using probabilistic measures were applied. This technique also furnished a large number of spectral profiles for each speaker, but these profiles were not used directly. Instead, they were transformed into a new set of parameters providing maximum differentiation among speakers. The transformation of the spectral measurements was determined by means of a multidimensional analysis of variance carried out during the learning mode of the technique.

Ramishvili (1966) investigated the distribution of the interval between adjacent extremal points in the speech waveform. The analysis of this parameter could be accomplished with relatively simple instrumentation and did not require any amplitude normalization of the speech signal. Lengthy speech

samples (2-3 min duration) from 15 speakers were differentiated, peak-clipped, and processed by a 20-channel interval discriminator. Decision rules using probabilistic measures were applied to the outputs of the discriminator. Perfect recognition was achieved when all 20 channels were used; when only the seven most useful channels were connected, there was a slight reduction in performance.

Among the more extensive searches for suitable parameters is a study by Edie and Sebestyen (1962). A set of 13 parameters was investigated, using probabilistic decision rules. The parameters included four formant frequencies, fundamental frequency, amplitude, and a measure related to the voiced-voiceless distinction. These parameters were extracted manually to insure their evaluation under optimal conditions. The results demonstrated that the selected parameters can lead to very accurate recognition in a controlled environment.

This study was continued by Floyd (1964), who investigated the effects of bandwidth limitations and additive noise on recognition performance. Two sets of parameters were examined. One set of 16 so-called spectral parameters included the frequencies and amplitudes of the first, second, and third formants, fundamental frequency, a measure related to the voiced-voiceless distinction, and a measure of the speech envelope. The other set of eight so-called rudimentary parameters included the average, maximum, and minimum values of fundamental frequency, and the duration of the voiced interval over which these measurements were taken. All parameters were extracted automatically by a combination of hardware and computer processing. A rudimentary parameter was typically associated with a much lower data rate than a spectral parameter.

Different probabilistic decision rules were applied to the two sets of parameters; the rules used with the spectral parameters were considerably more complex than the rules used with the rudimentary parameters. Nevertheless, recognition using the rudimentary parameters was found to be generally superior to recognition using the spectral parameters, especially with noisy or band-limited speech. The technique using the spectral parameters suffered primarily from large errors made in extracting the formant frequencies. This study thus demonstrates that parameters yielding acceptable recognition scores under controlled conditions may be totally inadequate under less favorable but perhaps more realistic conditions.

D. MACHINE FALLIBILITY

One of the reasons for pursuing speaker recognition by machine is the belief that this method is potentially less fallible than other methods because it excludes human error. While the methods of speaker recognition employing listeners and observers are undoubtedly influenced by the limitations of human perception, memory, and judgment, it is important to realize that machines also introduce errors. Excluding errors attributable to mechanical or electronic malfunctions, machine errors are typically due to design shortcomings and

inadequate programming. Specific examples of such problems are the dependence on an accurate temporal alignment of utterances of specific cue material, the extraction of inefficient parameters, and the application of arbitrary and inflexible decision rules. These problems reflect not only economic considerations but also a relatively poor understanding of the acoustical correlates of speaker identity. At the present time, the fallibility of speaker recognition by machine is far from negligible.

There are, however, operational situations in which the errors associated with a given technique are tolerable. For example, in a communication situation involving a cooperative speaker who is to be authenticated for security purposes, several available techniques providing discrimination may be applied successfully. If many prior utterances of a password (specific cue material) are available for the speaker, if other speakers are not likely to use this same password, and if the speaker does not mind repeating his password in case he is falsely rejected, the accuracy with which imposters are rejected may be very high. By varying only the decision threshold, the probability of a false rejection may be decreased, at the expense of increasing the probability of accepting an imposter, until a balance is reached that is best suited to the particular situation.

E. COMPARISON WITH SPEAKER RECOGNITION BY LISTENING

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. The corresponding machine scores are given in Tables 20 and 21. These 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 (63%) 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 entire long-term spectrum (83%) approaches the listener score. It may be concluded from these comparisons that 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.

As pointed out by Clarke and Becker (1969), human listeners do not necessarily use the same parameters that have been found advantageous for machine recognition. The nature of the decision rules used by listeners is also insufficiently understood. Analyses of the responses obtained on aural tests sometimes suggest that individual listeners use different acoustical clues and different

criteria in their evaluations of speech signals. A given listener may also modify his strategy as a test proceeds, especially if he is given some indication of his immediate performance. Other factors that may favorably influence listener scores include auditory memory from test item to test item, and prior experience in differentiating among speakers with perceptually similar voices. In speculating on possible reasons for the superiority of listeners over machines in recognizing speakers, it is well to remember that even the most naive listener has lived in a speech environment for a considerably longer period of time than any machine. The experience he has thus acquired cannot be readily defined and analogized.

Chapter VI

SUMMARY

The acoustical properties of a given word or phrase vary from speaker to speaker (interspeaker variability) and, for a particular speaker, from utterance to utterance (intraspeaker variability). Speaker variability in the speech signal reflects many differences in speech production, including differences in glottal-source characteristics, vocal-tract configurations, and articulatory transitions. Although speaker variability is difficult to quantify, it is possible to estimate the relative magnitudes of interspeaker and intraspeaker variability from several experimental studies. Interspeaker variability ordinarily exceeds intraspeaker variability. This fact is the basis for all methods of speaker recognition.

The oldest and most widely studied method is speaker recognition by listening. Several speakers are recorded reading selected speech material, the recordings are edited and presented to listeners, and the listeners carry out a recognition task. Many variables affect listener performance, including the size of the speaker group, the kind of speech material used, and the task assigned to the listeners. A test format provides control over these variables. The most common tests are the multiple-choice identification test, in which the listener matches a speech sample to one of a number of reference samples, and the discrimination test, in which the listener decides whether two speech samples were uttered by the same speaker or by different speakers.

Research on speaker recognition by listening is also concerned with the perceptual factors which underlie the ability of listeners to differentiate among voices. These factors are usually explored with the voice-attribute rating test, in which the listener rates a speech sample on many psychological scales. Another research objective is knowledge of the acoustical manifestations of speaker identity. To determine what features of the speech signal are speaker dependent, the speech signal is selectively modified and the effects on listener performance are noted. Speaker recognition by listening is also used to evaluate communication systems.

A second method is speaker recognition by visual comparison of spectrograms. This method of speaker recognition makes use of an instrument (the sound spectrograph) which provides a visual display of the speech signal (a spectrogram). Several speakers are recorded reading selected cue material, spectrograms of different utterances of the same cue material are prepared

and presented to observers, and the observers carry out a recognition task. Among the many variables affecting observer performance are the context in which the cue material was uttered, the type of spectrogram used, and the training of the observers. The most common tests are the multiple-choice identification test and the discrimination test. These tests resemble the same-named tests used in speaker recognition by listening. Experiments involving both methods of speaker recognition indicate that speaker recognition by visual comparison of spectrograms is less accurate. However, these experiments were conducted with minimally trained observers.

Speaker recognition by visual comparison of spectrograms is used as an investigative tool by law-enforcement agencies. Many published claims about the accuracy and reliability of this method of speaker recognition are not adequately supported by experimental data. For these reasons, the method must be studied further. The practice of referring to a spectrogram as a voiceprint is misleading; it suggests a direct analogy between spectrograms and fingerprints, whereas no such analogy exists.

A third method is speaker recognition by machine. There are two approaches: The machine can be designed to generate and examine amplitude-frequency-time matrices of specific cue material, or to extract and analyze speaker-dependent parameters of the speech signal. Many of the recognition techniques using specific cue material require, but do not provide, an accurate temporal alignment of the data matrices being compared. Those techniques that avoid the problem of temporal alignment are considered more promising. The recognition techniques using statistical analyses of speech parameters are handicapped by a lack of knowledge of efficient speaker-dependent parameters and by difficulties in extracting the selected parameters from the speech signal. In both approaches, the application of decision rules to the data that represent the different speakers is well understood. At the present time, speaker recognition by machine is considerably less accurate than speaker recognition by listening, but this performance gap is likely to close as research continues and new machines are developed.

REFERENCES

ALPERT, M., KURTZBERG, R. L., PILOT, M., and FRIEDHOFF, A. J., Comparison of the spectra of the voices of twins. *J. acoust. Soc. Amer.*, **35**, 1877 (A) (1963).

ANON., Voice print identification. In W. W. Turner (Ed.), *Criminalistics.* Rochester: Aqueduct Books (1965).

BELL, C. G., FUJISAKI, H., HEINZ, J. M., STEVENS, K. N., and HOUSE, A. S., Reduction of speech spectra by analysis-by-synthesis techniques. *J. acoust. Soc. Amer.*, 33, 1725-1736 (1961).

BOLT, R. H., COOPER, F. S., DAVID, JR., E. E., DENES, P. B., PICKETT, J. M., and STEVENS, K. N., Speaker identification by speech spectrograms: A scientists' view of its reliability for legal purposes. *J. acoust. Soc. Amer.*, 47, 597-612 (1970).

BORDERS, W., Voiceprint allowed as evidence; ruling called first of its kind. *New York Times* (April 12, 1966).

BRICKER, P. D., and PRUZANSKY, S., Effects of stimulus content and duration on talker identification. *J. acoust. Soc. Amer.*, **40**, 1441-1449 (1966).

CARBONELL, J. R., GRIGNETTI, M. C., STEVENS, K. N., WILLIAMS, C. E., and WOODS, B., Speaker authentication techniques. Rpt. 1296 prepared under Contract No. DA-28-043-AMC-00116(E) by Bolt Beranek and Newman, Inc., Cambridge, Mass. (1965).

CLARKE, F. R., and BECKER, R. W., Comparison of techniques for discriminating among talkers. *J. Speech Hearing Res.*, **12**, 747-761 (1969).

CLARKE, F. R., BECKER, R. W., and NIXON, J. C., Characteristics that determine speaker recognition. Rpt. ESD-TR-66-636, Electronic Systems Division, Air Force Systems Command, Hanscom Field (Dec. 1966).

COMPTON, A. J., Effects of filtering and vocal duration upon the identification of speakers, aurally. *J. acoust. Soc. Amer.*, **35**, 1748-1752 (1963).

DUNN, H. K., Methods of measuring vowel formant bandwidths. *J. acoust. Soc. Amer.*, 33, 1737-1746 (1961).

EDIE, J., and SEBESTYEN, G. S., Voice identification general criteria. Rpt. RADC-TDR-62-278, Rome Air Development Center, Air Force Systems Command, Griffiss AFB (May 1962).

EGAN, J. P., Articulation testing methods. *Laryngoscope*, **58**, 955-961 (1948).

EGAN, J. P., SCHULMAN, A. I., and GREENBERG, G. Z., Operating characteristics determined by binary decisions and by ratings. *J. acoust. Soc. Amer.*, **31**, 768-773 (1959).

FANT, G., *Acoustic Theory of Speech Production.* 's-Gravenhage: Mouton (1960).

FLANAGAN, J. L., Some properties of the glottal sound source. *J. Speech Hearing Res.*, **1**, 99-116 (1958).

FLANAGAN, J. L., *Speech Analysis Synthesis and Perception.* New York: Academic Press (1965).

FLOYD, W., Voice identification techniques. Rpt. RADC-TDR-64-312, Rome Air Development Center, Research and Technology Division, Air Force Systems Command, Griffiss AFB (Sept. 1964).

FUJIMURA, O., Analysis of nasal consonants. *J. acoust. Soc. Amer.*, **34**, 1865-1875 (1962).

GARVIN, P., and LADEFOGED, P., Speaker identification and message identification in speech recognition. *Phonetica*, **9**, 193-199 (1963).

GLENN, J. W., and KLEINER, N., Speaker identification based on nasal phonation. *J. acoust. Soc. Amer.*, **43**, 368-372 (1968).

HARGREAVES, W. A., and STARKWEATHER, J. A., Recognition of speaker identity. *Lang. Speech*, **6**, 63-67 (1963).

HECKER, M. H. L., and GUTTMAN, N., Survey of methods for measuring speech quality. *J. audio engr. Soc.*, **15**, 400-403 (1967).
HECKER, M. H. L., STEVENS, K. N., VON BISMARCK, G., and WILLIAMS, C. E., Manifestations of task-induced stress in the acoustic speech signal. *J. acoust. Soc. Amer.*, **44**, 993-1001 (1968).
HECKER, M. H. L., and WILLIAMS, C. E., On the interrelation among speech quality, intelligibility, and speaker identifiability. Proc. 5th Intern. Congr. on Acoustics, Liège (1965).
HEINZ, J. M., and STEVENS, K. N., On the properties of voiceless fricative consonants. *J. acoust. Soc. Amer.*, **33**, 589-596 (1961).
HEMDAL, J. F., Some results from the normalization of speaker differences in a mechanical vowel recognizer. *J. acoust. Soc. Amer.*, **41**, 1594 (A) (1967).
HEMDAL, J. F., and HUGHES, G. W., A feature based computer recognition program for the modeling of vowel perception. In W. Wathen-Dunn (Ed.), *Models for the Perception of Speech and Visual Form.* Cambridge: M.I.T. Press (1967).
HIRANO, M., and SMITH, T., Electromyographic study of tongue function in speech: A preliminary report. Working Papers in Phonetics 7, University of California, Los Angeles (Nov. 1967).
HOLMGREN, G. L., Speaker recognition. Rpt. AFCRL-63-119, Air Force Cambridge Research Laboratories, Office of Aerospace Research, Bedford, Mass. (May 1963).
HOLMGREN, G. L., Physical and psychological correlates of speaker recognition. *J. Speech Hearing Res.*, **10**, 57-66 (1967).
INGEMANN, F., Identification of the speaker's sex from voiceless fricatives. *J. acoust. Soc. Amer.*, **44**, 1142-1144 (L) (1968).
JAKOBSON, R., FANT, C. G. M., and HALLE, M., *Preliminaries to Speech Analysis.* Cambridge: M.I.T. Press (1963).
KENNEDY, H., Appeals court reverses state's first "voiceprint" conviction. *Los Angeles Times* (Oct. 11, 1968).
KERSTA, L. G., Voiceprint identification. *J. acoust. Soc. Amer.*, **34**, 725 (A) (1962a).
KERSTA, L. G., Voiceprint identification. *Nature*, **196**, No. 4861, 1253-1257 (1962b).
KERSTA, L. G., Voiceprint-identification infallibility. *J. acoust. Soc. Amer.*, **34**, 1978 (A) (1962c).
KERSTA, L. G., Environmental influence on the speech of family members shown by spectrographic speech matching. *J. acoust. Soc. Amer.*, **38**, 935 (A) (1965a).
KERSTA, L. G., Voiceprint classification. *J. acoust. Soc. Amer.*, **37**, 1217 (A) (1965b).
KERSTA, L. G., Voiceprint classification for an extended population. *J. acoust. Soc. Amer.*, **39**, 1239 (A) (1966).
KURTZBERG, R. L., ALPERT, M., and FRIEDHOFF, A. J., Identification from voice: Techniques for the reduction of trial-retrial variability. *J. acoust. Soc. Amer.*, **35**, 1877 (A) (1963).
LADEFOGED, P., and VANDERSLICE, R., The voiceprint mystique. Working Papers in Phonetics 7, University of California, Los Angeles (Nov. 1967).
LI, K. P., DAMMANN, J. E., and CHAPMAN, W. D., Experimental studies in speaker verification, using an adaptive system. *J. acoust. Soc. Amer.*, **40**, 966-978 (1966).
MATHEWS, M. V., MILLER, J. E., and DAVID, JR., E. E., Pitch synchronous analysis of voiced sounds. *J. acoust. Soc. Amer.*, **33**, 179-186 (1961).
MCDADE, T., The voiceprint. *The Criminologist*, No. 7, 52-60 (Feb. 1968).
MCGEE, V. E., Invariance of personal characteristics of voice over two vowel sounds. *Percept. mot. Skills*, **21**, 519-529 (1965).
MCGEHEE, F., The reliability of the identification of the human voice. *J. gen. Psychol.*, **17**, 249-271 (1937).
MCGEHEE, F., An experimental study in voice recognition. *J. gen. Psychol.*, **31**, 53-65 (1944).
MEEKER, W. F., Speaker authentication techniques. Tech. Rpt. ECOM-02526-F, U.S. Army Electronics Command, Ft. Monmouth, N.J. (Dec. 1967).
MILLER, J. E., Decapitation and recapitation, a study of voice quality. *J. acoust. Soc. Amer.*, **36**, 2002 (A) (1964).
MILLER, J. E., and MATHEWS, M. V., Investigation of the glottal waveshape by automatic inverse filtering. *J. acoust. Soc. Amer.*, **35**, 1876 (A) (1963).
MYSAK, E. D., Pitch and duration characteristics of older males. *J. Speech Hearing Res.*, **2**, 46-54 (1959).
NELSON, A. L., HERSCHER, M. B., MARTIN, T. B., ZADELL, H. J., and FALTER, J. W., Acoustic

recognition by analog feature-abstraction techniques. In W. Wathen-Dunn (Ed.), *Models for the Perception of Speech and Visual Form.* Cambridge: M.I.T. Press (1967).

OSGOOD, C. E., SUCI, G. J., and TANNENBAUM, P. H., *The Measurement of Meaning.* Urbana: Univ. Ill. Press (1957).

PETERS, R. W., Studies in extra messages: Listener identification of speakers' voices under conditions of certain restrictions imposed upon the voice signal. U.S. Naval School of Aviation Medicine, Joint Project NM 001-064-01, Rpt. 30, Pensacola, Fla. (Oct. 1954).

PETERS, R. W., Studies in extra messages: The effect of various modifications of the voice signal upon the ability of listeners to identify speakers' voices. U.S. Naval School of Aviation Medicine, Joint Project NM 001-104-500, Rpt. 61, Pensacola, Fla. (May 1956).

PICKETT, J. M., Recent research on speech-analyzing aids for the deaf. *IEEE Trans. audio Electroacoust.*, AU-16, 227-234 (1968).

POLLACK, I., PICKETT, J. M., and SUMBY, W. H., On the identification of speakers by voice. *J. acoust. Soc. Amer.*, **26**, 403-406 (1954).

POTTER, R. K., KOPP, G. A., and GREEN, H. C., *Visible Speech.* New York: D. van Nostrand (1947).

PRESTIGIACOMO, A. J., Amplitude contour display of sound spectrograms. *J. acoust. Soc. Amer.*, **34**, 1684-1688 (1962).

PRESTI, A. J., High-speed sound spectrograph. *J. acoust. Soc. Amer.*, **40**, 628-634 (1966).

PRUZANSKY, S., Pattern matching procedure for automatic talker recognition. *J. acoust. Soc. Amer.*, **35**, 354-358 (1963).

PRUZANSKY, S., and MATHEWS, M. V., Talker-recognition procedure based on analysis of variance. *J. acoust. Soc. Amer.*, **36**, 2041-2047 (1964).

PTACEK, P. H., SANDER, E. K., MALONEY, W. H., and JACKSON, C. C. R., Phonatory and related changes with advanced age. *J. Speech Hearing Res.*, **9**, 353-360 (1966).

RAMISHVILI, G. S., Automatic recognition of speaking persons. Rpt. FTD-TT-65-1079, Foreign Technology Division, Air Force Systems Command, Wright-Patterson AFB (Dec. 1965).

RAMISHVILI, G. S., Automatic voice recognition. *Engng. Cybernetics*, No. 5, 84-90 (Sept.-Oct. 1966).

SCHROEDER, M. R., Similarity measure for automatic speech and speaker recognition. *J. acoust. Soc. Amer.*, **43**, 375-377 (L) (1968).

SCHWARTZ, M. F., Identification of speaker sex from isolated voiceless fricatives. *J. acoust. Soc. Amer.*, **43**, 1178-1179 (L) (1968).

SEBESTYEN, G. S., *Decision Making Processes in Pattern Recognition.* New York: Macmillan (1962).

SHEARME, J. N., and HOLMES, J. N., An experiment concerning the recognition of voices. *Lang. Speech*, **2**, 123-131 (1959).

SILBIGER, H. R., Voice classification by hierarchical clustering. *J. acoust. Soc. Amer.*, **40**, 1282 (A) (1966).

SKALBECK, G. A., An experimental study of several factors in speaker recognition. Unpublished master's thesis, Univ. of Wash. (1955).

SMITH, J. E. K., Decision-theoretic speaker recognizer. *J. acoust. Soc. Amer.*, **34**, 1988 (A) (1962).

STARKWEATHER, J. A., Content-free speech as a source of information about the speaker. *J. abnorm. soc. Psychol.*, **52**, 394-402 (1956).

STEVENS, K. N., HECKER, M. H. L., and KRYTER, K. D., An evaluation of speech compression systems. Rpt. RADC-TDR-62-171, Rome Air Development Center, Griffiss AFB (March 1962).

STEVENS, K. N., and HOUSE, A. S., An acoustical theory of vowel production and some of its implications. *J. Speech Hearing Res.*, **4**, 303-320 (1961).

STEVENS, K. N., HOUSE, A. S., and PAUL, A. P., Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation. *J. acoust. Soc. Amer.*, **40**, 123-132 (1966).

STEVENS, K. N., WILLIAMS, C. E., CARBONELL, J. R., and WOODS, B., Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material. *J. acoust. Soc. Amer.*, **44**, 1596-1607 (1968).

STUNTZ, S. E., Speech intelligibility and talker recognition tests of Air Force communication systems. Rpt. ESD-TDR-63-224, Electronics Systems Division, Air Force Systems Command, Hanscom Field (Feb. 1963).

UNGEHEUER, G., Ein einfaches Verfahren zur akustischen Klassifikation von Sprechern. (A simple procedure for the acoustical classification of speakers.) Proc. 5th Intern. Congr. on Acoustics, Liège (1965).
VOIERS, W. D., Perceptual bases of speaker identity. *J. acoust. Soc. Amer.*, **36**, 1065-1073 (1964).
VOIERS, W. D., Performance evaluation of speech processing devices II. The role of individual differences. Rpt. AFCRL-66-24, Air Force Cambridge Research Laboratories, Office of Aerospace Research, Bedford, Mass. (Dec. 1965).
VOIERS, W. D., COHEN, M. F., and MICKUNAS, J., Evaluation of speech processing devices I. Intelligibility, quality, speaker recognizability. Rpt. AFCRL-65-826, Air Force Cambridge Research Laboratories, Office of Aerospace Research, Bedford, Mass. (July 1965).
WELCH, P. D., and WIMPRESS, R. S., Two multivariate statistical computer programs and their application to the vowel recognition problem. *J. acoust. Soc. Amer.*, **33**, 426-434 (1961).
WIDROW, B., and HOFF, M. E., Adaptive switching circuits. Tech. Rpt. 1553-1, Stanford Electronics Lab., Stanford Univ. (June 1960).
WILLIAMS, C. E., The effects of selected factors on the aural identification of speakers. Sect. III of Rpt. ESD-TDR-65-153, Electronic Systems Division, Air Force Systems Command, Hanscom Field (Dec. 1964).
WILLIAMSON, J. A., An investigation of several factors which affect the ability to identify voices as same or different. Unpublished dissertation, Univ. Edinburgh (1961).
WISE, C. M., *Introduction to Phonetics.* Englewood Cliffs: Prentice-Hall (1957).
YOUNG, J. R., and HECKER, M. H. L., Some observations on the problem of machine recognition of speech. Proc. 1968 National Electronics Conf., Chicago (Dec. 1968).
YOUNG, M. A., and CAMPBELL, R. A., Effects of context on talker identification. *J. acoust. Soc. Amer.*, **42**, 1250-1254 (1967).