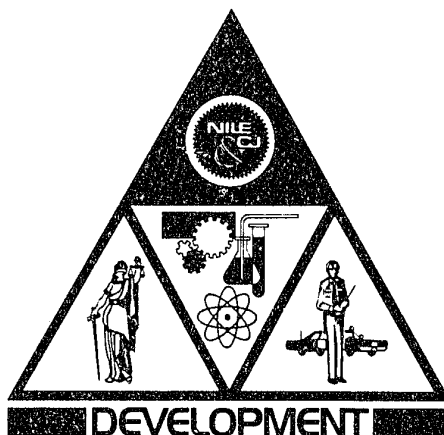


EQUIPMENT SYSTEMS IMPROVEMENT PROGRAM

APPLICATIONS OF SEMI-AUTOMATIC SPEAKER IDENTIFICATION TECHNIQUES

Law Enforcement Development Group

March 1975



Prepared for

NATIONAL INSTITUTE OF LAW ENFORCEMENT AND CRIMINAL JUSTICE

Law Enforcement Assistance Administration

U.S. Department of Justice

THE AEROSPACE CORPORATION



19966

DUP

EQUIPMENT SYSTEMS IMPROVEMENT PROGRAM

APPLICATIONS OF SEMI-AUTOMATIC
SPEAKER IDENTIFICATION TECHNIQUES

Law Enforcement Development Group
THE AEROSPACE CORPORATION
El Segundo, California

March 1975

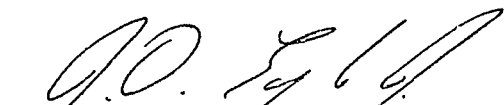
Prepared for
NATIONAL INSTITUTE OF LAW ENFORCEMENT
AND CRIMINAL JUSTICE
Law Enforcement Assistance Administration
U.S. Department of Justice

This project was supported by Contract Number J-LEAA-025-73 awarded by the Law Enforcement Assistance Administration, U.S. Department of Justice, under the Omnibus Crime Control and Safe Streets Act of 1968, as amended. Points of view or opinions stated in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice.

EQUIPMENT SYSTEMS IMPROVEMENT PROGRAM

APPLICATIONS OF SEMI-AUTOMATIC SPEAKER
IDENTIFICATION TECHNIQUES

Approved



John O. Eylar, Jr., General Manager
Law Enforcement and Telecommunications Division

ABSTRACT

This report documents a study of the potential uses of techniques for speaker identification through computer analysis of voice samples. Related efforts in automatic personal identification, using fingerprints, voice samples, etc., are surveyed and the fundamental techniques in use are discussed. The operation of the Semi-Automatic Speaker Identification System is presented and modifications to the system are described, which would be required to use its capabilities for automatic personal identity verification.

Specific application areas for the speaker identification technique are discussed, including: identification for criminal apprehension, identification for crime prevention (personal identification), and non-crime-related applications. Since the technical requirements for criminal apprehension were discussed in a previous report,* this report confines itself to the technical requirements for specific functions of personal identification for areas such as computer access security, area access security, identity verification for check and credit card usage, and remote identity verification by police in the field. The principal technical factors addressed are verification accuracy of the automatic identification equipment, and the requirements for generating and maintaining a data base of individual speech characteristics for use in identification applications.

*"Preliminary Investigation of Applications of the Computer-Aided Speaker Identification System," Aerospace Corp. Report No. ATR-74(7907)-1 (June 1974).

The results of the study indicate that the use of personal attributes (fingerprints, voice characteristics, hand geometry, etc.) for identification is becoming increasingly important in the areas of public safety, business, and industrial operations. Since the variety of applications is large, a number of different approaches will be needed to provide equipment of varying costs, accuracies, vulnerabilities, and communication requirements. The techniques developed for semi-automatic speaker identification have the capabilities to meet the personal identification requirements for many of the specified applications.

ACKNOWLEDGMENTS

The efforts described in this report were conducted by The Aerospace Corporation supported by members of the Autonetics Division of Rockwell International Corporation. Mr. P. K. Broderick is the principal author of the report; he was assisted in its preparation by Dr. C. Henderson, Mr. M. Lubofsky and Mr. K. Henrie of The Aerospace Corporation. Valuable assistance was provided in reviewing the report by Mr. R. Cox, Dr. D. P. Duclos, and Dr. A. Tescher.

SUMMARY

Individuals and organizations concerned with law enforcement and the security of valuable or dangerous assets are increasingly relying upon new methods for identifying people by means of personal attributes. Of particular interest are those personal attributes that can be readily detected by unattended equipment but cannot be easily duplicated by impostors. These personal attributes can be employed to: identify criminals, control access to valuable property, and restrict unauthorized persons from obtaining sensitive information or data.

Identification systems currently in use or under development are based upon detection and analysis of attributes related to fingerprint patterns, voice characteristics, hand size and geometry, and handwriting parameters. Fingerprint systems match the user's print with a file of authorized prints using either optical correlation techniques or a digital approach whereby only the distinctive parts of the print are coded and compared with the corresponding locations of the same distinctive characteristics on prints in the authorized file. Both techniques are currently implemented and operational systems for access security are being marketed.

Voice identification is based upon the observation that the variations between different speech sounds made by a given individual are measurably less than the variations between different individuals. The speech sounds, therefore, are relatively invariant for a given individual and can be considered as an identifying characteristic of that individual. Two specific

techniques used to identify individuals from speech are based on either a detailed spectral analysis of selected speech segments in the frequency domain or on time-domain analysis of selected speech signal elements. A third, but less accurate approach is to simply perform a spectrum analysis on the averaged spectrum of a long (20 seconds or more) speech sample. Such analyses can provide speaker discrimination under certain circumstances but are very dependent upon the quality and stability of the communications channel.

Identification using hand geometry measurements or handwriting characteristics have also been implemented. These techniques are limited in versatility and reliability, however, and are usually employed as supplementary aids in conjunction with other identification methods.

The Semi-Automatic Speaker Identification System is a computer-aided system developed to provide quantitative, objective measures of the similarity* between recorded voice samples. It is intended to improve the effectiveness of police in investigating crimes where voice recordings are involved. Although the system has been designed as an investigative tool using a semi-skilled operator, the system performance indicates that the technique has potential for use as a fully automatic identity verification system based on voice characteristics. The modifications to the system which would be required to produce this capability consist mainly of adding capabilities for: automatic phoneme recognition, sound waveform segmentation, and decision processing.

*The term "similarity" is used to denote a mathematical function which measures the relationship between two arbitrary signals or patterns.

The scope of the application areas for the speaker identification technique include: identification for criminal apprehension and prosecution, personal identification for crime prevention due to improved security techniques, and voice identification systems to allow remote conduct of business and financial transactions that currently rely on personal contact.

An analysis of the potential accuracy achievable by the semi-automatic speaker identification techniques indicated that, with the present configuration, both the false acceptance rate and the false rejection rate could be maintained below 0.2 percent when eight phonemic events were used. These figures were based on the system performance with 118 speakers. An increase in the size of the speaker population could tend to increase the error rates; however, since the 118 speakers were all using the same dialect (General American English), the results can be considered conservative when applied to a population having heterogeneous dialects. Analyses of the data base requirements for storing and retrieving speaker data from automated identification systems indicate that costs would be modest even for populations of thousands of speakers.

Several potential configurations for practical speaker identity verification systems are described for both law enforcement and commercial utilization. The systems are currently realizable with present technology and provide improvements in terms of operational effectiveness and economy in the application areas. The actual implementation of these concepts should be preceded by tests and analyses to provide information for detailed design.

CONTENTS

ABSTRACT	iii
ACKNOWLEDGEMENTS	v
SUMMARY	vi
I. INTRODUCTION	1-1
A. Background — Personal Identification	1-1
B. Existing System Approaches	1-3
1. Fingerprint approaches	1-3
2. Voice identification approaches	1-5
3. Hand geometry approaches	1-16
4. Handwriting scanner approaches	1-17
C. Semi-Automatic Speaker Identification System (Current Implementation)	1-18
D. Modifications to Current Implementation	1-28
II. POTENTIAL APPLICATION AREAS	2-1
A. Identification for Criminal Apprehension	2-1
1. Suspect identification	2-2
2. Criminal evidence	2-7
B. Personal Identification for Crime Prevention	2-10
1. Computer access security	2-11
2. Area access security	2-15
3. Commercial credit	2-18
C. Non-Crime-Related Applications	2-22

CONTENTS (Continued)

III.	TECHNICAL EVALUATION	3-1
A.	Accuracy	3-1
1.	Decision-making objectives and constraints	3-2
2.	Theoretical and demonstrated accuracies	3-7
B.	Data Base Storage Requirements	3-17
1.	Analysis	3-18
2.	Conclusions	3-22
IV.	POTENTIAL IMPLEMENTATIONS	4-1
A.	Police Remote Identity Verification Concept	4-1
B.	Automated Identification System Concept for Financial Transactions and Access Control	4-5
V.	CONCLUSIONS AND RECOMMENDATIONS	5-1
	NOTES	N-1
	APPENDIX A: ANALYSIS OF POTENTIAL VOICE IDENTIFICATION EFFECTIVENESS	A-1
	APPENDIX B: POTENTIAL CONCEPT FOR ESTABLISHING ARRESTEE VOICE SAMPLE FILE	B-1

TABLES

1-1	Spectrum of Security Levels for Personal Identification	1-2
1-2	Selected Phoneme Types	1-19
2-1	Potential Speaker Identification Applications in Local Law Enforcement Agencies	2-4
2-2	Summary of Applicable Cases in 1973	2-5

ILLUSTRATIONS

1-1	Linear Model of the Speech Process	1-6
1-2	Spectral Transformation	1-7
1-3	Slope Classification of Vowel Sound in the Word "Bed"	1-11
1-4	System Operation	1-21
1-5	Example of Alphaphonetic Transcription	1-22
1-6	Computer-Aided Voice Comparison	1-24
1-7	Semi-Automatic Speaker Identification System Process	1-27
3-1	Speaker Recognition/Verification Definitions	3-4
3-2	Classification/Recognition Errors	3-6
3-3	Average Classification Accuracy for 25 Speakers Using 6 Concatenated Phonemes; Features Derived from Special Features Plus LPC Spectral Estimates	3-8
3-4	False Identification Error-Rate Estimate, 118 Speaker Male Population for a Closed-Decision Test	3-11
3-5	Theoretical Calculation of Probability of Successful Identification of Voice Signal vs. Number of Possible Candidates	3-12

ILLUSTRATIONS (Continued)

3-6	Identification Error-Rate Estimates for Open-Decision Tests Using 118-Speaker Population	3-13
3-7	Example of Confidence Estimates on System Statistics	3-16
3-8	Simplified System Speaker Identification Mechanization	3-19
3-9	Representative Speech Characteristics Storage Blocks	3-20
4-1	Concept for Field Identification	4-2
4-2	Automated System Concept for Financial Transactions	4-8

CHAPTER I. INTRODUCTION

A. Background - Personal Identification


There is a growing need for foolproof and instantaneous verification of positive identity in a number of economic and social areas. These include: criminal apprehension, intelligence and security, banking and credit, merchandising, and drug abuse prevention. To meet the need, the technology for processing information regarding the unique human attributes that convey identity has been enhanced by development of automated and semi-automated systems. Such systems have resulted from advancements in the technologies of pattern recognition and data processing.

The various approaches to personal identification techniques are listed in Table 1-1.¹ The table presents the basis for identification, the relative level of security achieved, and the requirements for defeating the technique (i. e., its vulnerability).

With the possible exception of the genetic code technique, systems based upon each of the approaches listed in the table are currently under development or have been implemented.

In the most general sense, all personal identification systems consist of three elements: (1) a unique or rare attribute or artifact possessed by the individual, (2) a device or file which contains a list of the specific artifacts that can be identified as belonging to authorized individuals, and (3) a method for accurately testing whether a given artifact presented by an individual actually is contained in the authorized list. In the simplest case, the artifact

Table 1-1. Spectrum of Security Levels for Personal Identification

Basis for Identification	Security Level	Requirements for Security Breach
Code or Combination System		No forgery of identification necessary
Key System		
Card System		Some simple forgery needed
Card-Code System		
Personal Appearance System		Readily duplicated personal attributes
Handwriting		
Hand Geometry System		
Voice Identification System		Difficult to duplicate personal attributes
Fingerprint System		
System Based upon Genetic Codes	Maximum	No known methods for duplication

would be a key and the device would be the matching form of the key contained within the corresponding lock. The matching method would be the lock mechanism, which releases the latch when the correct key is inserted and turned. The other extreme would involve posting a guard at the access or entry point, and allowing entrance only to those individuals known to him. In this case, the personal appearance (facial features, mannerisms, voice, etc.) of the individual would be the artifact. The memory of the guard would be the file, and the test method would be the guard's ability to remember and recognize individuals. For most security applications, the lock and key approach is too vulnerable and the human guard too expensive for practical utilization.

For automatic and semi-automatic access control and identification, four basic approaches have been developed, combining the security associated with techniques based upon personal attributes with the capability for low-cost, reliable implementation. The techniques employed are: fingerprint scanners, voice analyzers, hand geometry sensors, and handwriting scanners. Each technique has been implemented in one or more equipment configurations.

B. Existing System Approaches

1. Fingerprint approaches. Fingerprints have been used as a positive means of identification for more than 100 years and are considered one of the most reliable means of personal identity verification in existence.

There are two basic techniques for fingerprint identification in use today. One technique uses correlation of the entire fingerprint image performed optically as the matching mechanism, and the other detects and matches only the minutiae characteristics of the fingerprint.* Systems based on the two techniques have been developed by KMS Technology Center, by the Calspan Corp., and others.

a. Automatic Personnel Verifier.² The Automatic Personnel Verifier is the first commercial secure-access optical correlation system designed to identify individuals from their fingerprints. It was developed by KMS Technology Center and has two modes of operation.

*Minutiae are the ridge endings and branches that occur in the fingerprint pattern and give a particular print the uniqueness useful for identification.

In the first mode, a card containing a prerecorded hologram* of an individual's fingerprint is inserted into the terminal device at the same time that the individual places the corresponding finger on the print reader portion of the terminal. The terminal matches the image of the impressed finger with the hologram on the card. If hologram and fresh print match according to a preset criteria, the individual is accepted.

In the second mode, the hologram cards are stored in the system and are not carried by the individual. The individual inserts his finger and enters an identification number. This number provides the address of the particular hologram card associated with the individual and the matching and decision process proceeds as before.

b. Fingerscan.¹ The Fingerscan system was developed by the Calspan Corp. and employs a solid-state photosensor array to transfer the fingerprint image to a computer in digital form for processing. In the computer, the type and relative locations of the minutiae contained in the fingerprint pattern are detected. As with the previous system, the fingerprint is entered simultaneously with an identification number. The number is used to locate an entry in the computer file containing the descriptive characteristics of the individual's fingerprint. These characteristics are compared with the characteristics of the print scanned by the terminal. If no match is made,

*A hologram is produced by recording, on photographic film, a pattern formed by optically transforming the fingerprint image. The recorded pattern is such that automatic matching of the original print with all prints can be accomplished.

another reading is requested. Subsequent unsuccessful attempts result in either an alarm or other operator notification.

2. Voice identification approaches. Speech is usable for identification because it is a product of the speaker's individual anatomy and linguistic background.³ When air is expelled from the lungs, it passes through the glottis which is the opening bounded on either side by the vocal folds. When the vocal folds are drawn together and air from the lungs is forced through them, they vibrate, making a buzzing sound. The waveform of the glottal source is a series of pulses, as shown in Figure 1-1. This sound is modified as it passes through the vocal tract, which is the tube formed principally by the pharyngeal cavity (throat) and the oral cavity. The sound emanating from the vocal tract will be distinctly different from the initial buzz and will have a complex waveform, as shown in Figure 1-1. The shape of the vocal tract serves to concentrate sound energy at certain frequencies and reduce it in others.

Figure 1-2 illustrates how the spectrum of the glottal source is modified by the vocal tract. The relationship between any input signal applied to the vocal tract and the resulting output signal can be mathematically described in terms of a transfer function. In transferring the acoustic energy from the glottis to the lips of the speaker, the vocal tract selectively emphasizes certain portions of the glottal spectrum in accordance with the particular transfer function it has at that point in time. During the speech, the shape of the vocal tract is continuously modified by movements of the tongue, lips, and other vocal organs. Thus, the quality of the speech sounds a speaker produces represents the sizes and shapes of his vocal organs and the manner in which he uses them. Speech characteristics obviously vary

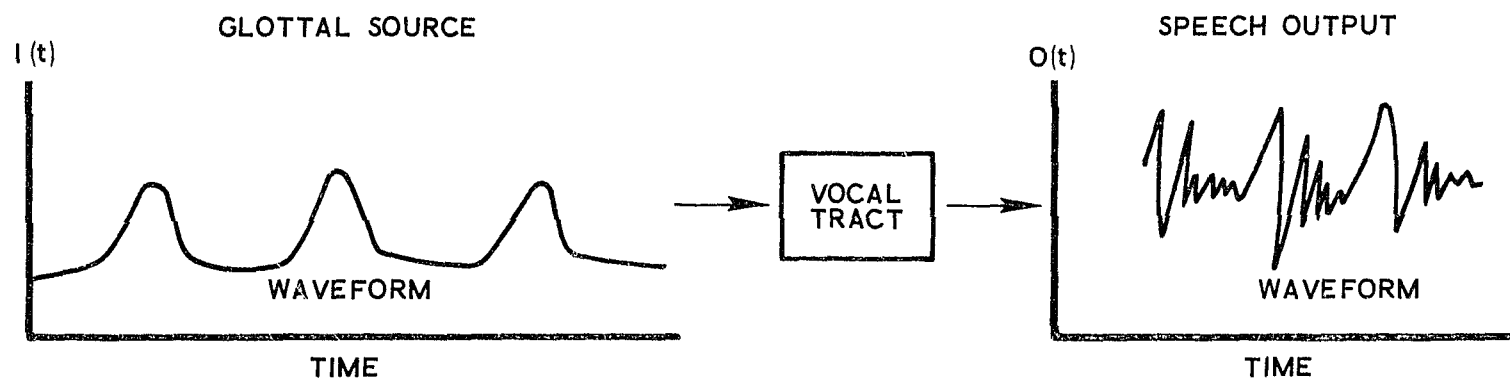
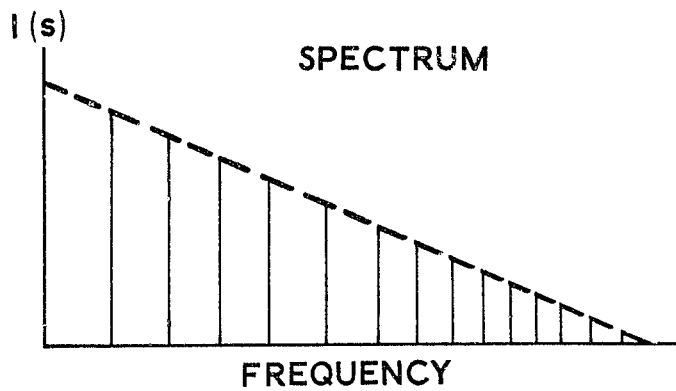
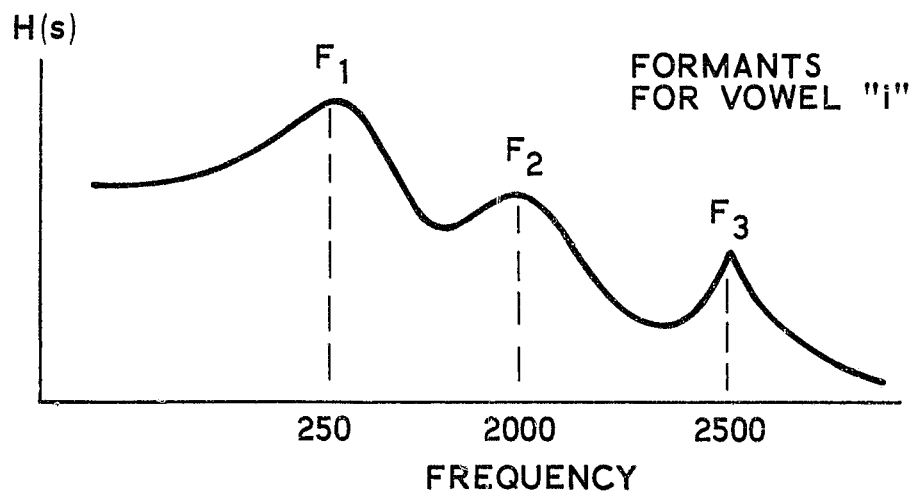


Figure 1-1. Linear Model of the Speech Process

GLOTTAL SOURCE



VOCAL TRACT TRANSFER FUNCTION



SPEECH OUTPUT

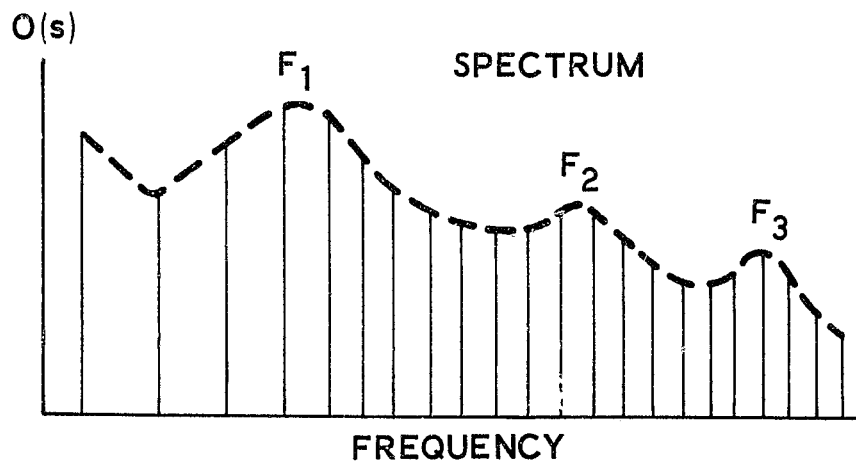


Figure 1-2. Spectral Transformation

between individuals. The effect is termed "interspeaker (between speakers) variability." Speech analysis also reveals variability when the same speaker utters a given sound several times, and this is called "intraspeaker (within speaker) variability." Intraspeaker variability arises because the physiological activity necessary to make speech sounds need not be exactly controlled to effect adequate communication.

The speech signal produced by a given individual is affected by both the organic characteristics of the speaker (in terms of vocal tract geometry) and learned differences due to ethnic or social factors.

From an individual's speech signal, a detailed spectral analysis can be performed to determine the shape of the vocal tract transfer function. Since the vocal tract transfer function exhibits several peaks that correspond to the natural frequencies of the vocal tract (formants), measurements of the properties of this function can provide indications of the unique manner in which a specific individual produces a given sound. Numerous such measurements, when properly combined, can provide information regarding the identity of the speaker.

The reliability of a speech identification approach is related to the degree to which the interspeaker variability can be maximized relative to the intraspeaker variability.

Most approaches to automatic personal identification through voice signal analysis include: (1) the capability to automatically recognize particular speech elements such as specific vowels, and (2) the capability to detect and measure carefully chosen speech parameters from these elements, which are speaker-dependent. The speech parameters should be principally

affected by the individual's physiological characteristics, such as the size and shape of the mouth and vocal cavities, and should be minimally affected by such things as emotional state, health, and other variable factors.

Voice verification techniques must also have the capability to compare these speech parameters to equivalent ones prestored as reference samples for the same individual. Based on these comparisons, accurate decisions must be made regarding whether the present speaker is the same individual who made the prestored reference sample.

A large number of organizations are currently conducting research and development in the area of automatic voice access and speech recognition systems, and several of these organizations have developed and are marketing equipment for secure access based on voice identification. The status of these efforts is discussed below.

a. Threshold Technology, Inc. The speaker verification technique developed by this manufacturer uses automatic speech recognition equipment to extract phoneme-like* elements from continuous speech and performs speaker verification on the basis of an analysis of these extracted elements.⁴ The automatic speech recognition equipment uses a 19-channel (14 channels for telephone grade signals) filter bank. The outputs of this filter bank are processed to obtain measures of the spectral energy, spectrum

* A phoneme is the smallest unit of speech that distinguishes one sound from another.

slopes, local maxima and minima, transitions, sequences, and simultaneous occurrences.

The initial processing is performed to identify specific speech elements (speech recognition) and to provide indicators of when a particular element begins and ends. These functions are performed by an array of networks, which automatically detect the occurrences of particular sound elements. The indicators of the beginning and ending of speech elements are used to segment selected elements from connected (conversational) speech. The selected elements are then analyzed to extract speaker identification characteristics.

The speaker identification procedure compares a quantized representation of the average spectrum slope for several samples of each selected speech element (phoneme) with corresponding stored data acquired during a training period. For speaker identification, the speaker whose stored sample best matches the unknown input sample is designated as the unknown speaker. Each of the 19 channels (or 14 channels) are treated separately. For a given phoneme, the slope of the spectrum in each channel is quantized into one of three classifications (-, o, +). Figure 1-3 illustrates the slope classification procedure of the vowel sound in the word "bed".

When several of the same types of phonemes are classified, a table is made of the relative frequency of occurrence of each slope classification in each channel. The tables constitute a measure of the average shape of the spectrum produced by an individual when he makes a particular sound.

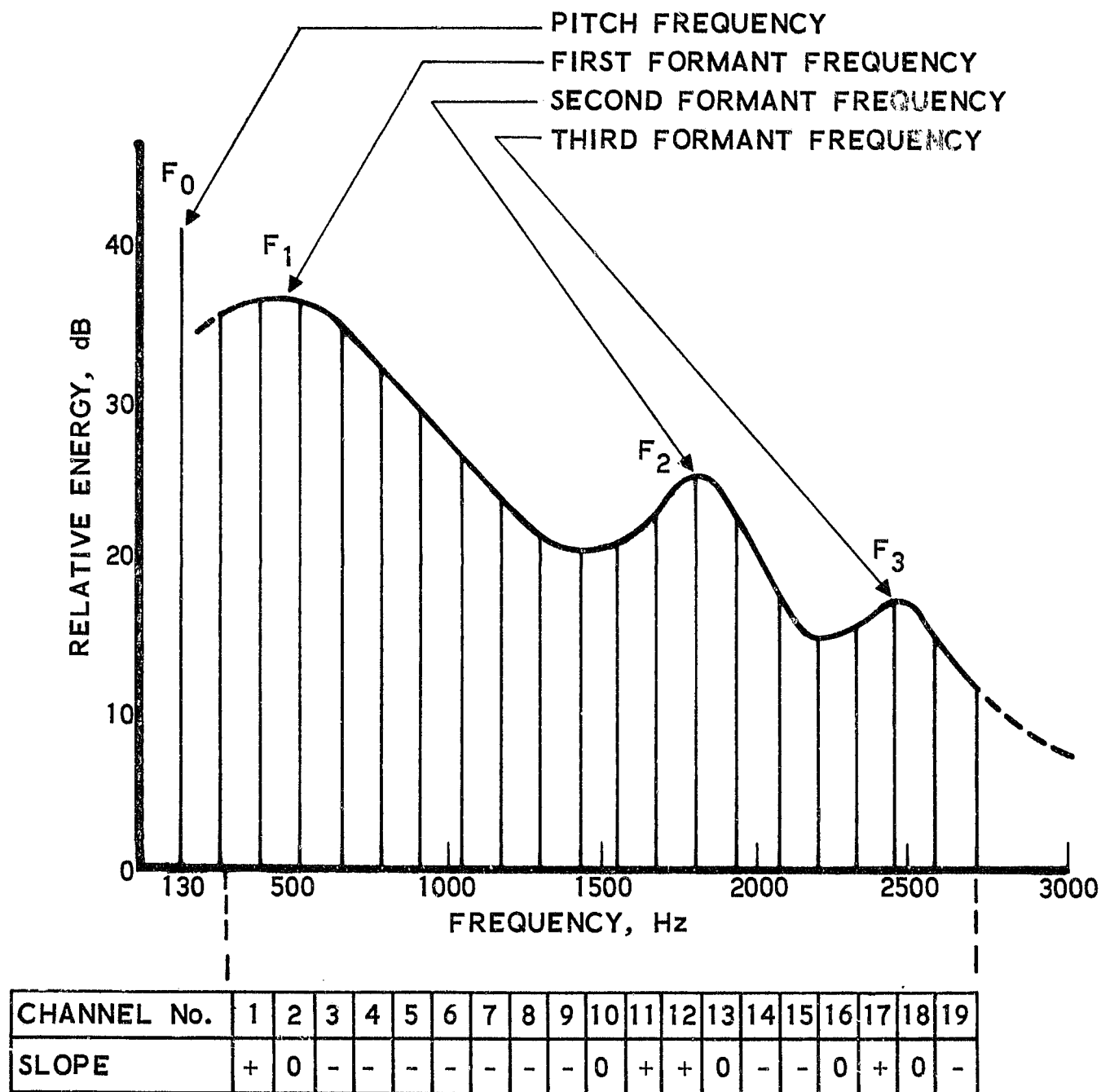


Figure 1-3. Slope Classification of Vowel Sound in the Word "Bed"

For identification of an unknown sample, the speech is processed in exactly the same manner, producing a table which measures the average shape of the spectrum for each type of sound in the unknown sample. This table will generally be based on fewer occurrences of each phoneme than are employed in the training phase. A single occurrence can be used, but better accuracy results if the unknown sample provides three or more occurrences of several phonemes.

To perform speaker identification by use of an individual phoneme, the sample table for that phoneme is compared to the corresponding training table for each speaker for whom tables have been stored. Each element in the sample table is subtracted from the corresponding element in the training table and the difference squared and summed over all slope classifications and channels to produce an overall comparison measure between the two tables. The speaker producing the test sample is identified as that speaker whose training sample table produces the smallest overall comparison measure. For improved accuracy, this measure may be derived for all the phonemes occurring in the unknown sample.

Tests have been performed to determine identification accuracy. Using 20 samples of each of four vowels (taken from continuous speech), relative-frequency-of-occurrence tables were prepared for each of 11 speakers. When these same samples were taken three at a time (three samples of each of the four vowels) and compared to the relative-frequency-of-occurrence tables using the procedure for speaker identification described above, 100-percent accurate classification resulted.

The same speech samples were used in a verification experiment. In this case, an acceptance threshold was set for each speaker, which would result in false rejection of each vowel one percent of the time. When the criterion for rejecting a speaker was the rejection of any one of the four vowels, the accuracy of rejection of impostors ranged from 90 to 100 percent. Each trial used three samples of each of the four vowels, generally obtained from five seconds or less of continuous speech. There were six trials for each speaker.

Experience with a large speaker population is necessary before the probability of correct identification can be clearly established. This experience has not been accumulated at this time.

b. Dialog Systems, Inc. The approach to speech processing taken by this organization is that the processing is done in the time domain rather than in the frequency domain, which is used in most of the other approaches. The time domain approach rests on the hypothesis that, during the production of a speech sound, a person's vocal tract can be characterized as a signal-carrying channel. Standard linear filtering techniques can be used to determine the parameters that define the characteristics of such channels when they are subjected to suitable test signals. In particular, if the test signal is either a short impulse or white noise, it has been shown that the channel can be characterized or identified by adjusting a linear filter to minimize the signal out of the filter-plus-channel combination.⁵

When a person is speaking, the vocal chord impulses and the motion of air during unvoiced speech are ready-made impulse and white noise test signals. Since the particular speech element being spoken by an individual is determined by the shape of his vocal tract, the determination of this shape by means of suitable linear filtering techniques will allow automatic recognition of the speech element.

The basic implementation of the technique utilizes time delay elements, which provide a linear representation of the speech signal, delayed by a fixed time interval. The individual delayed outputs are weighted by appropriate constant values and combined. A particular set of weighting constants is derived for each particular speech element to be recognized. The summed weighted outputs are fed to a decision-making unit, which identifies a speech element being received at a particular time. Since variations in a speech signal can be caused either by changes in the spoken speech element or by changes in the speaker, it is theoretically possible to utilize the technique for speaker identification as well as for speech element recognition.

The equipment currently being marketed by Dialog Systems, Inc. is specifically intended for speech recognition applications, and no capabilities for speaker identification or verification are described.⁶ The system operates by selecting approximately 12 short segments from each word. Each of these segments will fall into one of about 30 phonetic categories. After

selecting the speech segments, the word recognition process is completed by feeding the speech data to a processor. For each vocabulary word, the processor computes the relative probability that the unknown signal is a sample of that word. Final choice may be the word with the highest probability score; if every word has a low probability, however, the system rejects the input and requests a repeat.

Users must pause at the end of each word to inform the system that one word is finished and that the next is about to begin. The system can accept up to 1000 different telephone messages. It repeats the message and asks for verification. Lack of verification causes the system to modify the recognition algorithms so that the system is adjusted to the speaker's particular voice characteristic in subsequent messages. By this means, the error rate in word recognition can be reduced to 0.1 percent. Preliminary tests were performed to test a prototype system in the speaker identification function. The results for this application, however, were inconclusive.

c. Texas Instruments Corp. A system for speaker verification has been developed for use by the military in controlling access to bases and other sensitive installations. The equipment accepts predetermined phrases and compares voice characteristics derived from the reference phrases stored in the computer with equivalent characteristics from utterances by individuals requesting access. For each phoneme uttered by each speaker, vectors representing gain and pitch contours are derived.

Each speaker is represented in the system by a set of reference vectors derived from repeated utterances of the prescribed phrases. A test (acceptance or rejection of a speaker) consists of taking the test utterances (reduced to a set of vectors) and comparing them to each of the reference sets by calculating a numerical similarity measure. A threshold value for the similarity measure is chosen so that the system vulnerability (acceptance of unauthorized speakers) will be acceptably low, while the erroneous rejection of authorized speakers will not be unduly large.

Preliminary tests have shown that the use of voice characteristics derived from signal amplitude and pitch frequency produce vulnerabilities to the system when confronted by skilled mimics. This discovery led to the use of supplementary features derived from second formant frequency characteristics. Tests are currently underway to evaluate the system with use of these supplementary features.

3. Hand geometry approaches. Precise measurements made of anatomical features have been recognized as valuable for identification ever since Alphonse Bertillon created the system of anthropomorphic measurements for criminal identification. Since the human hand has an intricate shape, is readily accessible, and changes little with time; it is a logical appendage to use in an identification system based upon anatomical measurements.

The system, developed by Identimation Corp., utilizes an optical scanner to make measurements of an individual's hand dimensions. The user's hand is placed in a fixed position against the equipment's scanning

window. Measurements are made of the length of each finger, of the curvature of each fingertip and of the skin translucence between certain of the fingers. The coded measurements are recorded on a machine-readable plastic card that is carried by the user. The terminal device compares the measurements recorded on the card with the dimensions of the hand currently on the scanner surface. If the measurements match to within some predetermined threshold, the individual is accepted.

A drawback to the system is the fact that many of the measurements are highly correlated (i.e., the relative sizes of the fingers on a given hand are generally the same for all individuals). In order to achieve discrimination between a large number of users, the system must make very precise measurements. These measurements are subject to variation due to slight changes in the finger positioning on the scanner and the amount of pressure used by the individual. The equipment is simple, inexpensive, and easy to use, however, and applications have been found in controlling access to areas where occasional errors are not deemed critical.

4. Handwriting scanner approaches. The theory upon which handwriting is used for personal identification is that every time a person writes, he automatically and subconsciously stamps his individuality in his writing. Through a careful analysis and interpretation of the individual and class characteristics, it is usually possible to determine whether two samples of written text were in fact written by the same person.

At least two methods for automatic identification by means of handwriting analysis are currently under development. One method employs

pattern analysis, whereby handwriting samples are microfilmed and processed so that patterns characteristic of the particular writing style are produced for each sample. Specific points of comparison are extracted from the patterns, and identifications are performed by matching the corresponding points of different patterns. This method has not been implemented for real-time use.

A second method employs a form of signature analysis. In this case, a time-varying signal is derived by measuring the pressure exerted on the writing instrument when an individual writes his signature. The signal used for identification, therefore, is not derived from the signature pattern, but from the way in which the pattern was made. This concept would be invulnerable to even an expert forger, unless the forger were able to duplicate the manner in which the signature was written. This system is currently undergoing development for use in a military base and installation security system.

C. Semi-Automatic Speaker Identification System (Current Implementation)

The Semi-Automatic Speaker Identification System developed by Rockwell International⁷, supported by the Law Enforcement Assistance Administration, utilizes a general-purpose computer coupled with data processing and pattern recognition algorithms. By a combination of operator and computer functions, the parts of the speech sample that best contribute to speaker discrimination are selected and compared with other samples. The selected phoneme types processed by the system are listed in Table 1-2. On the basis of these comparisons, the computer can measure the degree of

Table 1-2. Selected Phoneme Types

Alphaphonetic Symbol	Class	Example
MX	Nasal	<u>m</u> oon
NX	Nasal	<u>n</u> o
NG	Nasal	sin <u>g</u>
EE	Vowel	<u>e</u> ve
IX	Vowel	<u>i</u> t
EH	Vowel	me <u>t</u>
AH	Vowel	<u>a</u> sk
AA	Vowel	<u>f</u> ather
AW	Vowel	<u>a</u> ll
UX	Vowel	pu <u>t</u>
UU	Vowel	bo <u>o</u> t
UH	Vowel	<u>u</u> p
ER	Vowel	bi <u>r</u> d

similarity between a criminal sample (e. g., from a bomb threat recording) and a sample from a suspect.

Figure 1-4 illustrates the overall operation of the system. Criminal speech samples from police station monitors, covert recordings, or authorized wire taps are processed and stored on digital magnetic tape. In the processing operation, specific phonetic events, which have been found to be good discriminators of speaker identity, are selected and labeled. When a suspect sample is obtained, the same phonetic events are selected for processing. In the comparison phase, each selected event from the criminal sample is compared with a similar event from a suspect sample. The points of comparison are well defined and yield quantitative results. The system is able to generate accurate and objective results on a repeatable basis.

Prior to using the equipment, the operator will listen to and write down the words spoken in the speech sample. He will then prepare an alphaphonetic transcription, which separates the words into their phonetic parts. Figure 1-5 presents an example of such a transcription. As shown in the figure, the word "six" is transcribed as /SX IX KX SX/, where each pair of letters in the transcription denotes a particular phonetic event. By examining the transcription, the operator will identify those phonetic events useful for comparison purposes. In the case of the word "six", the IX phonetic event is likely to be selected for comparison because as shown in Table 1-2 it is one of the selected phoneme types. In general, phonetic events representing vowel or nasal sounds have been found to be most useful for purposes of speaker identification.

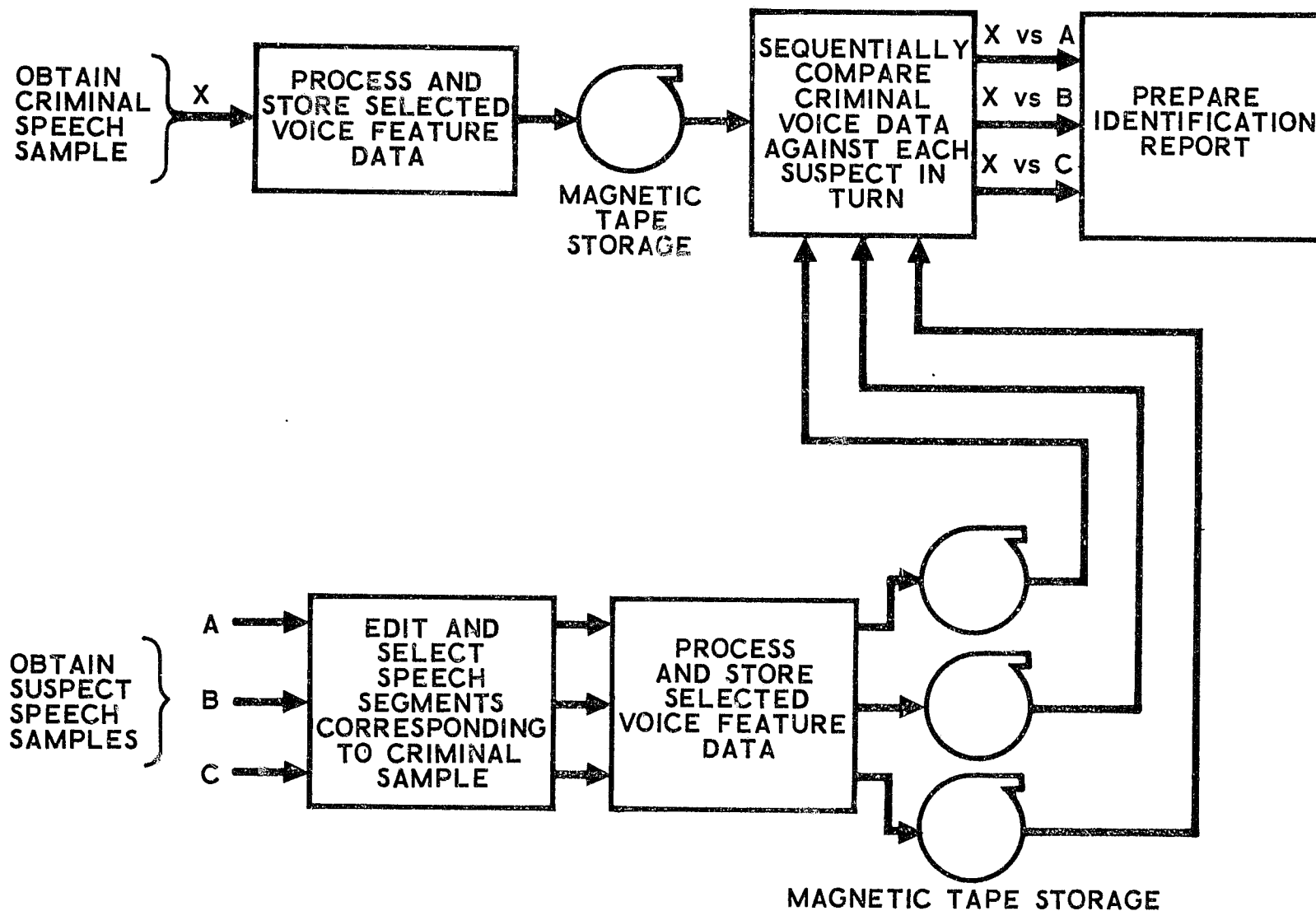


Figure 1-4. System Operation

ORIGINAL TEXT

HAVE

THE

MONEY

READY

BY

SIX

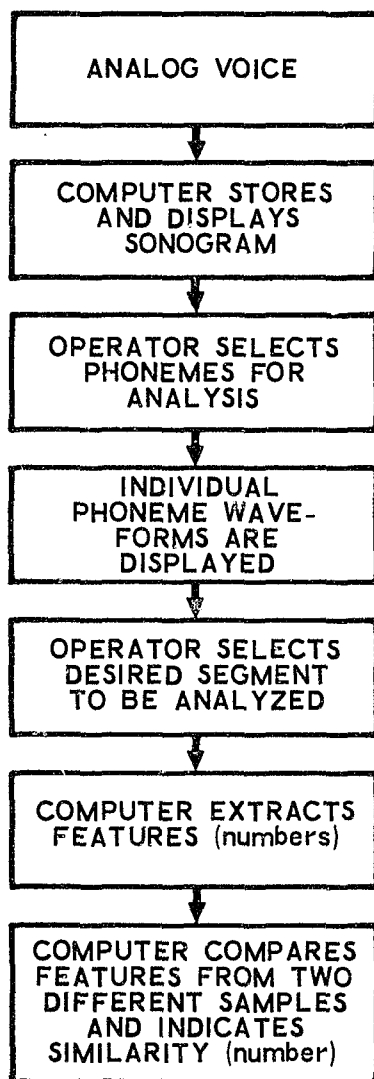
ALPHAPHONETIC TRANSCRIPTION

| HX AH VX | DH UH | MX UH NX EE | RX EH DX IX | BX AA IX | SX IX KX SX |

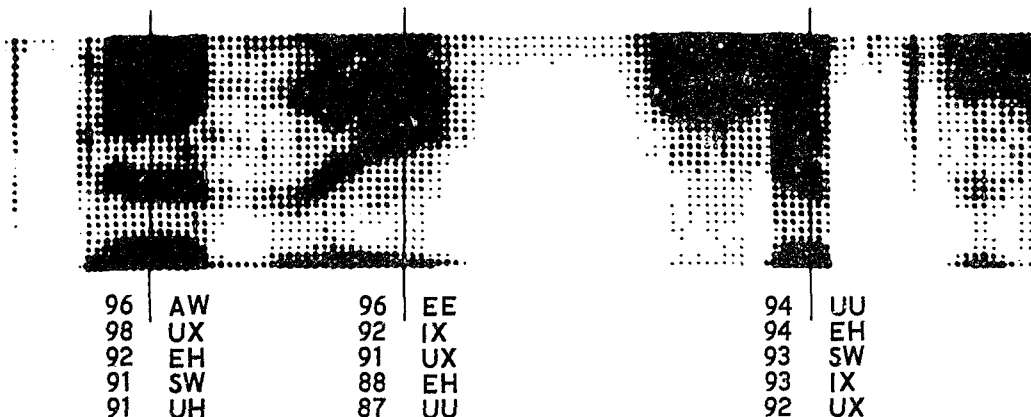
Figure 1-5. Example of Alphaphonetic Transcription

The operator first selects the phonetic events for computer analysis, then inputs appropriate speech segments for processing. When the operator is satisfied that the speech data have been correctly entered into the system, he then directs the system to transform the data into a representation showing the frequency distribution of the speech signal for a given time segment. This representation is called a sound spectrogram or sonogram. The operator can cause the sonogram to be displayed on the system graphics terminal. Each screen full, or frame, of the display represents 1.1 seconds of speech data and, in general, will contain from one to four phonetic events useful in the comparison. The upper portion of Figure 1-6 illustrates a typical sonogram display and is designated as the macrophase display.

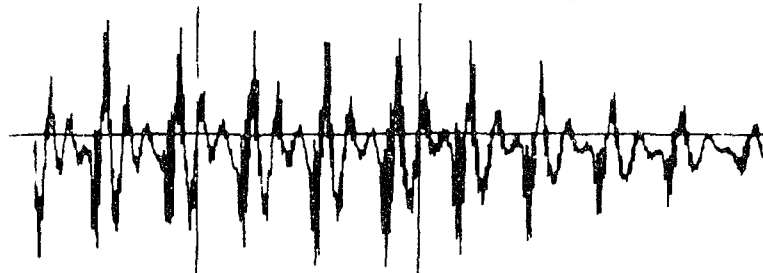
During the macrophase labeling procedure, the operator identifies and labels on the interactive terminal those phonetic events to be used in the speech sample comparison. The method for pointing to and identifying events is an interactive graphic cursor, or electronic crosshair, superimposed over the sonogram, which the operator positions by a thumbwheel control. Through keyboard input, the operator controls audio playback of sections of speech centered on the cursor. A further aid in identifying and labeling phonetic events is the capability to compare selected segments of the sonogram against a reference inventory. Upon operator command, the computer displays the alphaphonetic names of the five phonetic types whose spectra correlate best with the spectrum of the speech segment pointed to by the cursor. The correlation values are also shown, as numbers between



MACROPHASE DISPLAY



MICROPHASE DISPLAY

ARRAY OF NUMBERS

POWER AMPLITUDES, FORMANT FREQUENCIES, ETC
UP TO 30 NUMBERS FOR EACH PHONEME

SINGLE NUMBER

INDICATES LIKELIHOOD THAT BOTH SPEECH SAMPLES
WERE UTTERED BY SAME PERSON

Figure 1-6. Computer-Aided Voice Comparison

0 and 99. Figure 1-6 shows three phonetic events under consideration for labeling.

Since individual phonetic event characteristics have been found to be dramatically affected by the phonemes adjacent to the target event, the target event and the two adjacent events are labeled and subsequently used in the event comparison. The three events thus labeled are referred to as a phonetic triad. Tentative acceptance of an event is made by labeling and numbering the phonetic triad in which the desired event is centered. The operator will label all phonetic events of interest in a given sonogram frame. After he has finished the macrophase for a sonogram frame, he signals the fact to the computer through the keyboard.

The computer then automatically proceeds to the microphase of labeling. In the microphase, a 100-millisecond segment of the speech waveform is displayed for each of the events labeled in the macrophase. A typical microphase display is shown in Figure 1-6. In the microphase, the operator must use the graphic cursor to mark off three consecutive pitch periods of the speech waveform for each selected event. This is required to ensure that an adequate signal sample is used in the spectral analysis calculation that is subsequently performed.

After sequencing through the microphase display for each macrophase selection, the system returns to the macrophase and the operator may label another sonogram frame. The alternation between the two phases of labeling continues until the operator has labeled every event of interest in the speech sample.

After labeling, the computer proceeds to compute the measurements or features* on each labeled event, which will be used for comparison. For each of the 13 event types allowed in the prototype system, there is a unique set of 30 features. When the features have been calculated, the voice sample can be compared with any other voice sample similarly processed.

The detailed comparison process is diagrammed in Figure 1-7. Each speech sample goes through the same series of steps, in which the sample is digitized, sonograms and other displays are generated, and phonetic events are selected. For the speech sample of speaker A, the selected events could be designated 1A, 2A, 3A, etc.; each event will produce a set of 30 features. Event 1A will thus produce features 1A1, 1A2, 1A3...1A30, as shown. Event 1B from speech sample B will likewise produce feature set 1B1, 1B3, 1B3...1B30. The two feature sets are combined in a manner to produce a distance measure set. The algorithm used to derive the distance measure was optimized so that the widest separation between different speakers is achieved, while maintaining the smallest distance between different utterances by the same speaker. A distance measure is obtained for each pair of selected phonetic event triads in the sample. Only like events are compared. Finally, the various distance measures are combined to arrive at an overall similarity measure for the two samples. As before, the method of combination is selected to maximize the system's speaker discrimination capabilities.

*Features are numerical values for specific signal properties such as power amplitudes at certain frequencies, frequencies of specific formants, spectrum slopes, etc.

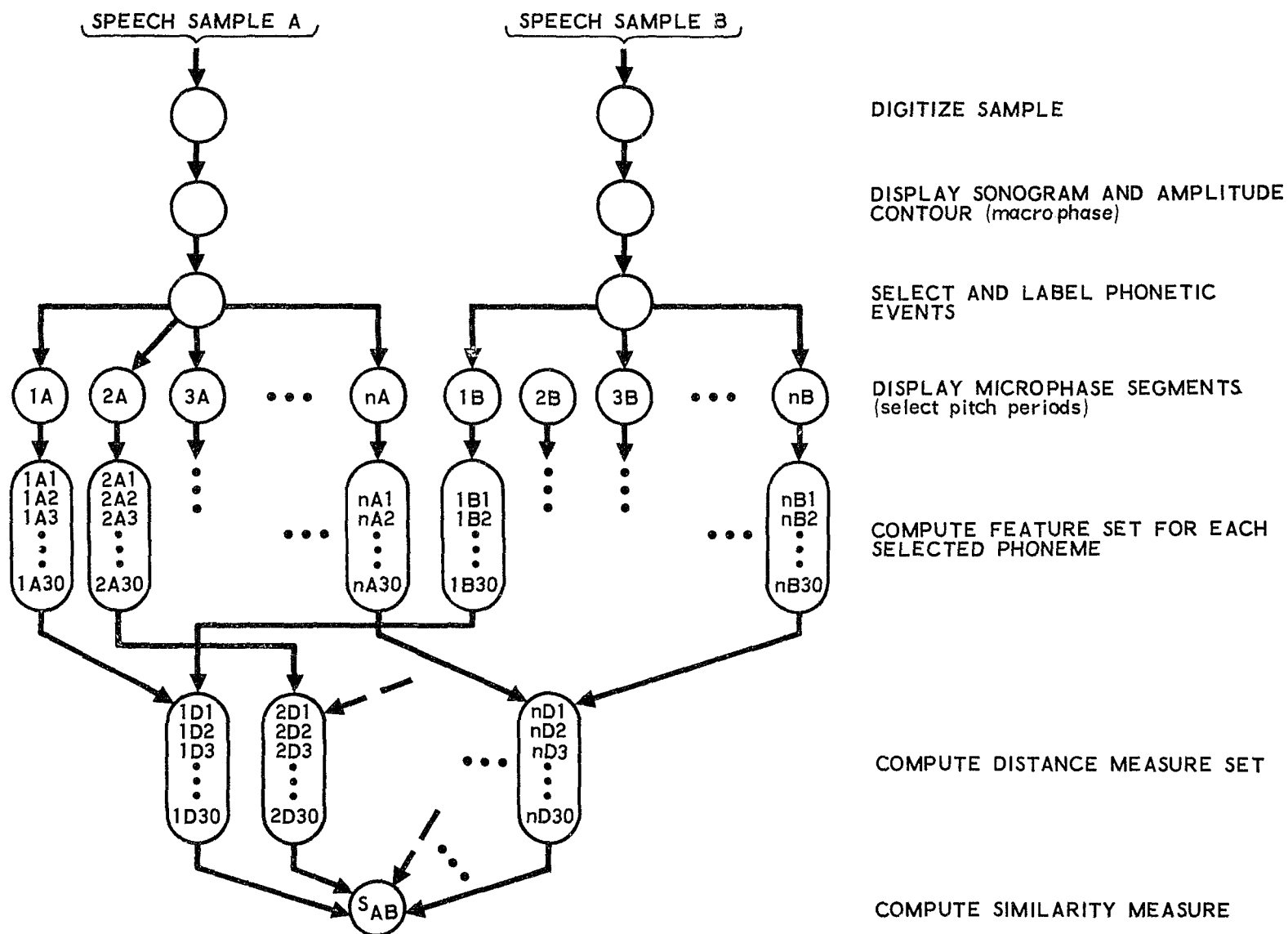


Figure 1-7. Semi-Automatic Speaker Identification System Process

In order to interpret the meaning of the similarity measure, the operator is supplied with a set of statistical performance data, which tabulate the similarity measures obtained when comparisons were made of speech samples from a representative sample of speakers. The comparisons were made between utterances made by the same person on different occasions (to obtain intraspeaker similarity measures) and between utterances made by different speakers (to obtain interspeaker similarity measures). Since the numerical values of the similarity measures obtained from voice sample comparisons are strongly dependent upon the types of phonetic events in the samples, the comparisons were made independently for every possible combination of selected phonetic events that could occur in a speech sample. The operator therefore consults the specific table of performance statistics corresponding to the set of events occurring in the speech samples being compared. The performance statistics indicate that, when 10 phonetic event categories are used in the comparison, use of the system would result in correct identification 97 percent of the time.

D. Modifications to Current Implementation

The Semi-Automatic Speaker Identification System, as currently implemented in its brassboard form, performs a rigorous and detailed analysis and subsequent comparison of speech samples. The algorithms developed for the system could be applied, with modification, to the problem of speaker identification in a non-forensic environment (e.g., speaker verification). In essence, the required modifications would be aimed toward

design of a fully automatic system independent of operator control or influence. These modifications would be made in the categories of speech recognition, phoneme partitioning, and decision-making procedures. In order to ensure that the basic comparison technique is maintained (i.e., the comparison of like phonetic events), some method is required to automatically detect and label specific events from the input speech signal. This modification would be facilitated by the presence of the correlation reference inventory in the brassboard system.

The correlation references are one of the labeling aids at the disposal of the operator during the macrophase. On request, a display of the top five correlations between the current speech element under consideration (as indicated by the position of the display cursor) and an inventory of reference event spectra are presented to the operator.⁷

In an automatic system, the correlation process would be performed on each segment of the speech signal. For each segment, a decision would be made relative to the type of phonetic event under consideration. Certain segments would be located at the transition between events and would therefore not be labeled as recognized events. Their position and relative correlation with the reference events would be noted, however, by the algorithm and used in the subsequent phoneme partitioning procedure.

One of the key features of the brassboard system is the manner in which selected portions of the speech waveform are isolated and analyzed. This isolation contributes a high degree of repeatability to the subsequent detailed analysis of the isolated waveform and consequently improves system accuracy.

In an automatic system, this feature could be retained by the addition of several processing steps. First, the speech segment with the highest correlation with a member of the reference inventory would be designated as the midpoint of the desired phoneme waveform to be isolated. Second, a cepstral* analysis would be performed to derive the basic pitch frequency for the designated segment (and possibly the adjacent segments) under consideration. Finally, the knowledge of the pitch frequency would be used to mask off a time interval (centered on the target segment) equal to three pitch periods. Software currently implemented in the system would be used to "fine-tune" the selected interval so that it begins and ends on zero crossings of the waveform. In this way, an isolated portion of the speech waveform could be partitioned for detailed analysis.

The last major area for modification of the brassboard system would involve automation of the decision-making process. It has been found in the brassboard development efforts that the values of the similarity measures obtained from voice sample comparisons are strongly dependent upon the types of phonetic events in the samples. In making a decision on the identity of an individual, the system must refer to the specific set of events used in the comparison process. Based on the brassboard system performance test

*Cepstral analysis is performed by first taking the Fourier Transform of the speech signal, then taking the logarithm of the resulting transform and finally taking the Fourier Transform of the logarithm. The resultant signal contains a greatly amplified version of the fundamental or pitch frequency.

data associated with that specific set of events, the similarity measure will be used to ascertain the probable false acceptance and false rejection errors associated with the similarity measure. At this point, the relative costs associated with the probable errors must be factored into the calculation to ensure that predetermined acceptable risks are not exceeded when a decision is made.

CHAPTER II. POTENTIAL APPLICATION AREAS

A. Identification for Criminal Apprehension

Much evidence exists regarding the need for reliable and valid procedures and equipment for speaker identification to apprehend and convict criminals. The Bell Telephone System verified that, during the first six months of 1969, 306,103 obscene or harassing telephone calls were reported in the United States. Since many such calls are unreported, the total number probably exceeds one million per year. During this same period, over 18,000 threatening calls (i. e., involving bribery, extortion, kidnap, etc.) were reported. It is also estimated that as many as 400,000 bomb threats are made to public buildings each year, the majority by telephone.

The telephone is also used to (1) report false fire alarms, (2) lure police to ambushes, (3) conduct illegal gambling operations, (4) organize the narcotics traffic, and (5) plan a wide variety of other criminal activities. Often the call itself is the crime, as in a bomb threat or extortion case. Increasingly, police and prosecutors are relying on the identification of criminals through the matching of their voice samples.

The interest of law enforcement investigators is not limited to telephone conversations. Undercover personnel utilize concealed voice recorders or transmitters (to remote receivers and recorders) in many narcotics, gambling, merchandise fencing, and vice investigations. The use of such recordings would be considerably enhanced if an effective means were available to identify accurately the speakers on these recordings. This application would

be particularly effective for narcotics investigations, where the use of facial disguises and code words by criminals make it difficult for police to determine the structure and the key members of a narcotics distribution and sales organization. By identifying individuals from voice samples recorded by undercover agents, investigators would gain valuable information pertinent to the apparatus of a criminal operation. There are indications that detailed analysis of a voice sample will provide information about a speaker unavailable from other identification techniques (such as fingerprints). This information could include: ethnic or geographic origin, physical size, age, state of health, emotional state, and other factors possibly useful in an investigation.

1. Suspect identification. The principal impact of speaker identification techniques will be improvement in the effectiveness of police investigative techniques in several crime categories. These categories include: bomb threats, false alarms, extortion, receiving stolen goods, gambling, and narcotics investigations. In conjunction with other technology, such as voice-actuated recorders, the system also could be used in armed robbery, vandalism, and burglary cases. The effects in investigative applications would be due to the improved ability to focus investigations on the most promising suspects. Suspects whose voice samples are dissimilar to the criminal sample would be eliminated from further consideration.

In a previous Aerospace study,⁸ it was estimated that approximately 1500 cases per year occurred in which suspect identification was established by use of voice identification techniques. The same study

estimated the potential application for this technique, assuming that certain of the drawbacks to conventional methods could be overcome. Table 2-1 presents the list of potential application areas uncovered in this study.

Table 2-2 presents a summary of the estimated caseload for a potential speaker identification system, obtained from representatives of the law enforcement agencies in the five cities shown in the table and surveyed in the study. Members of management (chief of police or staff), records division, and heads of each of the six crime-type investigative units were asked to estimate the number of cases in which use of voice identification could have a significant effect.

An additional Aerospace investigation concluded, based on economical considerations, that burglary and robbery in commercial establishments are prime areas for applying voice recording/identification techniques.⁹ The details of this investigation are presented in Appendix A. The investigation showed that if the probability of a successful single burglary can be reduced from 0.97 (current accepted value) to 0.95, the total number of burglars apprehended (assuming 25 burglaries per year per burglar) can be increased by about 35 percent.

At this time, there has been no evaluation of the specific nature of the crimes of burglary and robbery to provide answers to certain basic questions, such as:

- Is there enough conversion during the crime to obtain a recording?

Table 2-1. Potential Speaker Identification Applications
in Local Law Enforcement Agencies

MAJOR USE: Personal Identification Verification Through Voice Comparison

- Narcotics Stakeout
- Narcotics Buy
- Narcotics Soliciting Buy
- Prostitution Stakeout
- Prostitution Buy
- Prostitution Soliciting Buy
- Burglary Stakeout
- Armed Robbery Stakeout
- Assault Stakeout
- Kidnap Threat Monitor
- Kidnap Payoff
- Murder Contract Buy
- Murder Contract Monitor
- Vandalism Stakeout
- Vandalism Buy
- Secure Area Access
 - General Area
 - Police Computer Room (Records, etc.)
 - Communications Room (Command, Control, etc.)
 - Confession
 - Interview Interrogation
 - Bomb Threats
 - Bomb Extortion Payoff
 - Lewd/Obscene Telephone Calls
 - Suspicious Telephone Calls
 - Harassment Telephone Calls
 - Merchandise Fence Monitor
 - Merchandise Fence Payment
 - Extortion By Wire
 - Extortion Payoff
 - False Fire Alarm
 - Arson Threats
 - Arson Payoff Monitor
 - False Police Alarm
 - Riot/Insurrection Monitor
 - General Investigative Tool
 - Court Authorized Wire Taps
 - Court Authorized Eavesdropping
 - Attorney of Record Verification
 - Insurance Investigator/Other Verification
 - Control of Inmate Egress/Ingress - Jail/Other

Table 2-2. Summary of Applicable Cases in 1973

City Crime Type	Washington	New Orleans	St. Louis	Tulsa	Santa Ana
Bomb Threats	75	210	25	60	7
Kidnapping	64	1	-	-	-
Narcotics	150	1300	700	400	750
Disturbing Telephone Calls	-	171	-	432	240
Gambling	1000	270	350	63	10
Prostitution	550	100	350	35	100
Total	2039	2052	1425	990	1107

- Would the sound quality of the recording be sufficient for voice identification analysis?
- What is the optimum type of microphone placement?
- Would the voice record evidence be useful in the apprehension and prosecution of the criminal?

These questions cannot presently be answered quantitatively.

Discussions with representatives of police departments and the security industry, however, have revealed indications that improved voice analysis techniques could have significant impact on the apprehension and conviction of certain criminals. There is growing interest in police crime laboratories

and identification bureaus for developing better methods for recording the identify of arrestees. New approaches include color photographs and video tape recordings of arrested persons, in addition to the conventional mug shots. Also under consideration is the possible use of voice recordings to be retained in arrest files. Appendix B describes a possible method of obtaining useful voice samples along with an example of possible sample material. There is also an increasing awareness and employment of the concept of installing sensitive sound detectors in protected premises. Such equipment is connected to a central station by a dedicated leased telephone line. Any sonic disturbance transmitted to the central station signals the monitor to listen to the sounds, identify the cause of the disturbance, and take appropriate action. The chief reason for monitoring the sounds is to reduce the incidence of false alarms. However, since the sound disturbances are routinely recorded at the central station, the system provides a ready source of potential evidence if conversation occurs during the crime. This is generally the case for the crimes of robbery and vandalism.

The typical cost for security systems of this type is about \$3000 for a basic configuration, with a \$30 to \$40 monthly monitoring fee. Reduced costs could be expected for a dedicated voice recording system for burglary apprehension. Such a recorder could be set up to record and automatically erase in each 24-hour period. Economic considerations indicate that commercial establishments provide the greatest opportunity for implementation of such systems.

2. Criminal evidence. At the present time, relatively few cases involving voice identification result in courtroom testimony with such evidence presented at a trial. In the past three years, about 25 trial cases per year have involved voiceprint evidence. Of those cases that have been brought to court, approximately 50 percent have resulted in convictions, and 50 percent in other dispositions. In nearly all convictions there were no expert witnesses called by the defense to challenge the voiceprint evidence. These results imply the inherent weakness involved in the use of voiceprints for forensic applications and their vulnerability to challenge from knowledgeable witnesses.

The use of voiceprint evidence in a particular court case is affected by a number of considerations. Among these are: (1) the relative seriousness of the case, (2) the quality of the evidence (recorded voice samples), (3) the amount and quality of other evidence, and (4) the availability of a qualified examiner to testify. Other factors are the problems related to the admissibility of such evidence and the expense involved in its preparation.

One of the major problems in assessing the reliability of voice identification is that there is scant literature available to confirm or reject the fundamental premise upon which the technique is founded, namely, voice uniqueness. Nor are there available at this time any authoritative data on the effect on voice spectrograms of (1) nasal or oral surgical operation, (2) muffling of the voice, (3) mimicking, (4) use of dentures, tooth extractions, and for example, (5) effects of illness, colds, puberty, external influences,

background noise, emotional state, or (6) effect on spectrograms of isolated cue words of the preceding and following sounds in a sentence. Such data would appear to be a necessary prerequisite to scientific acceptance based on proven reliability.¹⁰

The identification of an individual by his voice, when made aurally and not by voiceprint, has long been held admissible in criminal trials. Aural voice identification has been held admissible by analogy to identifying techniques that involve bodily or physical examinations. Compelled aural voice identification is not protected by the self-incrimination privilege, even when a suspect is required to repeat the same words that a witness has indicated were used by the perpetrator of the crime.¹¹

The response of various state courts to the use of spectrographic identification has been uneven and contradictory. For example, the New Jersey State Supreme Court and the California Court of Appeals have ruled that voiceprint evidence is not admissible due to lack of scientific acceptance, while the Minnesota State Supreme Court has ruled that the technique may be used. A recent ruling has involved the U.S. Court of Appeals in Washington, D.C. In this case, Judge Carl McGowan ruled that voiceprint evidence is not sufficiently accepted by the scientific community to form a basis for a jury's determination of guilt or innocence.

The principal objections to the use of voice identification in court proceedings are: (1) the lack of data regarding the invariance of speech, (2) inadequacy of identification experiments conducted thus far,

(3) contradictions in results of experiments conducted by different investigators, and (4) the poor quality of the recordings used in the field for identification purposes.

As far as the criminal recordings are concerned, poor quality can be expected. Although some improvements are possible, the factors affecting the recorded signal quality for the most part are uncontrollable. An equally significant problem associated with the use of voiceprints results from poor quality of the exemplar material (i.e., the suspect recordings). These recordings are usually made by police interrogators, with no consideration given to the quality of the recorded speech or the presence of background noise, reverberations, etc. To minimize this problem, at least one agency has begun to obtain exemplars over the telephone. The situation is often crucial to an identification process, since the law requires a suspect to provide only one exemplar. If the recording process is faulty, this item of evidence could be irretrievably lost.

Spectrographic voice identification offers great potential as a reliable means of establishing identity, provided the claims by its proponents can be substantiated by reliable, unbiased research of the type demanded by its critics. At the present time, admissibility appears to hinge on whether the test meets the "general acceptance" standard for novel scientific methods. This acceptance has not yet occurred either for the principle of voice uniqueness or for the reliability of the art of comparing speech spectrograms.¹⁰ The results of the previous study⁸ indicated that use of voice identification techniques could produce savings in investigative and judicial areas and that

a typical large city police department such as Washington, D.C., New Orleans, or Tulsa could approximate cost savings of \$300,000 per year. (These savings are based on the processing of about 2000 cases each year.) The cost of such processing would be between \$60,000 and \$70,000 per year. On a simple cost comparison, therefore, the use of such techniques is worthwhile. However, the real benefit would result from increased effectiveness of law enforcement activities and deterrence of crimes that cause fear, anguish and harm to the general public.

B. Personal Identification for Crime Prevention

In a modern technological society, there is a continual evolution in the methods used by criminals, as well as in the types of victims and perpetrators and the types of losses. This evolution is in response to new developments in technology and the manner in which these new developments alter the traditional methods of conducting business, governmental, and other affairs. Changes are particularly evident in the areas of transportation, communication, credit cards, and firearms control. The growing use of falsification in the traditional procedures for establishing identity has become a serious problem in the control of crime related to narcotics traffic, confidence swindles, terrorism, and espionage.¹² The lack of effective procedures and safeguards in the issuance of birth certificates, driver's licenses, and other documents allows individuals to create multiple identities that can be used in illegal activities. These activities include obtaining government payments by fraud and creating false identities for narcotics distributors, illegal aliens,

espionage agents, and confidence men. There is a continual challenge for new technology to combat such elements.

1. Computer access security. The recent widely publicized attempt to defraud the Los Angeles City Treasury is one example of a continuing assault on the financial assets of individuals and institutions through manipulation of the computer systems that record and control these assets. The involvement of organized crime in this particular case lends credence to the view that this type of threat is part of a national pattern that can be expected to recur.¹³ A recent report on computer abuse¹⁴ contained the following observation:

Computer technology and data communication technology are subject to increasing abuse as they penetrate into sensitive areas of human activity. The exchange, transfer, and recording of wealth and information have traditionally been performed manually, using the media of paper, films, postal service, telephone, and speech and validated by handshakes, handwritten signatures, affixed seals, and witnessing. A transition from these methods to the use of computers and data communications involving the electronic/magnetic media is now taking place in the post-industrial age.

....(Physical theft of negotiable securities) will be an obsolete crime in a few years. Negotiable securities will be stored magnetically and electrically as data inside computers and transmitted over communication circuits from one computer to another.

Perpetrators of security thefts will use the skills, knowledge, and access associated with computer and data communications technology and will not be dealing with as simple a victim as the (traditional) messenger.

Recent studies indicate that about 5 percent of the present United States work force of 84 million work directly with computers, and another 15 percent work indirectly with them. In 1971 440,000 persons were transcribing data into computer-readable form on punch cards and magnetic media; 360,000 persons were doing systems analysis and writing programs for computers; and 200,000 others operated computers and handled data.¹⁵

All but the smallest business and government agency own, lease, or use computer services. Most large organizations have decided that they can function for only a few hours or few days at most without the correct functioning of their computers. At least 60 percent of all banks are automated and would be unable to function efficiently unless their demand deposit accounts were successfully processed on computers. It is estimated that computer manufacturing, data communications, and operation of computer systems will collectively represent 14 percent of the gross national product by 1980. The bulk of that equipment will be used directly in the processing of the wealth and information of our society.

Financial losses in computer-related crimes occur in certain types of crime areas. The 1967 Report of the President's Commission on

Law Enforcement and Administration of Justice estimated the following annual losses (in millions of dollars):

Embezzlement	200
Fraud	1350
Tax Fraud	100
Forgery	80

By comparison, in 65 reported cases of computer abuse over a 9.5-year period, losses averaged about \$1 million per year. A consensus of 30 experts recently indicated that losses, injuries, and damage directly associated with computers will exceed \$2 billion annually by 1982.¹⁶ It is apparent that this increase will be due in part to a transfer of traditional crime activity (embezzlement, fraud, etc.) into the computer field.

The theft of financial assets is not the only type of crime involving computer access. Unauthorized entry into and use of computer files containing confidential information have caused increasing concern among private and government groups. Abuse of personal privacy can be a byproduct of any kind of computer abuse where data associated with individual identity are compromised. Recognition of the problem, i.e., types of data that may be stored, legal rights of due process for full disclosure, and disputing and monitoring the uses of personal data, have received great attention, resulting in a high level of legislative activity. However, little action has been taken in the consideration of personal privacy issues in other computer abuses, such as theft, vandalism, and fraud, and possible exposure of private data - as evidenced in the minimal prosecution of perpetrators of such abuses.

Continuing support of research and the development of secure computer systems may result in control of computer abuse before it becomes a catastrophic problem. In this event, a reduction of all types of crime is possible where data related to vulnerable activities are inputted to the computer for processing.

Finally, human physical security and well-being are becoming increasingly dependent on the correct and reliable functioning of computers. For example, computer use is increasing in a variety of areas: (1) computer-controlled scheduling planned for the San Francisco Bay Area Rapid Transit system and the Seattle-Tacoma airport transit system, (2) computer monitoring of patients in hospital intensive care wards, (3) traffic lights controlled by computer in San Jose, California, and other cities, (4) computer-assisted air traffic control, (5) automated airliner landing, and (6) computer-regulated electrical power and water distribution. Abuses, both intentional and through negligence, are likely to occur.

Since much of the equipment used to access computer systems consists of remote terminals that are connected to the computer system over voice-grade telephone lines, identification of terminal users through verification of the users speech characteristics is a straightforward approach to improving security. There would be little or no cost impact at the computer terminals and the flexibility of the identification technique would provide a measure of protection currently unavailable. The use of access codes to provide security has been demonstrated as ineffective in the past for a

number of reasons. In the development of new technology to perform this task, the use of voice identification technique can be expected to be a major factor.

Revolutionary change may be needed but may not be perceivable, because we are in an intense and rapid period of transition from a manual, paper-based society to an automated society. A shift of trust from people to people-produced systems or even a shift to different concepts of suspicion and trust in business transactions could require replacements of present business law and practices with new concepts and new techniques which can relate traditional, manual methods to automated applications.

2. Area access security. The protection of valuable property from theft is a serious problem requiring large expenditures of money and effort by many segments of society. An equally important problem, and one which is becoming more critical as time goes on, is the protection of potentially dangerous materials or equipment from unauthorized theft or use. The theft of explosives or weapons by terrorist organizations has become a frequent occurrence. These materials may cause disruption, serious damage and death when they are exploded in public buildings.

Another cause for concern is the increasing use by technology of many toxic substances, particularly radioactive materials, which can be stolen by knowledgeable groups and used as blackmail or to attack the population.

Each of these problems can be alleviated by the development of access control equipment to enable discrimination of those individuals who

are authorized access from those who are not. The implementation of access control equipment must involve tradeoffs among a number of interrelated factors such as: the allowable cost of equipment, equipment reliability, degree of security required, number of individuals needing (and authorized for) access, and the acceptability of the identification technique to the user. The sophistication and expense associated with voice identification make it improbable that it would find widespread application to access security in private residences and general business operations. There are specialized situations, however, where such a security system may have significant value, and two are described in the following.

a. Power generating plants. The criticality of these facilities to the public safety makes them targets for terrorist attack. This is particularly true of the new nuclear power plants. The costs of a serious reactor malfunction, whether accidental or intentional, can be enormous. An Atomic Energy Commission¹⁷ report indicated that, in the event of a major reactor accident, people could be killed at distances up to 15 miles, and injured up to 45 miles. As many as 3400 could be killed and 43,000 injured (assuming the plant was relatively distant from a medium-sized city). Property damage from radioactive contamination of land could range as high as \$7 billion, the area affected being as great as 150,000 square miles. While the results of this study have been questioned, there is general agreement that nuclear power plants constitute a real potential danger. It is estimated that by the year 1990 there will be approximately 1000 nuclear power plants operating in the United States. There also will be numerous fuel refining

and processing plants distributed throughout the country. The theft of even small quantities of concentrated nuclear fuel from these plants would constitute a serious threat to public safety.

b. Correctional institutions. Significant manpower in correctional institutions is required to conduct inmate head counts and to perform surveillance. The counts are performed four to six times a day, and often involve inmates numbering in the thousands. Surveillance is a continual problem to determine the location of inmates on the grounds and in the buildings of correctional institutions.

Present technology used in inmate accountability ranges from crude to semi-sophisticated methods. Primarily, officers perform visual body-counts while maintaining checklists with hand-carried clipboards. This visual checking is assisted in some applications (e.g., Los Angeles County Jail) by color-coded identification wristbands worn by each inmate. Closed circuit television is used in some instances to monitor inmate mobility. One such application of television utilizing recording tapes is being investigated by the MITRE Corp. under separate grant funding in conjunction with the Illinois Department of Corrections.

Recent MITRE and Aerospace experience with local and national corrections personnel indicates that an urgent need for an automated inmate accountability system exists. The experience involved technical discussions at various correctional institutions within Illinois and California.

The urgency is prompted by the overcapacity situation that exists in most prison and jails, as combined with the time-consuming process

of accountability and manual body-counting. The problem is most acute in jails because of the inherent high volume and high turnover rate (75 percent of the jail inmates are housed for 24 hours or less). A reliable system is needed to automatically account for each prisoner included in a time and position varying inmate roster at prescribed or as-required inventory times. Three levels of sophistication may be assigned to such a system: (1) to account for the inmate presence within the institutional facility, (2) to identify specific location of the inmate, and (3) to maintain a temporary record of this location. In addition, the system can provide the capability to indicate passage of an inmate through restricted areas. Inmates are often restricted to cramped quarters or crowded areas because of lack of adequate accountability.

A properly designed and implemented voice identification system could monitor inmate presence from a remote location. The transmission medium would require only telephone channel quality. This concept would have the advantage of being independent of transmitting devices carried by inmates which could be broken or transferred. Positive identification could be made at a number of locations. The low cost of terminal devices for this approach would make them attractive, particularly if they could be used to replace or supplement guard posts. The high cost of one-man guard posts (\$35,000 to \$48,000 per year for 24 hours/day operation) is a substantial element of the cost of operating minimum security facilities.

3. Commercial credit. One of the most revolutionary developments in the area of economics has been the rapid growth of credit card use by

consumers. In 1969, almost six million travel and entertainment credit cards (American Express, Carte Blanche, etc.) were in use and \$3 billion was charged to them. In the same year, over \$5 billion was charged to the 50 million bank credit cards (Bankamericard, Master Charge, etc.). It is estimated that by 1980, 25 percent of consumer credit will be transacted by means of such cards.

Mr. Thomas La Forge, Systems Consultant to the Bank of America and a member of the Bank Card Standards Committee of the American Bankers Association, commented on the need for improved personal identification techniques in the banking industry:*

Currently, the banking industry basically relies on manual procedures such as signatures comparisons and additional identification, including bank courtesy cards, driver licenses, other credit cards and plastic cards containing customer photographs. Recently, with the introduction of unattended automated tellers and terminals located at bank teller stations, "secret" numbers have been issued to bank customers for personal identification. There remains a consensus within the banking industry, however, that these methods continue to fall short of reliable personal identification.

There is a recognized need in the banking industry for techniques that enable individual identification through verification of some personal characteristics, such as signatures, voice characteristics, or fingerprints.

*Private correspondence.

The technique must be acceptable to consumers and merchants and must also be technically, operationally, and economically feasible.

In order to conform to the established practices and standards in use by the banking industry, any identification technique must meet certain constraints:

- Compatibility of any plastic cards used in the transaction with the card standards currently published by the American Bankers Association. This, in turn, could impact on equipment for automatically reading information from a card.
- Compatibility with established procedures at the point of transaction (merchant or bank). The technique should permit operation in either a manual or automatic mode with access to the bank's customer files.
- Capability for fully automated operation. The technique should allow transactions to be carried out by means of unattended teller devices, which operate either in an on-line or off-line mode with respect to the bank's computerized customer data files.

In addition to its use in credit card transactions, automatic personal identification techniques have great potential usefulness in check-cashing operations. Reported losses due to bad checks are reported to exceed \$750 million annually and are increasing at 7 percent per year. In

addition, it is estimated that up to 80 percent of all bad-check cases are unreported to the police.

Numerous systems and equipment have been devised to reduce this loss, including thumbprint recorders, photographs, and check number verifiers. No system in current use, however, provides the real-time verification of identity to prevent losses before they occur. Even with the inherent limitations, the presently used techniques have resulted in reports of significant reductions in crime from bad-check passing. There is concern among merchants regarding the acceptability of these techniques (particularly thumbprints) to consumers. The long-term results of the current techniques also have not been assessed. In addition, "bad-check artists" are often skilled professionals who may discover unanticipated vulnerabilities in these systems.

It has been estimated that the cost of "ordinary" crimes against business (i. e., burglary, robbery, vandalism, bad checks, arson, credit card fraud, and employee theft) rose to \$20 billion in 1974.¹⁸ Much of this crime may be directly attributed to the inability of current equipment and procedures to accurately and efficiently provide information regarding the identity of individuals. The development of advanced voice analysis equipment could provide an important tool in reducing the incidence in a number of these crime areas.

C. Non-Crime-Related Applications

A number of interrelated technological factors are currently in effect which can be expected to produce many applications for automatic voice identification techniques. These factors include:

- Widespread existence of low-cost, voice grade communication channels in the form of telephone networks, radio equipment, microwave, and cable.
- Growth of the use of touch-tone telephones and other digitally oriented communication devices, which facilitate direct communication between computer facilities and remote users.
- Increasing sophistication of speech recognition equipment in terms of versatility, accuracy, and speed.

The initial use of voice identification may be as a verification device in authorizing the transfer of funds. Currently in existence are systems that allow consumers to pay bills by means of a phone call to a savings institution rather than by check. The system utilizes push-button telephones for direct input to the savings institution computer or uses a manual operator if the touch-tone telephone is not available. This system is attractive to savings and loan companies, which are restricted by law from offering checking accounts to their depositors. Under the present method, some means of manual authentication is needed to verify that the talker was actually authorized to approve payment. In the future, automatic identity verification by analysis of voice characteristics will be used to ensure that the transaction was legitimate. The technology developed for this application

will be transferable to other areas in which the purchasing of goods or services or the direction of business transactions can be performed without the need for personal contact. These transactions could conceivably include the action of casting ballots in political elections over the telephone. Remote automatic identification also could be coordinated with radio communication, now used by many land mobile services, to extend credit card privileges for taxis and other transportation-associated activities.

CHAPTER III. TECHNICAL EVALUATION

In order to assess the potential uses of the semiautomatic speaker identification technique for the purposes of personal identification, voice classification and identity verification, the brassboard system performance data were analyzed and evaluated. The following discussion is based upon measured data and results of the brassboard system development effort.

A. Accuracy

The accuracy¹⁹ that must be achieved by a speaker recognition system is dependent upon the specific application of the system. In general, the amount of error that can be tolerated in a decision-making system is an inverse function of the losses or the costs that might be incurred as a result of the errors. For most practical problems, these costs are difficult to define because they tend to be indirect costs, and the occurrence of a given cost or loss will depend upon many factors often remote to the decision process itself. For an example, consider the case of a secure entry system. If an authorized person is denied entry into a facility because of a misrecognition in the automatic gate system, it would be necessary to quantify the inconvenience incurred by that individual, due to the faulty system, in order to determine the cost of the misrecognition. This would be a formidable task in most instances. Similarly, if an impostor were given access to a secure facility, the ultimate cost of this misrecognition would depend upon the nature of the facility, the intent of the impostor, and the overall vulnerability of the facility as a result of unauthorized access. Again, the quantification of these factors would be extremely difficult.

Several different types of strategies have been developed from decision theory to permit decision-making under conditions of risk.²⁰ For the most part, these strategies require the losses to be expressed in monetary terms, although other measures of utility are sometimes used. The major disadvantage of these strategies lies in the subjectiveness that usually accompanies the estimates of losses or costs assigned to specific decision-making actions. For this reason, the required accuracies of recognition systems seldom reflect specific costs; instead, they are primarily based on what can be expected of state-of-the-art technology and what can be reasonably tolerated in the way of decision-making errors. Although subjectivity still exists in these criteria, one would expect the variance in estimating permissible errors to be small as compared to the variance in estimating, say, indirect costs of misrecognition.

In the discussion on speaker recognition system performance that is to follow, emphasis is placed on the implication of demonstrated accuracies in connection with the different types of recognition problems that might be addressed. That is, rather than specify required accuracies for the wide range of speaker recognition system applications, the expected accuracies that should be achievable are discussed in light of inferences that can be deduced from state-of-the-art technology. In this case, the state-of-the-art technology that is of primary concern is that associated with the development of the brassboard Semi-Automatic Speaker Identification System.

1. Decision-making objectives and constraints. The expressions "speaker recognition" and "speaker identification" are synonymous and can be defined as the general problem of relating a voice signal to the person who uttered it. The term "speaker verification" refers to the problem of authenticating a single speaker. This latter problem can be considered as a

subset of the speaker recognition problem as shown by the hierarchical structure in Figure 3-1.

When the objective of the speaker identification system is to classify voice samples from m speakers in a closed set, this system can be described as an m -choice, closed-decision system. Since the set of possible speakers is limited to m known speakers and no others, the decision-making task is considered to be a closed decision. For this reason, a closed-decision system would be used where it was not possible for an impostor to input his voice signal into the recognition process. It could be used as an aid in transcribing recordings from a closed meeting or from any closed environment where the total set of possible speakers is known beforehand. It would not be applicable for unknown speaker populations.

An m -choice, open-decision system is capable of classifying an unknown voice signal as belonging to m known speakers or to some person outside the set of known speakers. In this case, provisions are made for the possibility that the unknown voice signal came from an impostor. Thus, the voice signal under test must not only be more like that of one of the m speakers than any of the rest, but it must also pass a threshold test. If it fails the threshold test, it is classified as belonging to some person outside of the set of known speakers.

A speaker verification system is a special case of the open-decision recognition problem. In that case, the unknown voice signal is classified as belonging to the claimed speaker or to some other person outside of the set - a set consisting of only a single person for each

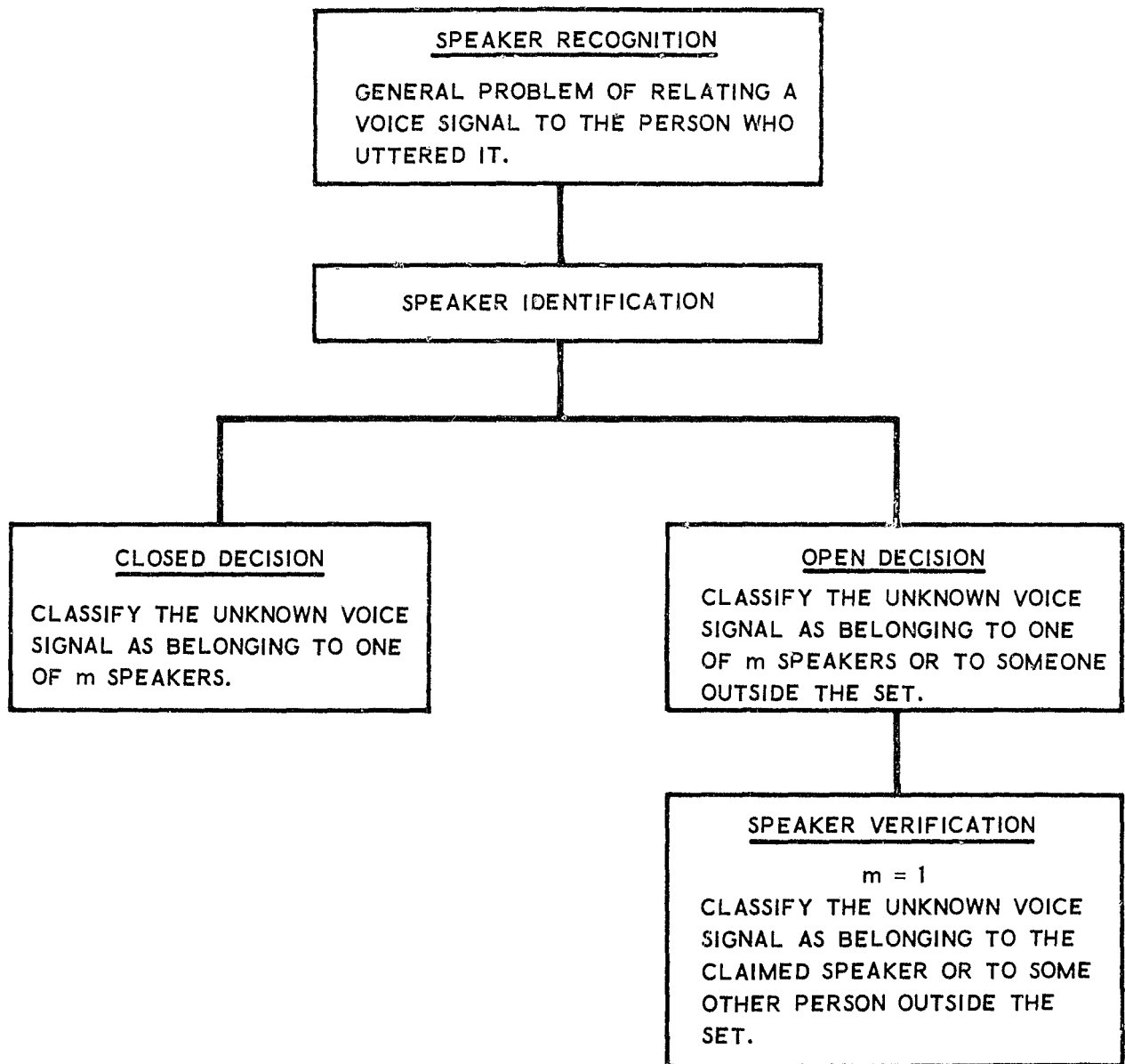


Figure 3-1. Speaker Recognition/Verification Definitions

classification decision. Again, the voice signal under test must pass a threshold test before it is accepted as belonging to the claimed speaker.

The threshold test in the open-decision system serves to discriminate between voice samples that are from the same speaker and those that are from different speakers. It is not necessary to know the identity of the different speakers. Thus, the problem of determining a decision criterion based on speakers and speaker characteristics not included in the classification data base has been circumvented, by merely requiring that the two voice signals under comparison match each other to within prescribed limits. The limits are determined from statistical analyses, and they influence the ultimate accuracy of the recognition system.

The types of errors that might be expected from the open- and the closed-decision systems are shown in Figure 3-2. It can be seen that in the closed-decision system there is only one type of error, i. e., false identification - the case where the unknown voice signal is matched with the wrong speaker. In the open-decision system there are three types of possible errors. The type IA error would occur if the unknown voice signal belonged to one of the m possible speakers, but was matched with the wrong one. The type II error would occur if the unknown voice signal belonged to one of the m possible speakers, but was not matched with any of them. Finally, the type IB error would occur if the unknown voice signal did not belong to anyone of the m speakers, but was matched with one of them. In the speaker verification problem, the type IA error is zero by definition because $m=1$. Therefore, there is only a single speaker for matching the unknown voice signal.

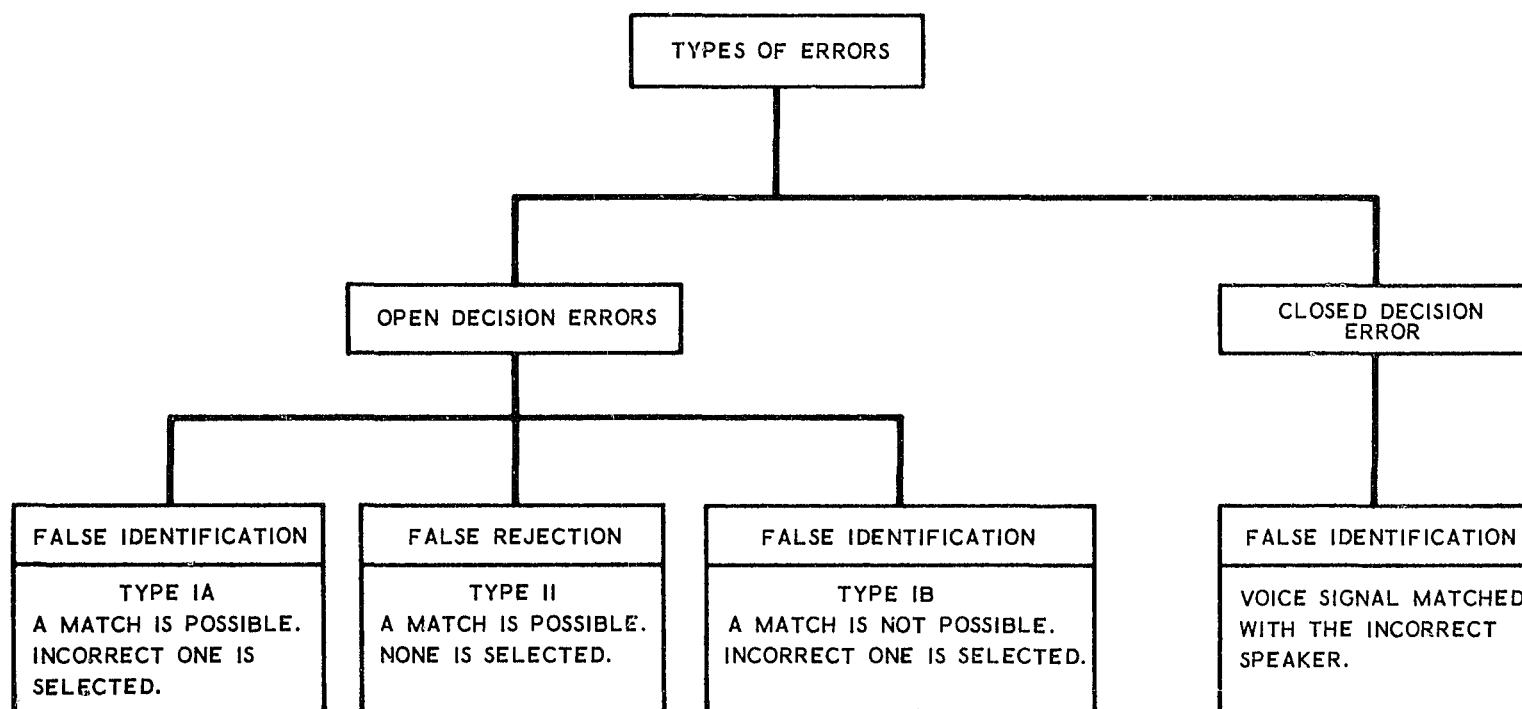


Figure 3-2. Classification/Recognition Errors

2. Theoretical and demonstrated accuracies. The feature selection process that was conducted during the brassboard system development phase provided sets of rank-ordered features that proved to be highly effective for the speaker identification task. The selected feature sets were found to be both efficient and consistent in the discrimination of voice samples from several speakers. Some indication of the effectiveness of these features in applications other than the above system can be seen by the following example.

Consider the case where a speaker identification system might be configured as an m -class, closed-decision classification system. Then, the results of the system feature extraction analysis show that for $m=25$, a linear classifier* could be designed to achieve average classification accuracies that approach 100 percent for feature vector sizes on the order of 50 features. These accuracies are shown in Figure 3-3 where features from only 6 phonemes are used; the best 6, a random set of 6, and the worst 6 phonemes in terms of their individual performance. With a vector comprised of, say, 10 to 14 phonemic events, one could expect the same accuracy

*A linear classifier consists of a linear combination of the feature vectors with a set of weighting functions. The weighting functions in this case are derived such that the separation between the sets of feature vectors representing different speakers is maximized.

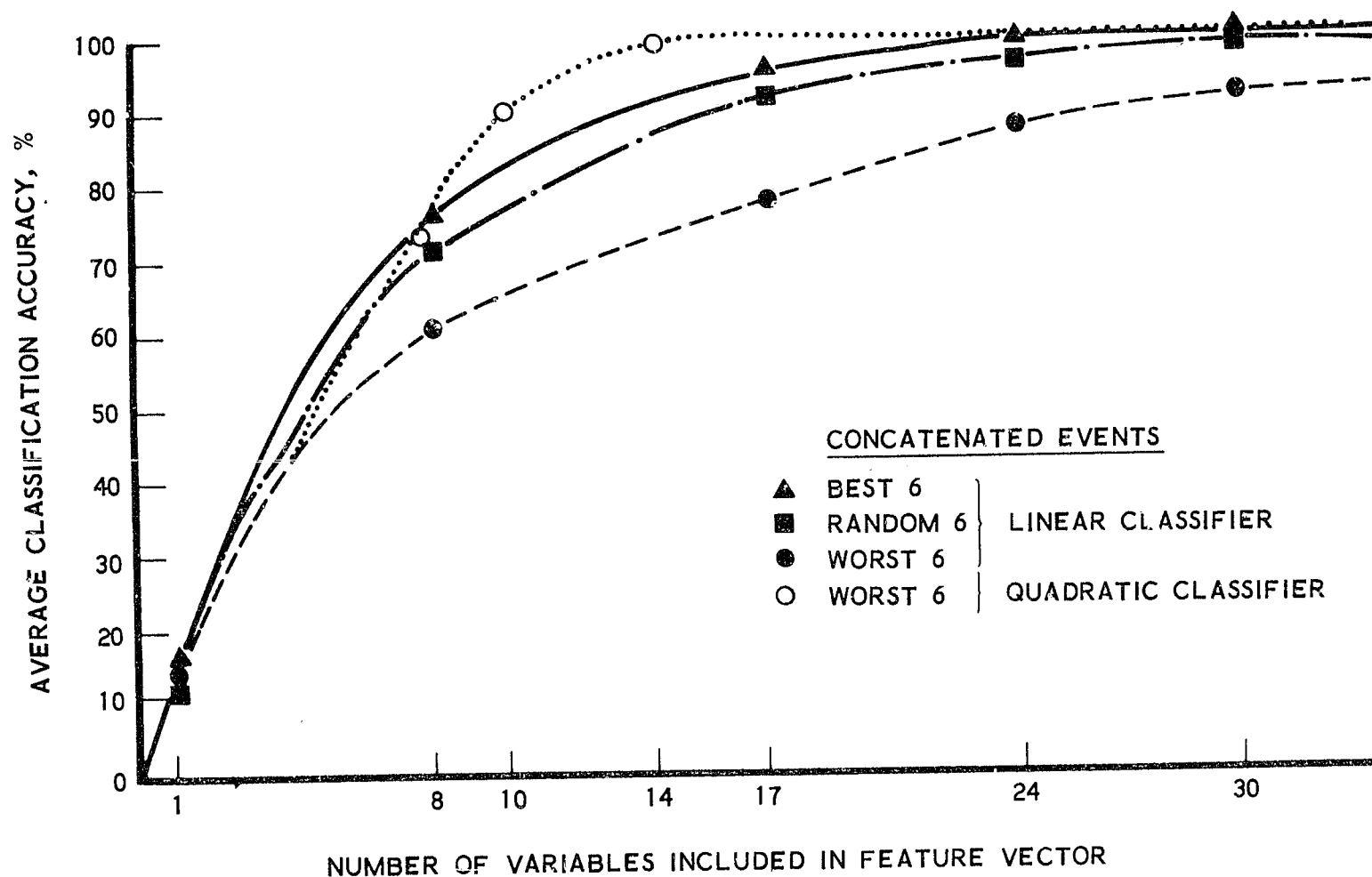


Figure 3-3. Average Classification Accuracy for 25 Speakers Using 6 Concatenated Phonemes; Features Derived from Special Features Plus LPC Spectral Estimates

with fewer total features. When a quadratic classifier* is used, then a classification performance that is comparable to the linear case can be achieved with only 14 total features. This indicates that the quadratic classifier is more nearly optimum for this particular recognition task than the linear classifier. However, in general, the former classifier requires a larger data base (collection of voice samples) to achieve statistical significance in the classification results.

The brassboard system is not a classifier; rather, it is more similar to a speaker verification system. It uses a similarity measure for purposes of determining whether two voice signals being compared come from the same person or from different persons. If the system were to be used in a closed-decision application, one can predict its recognition performance by means of the following theoretical equation:²¹

$$P_c = \sum_{j=1}^J h(j\Delta x)\Delta x \left\{ 1 - \left[\frac{g(j\Delta x)\Delta x}{2} + \sum_{k=1}^{j-1} g(k\Delta x)\Delta x \right] \right\}^{m-1} \quad (1)$$

where P_c is the probability of a correct decision, $h(j\Delta x)$ and $g(k\Delta x)$ are, respectively, the intraspeaker and interspeaker similarity measure

*A quadratic classifier results when a term, which is derived from a second-order function of the feature vector, is added to the linear classifier equation.

histograms, m is the number of speakers, J is some integer above which $h(j\Delta x)$ and $g(k\Delta x)$ are both zero, and Δx is the histogram bin width. The error rate is estimated as $P_e = 1 - P_c$.

Using the intraspeaker and interspeaker histograms for the system data where eight phonemic events were included in the feature vector, then the predicted error rate can be determined as shown in Figure 3-4. This curve shows the effect of increasing the number of speaker candidates for a close-decision test. It can be seen that, for the specific eight phonemic events shown in Figure 3-4, a recognition accuracy of greater than 95 percent can be expected with up to 30 speaker candidates. When the number of speakers that must be recognized in a closed-decision system is expanded indefinitely, the probability of correct recognition decreases according to the curve shown in Figure 3-5. Again, this curve applies to the case described in Figure 3-4.

In most applications of computer-aided speaker identification technology it will not be possible to operate a close-decision system. Hence, the expected accuracies for the two-choice, open-decision system should be examined. The histograms which were described above actually approximate probability-density functions and when they are integrated in the appropriate directions, the cumulative probability distributions shown in Figure 3-6 are obtained. The curves show the probability of a particular type of error if a decision threshold was placed at some value of the similarity measure. The decision criterion would be to say that two voice samples being compared are different if they resulted in a similarity measure greater than the threshold value. If two decision thresholds are used so that the overlap regions between

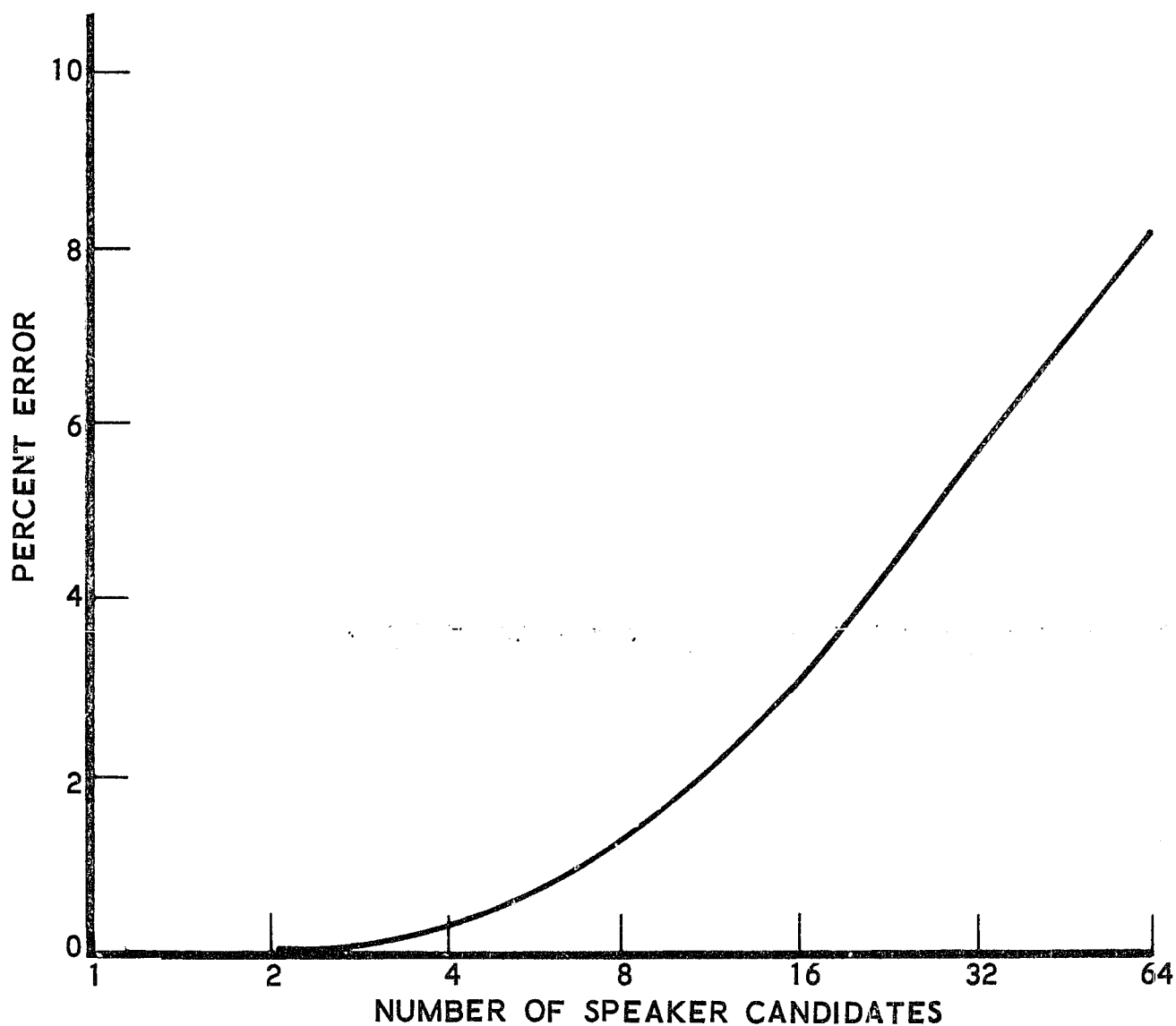


Figure 3-4. False Identification Error-Rate Estimate, 118 Speaker Male Population for a Closed-Decision Test. Feature vectors consist of spectral values from each of 8 phonetic events: $[\eta/]$, $[i/]$, $[I/]$, $[\Lambda/]$, $[3'/]$, $[\xi/]$, $[a/]$, $[a/]$, $[o/]$, $[U/]$, $[u/]$.

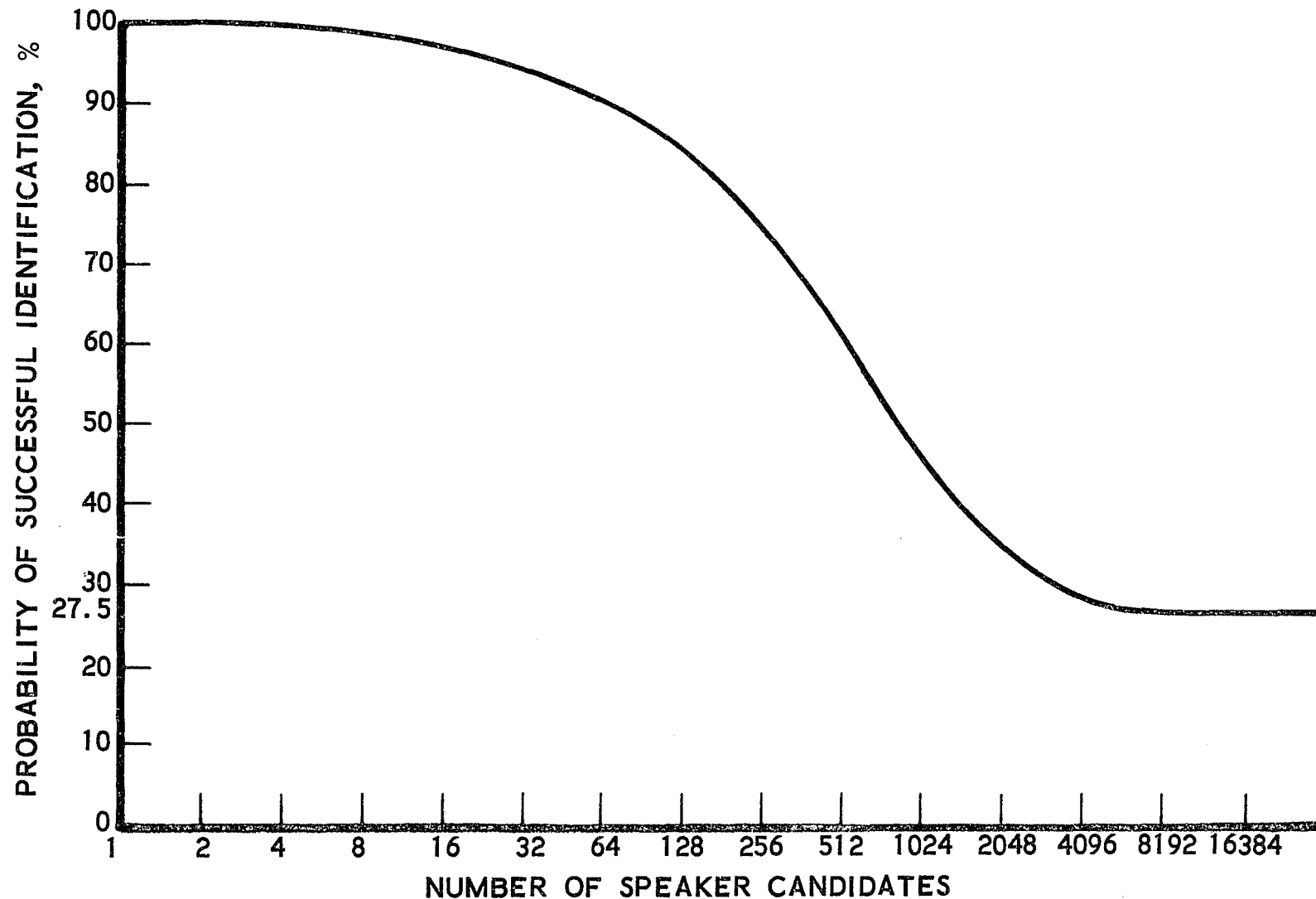


Figure 3-5. Theoretical Calculation of Probability of Successful Identification of Voice Signal vs. Number of Possible Candidates (Close-decision test based on 118-speaker population and 8 phonetic events shown in Figure 3-4.)

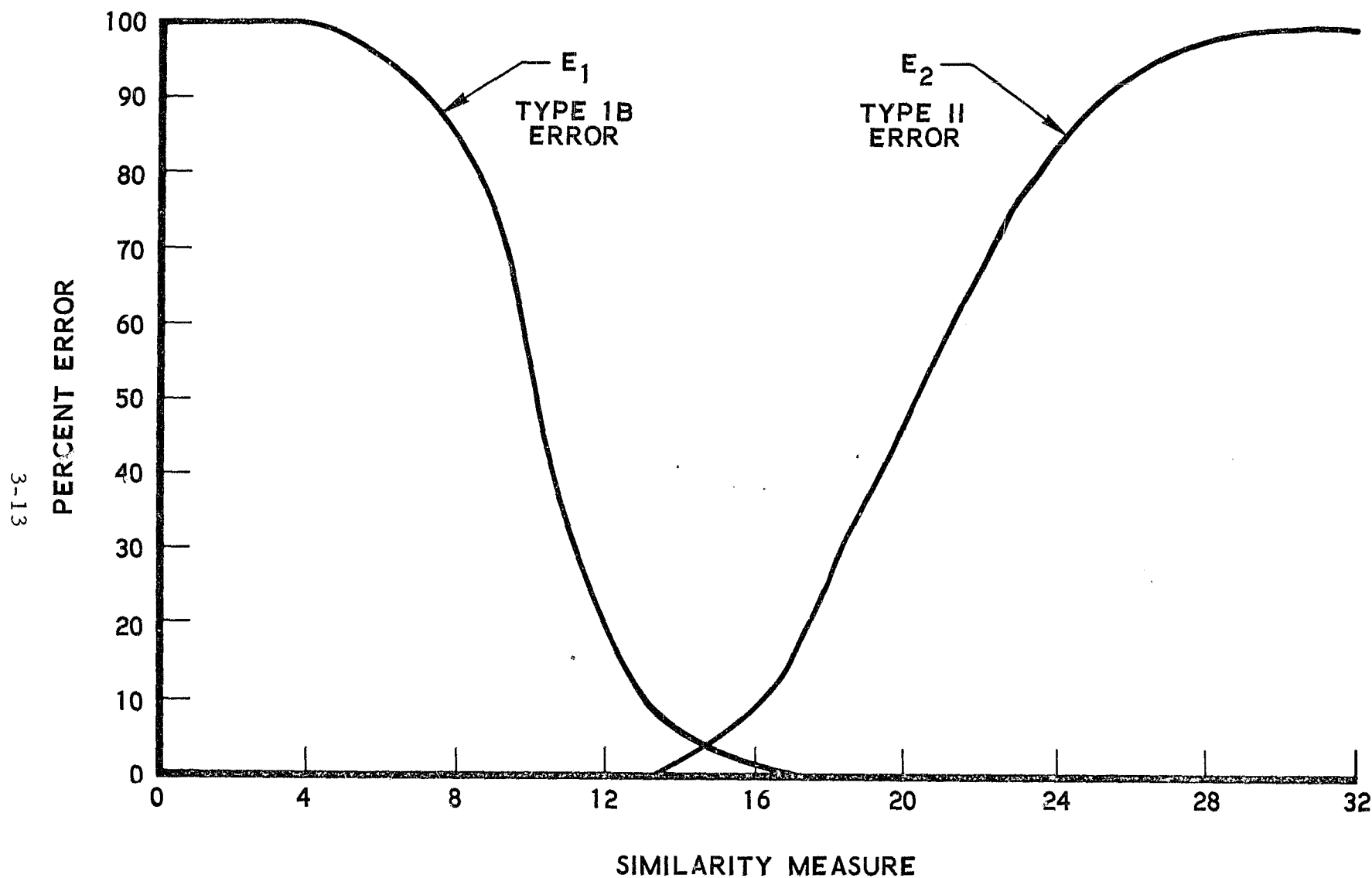


Figure 3-6. Identification Error-Rate Estimates for Open-Decision Tests Using 118-Speaker Population (Feature vector consists of 8 phonemic events shown in Figure 3-4.)

$S = 13$ and $S = 17$ are avoided, there would be essentially no probability of a false rejection of a false identification by the system. That is, the probability of error would be negligibly small for this case. However, it would not be possible to make a decision 10 percent of the time if the speaker was, in fact, who he claimed to be, and 17 percent of the time if the speaker was an impostor. The a priori probabilities associated with these two situations would influence the expected percentage of time in which the system could not make a decision.

In an operational system it is possible to allow the system to operate in multiple-attempt mode. That is, the system will allow a speaker some pre-determined number of attempts to gain access before making a final decision. For example, if an individual requested identification by means of a voice sample and was unsuccessful because his sample was separated from the reference by more than the allowable threshold distance, the system may allow him to make a second or a third attempt.

Operation in the multiple-attempt mode would allow a single threshold to be set to a level which would preclude false identifications (i.e., at $S = 13$ in Figure 3-6.) If it is assumed that a typical speaker would have the intraspeaker variance distribution equal to that for the population (curve E1 in Figure 3-6), he could be expected to experience false rejection approximately 10 percent of the time. The probability that he would be rejected again would be only one percent, however, and the probability that he would be rejected three times in succession would be one in 1000. In particular, the performance data indicate that for a threshold of 12.5, the

false acceptance error would be 0.2 percent, and the false rejection error after three attempts is also 0.2 percent. In practice, the probability would be significantly less than this, since Figure 3-6 depicts the intraspeaker variation for randomly selected pairs of utterances. In a practical system, the speech reference used in the system would likely be some characteristic utterance averaged over a number of repetitions and would thus represent a type of "mean value" for that speaker. Variations away from this mean would thus be generally less than indicated in Figure 3-6.

Even a multiple-attempt mode system would have to have some limiting element in the number of allowable attempts to protect the system against impostors. The signaling of an alarm following three unsuccessful attempts would provide protection for the system and allow manual intervention in the unlikely event that an authorized individual was still being denied access.

It was mentioned above that there was a negligible probability of error expected for decisions on either side of the threshold similarity values. The degree of negligibility depends upon how well the curves shown in Figure 3-6 actually describe the population, since they are based on 118 different speakers. The confidence that can be placed in the error estimates E1 and E2 can be determined by means of the curves shown in Figure 3-7. Thus, the upper and lower bounds on these estimates may be established for a given level of desired confidence, and one can observe how these new values affect the location of the decision threshold.

The important consideration that should be noted is that whereas the brassboard system does not have control over the number or the types of

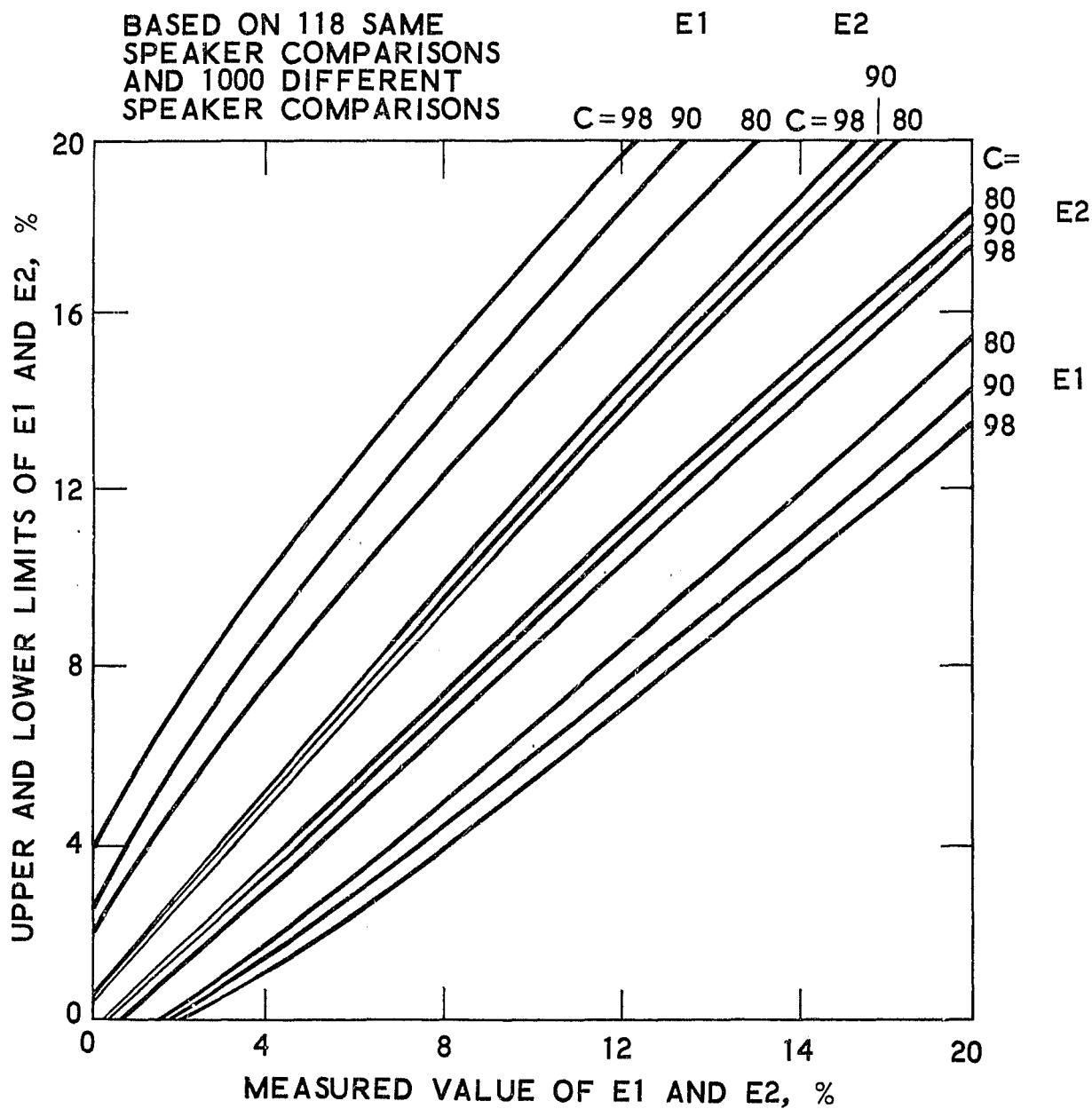


Figure 3-7. Example of Confidence Estimates on System Statistics

CONTINUED

1 OF 2

phonemes that are uttered, most other applications would have this control. Therefore, since previous analyses show that recognition accuracy improves up to a point with a larger set of phonemes, one can expect that systems using this larger set, such as speaker verification systems, should achieve speaker recognition accuracies that are at least as good as the brassboard system results. Hence, recognition errors of less than 1 percent for 95 percent to 99 percent of the time that a decision is made is a reasonable performance to expect from other applications of the Semi-Automatic Speaker Identification technology.

B. Data Base Storage Requirements

The following analysis presents, in parametric terms, the data base size for a cooperative-population speaker identification system.²² In turn, the data base size specifies the direct access memory (bulk storage) required by such a system. The main component of the data base is the speech characteristics data set (file), which is developed from cooperatively vocalized expressions of the population's members. The system has applications where automated entry into restricted areas is desired, or, where computerized speaker identification offers an enhanced capability to differentiate one individual from all others. An example would be a system where an individual, to gain entry into a closed area, speaks his social security number into a microphone. The system receives this utterance, extracts the necessary speech characteristics, and compares them with those contained in the data base for the same social security number.

1. Analysis. The speaker identification portion of the system is shown in Figure 3-8, in which the relationship of the data base storage medium to the computer is clarified. The speech characteristics are stored in feature vectors, with one feature vector defined for each phoneme to be considered in the similarity calculations. The elements of the feature vector are speech features, such as the power spectral components of the associated phoneme. The data set size, which contains all the feature vectors from the cooperative utterances of all the population members, can be defined parametrically as

$$S = n[m(l + k) + j]$$

where

S = data set size in bytes (8 bits)

n = population size

m = number of feature vectors (phonemes) stored for the cooperative utterance

l = number of bytes for overhead per feature vector

k = number of bytes for the elements of each feature vector

j = number of bytes for storage block identification.

Two constructions of the storage block that can be used for each member's speech characteristics are shown in Figure 3-9. Version A, for a fixed number of feature vectors (phonemes) per member, also implies that the

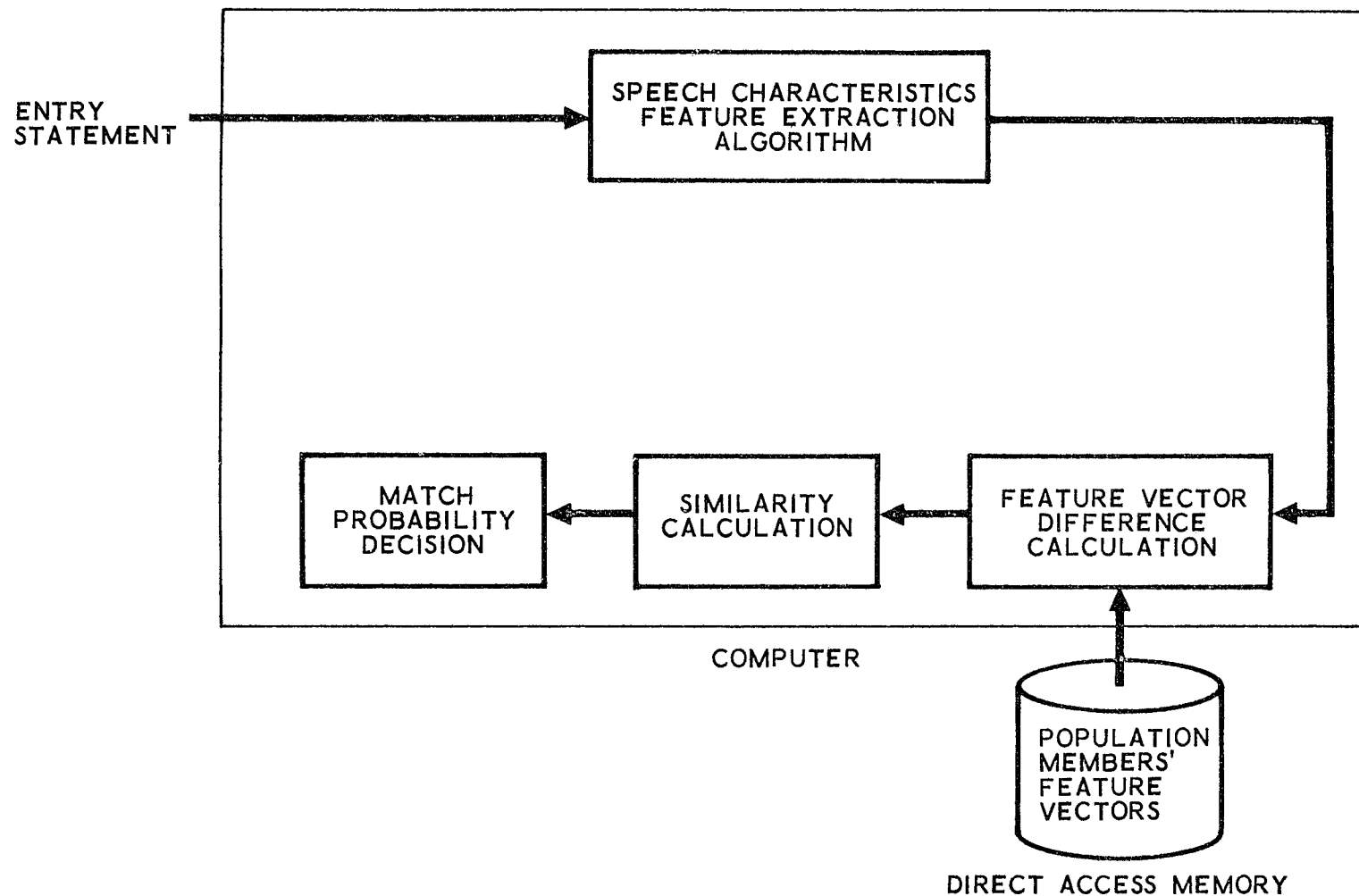
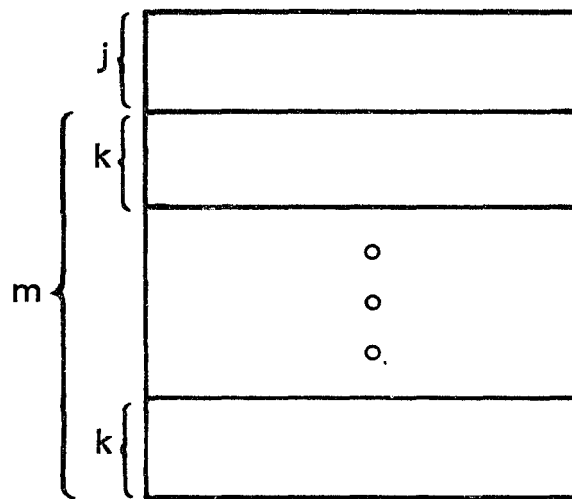
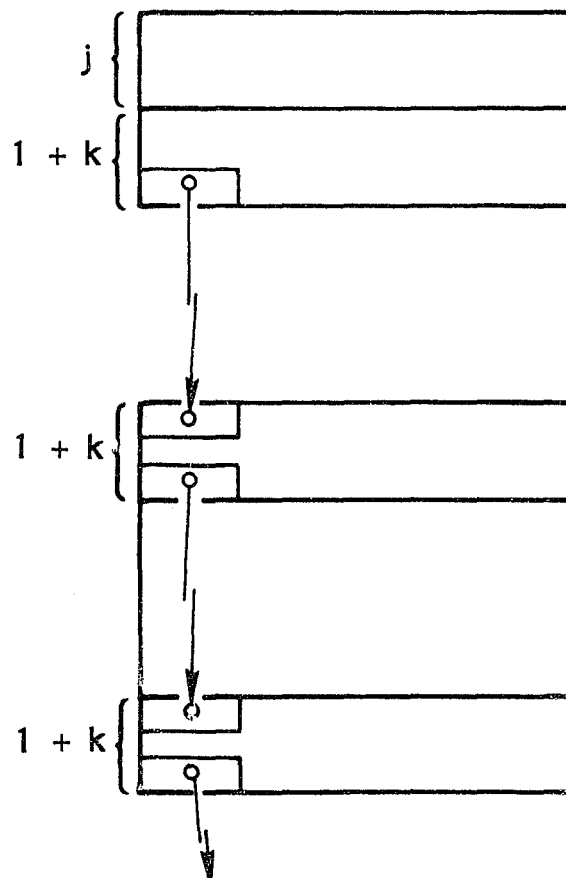


Figure 3-8. Simplified System Speaker Identification Mechanization



(A) FIXED m



(B) VARIABLE m

Figure 3-9. Representative Speech Characteristics Storage Blocks

phonemes are in a fixed order so that there is no feature vector overhead. Version B allows for a variable number of feature vectors, which may be entered in any order. It provides greater flexibility at the possible cost of additional storage. There will be a feature vector overhead consisting of pointers and phoneme identification.

A representative example of the speech characteristics data set size, when Version A data structure is used, is presented below for the following conditions:

- Population size, $n = 10,000$
- Number of feature vectors per population member, $m = 7$
- Number of bytes for the elements of each feature vector (at four bytes/element), $k = 400$
- Number of bytes for storage block identification, consisting of member's name (30 bytes) and social security number (five bytes), $j = 35$ bytes.

Data set size for the example above is

$$\begin{aligned} S &= 10^4 [7(0 + 400) + 35] \\ &= 2835 \times 10^4 \\ &\approx 28 \times 10^6 \text{ bytes} \end{aligned}$$

Another example, if only 25 feature vector elements are sufficient but 12 phonemes are required for each population member,

$$\begin{aligned}
S &= 10^4[12(0 + 100) + 35] \\
&= 1235 \times 10^4 \\
&\approx 12 \times 10^6 \text{ bytes}
\end{aligned}$$

Use of Version B data structure would add only four to five bytes per feature vector, so that any size change over Version A would be most sensitive to the average number of feature vectors stored per member.

As previously stated, the speech characteristics data set is the major component of the system data base. The storage required for any other components will depend on the memory size of the computer used, but will not add significantly to the amount of bulk storage required. One of the components could be the tables of probability distributions used in the comparison algorithm.

2. Conclusions. The direct access storage required for a cooperative-population speaker identification system can be approximated by

$$S \approx nmk$$

where

S = direct access storage (bytes)

n = population size

m = the number of feature vectors (phonemes) for each cooperative utterance

k = number of bytes required for the elements of each feature vector.

This simplification derives from the examples shown above.

An upper limit to the data set size appears to be when 13 phonemes are required with 100 elements in each feature vector. For this case,

$$\begin{aligned} S &\approx 10^4 \times 13 \times 400 \\ &\approx 52 \times 10^6 \text{ bytes} \end{aligned}$$

While this is a considerable amount of data, it is worthwhile to put it into some perspective with respect to cost. A data disk storage drive, which provides approximately 25 million bytes of storage can be purchased and installed for \$22,000. This would put the worst case cost for system data base storage from \$44,000 to \$66,000.

CHAPTER IV. POTENTIAL IMPLEMENTATIONS

A. Police Remote Identity Verification Concept

The widespread use of false or stolen identification among criminals has made many standard techniques for identifying suspects inadequate. This is particularly true in the field, where patrolmen are able to determine little more than sex, approximate height, weight, complexion, etc., for ascertaining whether an individual in custody is wanted. When a suspect is apprehended, he is routinely fingerprinted and photographed. Unfortunately, current methods for transmitting and processing fingerprint and pictorial data are very slow and cumbersome, delaying both the identification of actual criminals and the release of innocent suspects.

The particular attributes of speech information, and the ease with which such information can be transmitted efficiently over standard telephone and radio channels, offer significant potential for implementation of accurate, positive identity verification from remote field locations.

The basic concept is illustrated in Figure 4-1. When a suspect is stopped by a patrolman, he would be required to produce identification such as a driver's license. He would also be asked to read certain words such as his name, license number, etc., into the microphone of a special mobile terminal device. A typical statement might be as follows:

"My name is Jack Smith and my California driver's license number is R123456."

This statement would be recorded and information pertaining to identity would be derived for subsequent processing at the central dispatching station

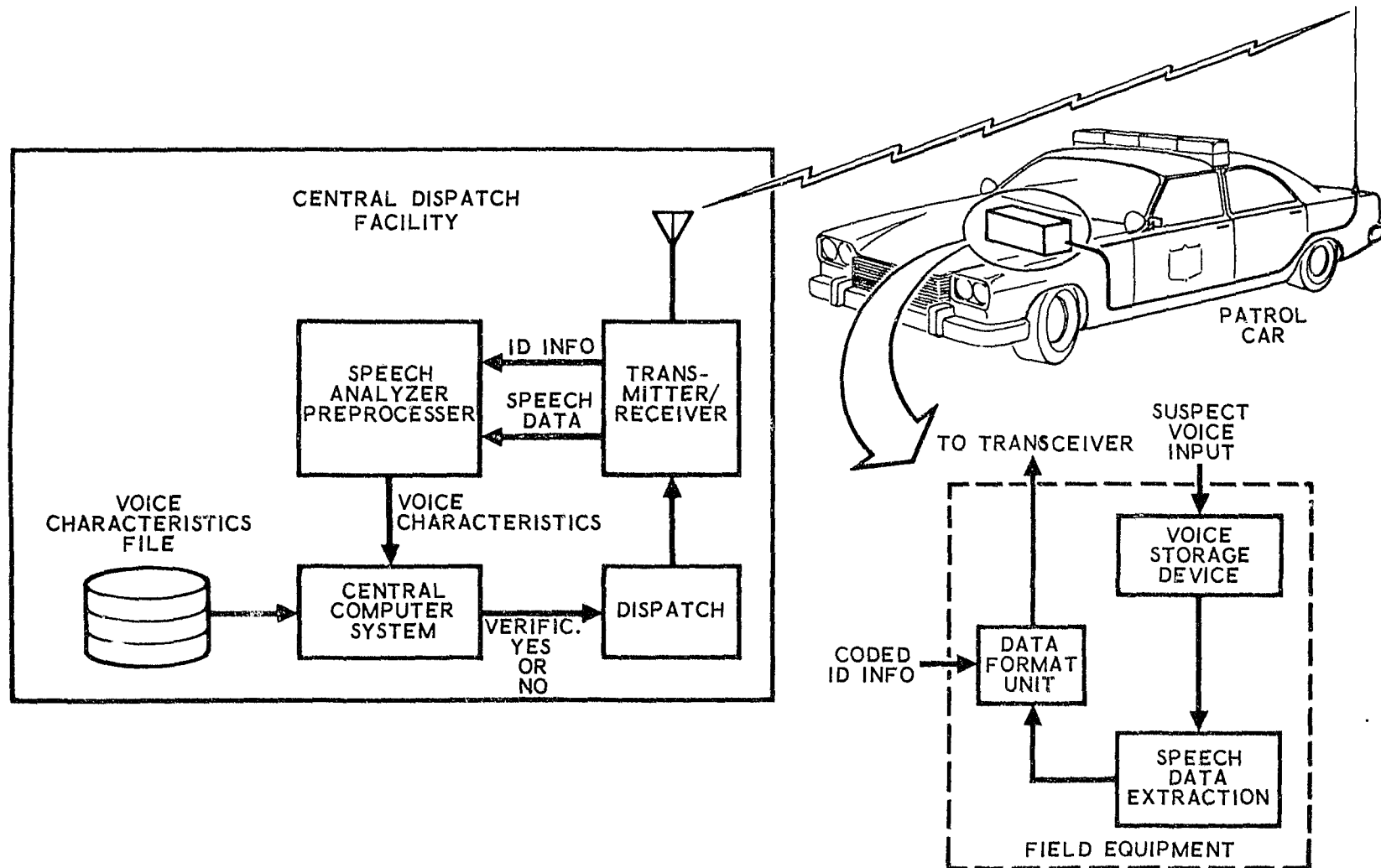


Figure 4-1. Concept for Field Identification

or some other processing facility. At this facility, a unique classifier based on voice characteristics would be derived and matched against a corresponding classifier stored in a data file for an individual with the stated name and driver's license number. A match would signify that the individual was using identification legally obtained. Simultaneous with the foregoing matching process, a search could be conducted of listings of wanted persons whose voice characteristics were on file. The advantage of this concept is that it would provide in real time a positive verification of identity based upon physical attributes.

Figure 4-1 describes the general functions performed by the system. In an actual implementation, numerous options exist and would be selected by a detailed tradeoff analysis. One of the most crucial factors involved is the impact of such a system on radio channel usage.

The simplest implementation would involve direct transmission of the suspect's speech sample to the dispatch center via radio, but this presents a significant detriment to the system. When speech is transmitted over a mobile radio link, the distortions and interference contributed by the link seriously degrade the speech signal. Such degradation would probably interfere with and possibly invalidate the extraction of voice characteristics suitable for identification.

An alternative approach would be to record the analog speech sample in the police car on a cassette tape recorder and relay the speech in digital form back to the central facility. This relay could be in non-real time and could be designed to preserve good speech quality. For example, assume that such quality requires a channel capacity of about 24,000 bits per second

(equivalent to the theoretical telephone channel capacity); this could be achieved by playing back the recorded speech at a speed reduction of 10 to 1 and transmitting at about 2400 bits per second. The speed reductions would be compensated for at the central facility. This technique would result in a high-quality speech sample at the central facility without unduly complicated equipment in the patrol car. The disadvantage lies in the need for a substantial increase in the requirements for radio channel occupancy.

A third alternative is to use more sophisticated equipment in the patrol car to perform certain of the processing tasks and in this way reduce the amount of speech data required to be transmitted. This initial preprocessing could perform the functions of segmenting the speech data into their constituent parts and labeling these parts with their appropriate phonetic identification. This process is equivalent to basic speech recognition and the equipment could, in fact, be coupled with a voice-to-digital transmission system to facilitate efficient mobile channel usage in normal dispatch communications.

After the patrol car equipment has segmented and labeled the desired speech elements, these elements and their associated labels would be transmitted in digital to the central facility. In the speech sample presented above (containing the subject's name and driver's license number), a total number of 10 to 15 usable phonetic events could reasonably be expected to occur. Each of these events would be approximately 0.1 second in duration and, in digitized form, would require approximately 1.0 second of transmission time. Label tag information and other data would also add to the transmission time. A reasonable estimate of the total time required to send the preprocessed sample would be 15 to 20 seconds. This is a relatively long message duration compared to the regular dispatch traffic. Offsetting advantages,

however, in terms of improved identification accuracy and the consequent enhancement of patrol unit effectiveness would more than compensate for the radio channel effects. In addition, the previously mentioned capabilities for direct voice-to-digital conversion in the patrol car, which would be an added fallout of this concept, would more than compensate for any loss of channel availability due to the identification system.

B. Automated Identification System Concept for Financial Transactions and Access Control

Automated identification systems have numerous application areas and specific requirements but possess many functional similarities. In general, an access control system would be functionally simpler than a system which controls and records financial transactions. For this reason the system concept discussion is limited to the financially oriented application. The techniques and requirements, however, are equally applicable to access control.

There is an increasing trend in the financial and retail sales community to the use of on-line credit authorization data entry systems.²³ It is estimated that automatic teller devices and cash dispensing machines will grow from 2500 units in 1974 to more than 25,000 units by the end of 1980.²⁴ Automated terminals, controlled by a minicomputer can even become a branch bank. The tellers accept paper transactions and can dispense cash as well. They are operated by entering an ID card and keying a confidential number to guard against unauthorized use of stolen or lost cards. The systems also include point-of-sale terminals at one or more retail stores, which are directly

on-line with the central processing unit at the bank center. There are several advantages to such on-line systems. First, the "hot" list - a list of known credit cards in fraudulent use - can be automatically maintained at the bank center and can be used to check each transaction. Currently, the list is updated periodically and mailed to merchants, who are obligated by agreement to check the list each time a credit sale is made. This list is difficult to maintain as current. Second, with an on-line system, all transactions can be evaluated by the credit authorization system. With an off-line system, only about 20 to 30 percent of all transactions (the estimated amount above the floor limit) can be evaluated. Third, an on-line system would greatly expedite the present slow manual method, which entails delays both at the sales point and at the bank center. Fourth, a fully automated on-line system would reduce the time lag between the updating of the authorization file at the bank center and the consumer's actual current credit position.

The principal method for providing security for on-line terminals is the use of a personal identification number. This number, typically including four to six digits, must be keyed in by the consumer in the unattended terminal after the magnetically encoded data on the credit card are read by the terminal. Subsequently, the account number and the identification number are transmitted in encrypted form from the terminal to the bank center.

Even with encryption, the system is vulnerable due to possible compromising of the identification code. The content of the credit card magnetic stripe can be transferred from one card to another, or modified completely, when the proper facilities exist to effect the change. These factors make an

unattended cash dispenser vulnerable to fraud and provide impetus on the part of the banking industry to seek better methods for identification using remote terminals.

Figure 4-2 shows a possible system concept for an automated, on-line terminal for processing financial transactions. These transactions could include cash dispensing, credit approval or direct purchasing (of an airline ticket, for example).

In using the system, the customer would speak his identification number into the terminal and at the same time enter the number via the keyboard. Correct entry would be verified by a terminal-generated display to the customer. The customer (or a sales clerk) would then enter the data describing the requested service into the terminal with the keyboard.

The terminal would extract the speech element sample data for transmission with the speech label data, the account number, and the request information, to the bank center. Encryption would be used to deter the misuse of the terminal via the transmission link.

At the bank center, the speech data would be processed to derive a set of voice identity characteristics that can be matched against the equivalent characteristics on file in the data base. The account number would be used as the file address key and the customer's identity would be verified in accordance with an established acceptance criteria.

The bank center system would either authorize or not authorize the transaction, based upon the results of the personal identification procedure,

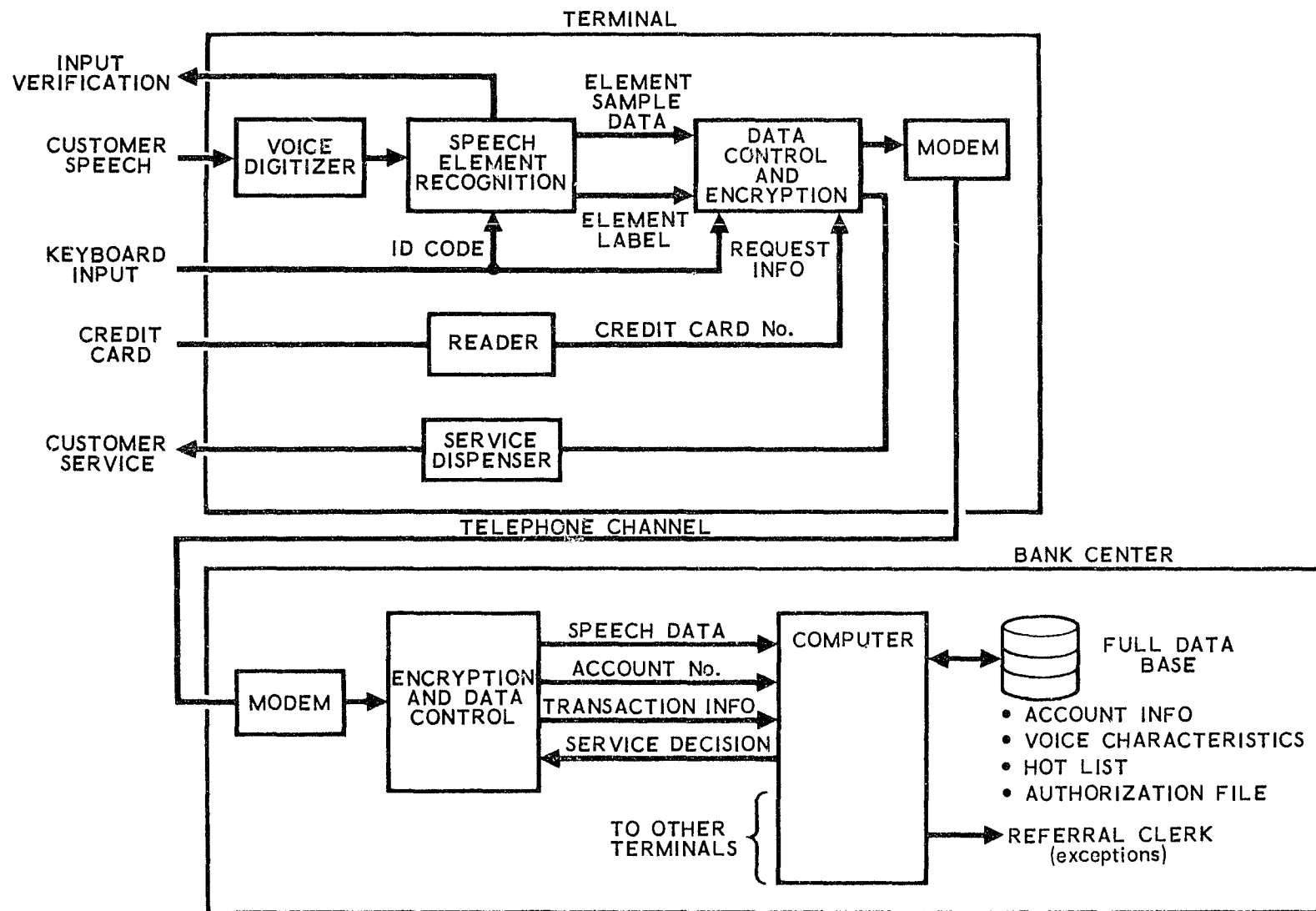


Figure 4-2. Automated System Concept for Financial Transactions

checking of the hot list, and evaluation of the customer's current credit position.

An authorization would result both in completing the requested transaction and in updating the consumer's checking or savings account records. Any number of remote terminals could be connected directly to the bank center system.

CHAPTER V. CONCLUSIONS AND RECOMMENDATIONS

In the report that documented the preliminary investigations of the applications study, tentative conclusions were drawn, which indicated (1) that the use of voice identification was becoming increasingly important in police work and (2) that the traditional methods of identification through voiceprint examination needed to be supplemented by more quantitative and objective techniques. The applications study performed subsequent to the preliminary investigation has substantiated and reinforced these tentative conclusions. In-depth discussions with voiceprint examiners and with representatives of police crime labs and identification bureaus has revealed a high level of interest in the concept of computer-aided voice identification. A number of federal and local law enforcement officials have expressed a desire to participate in the development and test program, in order to operationally implement the technique at the earliest time period.

The aim of the applications study was to assess the relevance of computer-aided speaker identification techniques to the general problem of identification. Since the preliminary report addressed the applications specifically related to investigative and forensic use, the latter part of the study focused on the use of voice identification to improve access security and improved identity verification concepts for financial transactions and for routine police activities. The study of these areas resulted in the following major conclusions:

- There is a need for better personal identification systems. The continuing increase in the complexity of society increases the

amount of damage that can be committed against society by criminal individuals and groups. Damage can take the form of physical violence to persons or property or harm done by improper manipulation of information relating to finances or personal data. The only feasible way for society to counteract this danger is to identify those areas of high potential risk and to institute safeguards to ensure that only those individuals who have legitimate access are admitted to these areas. Thus the problem of accurately and reliably identifying individuals will become an increasingly important factor in social relationships.

- There is a general lack of information regarding the technical requirements of automatic systems for access control or personal identification. This is particularly true with regard to the most essential requirement, i.e., accuracy. Clearly, before any system design for a personal identification system can take place, the fundamental requirements for the allowable rates for false acceptance and false dismissal must be ascertained. With these figures as benchmarks, various approaches can be investigated in terms of cost, user acceptance, response time, etc., and tradeoffs can be conducted to arrive at the optimum system configuration. Ideally, perfect accuracy is desirable, but also inherently unrealizable. Each area of application must be independently evaluated in terms of a cost-benefit analysis to determine the criteria best suited to its individual needs.

- The use of voice characteristics for voice identification is feasible and practical. The tests performed and the statistics generated in the development of the computer-aided speaker identification system and in other programs has demonstrated that there are measureable differences in the voice characteristics of individuals (interspeaker). Further, it has been shown that the differences between repeated utterances by the same person (intraspeaker) are significantly less than the interspeaker differences.
- The existence and use of a voice data arrest file would improve the effectiveness of local and state identification bureaus. New technology associated with burglar, robbery and vandalism alarm systems as well as with advanced 911 emergency telephone systems will increase the quantity and quality of criminal voice samples obtained in several crime categories. The ability of investigators to search arrest files for a perpetrator's identity using new voice analysis and classification methods can be expected to increase the effectiveness of the police in apprehending criminals.
- The techniques developed for the Semi-Automatic Speaker Identification System can be modified and extended for use in an automatic personal identification system. The modifications are well defined and conform to well established techniques. The analysis based on the statistics accumulated thus far on the system indicate that the system accuracy would meet the requirements posed by

most security operations. The basic design of the system for use with telephone grade voice data makes the technique applicable for a wide variety of implementations.

The development and evaluation of the Semi-Automatic Speaker Identification System has revealed a high level of flexibility and performance on the constrained test samples used to date. Refinements and modifications to the system can be expected, however, as new sample material containing female voices, disguise, etc. is presented for analysis. Full evaluation of the technique will only be possible when the system is exposed to the full range of possible speech samples and under the full range of operating conditions.

If this evaluation proves the effectiveness of the computer-aided technique, a number of ancillary techniques will become practicable. These include: (1) voice classification and the searching of voice data arrest files, (2) the analysis of speech samples to determine factors relating to suspect size, background, health, etc., and (3) the extension of voice analysis techniques to facilitate access control, speech recognition, and identity verification.

Specific recommendations based on relative priorities, funding limitations, momentum of current efforts and the potential for maximum long-range effect include:

- Complete the laboratory testing of the Semi-Automatic Speaker Identification System to evaluate the system effectiveness with female speakers, with speakers under emotional stress conditions, and with speakers using disguise. The testing should also include

evaluation of system performance with actual voice samples obtained in the field. Improvements to the system where needed should be made in response to these test results.

- The system should be pilot-tested in cooperation with selected user agencies to determine its effectiveness in an actual operating environment.
- Detailed plans for formal field tests and evaluations of the system should be prepared. The scope and duration of these tests should be such that a valid statistical conclusion regarding the system capability can be drawn.
- Close coordination and communication pertaining to the development and system test activities should be established and maintained with members of the speech science community and in particular with those involved in voice identification. Where possible, the active cooperation and participation of key individuals both in private security and in law enforcement, should be solicited.
- Law enforcement agencies should be encouraged to utilize new techniques for identification, such as voice analysis and comparison. Guidelines for recording speech samples and for utilizing equipment should be provided on request and suggestions regarding the implementation of new programs, such as the establishment of arrest data bases of speech samples, should be furnished.
- The speaker data base material developed to date in the program should be utilized to investigate and develop methods for analyzing

speech samples to extract descriptive information regarding speaker characteristics. Efforts should include the development of methods whereby speech characteristics can be used to classify voice samples for subsequent filing and searching operations.

- Local and federal committees responsible for establishing policy and standards for personal identification should coordinate efforts with industry to arrive at accepted standards and goals for the performance of automated identity verification systems.

NOTES

1. C.W. Swonger, "Applications of fingerprint identification technology to criminal identification and security systems," Proceedings of 1st International Conference on Electronic Crime Countermeasures, July 1973.
2. M. Eleccion, "Automatic fingerprint identification," IEEE Spectrum, September 1973.
3. M.H.L. Hecker, "Speaker recognition - An interpretive survey of the literature," American Speech and Hearing Association Monograph Number 16, January 1971.
4. M.B. Herscher, T.B. Martin, and W.F. Meeker, "Automatic speaker identification," Proceedings of the 1970 Carnahan Conference on Electronic Crime Countermeasures, April 1970.
5. S.L. Moshier, "Some algebraic properties of speech signals," Proceedings of the IEEE, Vol. 58, No. 2, February 1970.
6. "Computer comprehends verbal telephone messages with 99% accuracy," Computer Design, December 1974.
7. J.E. Paul, et al, "Semi-Automatic Speaker Identification System, - Analytical Studies Final Report," Rockwell International Report No. C74-1184/501 (December 1974).
8. "Preliminary Investigation of Applications of the Computer-Aided Speaker Identification System," The Aerospace Corporation Report No. ATR-74(7907)-1 (June 1974).

9. Memorandum, "An Investigation of the Use of Voice Recording/ Identification in Crime," K.L. Henrie to J.C. Daly, The Aerospace Corporation (September 25, 1974).
10. A.A. Moenssens, et al., "Scientific Evidence in Criminal Cases," The Foundation Press, Inc., Mineola, New York (1973).
11. "Voiceprint identification," Georgetown Law Journal, Vol. 61, Issue 3, February 1973.
12. "False I.D. committee begins search for control methods," Crime Control Digest, Vol. 8, No. 46, November 1974, p.1.
13. "L.A. city computer manipulated by Mafia for \$2.5 million payoff," Security Systems Digest, Vol. 5, No. 26, December 18, 1974, p. 1.
14. D.B. Parker, S. Nycum, and S. Oura, "Computer abuse," The National Science Foundation/RANN NSF/RA-S-73-017, Washington, D.C., 20530, November 1973.
15. "The state of the computer industry in the United States," AFIPS and the National Science Foundation (GJ-996), 1973, p. 28.
16. G. Salancik, T. Gordon, and N. Adams, "On the nature of economic losses arising from computer-based systems in the next fifteen years," Institute for the Future, R-23, March 1972.
17. R. Curtis and E. Hogan, "Perils of the Peaceful Atom," Doubleday and Company, New York (1969).
18. "The Cost of Crimes Against Business," U.S. Department of Commerce, Superintendent of Documents, Washington, D.C.

19. Memorandum, "Accuracy Requirements for Semi-Automatic Speaker Identification Technology," C. Henderson to P. Broderick, The Aerospace Corporation Report, No. ATM 75(7907)-3 (17 January 1974).*
20. H. Chernoff and L.E. Moses, "Elementary Decision Theory," John Wiley and Sons, New York (1959).
21. G.D. Hair and T.W. Rekieta, "Speaker Identification Research," Texas Instruments, Inc. Final Report (August 1972).
22. Memorandum, "Data Base Storage Requirements for a Cooperative-Population Speaker Identification System," M. Lubofsky to H.W. Nordyke, The Aerospace Corporation Report, No. ATM 74(7907)-11 (11 February 1974).*
23. J. Sroigals and H. Ziegler, "Magnetic-stripe credit cards: big business in the offing," IEEE Spectrum, December 1974.
24. "Automated Tellers on the Rise," Modern Data, vol. 8, no. 2, p. 12, February 1975.

*Not available for external distribution.

APPENDIX A. ANALYSIS OF POTENTIAL VOICE
IDENTIFICATION EFFECTIVENESS

In order to obtain an indication of the effectiveness of a voice recording and identification concept, it was assumed that voice recorders were available to the full population in each area of crime.

1. Basic Considerations

CASE 1: OPTIMISTIC CASE

Assume a number of people (N) who are going to embark on a career of crime. They have selected a specific area of crime to pursue, an area with a high probability of success (P_s).

The first crime event by all results in failure (capture) for some, and success for the rest, i.e.:

$$\text{Number of failures} = N(1-P_s)$$

$$\text{Number of successes} = NP_s$$

After a failure the criminal is removed from the system ever after, hence the title "Optimistic Case."

After the first several events by all, the total number of failures and successes is:

Total number of failure events over n attempts =

$$F = N(1-P_s) + NP_s(1-P_s) + NP_s^2(1-P_s) + \dots + NP_s^{n-1}(1-P_s)$$

$$= N(1-P_s) \frac{1-P_s^n}{1-P_s}$$

$$= N(1-P_s^n)$$

Total number of success events over n attempts =

$$\begin{aligned} S &= NP_s + NP_s^2 + NP_s^3 + \dots + NP_s^n \\ &= N(P_s + P_s^2 + P_s^3 + \dots + P_s^n) \\ &= NP_s \frac{1 - P_s^n}{1 - P_s} \end{aligned}$$

The above represents a binomial series converging on the value $N\left(\frac{1}{1-P_s} - 1\right)$ in the case of total number of successes.

Let R_s = success ratio and be equal to the probability of committing n successful crimes without being caught, that is

$$R_s = P_s^n$$

See Figure A-1 for plotted values of R_s as a function of n and P_s .

R_s identifies the degree of success over n events. A cross plot of these data is presented in Figure A-2 where percentile is plotted as a function of P_s for various values of crime rate.

CASE II: PESSIMISTIC CASE

Under this assumption, apprehending a criminal does not keep him from committing additional crimes or alter his chances of being caught. Alternatively, this also models the situation where a new criminal replaces each one removed from the system. In this case,

$$\text{Number of failures} = Nn(1 - P_s)$$

$$\text{Number of successes} = NnP_s$$

The success ratio R_s remains the same as the first case.

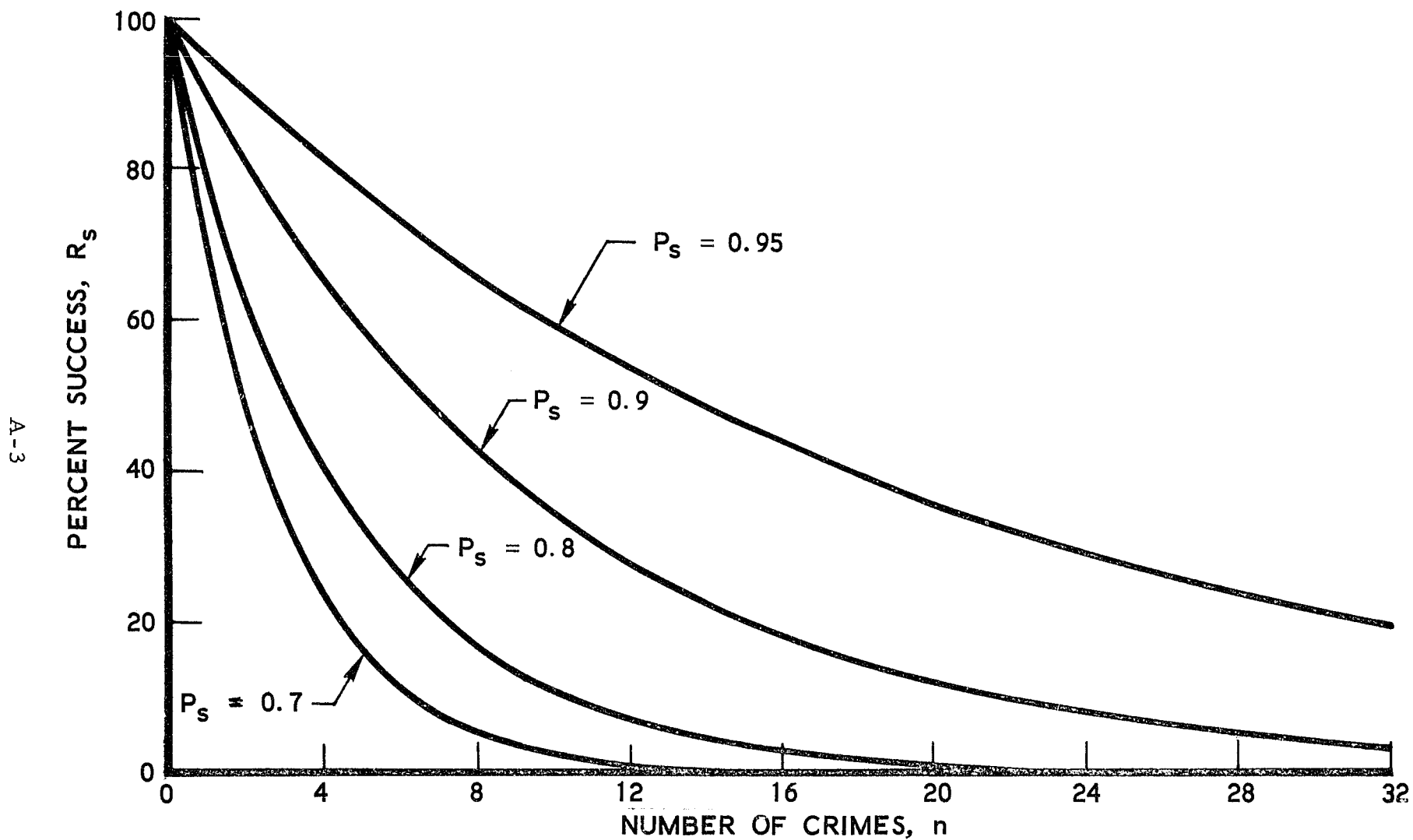


Figure A-1. Success Ratio for Crime Perpetration vs. Number of Crime Events for Various Values of Success Probability

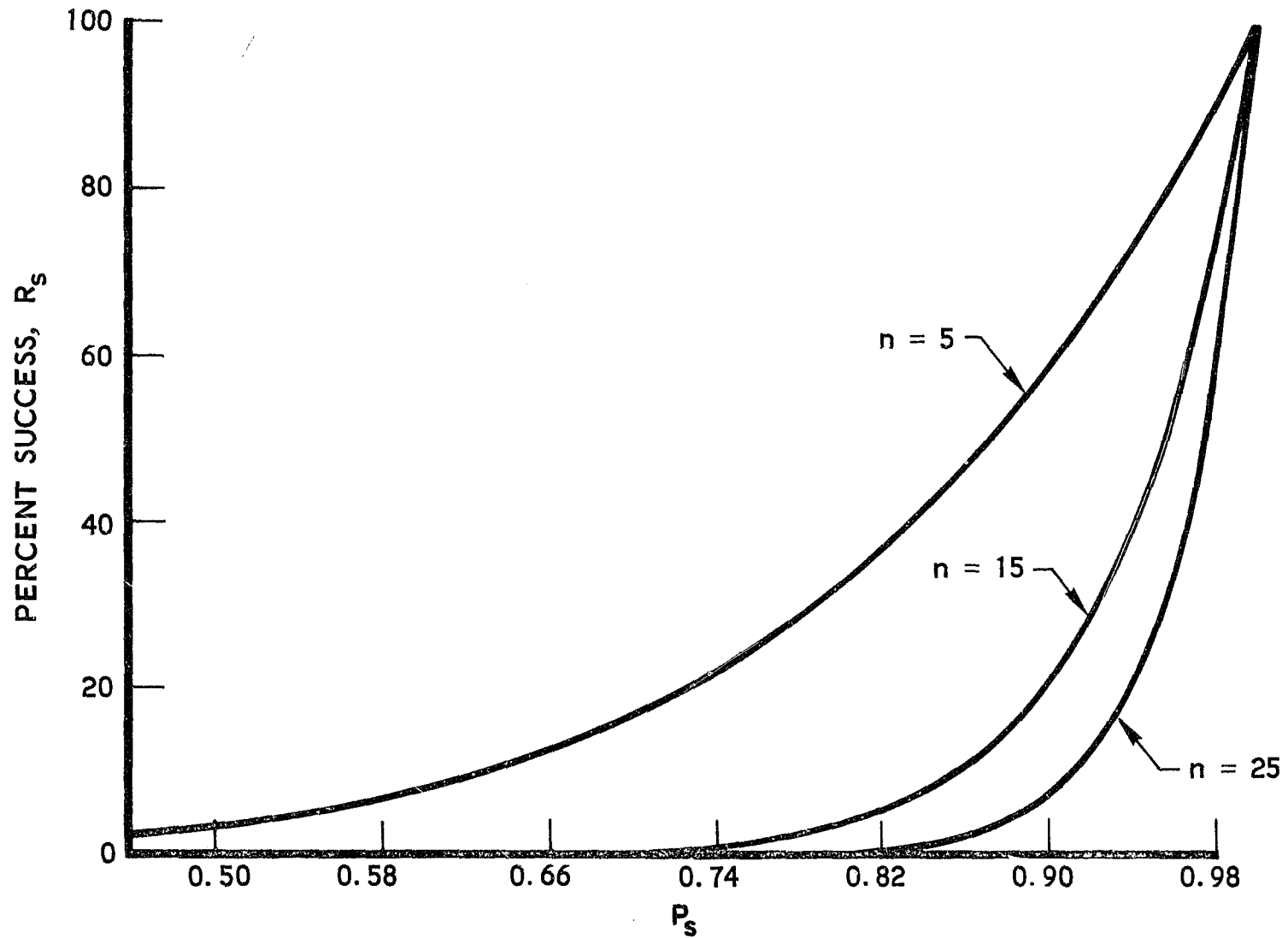


Figure A-2. Success Ratio for Crime Perpetration vs. Success Probability for Various Values of Crime Rate

2. Data Application

Consider the crime of burglary applied to commercial establishments in the five largest cities in the United States for the year 1972. The total number of establishments was 969,000; the total number of completed burglaries was 264,000.

Assume:

1. $P_s = 0.97$ (currently accepted value)
2. Burglar crime rate of (a) $\frac{25 \text{ burglaries}}{\text{yr}}$
 (b) $\frac{15 \text{ burglaries}}{\text{yr}}$
3. Two burglars per burglary (in actuality 70 percent of all burglaries are committed by two or more persons).

Conclude:

- a. Based on the above data and assumptions:
1. Number of burglars (a) $\frac{264,000}{25} \times 2 = 21,120$
(b) $\frac{264,000}{15} \times 2 = 35,200$
2. Percent success (see Figure A-2) (a) 46.7
(b) $63.3 (1 - P_s^n)N$
3. Number of burglars caught (a) $(1 - 0.467) \times 21,120 = 11,257$
(b) $(1 - 0.633) \times 35,200 = 12,918$

b. Effect of adding voice identification capability. Consider the effect of voice identification techniques used to provide additional data in the prosecution of the burglar. If the net effect is to drive the P_s from 0.97 to 0.95, the percentile success changes (see Figure A-2) are:

(a) from 46.7 to 27.3

(b) from 63.3 to 46.3

and the number of burglars caught increases from:

(a) 11,257 to 15,270 or a 35.6 percent increase in apprehensions

(b) from 12,918 to 18,902 or a 46.3 percent increase in apprehensions

Recognizing the significant increase in apprehensions of burglars as indicated above resulting from a 2-percent reduction in P_s , one asks the question: Is a change in P_s from 0.97 to 0.95 possible with voice identification techniques?

Let us assume:

1. All 969,000 establishments have voice recorders installed.
2. At the time of burglary 80 percent of the recorders were working, but only 50 percent of the working recorders recorded any conversation. Further, assume that 50 percent of the recordings with any conversation are sufficient for meaningful voice identification use. Therefore, the number of useful recordings = $264,000 \times 0.80 \times 0.50 \times 0.50 = 52,800$, i.e., 20 percent of the burglaries had useful recordings.

3. For 25 percent of the burglaries committed, one or more suspects are arrested or held for cross examination.
4. The number of useful voice recordings used in cross-examination are related directly to the percent of all burglaries producing one or more suspects, i.e., 20 percent of all burglaries with useful recordings times 25 percent of all burglaries producing suspects = 5 percent of all burglaries where voice evidence can be applied.

As previously mentioned, in the cases where voice identification evidence is used in a trial, convictions result approximately 50 percent of the time. If voice identification could be used in 5 percent of all burglaries, the consequent reduction in the success rate for this crime could be expected to be about 2.5 percent.

APPENDIX B. POTENTIAL CONCEPT FOR ESTABLISHING ARRESTEE VOICE SAMPLE FILE

In obtaining voice samples of arrestees for use in a machine-assisted voice identification system, a number of conditions must be met. These conditions involve the obtaining of material (speech sounds) containing usable events of good quality, spoken in a normal voice. The material used in the speech sample should be (1) simple, (2) should be easy to enunciate, (3) should not be incriminating to the speaker, (4) should contain a maximum number of those sounds useful for voice discrimination, and (5) should be such that an arrestee will not be antagonized by the process.

Table B-1 contains a possible set of sentences which could be used in a speech data file. The examining officer could solicit information from the arrestee in order to complete the sentences and could then request the subject to repeat the sentences. In order to minimize distortion effects, the arrestee could be interrogated in a quiet room with ample furniture to suppress echoes. The arrestee could either use a microphone or a standard telephone headset, and the recording could be made on a good-quality tape recorder.

Table B-1. Possible Voice Sample Sentences

1. My full name is _____.
|MX AA IX | FX UX LX | NX EH IX MX | IX ZX | _____.
2. I live at _____.
|AA IX | LX IX VX | AH TX | _____.
3. I was born in _____.
|AA IX | WX UH ZX | BX AW RX NX | IX NX | _____.
4. My occupation is _____.
|MX AA IX | AA KX UU PX EH IX SH UH NX | IX ZX | _____.
5. I am working for _____.
|AA IX | AH MX | WX ER KX IX NG | FX AW RX | _____.
6. I am _____ married.
|AA IX | AH MX | _____ | MX AH RX EE DX | .
7. I have _____ children.
|AA IX | HX AH VX | _____ | CH IX LX DX RX EH NX | .
8. My father's name is _____.
|MX AA IX | FX AA DH ER ZX | NX EH IX MX | IX ZX | _____.
9. My mother's name is _____.
|MX AA IX | MX UH DH ER ZX | NX EH IX MX | IX ZX | _____.
10. I have _____ been arrested before.
|AA IX | HX AH VX | _____ | BX EH NX | AH RX EH SX TX EH DX |
|BX EE FX AW RX | .

11. I have been living in _____ since _____ .

|AA IX | HX AH VX | BX EH NX | LX IX VX IX NG | IX NX |
|SX IX NX SX | _____ .

12. The full recording of my voice will be used for identification.

|DH UH | FX UX LX | RX EE KX AW DX IX NG | UH VX | MX AA IX |
|VX AW IX SX | WX IX LX | BX EE | JX UX ZX DX | FX AW RX |
|AA IX DX EH NX TX IX FX IX KX EH IX SH UH NX | .

13. The police department will choose the parts of my speech to keep on file.

|DH UH | PX OU LX EE SX | DX EE PX AA RX TX MX EH NX TX |
|WX IX LX | CH UU ZX | DH UH | PX AA RX TX SX | UH VX | MX AA IX |
|SX PX EE CH | TX UU | KX EE PX | AH NX | FX AA IX LX | .

END