Acoust. Soc. Amer., vol. 46, pp. 442–449, Aug. 1969.
[35] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367–377, Dec. 1972.
[36] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pt. 2, pp. 637–655, Aug. 1971.
[37] F. Itakura and S. Saito, "An analysis-synthesis telephony system based on maximum likelihood method," *Electron. Commun. Japan*, vol. 53A, pp. 36–43, 1970.
[38] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.
[39] J. D. Markel, A. H. Gray, Jr., and H. Wakita, "Linear prediction of speech: Theory and practice," Speech Communications Res. Lab., Santa Barbara, CA, SCRL Monogr. 10, Sept. 1973.
[40] R. W. Schaefer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, pp. 634–648, Feb. 1970.
[41] J. Olive, "Automatic formant tracking by a Newton-Raphson technique," *J. Acoust. Soc. Amer.*, vol. 50, pt. 2, pp. 661–670, Aug. 1971.
[42] J. D. Markel, "Digital inverse filtering—a new tool for formant trajectory estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 129–137, June 1972.
[43] S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 135–141, Apr. 1974.
[44] Lo-Soun Su, K. -P. Li, and K. S. Fu, "Identification of speakers by use of nasal coarticulation," *J. Acoust. Amer.*, vol. 56, pp. 1876–1882, Dec. 1974.
[45] K. P. Li, G. W. Hughes, and A. S. House, "Correlation characteristics and dimensionality of speech spectra," *J. Acoust. Soc. Amer.*, vol. 46, pt. 2, pp. 1019–1025, Oct. 1969.
[46] G. R. Doddington, J. L. Flanagan, and R. C. Lummis, "Automatic speaker verification by nonlinear time alignment of acoustic parameters," U. S. Patent 3, 700, 815, issued Oct. 24, 1972.
[47] A. E. Rosenberg, "Evaluation of an automatic speaker verification system over telephone lines," *Bell Syst. Tech. J.*, to be published in 1976.
[48] A. E. Rosenberg, "Automatic speaker verification systems: a review," this issue, pp. 475–487.
[49] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, Feb. 1975.
[50] T. Y. Young and T. W. Calvert, *Classification, Estimation, and Pattern Recognition.* New York: American Elsevier 1974, pp. 24–26.
[51] M. R. Sambur, "Speaker recognition and verification using linear prediction analysis," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, Sept. 1972.
[52] A. E. Rosenberg and M. R. Sambur, "New techniques for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, Apr. 1975.
[53] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 221–226, June 1968.
[54] J. R. Ragazzini and G. F. Franklin, *Sampled-Data Control Systems.* New York: McGraw-Hill, 1958.
[55] S. Furui and F. Itakura, "Talker recognition by statistical features of speech," *Electron. Commun. Jap.*, vol. 56A, pp. 62–71, Nov. 1973.
[56] J. E. Miller, "Decapicitation and recapicitation, a study of voice quality," *J. Acoust. Soc. Amer.*, vol. 36, p. 1876 (A), Oct. 1964.
[57] Reference 13, pp. 24–49.
[58] F. McGehee, "An experimental study in voice recognition," *J. Gen. Psychol.*, vol. 31, pp. 53–65, 1944.
[59] G. L. Holmgren, "Physical and psychological correlates of speaker recognition," *J. Speech Hearing Res.*, vol. 10, pp. 57–66, 1967.
[60] W. D. Voiers, "Perceptual bases of speaker identity," *J. Acoust. Soc. Amer.*, vol. 36, pp. 1065–1073, June 1964.
[61] P. D. Bricker and S. Pruzansky, "Effects of stimulus content and duration on talker identification," *J. Acoust. Soc. Amer.*, vol. 40, pp. 1441–1449, June 1966.
[62] K. N. Stevens, C. E. Williams, J. R. Carbonell, and B. Woods, "Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material," *J. Acoust. Soc. Amer.*, vol. 44, pp. 1596–1607, Dec. 1968.
[63] F. R. Clarke and R. W. Becker, "Comparison of techniques for discriminating among talkers," *J. Speech Hearing Res.*, vol. 12, pp. 747–761, 1969.

Automatic Speaker Verification: A Review

AARON E. ROSENBERG, MEMBER, IEEE

Abstract—**The relation of speaker verification to other pattern-recognition problems in speech is discussed, especially the distinction between speaker verification and speaker identification.**

The prospects for automatic speaker verification, its settings and applications are outlined. The techniques, evaluations, and implementations of various proposed speaker recognition systems are reviewed with special emphasis on issues peculiar to speaker verification. Two large-scale operating systems using different analysis techniques and applied to different settings are described.

I. Introduction

PATTERN-RECOGNITION problems are among the most challenging and fascinating areas in speech research. The speech pattern recognition facility of human beings is remarkable. It is easily taken for granted and is only appreciated when one attempts to make machines perform similar tasks. Some of the speech-pattern-recognition problems of current interest are speech recognition, speaker recognition, language identification, diagnosis of speech pathologies, and even characterizing emotional state and attitude by voice analysis. Of these, by far the most attention has been given to speech recognition. This problem has had as much fascination and potential payoff for the speech researcher as the conversion of lead to gold had for the alchemist. (With intelligent circumscription of the problem, the chances of success for the speech researcher seem much better than those of the alchemist (cf. papers by Martin, Reddy, and Jelinek in this issue).)

Speaker recognition has also received a great deal of attention among speech researchers. It seems to be a problem which is an order of magnitude less difficult than speech

Manuscript received September 15, 1975; revised October 31, 1975.
The author is with Bell Laboratories, Murray Hill, NJ 07974.

recognition because it is not necessary to extract the message-bearing information from the speech waveform to be able to characterize the speaker of the message. It is nevertheless a difficult and challenging problem and like speech recognition has more chance of success when it is intelligently circumscribed. The speaker-recognition[1] problem in general is treated in the paper by Atal in this issue. The particular circumscription of the speaker-recognition problem which has already achieved a modicum of success and which has a considerable potential practical payoff is speaker verification.

The more general problem, speaker identification, may be stated as follows. Out of a total population of N speakers, find that speaker whose reference pattern is most similar to the sample pattern of an unknown speaker. Since the sample pattern is compared to each of the N reference patterns and since there is a finite probability of an incorrect decision for each comparison, it is apparent that the overall probability of an incorrect decision must be a monotonically increasing function of N.

The speaker-verification problem may be stated as follows. The sample pattern of an unknown speaker together with a claimed identity is given. Determine whether the sample pattern is sufficiently similar to the reference pattern associated with the claimed identity to accept that claim. In this case just one comparison of patterns is required regardless of the size of the population. Thus the probability of an incorrect decision is generally independent of the population size.

It is possible to treat the distinction between speaker verification and speaker identification more formally and arrive at this same conclusion. However, the assumptions generally made are that the impostor is drawn from the same population as the true speaker and that all speakers are equally likely to be an impostor. In fact, it may be that a speaker with a high probability of acceptance is more likely to be an impostor, i.e., a determined mimic, and that the *a priori* probability of such a speaker grows as the population of speakers is allowed to increase.

With the exception of the above *caveat*, it is seen that the speaker verification is a more tractable problem than the more general speaker identification problem in that the error rate is independent of the population size. Also, only a single comparison of unknown pattern to reference pattern and a simple accept or reject decision is required. This allows faster computation and less complexity than the compound comparisons and decisions required for speaker identification.

Doddington [13] has provided a formal treatment of the distinction between speaker identification and speaker verification including a simulation for normally distributed speaker measurements. Expected error rate is computed as a function of population size. An example of these computations is shown in Fig. 1.

Thus speaker-verification systems are theoretically capable of handling large as well as small speaker populations, with limitations only with regard to storage capacity and speed of access to reference patterns. For this reason there has been a good deal of impetus recently towards developing commercially practical speaker-verification systems. Some of the possible applications include banking and credit authorizations, access to secure information or premises, and carrying out transactions from remote locations by telephone or other voice communication links.

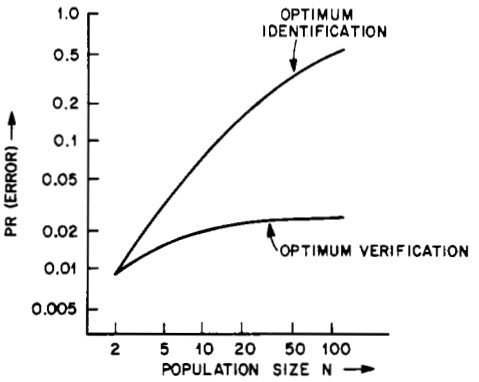

Fig. 1. Expected error rate for speaker verification and speaker identification as a function of population size; results from a computer simulation by Doddington [13].

This paper deals almost exclusively with automatic techniques for speaker verification. There are two other techniques for speaker verification which are dependent on subjective abilities. These are auditory comparison of speakers and visual comparison of spectrograms or other speech features which can be displayed. Visual comparison of spectrograms (voiceprint analysis), is especially controversial because of the wide claims of applicability put forth by its advocates and the large variation in performance found in various evaluations of the technique. References to both auditory and visual techniques will be found in the bibliography.

A basic assumption for most automatic speaker verification systems is that the speakers in the customer set are cooperative. That is to say, they make no overt attempts to alter their speaking behavior from trial to trial. With this restriction, present automatic speaker verification methods have little forensic application.

II. Approaches to Speaker Verification

Much of the theory and techniques of speaker verification is common to speaker identification. These are adequately covered in the paper by Atal in this issue. However, we will review them again here with emphasis on techniques and issues peculiar to speaker verification.

One dichotomy for speaker-recognition systems is whether there is a text prescribed for speakers to utter or not. In most investigations of automatic speaker-recognition systems there is a prescribed text. In those systems which are stated to be text independent there are still some restrictions on the nature of the material, usually the length [3], [16], [23], [30]. Another class of systems, although strictly speaking not text independent, are often stated to be potentially text independent if the speech material used as input is general enough to contain certain specified speech events which can be extracted for analysis [11], [19], [46], [50], [52]. In this class of systems a certain amount of recognition and segmentation of speech events is required prior to analysis for purposes of speaker recognition.

Text independent speaker recognition, with restrictions on the length and context of the speech material, is the operating mode for speaker recognition by audition or by visual analysis of spectrograms. In such cases the speakers are not required to be cooperative and in many forensic applications, in fact, are not expected to cooperate.

[1] The term speaker recognition in this paper refers to the general category of problems which includes speaker identification, speaker verification, speaker classification, etc.

As indicated in the introduction, for many applications the speakers are expected to be cooperative so that a prescribed text is perfectly feasible. Text-dependent systems are the subject of most investigations in speaker recognition and are the ones which are the closest to practical implementation in a speaker verification mode. The prescribed texts are usually sentence-long utterances or isolated words.

A. Speech Signal Processing

The most frequently used preprocessor for speaker recognition systems is the filter bank. The filter bank, widely used in other areas of speech analysis, is a series of contiguous bandpass filters spanning the useful frequency range of speech. The output of each filter is rectified, smoothed, and sampled say every 10 ms to obtain the energy of the signal over the specified passband. The summed output of the filter bank can be considered as the short-time power spectrum of the input speech signal. Filter banks can provide a highly efficient and comprehensive analysis of speech signals. Although digital representations of filter banks are potentially computationally superior and provide reproducible measurements, analog configurations provide an on-line analysis of speech which can be input to a digital computer for further processing. The tandem combination of analog filter bank and digital computer is capable of providing a fast responding speaker recognition system. A disadvantage of filter bank analysis is an inherent inflexibility since all subsequent processing is tied to combining measurements of the individual filter outputs. Pitch and formant analysis, for example, is difficult to achieve from filter bank processing alone. Very often filter bank processing is used in combination with other analyses such as pitch, formant, and overall energy [11], [52]. The few systems that do not rely on filter-bank techniques generally make use of pitch, intensity, formant, and LPC analysis [2], [32], [44], [46].

B. Analysis Techniques

Analysis and possibly feature selection follow the basic processing. Analysis in this paper is considered to be the process in which the raw measurements, from a filter bank for example, are combined or restricted to certain speech segments in such a way as to reduce the dimensionality of the original measurement space while at the same time preserving or enhancing speaker discriminability according to a prescribed system design. It is the central process in a system, the process which makes the most of the designer's intuitive ideas or theoretical knowledge. It can be considered a feature selection process. But in this paper feature selection is considered to be a statistical process in which a subset of the entire set of analyzed features which is most useful in discriminating speakers is selected by a statistical technique.

In many filter bank systems, analysis does not proceed beyond the processing of raw spectral data. In these systems spectral energy matrices are operated on directly by statistical decision techniques to carry out the speaker-recognition tasks [5], [29], [38], [39]. The simplest treatment of the raw spectral data by prescribed design is the formation of long-term spectral averages by summing and averaging the spectral energies over the utterance time interval [5], [16], [30], [41]. Other transformations of filter bank data include conversion to cepstral measurements [31] and extraction of formants [12], [32], [52].

Perhaps the most sophisticated and potentially powerful reductions of raw analysis data are those in which some segmentation of the input speech is required. There are two goals for speech segmentation. The first goal, where prescribed texts are used as speech input, is to precisely align corresponding speech events in reference and test utterances. This alignment allows the comparison of equivalent events and is required to compensate for expected variations in the occurrence of speech events in the repetition of a given utterance by a given speaker. This type of segmentation is best described as time registration. Some successful automatic time registration techniques have been investigated [12], [13], [26], [32], [43].

The second goal for segmentation is to locate and isolate acoustic features or events which insight and investigation have shown to be powerful discriminants for speakers. Ideally, as Wolf [52] remarks, such cues should occur naturally and frequently in normal speech, be easily measurable, not change over time or be affected by the speaker's health, background noise or transmission characteristics, and not be consciously modifiable by the speaker. Two studies [46], [52], investigated a large set of extracted acoustic features and tested their relative effectiveness. Feature extraction also played a central role in the investigations of Das and Mohn [11] and Hair and Rekieta [21], [22]. Although the extraction of specific speech events for speaker recognition purposes is potentially very powerful, segmentation is an extremely difficult problem. Most of the investigations have used manual segmentation schemes. An automatic segmentation scheme was reported by Das and Mohn [11], but it is quite complex. In addition it was not entirely reliable since 10 percent of the test utterances were rejected because of segmentation failure.

As a specific example, the segmentation and analysis of nasal consonants has attracted considered attention due to the conjecture that the acoustical properties of nasal consonants are strongly speaker dependent and that there is little movement of the articulators during phonation [19], [46], [50], [52]. Most of these investigations measured spectral characteristics of individual nasal consonants. However, one investigation carried this type of analysis one step further (Su *et al.* [50]). It was hypothesized in this study that differences in the spectra of a particular nasal consonant due to coarticulation with following vowels is also strongly speaker dependent and even less likely to be subject to conscious modification.

Another distinct mode of analysis is the transformation of raw data into functions of time or contours. Most often the speech input is prescribed sentence-long utterances. The hypothesis is that the time functions of many acoustic features have strong speaker dependent characteristics. Some of the features that have been investigated as time functions are pitch, intensity, formants, and predictor coefficients [2], [12], [32], [42], [44].

C. Statistical Feature Selection and Decision Techniques

Statistical feature selection and decision techniques are important elements of any speaker recognition scheme. The most common statistical feature selection technique is that of the F-ratio or analysis of variance [2], [5]. For this purpose several statistics are computed over the set of training of reference utterances provided by each speaker. These include the mean feature vector and covariance matrix of the elements of the feature vector. To compute the F-ratio a between-talkers covariance matrix B, which is a measure of the disper-

sion of the feature vector means, and a within-talkers covariance matrix W, which is the average of the individual talker covariance matrices, are computed. The F-ratio given by

$$F = a^t Ba / a^t Wa$$

is maximized to find a set of eigenvectors $a_1, a_2, \cdots$, which linearly transform the original feature vector into a space in which the speaker feature vector means are maximally separated relative to the within-talkers covariance. As a special case, if the correlations between the components of the feature vector are ignored, i.e., if the off diagonal elements in B and W are set equal to zero, the F-ratio provides an ordering of the separability between speaker means in terms of the original individual coordinates [11], [21], [22], [39], [52]. A subset of these coordinates with the largest F-ratio is selected as the new feature vector. Mohn [34] describes the distinction between the general technique and the special case.

A different approach to statistical feature selection is to select those features which from processing a set of training or reference utterances through the designated speaker recognition system provide the lowest error rate [44], [46]. If the original feature vector has a relatively high dimensionality, it is important to adopt a systematic and efficient procedure for investigating as many possible lower order subsets of this vector. One such approach is that of the "knockout" tournament [44], [46].

Decision techniques are all based on the computation of a distance which quantifies the degree of dissimilarity between the feature vectors associated with pairs of utterances. There are many distance metrics that have been investigated. Let r_i be the ith reference vector and x an unknown test vector. The most common distance metric is the simple Euclidean distance.

$$\|x - r_i\| = ((x - r_i)^t (x - r_i))^{1/2} = \left(\sum_j (x_j - r_{ij})^2 \right)^{1/2}$$

where x_j and r_{ij} are the jth components of x and r_i, respectively. Other possibilities are $\Sigma_j |x_j - r_{ij}|$, and $x \cdot r_i / \|x\| \, \|r_i\|$ which is the cosine of the angle between x and r_i. There have been several investigations in which the various distance metrics have been compared [8], [16], [30].

The simplest decision rule is that of the "nearest neighbor." For speaker identification this means that distances are calculated from the unknown vector to each reference vector and the speaker corresponding to the minimum distance is designated the identified speaker. That is if $D(x, r_{jn})$ represents the distance between an unknown test vector and the vector corresponding to the nth utterance of the jth speaker we select J if

$$\min_{j,n} D(x, r_{jn}) = D(x, r_{JU})$$

where r_{JU} is the vector corresponding to a particular utterance U of speaker J. The nearest neighbor principle has been used by several investigators [16], [31], [38]. For speaker verification, the minimum distance for the claimed identity J, $D(x, r_{JU})$, is compared against a threshold to decide whether to accept or reject.

At the next level of sophistication explicit account is ken of the distribution of the feature vectors for each class. For example, prototype vectors can be calculated by computing the mean feature vector for each speaker

$$m_j = \frac{1}{N} \sum_n r_{jn}.$$

The nearest neighbor procedure is then the determination of

$$\min_j D(x, m_j).$$

This decision is obviously more efficient than the more general search over both speakers and utterances, but it assumes that the set of m_j are good prototypes, i.e., that the r_{jn} form well-behaved clusters by speakers. This nearest neighbor prototype approach has been used by Bunge [8], Furui *et al.* [16], Glenn and Kleiner [19], and Hair and Rekieta, [21], [22]. Still another level of sophistication is added to this approach by weighting the components of the feature vector inversely by the calculated variances of these components over each individual speaker's utterances. This weighting has the effect of giving more influence in the distance calculation to those components of the feature vector which are more strongly clustered. This approach has been used in [8], [16], [21], [22], [32], [39], [42]–[44], [52].

Since this latter technique provides a weighting of the components of the feature vector it is, in effect, providing statistical feature selection. In fact, the feature vector can be weighted by the inverse of the entire covariance matrix of the feature vector over each individual speaker's utterances provided the number of utterances available is sufficiently larger than the number of components (Furui *et al.* [16]). Another possible weighting is the inverse of the pooled covariance matrix over all speakers W, if the individual covariance matrices are not too disparate [3], [16], [46]. Finally, consideration can be given to the between-speakers covariance matrix B as well as W using the weighting provided by the F-ratio procedure described earlier.

The techniques discussed above all provide a linear weighting of the feature vector using moments of the distribution of reference or training vectors. Another class of procedures establishes linear weightings of the feature vectors by so-called nonparametric techniques, such as ADALINE or linear threshold elements (Nilsson). The weights are calculated by operating iteratively on the set of training feature vectors so that the vectors associated with each speaker are optimally separated by hyperplanes which are specified by the weights. Nonparametric decision procedures have been investigated by Das and Mohn [11], Hargreaves and Starkweather [23], and Li *et al.* [29].

An important extension of decision techniques is the use of sequential strategies. In a sequential strategy, each identification or verification trial is transformed into a series of trials. At each trial there is an option to defer decision to the next trial if the distance associated with the trial is too close to a predetermined decision threshold. The procedure terminates on the trial in which a decision is made. In practice, a limit is placed on the number of trials, with a special decision category if the limit is exceeded. A sequential procedure is useful if trial-to-trial measurements are not strongly correlated. Sequential strategies are especially useful in on-line speaker verification implementations. Doddington has demonstrated the effectiveness of such strategies in reducing error rate [13]. Some of these results are described later in this paper.

III. Evaluations and Implementations

In this section, we will briefly describe the evaluation and results of some of the systems whose techniques are sketched in the previous section. We will also discuss some of the factors which affect the evaluation and implementation of systems.

A. Large Populations

There are not many systems in which a significantly large sample of speakers and utterances has been used for an evaluation, or in which the collection and recording of the sample has been accomplished under "real world" conditions. Most studies, in fact, can be considered preliminary investigations of particular speaker-recognition techniques with no special claims to be "real-world" systems operating outside the laboratory. Since a goal of this paper is to emphasize practical systems, particular attention is paid in the next section to two systems which come closest to operating in the "real world". These are the Texas Instruments entry control system which at this date has made over 150 000 verifications over a period of a year on a population of 180 users and the Bell Labs telephone system which made over 4500 verifications over a five-month period on a population of approximately 100 users.

Other large test populations have been reported by Bricker *et al.* [5], Das and Mohn [11], and Hair and Rekieta [21], [22]. The population tested by Bricker *et al.* consisted of 172 talkers each of whom recorded five replications of digit names in a single session. The recordings were made with high quality equipment but in a somewhat uncontrolled environment. In the speaker identification experiment which was carried out accuracies as great as 94 percent could be achieved if information from more than one word per speaker is included in the analysis.

Das and Mohn made use of a population of 118 male talkers including 50 "customers" who provided 20 replications of a test phrase in five sessions held once per week and 68 impostors who provided 20 replications in a single session. These were high-quality recordings made in a sound booth. In the speaker verification experiment carried out by these experimenters an equal-error rate of 1 percent was obtained although decisions could not be made on 10 percent of the utterances submitted.

Hair and Rekieta utilized single-word recordings from approximately 40 speakers, male and female, collected in eight or nine sessions held once per week. These were also high-quality recordings obtained in a sound booth. Equal error rates of less than 1 percent were reported. An unreported number of utterances were excluded due to "bad colds."

B. Intersession Variability and Reference File Maintenance

In addition to the desirability of testing a large population of speakers there are other considerations to be examined with regard to the experimental data base. One of the most important considerations is the time period over which utterances are collected and the methods used to establish and maintain reference patterns over this period. Intersession variability for a given speaker was recognized as a significant effect in some of the earliest investigations reported [11], [23], [31]. In the speaker verification experiment reported by Luck [31], it was found necessary to include speech samples taken over a 5-week period in the reference data for an adequate representation of the speaker's voice. The problem of long-term variation has been examined in some detail in the investigations of Furui *et al.* [16]–[18]. With their data base they were able to examine the effects of intervals of up to 18 months between samples. In [16], the best identification and verification rates were obtained when the reference data was calculated from 4 samples taken at intervals of 3 months. The data base used by Sambur also extended over a long period—up to 3.5 years. The only feature reported by Sambur to be strongly affected by this interval was average pitch.

In any practical speaker verification system the problem of maintaining reference files over extended periods of time is an immediate concern. It is quite clear that a verification system which requires several reference or training samples collected over an extended period of time before the system can be accessed for actual use has little practical value for most applications. This problem and methods to deal with it are discussed again in the descriptions of the Texas Instruments an Bell Labs systems.

C. Speaker Characteristics

The composition and characteristics of the speaker population are other important considerations. Many evaluations have included only male talkers. (Two investigations, Atal [2] and Hargreaves and Starkweather [23], employed only female talkers.) The difficulties associated with analysis of female speech are well known. The fundamental problem is the loss of spectral resolution compared with analysis of male speech.

Temporary and chronic speech irregularities are important characteristics of a speaker population and must be considered in the light of each type of analysis. For example, nasal congestion associated with upper respiratory disorders is likely to modify the spectral characteristics of many sounds especially nasal consonants. Laryngeal inflammation can have a profound effect on pitch measurements. Diplophonia, a condition associated with a husky or raspy voice quality, will also disturb pitch measurements. Any overt speech pathology is likely to have a serious effect on almost every type of analysis. Especially severe effects can be expected from one of the most common types of pathology, that of stuttering.

There is also a subtle change in speaking behavior to be expected when speakers are removed from the formal experimental environment of a sound booth. In their normal home or work environment speakers are likely to lose their experimental "set" or attitude. They may become less attentive, impatient, easily distracted or, generally speaking, less "cooperative."

D. Recording Environment and Transmission Conditions

Factors which can loom large over the implementation of a speaker-recognition system are the recording environment and the conditions governing the transmission of the speech signal to the processor. Until recently these factors have received little attention. Most evaluations have been carried out in the hothouse atmosphere of the sound booth and high-quality recordings. Eventually, however, one must consider whether these conditions represent a fair approximation to conditions that are expected in a practical application. For an application such as entry control as in the Texas Instruments system, this may be the case. Even here, however, one might find it difficult to control such conditions as angle of incidence and distance from the microphone and certain kinds of background noise.

The most severe effects can be expected when a recognition transaction is carried out over remote communication links where it is possible that either or both the recording environment and the transmission conditions will vary from transaction to transaction. The most obvious situation of this type is recognition transactions over dialed-up telephone lines. This is the principal setting proposed for the Bell Labs speaker-verification system, and its evaluation over dialed-up telephone lines is described in a later section. Dialed-up telephone lines have been used in the evaluation of one other speaker verification system, that of Li *et al.* [29]. In the design of such systems, careful allowance should be given to the effects of background noise and room reverberation at the source and the reduced bandwidth, distortion, and line disturbances over a dialed-up line. In addition, the possibility of variation of these conditions from one transaction to another must be considered. Spectral amplitude measurements of a critical nature are especially susceptible under such conditions and for that reason are avoided in the Bell Labs system. Some spectral normalization techniques have been suggested to compensate for some of these conditions. Doddington [13] and Glenn and Kleiner [19] normalize the spectral data vector obtained from their filter bank output by the sum of the filter outputs for each measurement. This normalization has the effect of stabilizing the measurements with respect to overall changes in signal level. Furui [18] has employed an inverse filtering technique. Long-time spectral measurements calculated over a given speaker's speech sample are used to construct a second-order inverse filter which characterizes the gross spectral distribution of the input signal. This filter is applied to the instantaneous input signal of the speaker. This approach has also been used successfully by Itakura [26] in a speech-recognition system. This type of normalization will tend to cancel out the broad characteristics of the source and radiation spectrum as well as that of the transmission system.

E. Mimic Resistance

An important question which must be answered for a speaker recognition system is how well can the system resist the effects of determined mimics. A related question is, can the system discriminate among closely related members of a family such as siblings, especially indentical twins. These questions are especially important for speaker-verification systems since the intended applications of these systems are usually ones in which there could be a large payoff if the system is successfully defrauded.

Four mimic investigations have been reported. The most extensive investigation was undertaken at Bell Labs by Lummis and Rosenberg [33]. Four professional mimics were employed to mimic eight customer voices. The initial acceptance rate for these mimics was high but a considerable degree of mimic protection was achieved in later versions of the system. The Bell Labs mimic experiment is described in more detail in a later section. Doddington [13] employed two of these mimics to test the Texas Instruments system. They provided imitations of 6 speakers. Doddington, too, found that mimic acceptance was significantly greater than acceptance of casual impostors.

A mimic experiment was reported by Hair and Rekieta [22]. A professional performer specializing in impersonations provided imitations of 6 speakers. Although some increase in similarity was observed for individual features in their verification scheme, the attempt at acceptance when all features were combined was unsuccessful. Luck [31] also performed a mimic experiment. The mimics were not professionals but were selected from the population of casual impostors used in the evaluation. Three mimics were afforded the opportunity to imitate a single speaker out of the customer set. Only the mimic who was the worst casual impostor significantly improved his performance (but not enough to be accepted).

The evaluation of a system's resistance to mimics depends on the definition of a skilled mimic. Since the skills required of a mimic vary from system to system and since subjective impressions can be quite faulty, the most reliable definition of a skilled mimic is not based on a priori appraisal. It is simply one who can significantly increase his acceptance as an impostor by deliberate imitation. This makes mimic evaluation difficult since one cannot be certain in advance who is a skilled mimic. One generalization seems clear however. This is that mimics who depend on caricature for their imitations are poor candidates since most effective systems will not tolerate exaggeration of speaker characteristics.

The systems that are more likely to be mimic resistant are those which make strongly physiologically correlated measurements rather than measurements correlated with behavior or learned characteristics. This is discussed in the description of the Bell Labs systems since many of its measurements are correlated with behavior. On the other hand, systems whose measurements are physiologically oriented may not be able to discriminate among close family members. An experiment with an identical twin impostor in the Bell Labs systems showed that even though the voices were physiologically identical, the twin impostor was not able to imitate his brother's speaking behavior well enough to be accepted by the system even when he was given every opportunity to do so. More study of identical twins is merited.

IV. Detailed Large-Scale Systems Descriptions

A. Texas Instruments Entry Control System

A practical method of automatic speaker verification has been under development at Texas Instruments under the direction of Doddington. The speech input is sentence-long utterances and the analysis provides spectral amplitude information on precisely located sections of the utterance. A block diagram of the analog processing is shown in Fig. 2. Each bandpass filter is composed of two-pole-pair sections with an approximate 220-Hz bandwidth. The center frequencies are spaced uniformly between 300 and 3000 Hz. The low-pass filters cutoff at 20 Hz. The filter outputs are sampled 100 times per second and digitized. (The most recent version uses a digital filter bank with 14 instead of 16 channels.) Each filter output is quantized to 3 bits with the levels chosen so that each is approximately equiprobable. For reference data storage in the most recent versions an additional 4 bits per filter output is required for adaptation purposes so that each time sample requires $14 \times 7 = 98$ bits. Succeeding stages of processing take place in a digital computer.

Within each test utterance several key points are specified to provide time registration and for data comparison. In the most recent implementation the test material consists of 4-word phrases each word being a CVC monosyllable. The key points are chosen at regions of maximum energy in the vowel portion of each word. A reference pattern consists of 6 spectral amplitude samples taken every 20 ms through a 100-ms interval around the key point. Thus each phrase is associated

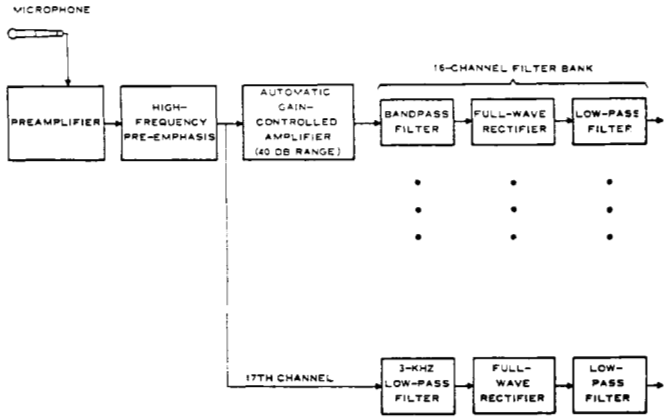

Fig. 2. Functional block diagram of the analog processing of the Texas Instruments system [13].

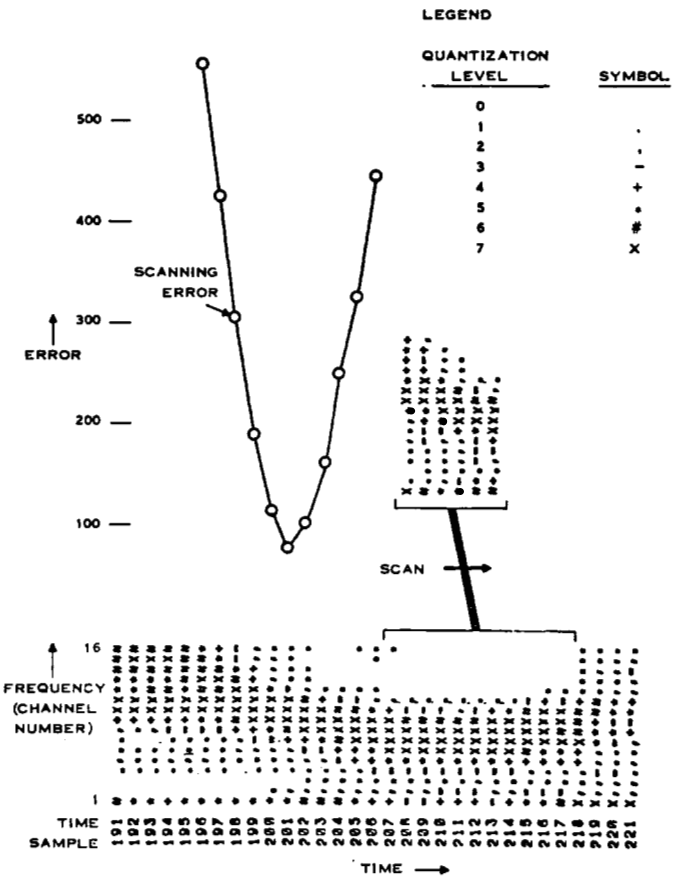

Fig. 3. Key point location by error function scanning [13].

with four reference patterns, each represented by $6 \times 98 = 588$ bits of information.

In the verification process each reference pattern is scanned across the spectral amplitude sample vectors of an incoming test phrase to locate the corresponding key points in the test phrase. In the scanning process a squared error difference is computed every 10 ms between the incoming spectral data pattern and each corresponding reference pattern. An error function is obtained for each reference pattern and the minimum of each error function determines the location of each key point in the incoming utterance. This scanning process is illustrated in Fig. 3. Once all the key points are located in the input sample utterance no further measurements are necessary. The decision function is simply the average scanning error at these points normalized by an estimate of the expected scanning error for the speaker under consideration. If all the key points cannot be located in an input phrase the strategy is to request an additional phrase. The estimate of the expected scanning error is computed during a series of initial training sessions. (In the most recent implementation a single enrollment session is required.)

Doddington has reported two evaluations. In both evaluations the data was collected in a sound booth using a dynamic microphone. In the first evaluation 10 females and 40 male "customers" provided 100 recording sessions of five phrases over a span of 2 months. Approximately 70 impostors, with approximately the same female–male ratio, provided 20 sessions each. Although a verification response was provided during data collection the results reported by Doddington represent an evaluation from digital recordings after the recording sessions were completed. In this evaluation the first 50 sessions were designated training sessions.

Using the recordings from the 70 member impostor population and customer recordings from the 50 test sessions a verification equal-error rate of approximately 1.6 percent was obtained for a single phrase with one phrase held in reserve. In this strategy, in the event that the key points in the primary phrase cannot be registered a second phrase is scanned to locate key points. If registration fails for both phrases the speaker is unconditionally rejected. The number of unregistered customer utterances is very small, less than 1 percent, but the number of unregistered impostor utterances is quite high, approximately 50 percent.

The error rate decreases dramatically with the use of 2 or more phrases. The decision function is simply the average of the decision function for each phrase used. For 2 phrases the equal error rate falls to 0.42 percent, for three phrases the error rate falls to 0.23 percent.

A sequential decision strategy using only as many phrases as are necessary to achieve a given level of confidence is a more efficient multiple-phrase strategy. In the sequential decision strategy examined by Doddington the speaker is accepted after the first phrase if the decision function is less than some threshold. Otherwise additional phrases are requested combining the decision functions at each stage. For this strategy the customer reject rate was established at 1 percent and the concurrent accept impostor rate was measured as 0.01 percent. The average number of phrases required of customers was less than 1.5.

Doddington tracked the average customer decision function (or reject rate) over the 6-week period of the test and observed a distinct increasing trend over the entire period as seen in Fig. 4. It is obvious then that some sort of periodic reference file adaptation is required as the customer accesses the system.

The influence of upper respiratory infections was also investigated. In 5.5 percent of the true speaker trials, the speakers reported that their voices were "not normal." At the decision level for which normal voice speakers were rejected at a 1-percent rate, "not normal" speakers were rejected at a 2-percent rate. The increase in rejection rate is significant but not alarming since the probability of acceptance is still quite high.

Also, two of the professional mimics used in the Bell Labs study were engaged to attempt to defeat the system. Each mimic chose 5 male speakers to mimic after listening to recordings from a selection of 35. The mimics were then pro-

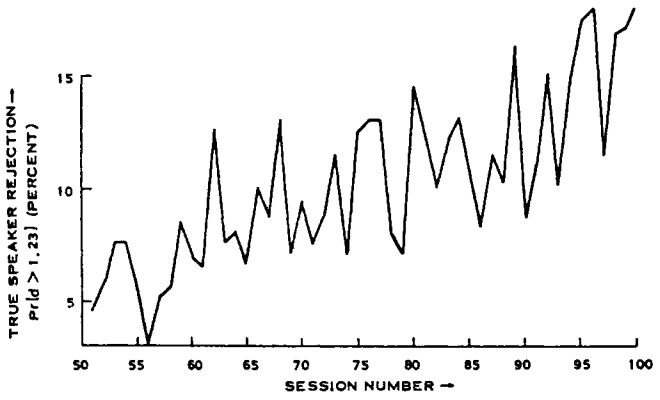

Fig. 4. True speaker rejection rate as a function of session number (adjusted to a nominal 10-percent reject rate).

vided with repetitive recordings of each of his 5 chosen speakers and were also allowed to listen to each voice in person. The mimic trials were performed on-line with immediate feedback for degree of success. For a one-phrase decision function it was found that the probability of mimic acceptance was approximately twice that of casual impostors.

In the second evaluation reported by Doddington the speaker set consisted of 63 male speakers who over the span of one month provided an average of 25 verification sessions. The impostor set consisted of 2 sessions each from 60 male speakers. The sample utterances consisted of 16 monosyllabic words organized in a set of 4-word sentence-like nonsense phrases. There were 32 allowable phrases in the set. At each trial the speaker was instructed by means of voice prompting to utter a phrase selected at random. A list of the words used to compose the phrases is given in Table I.

The initial training for this evaluation consisted of an enrollment session in which enough phrases were collected so that each of the 16 words was presented 5 times. A reference pattern was constructed for each word based on the location of key points from at least five repetitions of each word. Initial training was extended by 4 post-enrollment sessions to establish an adapted set of reference patterns and to estimate speaker variability.

Each verification session consisted of the utterance of 4 test phrases. Adaptation was allowed if the decision function averaged over all 4 phrases was less then some threshold. Adaptation consisted of refreshing the reference pattern by averaging it with a small fraction of the corresponding input sample pattern. The estimated scanning error and the timing between key points in the phrase were also adaptively updated.

For this evaluation the equal error rate obtained for a single phrase was approximately 4 percent. Again there is marked improvement with the use of more than one phrase. With two phrases the error rate decreased to 1.5 percent and with the use of all 4 phrases the error rate is 0.5 percent. With the use of a sequential strategy in which an average of 1.3 phrases is required for verification the true speaker reject rate is 0.3 percent with an accompanying 1 percent accept impostor rate.

It was found that scanning error when plotted as a function of session number decreases more or less uniformly to an asymptote after an initial elevation. This finding is similar to the trend for the threshold function and error rate for the Bell Labs system. This is shown in Fig. 5 for a subset of 17 speakers who completed at least 30 sessions. Doddington attributes this trend to the use of session-by-session adaptation of reference patterns. This behavior is in contrast to the trend shown in Fig. 4 where no adaptation was employed.

TABLE I
MONOSYLLABIC WORDS USED TO COMPOSE 4-WORD NONSENSE PHRASES (DODDINGTON [13])

Cool	birds	stopped	west
Small	bugs	sing	down
Huge	twigs	sang	deep
Strange	toads	stood	wild

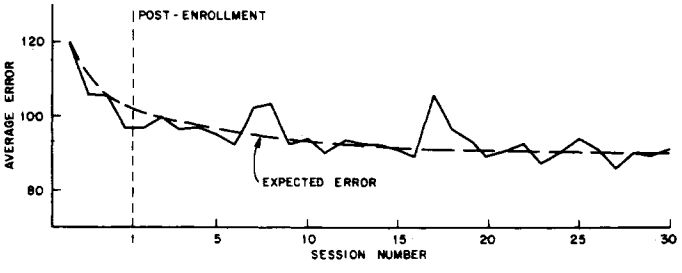

Fig. 5. Average scanning error a function of session number when reference patterns are adapted session by session.

At this writing, Doddington has installed an operational entry control system for use at a Texas Instruments computer facility. The verification data from this installation are available for evaluation. Approximately 180 users, 13 percent female, make an average of more than 400 entries per day. Using a 4-phrase sequential decision strategy, an impressive customer reject rate of 0.3 percent is achieved with a 1-percent impostor accept rate. (The impostor accept rate is calculated from off-line cross-comparisons.) The average number of utterances required per transaction is 1.3. The average verification time is 5.8 s.

In summary, the Texas Instrument automatic speaker verification system achieves a high level of performance making efficient use of computer processing with respect to both execution time and storage. The fast execution time allows the comfortable use of a multiphrase sequential decision strategy which provides impressive error rates.

The use of a filter bank preprocessor to provide the basic measurements on the speech signal is largely responsible for the fast processing times and high level of accuracy. It seems likely, however, that to maintain the effectiveness of such measurements a relatively large signal-to-noise ratio (Doddington suggests 30 dB) and uniform transmission conditions are required. The system in its present configuration is particularly well suited to the principal application for which it was designed—access control. In this type of application the speaking environment and transmission conditions are the same from trial to trial. A sound booth is used to provide a uniform low-noise speaking environment.

B. Bell Labs Systems

The Bell Labs automatic speaker verification system has been under development since 1970 under the aegis of several investigators. The original implementation was carried out by Doddington, presently at Texas Instruments, for his doctoral dissertation based on a suggestion by J. L. Flanagan [12]. It has since been studied and improved by Lummis [32], Sambur, and Rosenberg [42]-[44].

From the outset it was hoped to provide a system which would tolerate variable and degraded environmental and trans-

mission conditions as might be experienced by use of the system over dialed-up telephone lines. Perhaps the most important distinguishing characteristics of the Bell Labs system are the choice and method of measurements which were deemed most likely to meet these goals. Measurements are chosen which are largely insensitive to the phase and/or spectral amplitude distortions likely to be encountered in the telephone plant. Examples of measurements that have been studied are pitch, overall intensity, formant frequencies and linear predictor coefficients. The measurements are calculated as a function of time over prescribed sentence-long utterances so that the emphasis is on the variations of these features rather than on absolute measurements at fixed instants of time or averages taken over long intervals of time. To a large extent, then, measurements obtained in this manner are strongly correlated with the inflectional or prosodic behavior of the speaker. To the extent that absolute measurements of pitch, formants, and predictor coefficients are obtained the analysis is partially correlated with physiological characteristics. There is no question, however, that there is a strong dependence in this system on the analysis of speaking behavior simply because this information is relatively slowly time varying, phase insensitive, and broad band.

A simplified block diagram of the Bell Labs system is shown in Fig. 6. A marked interval is provided for input of a sentence-long sample utterance. The input signal is digitized at a 10-kHz sampling rate. For pitch and intensity analysis the input is low-pass filtered typically at 900 Hz. For predictor coefficient or formant analysis the input is subject to 3000-or 4000-Hz low-pass filtering. For each feature analyzed a measurement is made every 10 ms (or 20 ms for predictor coefficients). Calculations on the intensity function are used to delimit the end points of the actual utterance within the recording interval. The time functions for each analyzed feature, subjected to 16-Hz low-pass smoothing, between the delimited end points are designated contours and are the basic patterns for the verification process. A typical set of contours is shown in Fig. 7. The intensity or gain contour is normalized to the peak intensity measurement resulting in a contour of relative intensity values. The pitch contour in the more recent implementations is extracted by means of the time-domain parallel processing technique of Gold and Rabiner [20]. Predictor coefficient measurements are obtained by the covariance method (Atal and Hanauer [1]). Following analysis, the set of sample contours are compared with a set of reference contours associated with the claimed identity. The reference contours are obtained by averaging and combining sets of contours previously obtained from sample utterances of the person whose identity is claimed. Before comparing the sample and reference contours an additional operation, time registration, is carried out. In this operation, the events of the sample contours are brought into the best possible time registration with the corresponding events of the reference contours. This operation attempts to compensate for normal expected variations in speaking behavior which may displace corresponding events from repetition to repetition of a given utterance. The intensity contour is used as the guide contour for this operation. In the latest implementation a dynamic programming algorithm is utilized. First, the sample intensity contour is linearly stretched or compressed to the standardized length of the reference intensity contour. Then a distance is calculated between the ith point of the sample contour and the jth point of the reference contour for each i and j. The dynamic programming algorithm is used to find the path of least accumulated distances through the matrix of distances (d_{ij}). The optimal path establishes the warping function required to replot the sample contour registered to the reference contour. The warping function obtained for the guide intensity contour is applied to all other contours analyzed for the sample utterance. Time registration of the intensity contour is illustrated in Fig. 8.

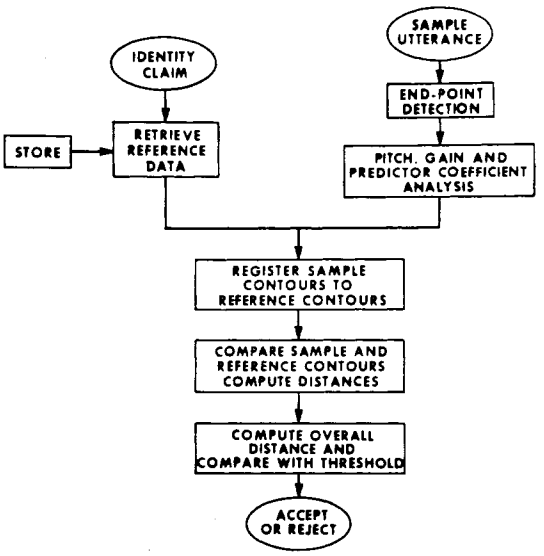

Fig. 6. Functional block diagram of the Bell Labs system.

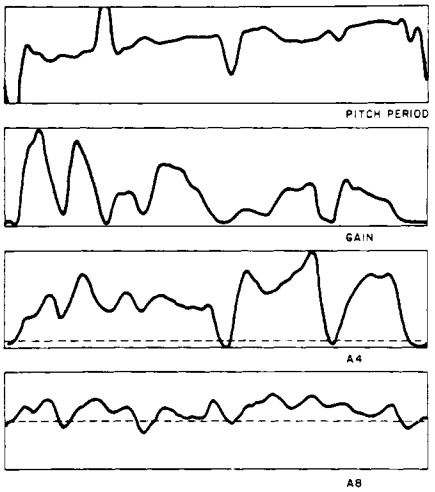

Fig. 7. A typical set of sample contours plotted as a function of time through a sample utterance. Included are pitch period, gain, and A4 and A8, the fourth and eighth coefficients of an eighth-order linear predictor.

Following registration, the contours are divided into 20 equal length segments as shown for intensity in the bottom panel of Fig. 8. In each segment, a set of measurements is applied to both the sample and reference contours and a squared difference is calculated specifying the dissimilarity between the contours for a particular measurement. The squared difference for each measurement and segment is weighted inversely by a variance which is calculated from the set of sample contours used to construct the reference. The effect of the variances is to weight those segments in which a particular measurement is

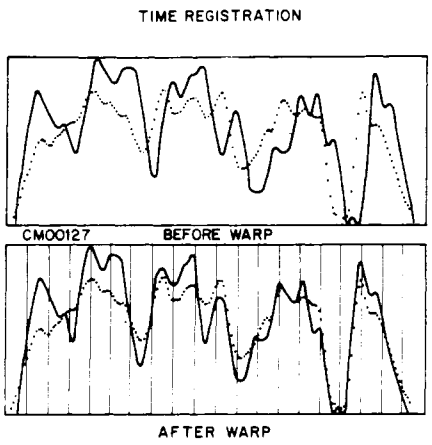

Fig. 8. Contour time registration. The dotted curve in each panel is a reference gain contour; the solid curve is a sample gain contour. The top panel shows the contours with end-points aligned, but before internal registration. The bottom panel shows the results of registering the sample contour to the reference contour.

more consistent over the set of sample contours. The various segment-by-segment measurements characterize the shape of the contours. In addition, there are distances based on the overall cross correlation of sample and reference contours and distances based on the amount of warping required to register the sample contours to the reference contours. Detailed descriptions of the measurements and distances can be found in [32], [42]-[44].

An overall distance is obtained by a simple average over the entire set of individual distances or over a subset of these distances. In recent implementations, speaker-dependent subsets have been calculated which select those measurements which are most effective in separating the overall distance distributions of customer and impostor sample utterances for each customer. The overall distance is compared with a speaker-dependent threshold distance to determine whether to accept or reject the identity claim. The threshold distance is estimated from the overall distance distributions of a set of customer sample utterances and a set of impostor sample utterances. For the purposes of evaluation an equal-error threshold is estimated.

Four distinct evaluations of the system have been carried out, the original evaluation and succeeding ones associated with major modifications. In addition, a study was carried out to determine the sensitivity of the system to the effects of professional mimics and a parallel study was undertaken to examine the ability of human listeners in a speaker verification task using the same speech samples used to evaluate the original system.

The first evaluation was carried out by Doddington [12] using recordings collected from a population of 40 male speakers. High-quality recordings were obtained from the speakers seated in a sound booth. At least two days elapsed between consecutive utterances. Eight of the speakers were designated customers and provided 15 utterances each. The remaining 32 speakers were designated impostors, providing just one utterance each.

The feature analysis employed by Doddington was a complete formant analysis (Schafer and Rabiner [47]) resulting in 3 formant contours plus a pitch and intensity contour. For each customer 3 reference files were constructed from 3 different subsets of 10 of the 15 sample utterances. The remaining 5 utterances for each customer were reserved for test utterances together with the 32 impostor utterances. An overall average equal error rate of approximately 1.5 percent was obtained. The importance of time normalization was underscored by the result that the error rate increased by a factor of 4 if only end-point alignment of sample and reference utterances were performed. Lummis provided some modifications of the system including time registration based on the intensity contour instead of the second formant contour. Using the same data base of utterances essentially the same average error rate of the order of 1 or 1.5 percent was obtained. Lummis also demonstrated that the distances extracted from the formant contours seemed to provide a negligible contribution to the performance of the system.

Two parallel studies were undertaken using the same 40-speaker data base. In the first study [33], professional mimics were subjected to intensive listening and training sessions prior to providing sample recordings. Mimic utterances provided by the four best mimics were processed by the system in the same manner as the original set of utterances. With threshold distance set for a customer reject rate of approximately 1.5 percent, it was found that mimic utterances were accepted at a 27-percent rate, compared to the 1.5-percent rate for casual impostors. In this study, in contrast to the earlier observations, it was found that formant information was a significant factor in the performance of the system. It is worth noting that the mimicking required a great deal of skill. The group of four chosen for evaluation were culled from an initial group of some 25. None of the mimics in the final group was primarily an entertainer. In fact, any mimic utterance that smacked of caricature was readily rejected by the system. In addition, one of the original eight customers had an identical twin. Even when this twin was afforded the same opportunity which the mimics had, to listen and rehearse his brother's utterances, his sample utterances were rejected by the system.

The second study undertaken in parallel with the original evaluation investigated the ability of human listeners to perform a speaker verification task [42]. The subjective tests were of the paired-comparison type in which each test presentation consisted of a challenge and comparison utterance. Listeners were required to respond whether the challenge and comparison utterances were from the same or different speakers. The comparison utterances were all drawn from either the set of impostor utterances or from a distinct set of utterances provided by the same speaker of the challenge utterance. Human listeners performed the speaker verification task with an average error rate of approximately 4 percent, with the false-alarm rate (customer rejection) about equal to the miss rate (impostor acceptance), a significantly higher rate than the automatic system. When the professional mimic utterances were included in the test presentations the miss rate increased to 22 percent, about the same level of performance as the automatic system.

A third modification and evaluation was undertaken by Rosenberg and Sambur to investigate the possibility of replacing the rather difficult and time consuming analysis of formants and improving the effectiveness of the system with respect to the efforts of professional mimics. In addition, a new and larger data base was obtained to provide a more statistically reliable evaluation of performance. The new data base was extracted from 50 recording sessions for each of 22 male speakers designed customers and one recording session for each of 55 male speakers designed impostors. The recording sessions were held over a period of two months with no more

TABLE II
ERROR RATES (ROSENBERG AND SAMBUR [44])

	Pitch + Intensity	Pitch + Intensity + 2 Pred. Coeff.
All measurements (equal-error)	6%	4%
Selected (speaker dependent) Measurements (equal-error)	3%	1.5%
Mimic acceptance rate (selected measurements, old data base)	16%	4%

than two sessions per customer per day. Two sentences were processed from each recording session, the all-voiced sentence "We were away a year ago" used in the previous evaluation and another all-voiced sentence, "I know when my lawyer is due."

Linear predictor coefficient analysis was chosen to replace the formant analysis because of its relative simplicity of computation. In a sense, formant analysis can be considered as an extension of linear predictor analysis, but for the purpose of obtaining speaker characterizing features there is no essential reason to extend the analysis to the computation of formants. An eight-order predictor analysis was employed with the fourth and eighth coefficients plotted as a function of time to provide contours. (Contours provided by two widely spread predictor coefficients were found to provide effective speaker discrimination.) In addition, a speaker-dependent measurement selection procedure was investigated which selected those measurements which best discriminated each customer from imposters. The results of this evaluation are summarized in Table II.

"All measurements" refers to the speaker-independent measurement technique of earlier evaluations. Distances calculated from these measurements provide an error rate of the order of 4 percent, somewhat greater than what was obtained in the earlier evaluations. The speaker-dependent selected-measurement technique provides error rates of the order of 1.5 percent. In both cases the inclusion of predictor coefficients in the analysis results in a significant reduction in error over the use of just pitch and intensity. The mimic data base was reprocessed using these techniques and in this case there was a striking reduction in error with the inclusion of predictor coefficients. Mimic acceptance was reduced to the tolerable level of 4 percent.

The fourth evaluation of the system has brought access of the system and recording of utterances into the "real world" [42]. Over 100 "customers" accessed the system and recorded their test utterances from their own telephones via ordinary dialed-up lines to a data set interfaced with a computer. Identity claims were made by keying in an identification number on a Touchtone® dial. Instructions and responses to the customer were made by means of a programmed voice-response system (Rosenthal *et al.* [45]). To provide a more tractable system for this evaluation, only pitch and intensity were included in the analysis. Initial reference files were established for each customer by processing a set of 5 sample utterances collected in a single session. Following this initial session, the customers were instructed to call the system nominally once each work day for a verification trial. The data for each verification trial in which the customer's identity was accepted was used to periodically refresh the customer's reference file and adaptively update the accept/reject thresholds.

Over the five-month period of evaluation an average of approximately 50 trials per customer were collected. Tabulation of errors over all the customer trials was used to calculate the customer reject rate. The accept impostor rate was calculated by tabulation of errors in periodic off-line cross comparison of customer sample files with the reference files of other customers of the same sex. Overall average error rates of 9 or 10 percent were obtained for both reject customer and accept imposter. (Estimated equal-error thresholds were used.) It was found that both the reject customer and accept impostor error rates were high for the customer's initial calls to the system, of the order of 15 percent, and declined to stable values of the order of 4 percent after a settling-in period of some 20 trials.

Thus when the customer's speaking behavior is stabilized and when the reference file has been sufficiently adapted by periodic refreshing, the error rate is of the order of 4 percent, which is approximately the benchmark achieved in the previous laboratory evaluation for pitch and intensity analysis. Although an extensive survey of the telephone plant environment was not possible in this evaluation it was encouraging to find no special adverse effects on performance due to either telephone transmission or customer speaking environment.

In agreement with Doddington, the results indicated that a considerable reduction in error rate can be obtained with a deferred decision strategy.

The current activity for this system includes reintroduction of LPC analysis for telephone evaluation and the integration of the system into generalized man–machine communication by speech systems.

V. CONCLUSION

An immediate conclusion of this paper is that speaker verification is emerging as a practical technology. That the future will bring new and improved systems is certain. There is a ready market for practical systems in many areas where it is desired to extend and automate valuable informational and transactional services to larger and larger groups of customers, while at the same time ensuring that these services are provided only to those who are authorized to receive them.

However, one should not lose sight of the fact that there are still large gaps in our knowledge with respect to speaker recognition techniques. The most important issue, and likely to be a long-standing one, is the specification of the most effective measurements to apply to the speech signal for the purposes of speaker discrimination. This really is a two-part issue since an intrinsically effective feature may prove to be impractical if it cannot be reliably extracted from the speech signal, especially when confronted with variable or degraded recording environments and/or transmission links. For example, measurement of nasal coarticulation as proposed by Su *et al.* [50], is a very sophisticated approach to speaker discrimination but the measurement may be difficult to carry out reliably in a practical system. Thus there are two needs. The first need is for more insight and more theoretical study on intrinsically effective speaker discrimination measurements. The second need is for more thorough investigations of the effect of recording environment and transmission characteristics on proposed speaker discrimination measurements to ensure a good match between the type of analysis and the type of application for the system.

Another issue remaining is the choice of statistical feature selection and decision techniques. Pattern-recognition theory is well advanced so that there is a good selection of these techniques from which to choose. The question is how sophisticated must these techniques be to be effective. A good bet is that if the basic analysis provides measurements which are sufficiently reduced in dimension and suitably normalized, then the statistical operations can be quite simple and computationally effective.

An issue which has not been discussed at all in this paper is that of the economic feasibility of speaker verification. A central concern here is the cost of customer reference data storage. Reference file storage for the Texas Instruments system has been considerably optimized and requires approximately 1200 bytes per customer [14]. There exists a 90-Mbyte disk pack drive with a price tag of approximately $30 000. Such a pack could serve some 75 000 customers at a cost of 40¢ per customer. Processing costs are harder to estimate since it depends on how many transactions are required per unit time, the time required for each transaction, multiplexing capabilities, the cost of special purpose hardware, and transmission costs.

In another paper in this issue, Flanagan has described the integration of automatic speaker verification into a 3-mode man–machine voice communication system. The other modes in this system are computer voice response and automatic recognition of words spoken in isolation by designated speakers. Although the system is primitive in comparison with the sophistication of the HAL-3000 of "2001" fame, it is an augury of things to come. The goal of communication with machines by means of man's most natural mode of communication–speech–is enormously significant. Automatic speaker verification will play an important role in this new technology. We can imagine that if automatic voice dialing becomes a reality, the same voice-dialing utterances might be used for authorization for billing!

References

[1] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.

[2] B. S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoust. Soc. Amer.*, vol. 52, pp. 1687–1697, 1972.

[3] —, "Effectiveness of linear prediction characteristics of the speech waves for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, 1974.

[4] R. H. Bolt, F. S. Cooper, E. E. David, P. B. Denes, J. M. Pickett, and K. N. Stevens, "Speaker identification by speech spectrograms: some further observations," *J. Acoust. Soc. Amer.*, vol. 54, pp. 531–534, 1973.

[5] P. D. Bricker, R. Gnanadesikan, M. V. Mathews, S. Pruzansky, P. A. Tukey, K. W. Wachter, and J. L. Warner, "Statistical techniques for talker identification," *Bell Syst. Tech. J.*, vol. 50, pp. 1427–1454, 1971.

[6] P. D. Bricker and S. Pruzansky, "Effects of stimulus content and duration on talker identification," *J. Acoust. Soc. Amer.*, vol. 40, pp. 1441–1449, 1966.

[7] —, "Speaker recognition," in *Contemporary Issues in Experimental Phonetics*, N. J. Lass, Ed. Springfield, IL: Charles C. Thomas, 1975.

[8] E. Bunge, "Automatic speaker recognition by computers," in *Proc. Carnahan Conf. Crime Countermeasures*, 1975.

[9] R. Carré, "A summary of speech research activities in France," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-22, pp. 268–272, 1974.

[10] R. O. Coleman, "Speaker identification in the absence of inter-subject differences in glottal source characteristics," *J. Acoust. Soc. Amer.*, vol. 53, pp. 1741–1743, 1975.

[11] S. K. Das and W. S. Mohn, "A scheme for speech processing in automatic speaker verification," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 32–43, 1971.

[12] G. R. Doddington, "A method of speaker verification," Ph.D. dissertation, Univ. Wisconsin, Madison, 1970.

[13] —, "Speaker verification–Final report," Rome Air Development Center, Griffiss AFB, NY, Tech. Rep. RADC 74-179, Apr. 1974.

[14] —, "Speaker verification for entry control," presented at WESCON, 1975.

[15] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd ed. New York: Springer-Verlag, 1972.

[16] S. Furui, F. Itakura, and S. Saito, "Talker recognition by long-time averaged speech spectrum," *Electron Commun. Jap.*, vol. 55-A, pp. 54–61, 1972.

[17] S. Furui and F. Itakura, "Talker recognition by statistical features of speech sounds," *Electron. Commun. Jap.*, vol. 56-A, pp. 62–71, 1973.

[18] S. Furui, "An analysis of long-term variation of feature parameters of speech and its application to talker recognition," *Electron. Commun. Jap.*, vol. 57-A, 1974.

[19] J. W. Glenn and N. Kleiner, "Speaker identification based on nasal phonation," *J. Acoust. Soc. Amer.*, vol. 43, pp. 368–372, 1968.

[20] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, pp. 442–448, 1969.

[21] G. D. Hair and T. W. Rekieta, "Automatic speaker verification using phoneme spectra," *J. Acoust. Soc. Amer.*, vol. 51, p. 131(A), 1972.

[22] —, "Mimic resistance of speaker verification using phoneme spectra," *J. Acoust. Soc. Amer.*, vol. 51, p. 131(A), 1972.

[23] W. A. Hargreaves and J. A. Starkweather, "Recognition of speaker identity," *Language and Speech*, vol. 6, pp. 63–67, 1963.

[24] M. Hecker, "Speaker recognition: An interpretive survey of the literature," *ASHA Monogr.* 16 (Amer. Speech and Hearing Assoc., Washington, DC, 1971.

[25] Information Center for Hearing, Speech, and Disorders of Human Communication, *Bibliography: Speech identification by eye, ear, and machine*, Johns Hopkins Medical Inst., Baltimore, MD, 1975.

[26] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, 1975.

[27] L. P. C. Jansen, "Speech identification by means of some relations between the parameters of vowel sounds," National Physical Research Laboratory, CSIR, Pretoria, South Africa, Rep., 1973.

[28] L. G. Kersta, "Voiceprint identification," *Nature*, vol. 196, pp. 1253–1257, 1962.

[29] K. P. Li, J. E. Dammann, and W. D. Chapman, "Experimental studies in speaker verification using an adaptive system," *J. Acoust. Soc. Amer.*, vol. 40, pp. 966–978, 1966.

[30] K. P. Li and G. W. Hughes, "Talker differences as they appear in correlation matrices of continuous speech spectra," *J. Acoust. Soc. Amer.*, vol. 55, pp. 833–837, 1974.

[31] J. E. Luck, "Automatic speaker verification using cepstral measurements," *J. Acoust. Amer.*, vol. 46, pp. 1026–1031, 1969.

[32] R. C. Lummis, "Speaker verification by computer using speech intensity for temporal registration," *IEEE Trans. Audio Electroacoust.* vol. AU-21, pp. 80–89, 1973.

[33] R. C. Lummis and A. E. Rosenberg, "Test of an automatic speaker verification method with intensively trained mimics," *J. Acoust. Soc. Amer.*, vol. 51, p. 131(A), 1972.

[34] W. S. Mohn, "Two statistical feature evaluation techniques applied to speaker identification," *IEEE Trans. Computers*, vol. C-20, pp. 979–987, 1971.

[35] K. O. Mead, "Identification of speakers from fundamental frequency contours in conversational speech," Joint Speech Research Unit, Rep. 1002, 1974.

[36] National Academy of Sciences-National Research Council, Committee on Hearing, Bioacoustics, and Biomechanics, Research on speaker verification, Report of working Group 53, 1971.

[37] N. J. Nilsson, *Learning Machines*. New York: McGraw-Hill, 1965.

[38] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," *J. Acoust. Soc. Amer.*, vol. 35, pp. 354–358, 1963.

[39] S. Pruzansky and M. V. Mathews, "Talker-recognition procedure based on analysis of variance," *J. Acoust. Soc. Amer.*, vol. 36, pp. 2041–2047, 1964.

[40] S. Pruzansky and B. A. Stevens, "Speaker recognition, 1937–1973," Bibliography 266, Bell Lab., 1974.

[41] G. S. Ramishvili, "Experiments on automatic verification of speakers," in *Proc. 2nd Int. Joint Conf. Pattern Recognition* (Copenhagen, Denmark), pp. 389, 393, 1974.

[42] A. E. Rosenberg, "Listener performance in speaker verification tasks," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 221–225, 1973.

[43] —, "Evaluation of an automatic speaker verification system over telephone lines," to be published in *Bell Syst. Tech. J.*, 1976.

[44] A. E. Rosenberg and M. R. Sambur, "New techniques for automatic speaker verification," *IEEE Trans. Acoustics, Speech Signal Processing*, vol. ASSP-23, pp. 169–176, 1975.

[45] L. H. Rosenthal, L. R. Rabiner, R. W. Schafer, P. Cummiskey,

and J. L. Flanagan, "A multiline computer voice response system utilizing ADPCM coded speech," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-22, 339-356, 1974.
[46] M. R. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-23, pp. 176-182, 1975.
[47] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, pp. 634-648, 1970.
[48] J. N. Shearme and J. N. Holmes, "An experiment concerning the recognition of voices," *Language and Speech*, vol. 2, p. 123, 1959.
[49] K. N. Stevens, C. E. Williams, J. R. Carbonell, and B. Woods, "Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material," *J. Acoust. Soc. Amer.*, vol. 44, pp. 596-1607, 1968.
[50] L-S Su, K-P Li, and K. S. Fu, "Identification of speakers by use of nasal coarticulation," *J. Acoust. Soc. Amer.*, vol. 56, pp. 1876-1882, 1974.
[51] O. Tosi *et al.*, "Experiment on voice identification," *J. Acoust. Soc. Amer.*, vol. 53, p. 2030, 1972.
[52] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 51, pp. 2044-2055, 1972.
[53] M. A. Young and R. A. Campbell, "Effects of context on talker identification," *J. Acoust. Soc. Amer.*, vol. 42, pp. 1250-1254, 1967.

Practical Applications of Voice Input to Machines

THOMAS B. MARTIN, MEMBER, IEEE

Abstract—**Voice input to machine is the most natural form of man-machine communications. In this type of system the machine responds to the mode of communications preferred by the user, rather than vice versa. Many practical applications exist today for limited capability voice input systems. The first operational voice input systems have taken place with limited vocabulary, isolated word voice input systems. Most of these initial systems were for industrial applications in which the users' hands or eyes were already busy with their normal work requirements. Future developments in both new applications and increased capability voice input systems can be expected to considerably expand the usage of this form of man-machine communications.**

I. Introduction

SPEECH RECOGNITION is the ultimate step towards simplifying communications between man and machine. It is the process whereby a human operator can use ordinary spoken commands that can be recognized and interpreted by an automatic speech recognition (ASR) system. Historically, man's communications with machines or computers have been according to the operational requirements of the machine. To control machines or computers required learning the "language" of the machine or the manipulation of special dials or keys in the proper sequence and format. Any deviation from this unnatural machine language produced errors which were not easily detectable because of the complexities of the rules for man-machine communications.

The development of limited capability voice recognition systems has made it possible for the first time for humans to "talk" information directly into a computer, with no intermediate keying or handwritten steps involved, or to control mechanical systems with voice commands. Input to machines is simplified since the operator provides instructions in his natural language. The machine, therefore, adapts to the requirements of the human and greatly simplifies the task of man-machine communications.

The purpose of this paper is to describe practical limited vocabulary recognition systems. The discussion will be narrowed to a consideration of operational performance capabilities that can be achieved with limited-vocabulary ASR systems and of what practical value are such systems. It is hoped that this discussion will provide a base line of insight upon which future developments can be referenced.

II. Types of Voice Recognition Systems

All ASR systems can be considered as belonging to one of two categories: continuous (connected) speech systems or isolated (discrete) speech systems. The differences between these two types of systems can become obscure and overlapping when attempting to classify a particular approach as either isolated or continuous. Isolated speech systems can be defined as those systems that require a short pause before and after utterances that are to be recognized an entities.

The minimum duration of a pause that separates independent utterances is on the order of 100 ms. Anything shorter than 100 ms can be confused with the closure of stop consonants in the midst of continuous speech that can produce stop gaps approaching 100 ms in duration. In actuality, a stop gap can exceed a 100-ms duration. For example, the word "rapid" can be spoken with a relatively long silence interval after "ra-." For a cooperative speaker, however, a 100-ms minimum separation between words is a reasonable compromise value.

The speaking rate that can be achieved with isolated speech recognition systems is naturally much less than for connected

Manuscript received August 1, 1975; revised September 19, 1975.
The author is with the Threshold Technology, Inc., Delran, NJ 08075.