York: Harper & Row, 1968.

[12] G. Fant, "Distinctive features and phonetic dimensions," in *Applications of Linguistics* (selected Papers of the 2nd Int. Congr. Appl. Linguistics, Cambridge, England, 1969). Cambridge, England: Cambridge Univ. Press, 1971.

[13] N. Umeda, "Linguistic rules for text-to-speech synthesis," this issue, pp. 443-451.

[14] N. Umeda and C. H. Coker, "Allophonic variation in American English," *J. Phonetics*, vol. 2-5, pp. 1-5, 1974.

[15] C. H. Coker and N. Umeda, The importance of spectral detail in initial-final contrasts of voiced stops," *J. Phonetics*, vol. 3-1, pp. 63-68, 1975.

[16] J. L. Flanagan, K. Ishizaka, and K. Shipley, "Synthesis of speech from a dynamic model of the vocal cords and vocal tracts," *Bell Syst. Tech. J.*, vol. 54, pp. 485-506, Mar. 1975.

[17] L. R. Rabiner, L. B. Jackson, R. W. Schafer, and C. H. Coker, "A hardware realization of a digital formant synthesizer," *IEEE Trans. Commun. Technol.*, vol. COM-19, pp. 1016-1020, Nov. 1971.

[18] M. M. Sondhi, "Model for wave propagation in a lossy vocal tract," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1070-1075, 1974.

[19] H. K. Dunn, "Methods of measuring vowel formant bandwidths," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1737-1746, 1961.

[20] C. H. Coker, Speech synthesis with a parametric articulatory model," presented at a Talk at Kyoto Speech Symposium, 1968.

[21] C.H. Coker, N. Umeda, and C. P. Browman, "Automatic synthesis from ordinary English text," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 293-298, Feb. 1973.

Automatic Recognition of Speakers ᶠrom Their Voices

BISHNU S. ATAL

Abstract—This paper presents a survey of automatic speaker recognition techniques. The paper includes a discussion of the speaker-dependent properties of the speech signal, methods for selecting an efficient set of speech measurements, results of experimental studies illustrating the performance of various methods of speaker recognition, and a comparison of the performance of automatic methods with that of human listeners. Both text-dependent as well as text-independent speaker-recognition techniques are discussed.

I. Introduction

MOST OF US ARE aware of the fact that voices of different individuals do not sound alike. This important property of speech—of being speaker-dependent—is what enables us to recognize a friend over a telephone. The ability of recognizing a person solely from his voice is known as speaker recognition. Speaker recognition by human listeners is a common experience and has been known for a long time [1], [2]. More recently, with the availability of digital computers, speech scientists have wondered if automatic and objective methods can be devised to recognize a speaker uniquely from his voice [3]-[10]. In many speech applications, it is often difficult to duplicate human performance by machines. In the case of speaker recognition, fortunately, this is not true. Not only successful speaker recognition by machines is possible, presently available experimental evidence suggests that the performance of machines in many instances exceeds that of human listeners [11].

The early work on speaker recognition was almost completely limited to human listening. A considerable part of this research was motivated by the desire to produce natural sounding speech from vocoders and similar speech processing devices. Although vocoder-generated synthetic speech was quite intelligible, it was often deficient with respect to the recognizability of the speakers. This particular problem led to a search for the factors which convey the speaker-dependent information in speech. The principal aim of these studies was not just to determine the accuracy with which listeners could identify speakers but to answer the fundamental question: how do listeners differentiate among speakers? Unfortunately, it is not easy to find a satisfactory answer to this question. The question about perceptual bases of speaker recognition and their acoustical correlates remains largely unresolved.

The availability of digital computers for processing of speech signals probably provided the greatest impetus to research on automatic speaker recognition. The motivation for such research stemmed from both a curiosity to see if human performance could be duplicated by machines and a promise of providing new and indeed revolutionary services in many diverse fields. Efficient banking and business transactions, controlled access of a facility or information to selected individuals, and a new tool for the law-enforcement agencies are among the many possible applications of automatic speaker recognition.

This paper presents a survey of the progress achieved towards automatic speaker recognition.[1] We will review in the remainder of this introductory section some of the fundamental aspects of the problem. Why is speaker recognition possible? What are the sources of interspeaker variability and how are

Manuscript received August 15, 1975; revised October 8, 1975.
The author is with Bell Laboratories, Murray Hill, NJ 07974.

[1] The term speaker recognition as used in this paper refers to any decision-making process that uses some features of the speech signal to determine if a particular person is the speaker of a given utterance which will include tasks such as identification, verification, discrimination, and authentication of speakers.

they manifested at the acoustic level? We will also discuss the differences among the various tasks in speaker recognition: speaker identification versus speaker verification and text-dependent versus text-independent speaker recognition. In Section II, we describe different acoustic parameters of speech and present techniques for selecting an efficient set of parameters for speaker recognition. Speaker-recognition procedures employed in various experimental studies on speaker recognition are presented in Section III. The performance of different speech parameters for speaker recognition is discussed in Section IV. In Section V, some results concerning the performance of human listeners are presented and compared with machine performance. Finally, in Section VI, we present some thoughts on "what have we achieved so far?" and "where do we go from here?"

Speech is produced as a result of a complex sequence of transformations occurring at several different levels: semantic, linguistic, articulatory, and acoustic. In general, differences in these transformations are likely to show up as differences in the acoustic properties of the speech signal. Speaker-related variations in speech are caused in part by the anatomical differences in the vocal tract and in part by the differences in the speaking habits of different individuals [12]. The anatomical differences relate to the fixed structural differences in the shape or size of the vocal tract which can vary considerably from one person to another. The differences in the speaking habits, on the other hand, result from the manner in which persons have learned to use their speech mechanism. Such differences show up in the temporal variations of speech characteristics of different individuals. Intonation patterns of individuals represent a good example of such variations. In speaker recognition, one attempts to exploit both anatomical as well as learned differences to distinguish speech of one speaker from that of another.

So far, we spoke primarily of interspeaker variations in speech. We must also consider intraspeaker variations—those occurring within different speech utterances of a single speaker [13]. Intraspeaker variations occur even though the texts of two utterances are identical. Such variations are caused by many factors such as the differences in the speaking rates, the emotional state of the speaker, his health, etc. The differences in the speaking rates can often be minimized by nonlinear deformation of the relative time scales of two utterances. It is desirable to select for speaker recognition those acoustic parameters of speech which show low intraspeaker but high interspeaker variability.

Often, in many speaker-recognition applications, it is possible to reduce the intraspeaker variability by requiring that the text of the unknown and reference utterances of a speaker be the same. It is still interesting to know if such a restriction is really necessary for successful speaker recognition. After all, human listeners do not need this restriction in order to recognize a person. Indeed, in most practical situations, the spoken text is different from one occasion to another. Although a large number of experimental studies in speaker recognition have dealt with a text-dependent format, a few recent studies have described methods suitable for both the text-independent and the text-dependent situations [14]-[16]. Generally speaking, text-independent speaker recognition is the more difficult of the two, since one must now cope with the additional variability due to the differences in the texts of the unknown and the reference utterances. We will be discussing the results of both kinds of studies in this paper.

One can think of two different applications of speaker recognition. In the first one, speaker identification, the task is to assign an unknown utterance to one person in a group of several known speakers. In the second application, speaker verification, the task is to verify if the unknown utterance was spoken by the claimed speaker. Although these two problems have quite a lot in common, the recognition procedures employed in each problem can be very different. In some sense, the verification procedure is much simpler than that of identification. Verification requires a binary decision, namely, that of accepting or rejecting the claimed identity of an utterance. In practice, it means comparing the unknown utterance with a reference utterance of the claimed speaker and deciding if the two are similar enough. Superficially, the task of identification requires comparisons with reference utterances of all speakers, which is likely to be impractical with a very large number of speakers. For applications involving a large number of speakers, which is of great practical interest, some method which limits the comparisons to only a few reference utterances must be used [7], [17]. For example, by using proper tree-structures algorithms, the number of comparisons involved in the search can be made independent of the number of speakers. Unfortunately, not much work has been done in this respect. Nevertheless, successful development of such techniques is essential for satisfactory speaker identification in large populations.

The main emphasis of this paper is on automatic methods of speaker recognition. We will present some results with respect to the performance of human listeners in situations comparable to those encountered by the machine, thus allowing us to compare the performance of the automatic methods with that of human listeners. This is not to suggest that the human performance should serve as a benchmark of achievement for the automatic methods. On the contrary, we wish to show that automatic methods are, in many cases, more accurate than human listening.

II. Selection of Speech Parameters for Speaker Recognition

A. Desirable Properties for Speaker-Recognition Parameters

The problem of speaker recognition may be divided into two parts: measurement and classification. In the first part, a number of measurements are made on the speech utterance to provide a set of parameters representing the speaker-dependent information of speech. In the second part, appropriate decision rules are used either to assign the measurements to one of the speakers or to verify if the measurements have originated from the speech of the claimed speaker. We will discuss the measurement problem in this section.

A large variety of parameters can be extracted from the speech signal either directly from the waveform or after spectral transformation to the frequency domain [18]. One of the most important steps towards achieving successful speaker recognition is the selection of speech parameters capable of efficiently representing the speaker-dependent information in speech. Procedures for selecting an efficient set of acoustic parameters have been discussed extensively in the literature [19]-[22]. Wolf, in a recent paper [19] on this subject, has outlined a set of desirable characteristics for suitable speaker-recognition parameters. Ideally, the chosen speech parameters should be as follows:

1) efficient in representing the speaker-dependent information;

2) easy to measure;
3) stable over time;
4) occur naturally and frequently in speech;
5) change little from one speaking environment to another; and
6) not be susceptible to mimicry.

B. Parameter Evaluation

Let us consider the problem of selecting a set of suitable speech parameters. Obviously, we must define first a reasonable criterion of effectiveness. We can then rank the different parameters in order of their measured effectiveness. The ideal measure of effectiveness is the probability of error in recognizing a speaker. This choice does, however, create some practical problems. The error probability depends not only on the parameters but also on the nature of classification rules. Thus the error probability cannot be determined without specifying the classification rules. The computing of error probability is generally quite time-consuming, particularly with a large number of speakers. It is, therefore, desirable to choose a measure of effectiveness which reflects the basic properties of the parameters alone in discriminating one speaker from another and is not dependent on the choice of the classification rules. Theoretically, one could avoid this problem by using optimal classification rules. However, it is not a practically sound idea since one would then be required to rank order not only the different speech parameters but also the different classification rules. Of course, the error probability is an ideal measure of effectiveness if the parameters are to be evaluated for a specific classification rule and indeed experimentally determined error probabilities have been used extensively for evaluating the performance of practical speaker-recognition systems.

A set of measurements made on an utterance may be thought of as mapping the utterance into a point in a multidimensional parameter space. Different utterances of the same speaker will generate a set of points in the parameter space whose distribution can be described by a multivariate probability density function. Roughly speaking, a set of measurements would be effective in discriminating between speakers if the distributions of different speakers are concentrated at widely different locations in the parameter space. For a single measurement parameter, this amounts to saying that a good measure of effectiveness would be the ratio of interspeaker to intraspeaker variance, often referred to as the F ratio [19], [21]-[22]. The F ratio is defined as

$$F = \frac{\text{variance of speaker means}}{\text{average intraspeaker variance}}$$

$$= \langle [\mu_i - \bar{\mu}]^2 \rangle_i / \langle [x_\alpha^{(i)} - \mu_i]^2 \rangle_{\alpha, i} \tag{1}$$

where $x^{(i)}$ is the parameter value from the αth repetition of an utterance spoken by the ith speaker, $\langle \, \rangle_i$ indicates averaging over the speakers, $\langle \, \rangle_\alpha$ indicates averaging over the different utterances of a speaker,

$$\mu_i = \langle x_\alpha^{(i)} \rangle_\alpha \tag{2}$$

is the estimated mean value of the parameter for the ith speaker, and

$$\bar{\mu} = \langle \mu_i \rangle_i \tag{3}$$

is the estimated overall mean value of the parameter averaged over all speakers.

For normal variables (having Gaussian probability density function), the F ratio has the nice property that the probability of misclassifying a speaker i to speaker j is a monotonically decreasing function of F. In general, however, the F ratio has no simple relationship to the probability of error. Still, it has been found very useful in evaluating the effectiveness of parameters. Pruzansky and Mathews [21] used this technique to select a small subset of features from a much larger set consisting of the quantized spectrographic data. They compared the performance of a number of parameters with large F ratios with the performance of the same number of parameters with small F ratios. As an example, with a subset consisting of 10 percent of the total number of parameters, the probability of error increased from approximately 18 percent for the largest F ratios to approximately 50 percent for the smallest F ratios, thus demonstrating the utility of the F ratio for selecting efficient speaker-recognition parameters.

The major disadvantage of evaluating parameters individually is that interparameter correlations are not taken into account. For example, if the parameter is highly correlated with another parameter, the performance of the two together would not be much better than either of the parameters alone. The individual evaluation of parameters can thus result in the selection of redundant parameters. It is advantageous to consider the parameters together as a vector. The concept of F ratio can be extended to the multidimensional case. We define two covariance matrices: an interspeaker covariance matrix B and an intraspeaker covariance matrix W. These matrices are defined as

$$B = \langle [\mu_i - \bar{\mu}] \, [\mu_i - \bar{\mu}]^t \rangle_i \tag{4}$$

$$W = \langle [x_\alpha^{(i)} - \mu_i] \, [x_\alpha^{(i)} - \mu_i]^t \rangle_{\alpha, i} \tag{5}$$

where $x_\alpha^{(i)}$ is a vector representing the measurements from the αth repetition of an utterance spoken by the ith speaker,

$$\mu_i = \langle x_\alpha^{(i)} \rangle_\alpha \tag{6}$$

is the estimated mean vector for the ith speaker, and

$$\bar{\mu} = \langle \mu_i \rangle_i \tag{7}$$

is the estimated overall mean vector averaged over all speakers. A suitable measure of effectiveness of a set of parameters for discriminating between different categories is "divergence," originally defined by Kullback [24] as a measure of uncertainty in the data. The divergence is given by [24], [25]

$$D = \langle [\mu_i - \mu_j] \, W^{-1} \, [\mu_i - \mu_j]^t \rangle_{i, j} \tag{8}$$

where W^{-1} is the inverse of the matrix W and the superscript t indicates the transpose of a vector. Equation (8) can also be written as

$$D = \operatorname{Tr} W^{-1} \langle [\mu_i - \mu_j] \, [\mu_i - \mu_j]^t \rangle_{i, j} \tag{9}$$

$$= \operatorname{Tr} W^{-1} B \tag{10}$$

which is an obvious generalization of the concept of the F ratio to the multidimensional case [6], [25]. For a one-dimensional case, (10) reduces to $D = F$.

The divergence was used by the author to find an efficient representation of pitch contours of different speakers [6], [25]. Originally, the pitch contours were specified by 40 equally-spaced samples along the time scale. These 40 pitch values were converted to 40 Karhunen-Loève components [26] and partitioned into 4 groups of 10 each in order of decreasing variance. Each group of parameters was used to

TABLE I
THE DIVERGENCE AND THE PERCENTAGE OF CORRECT IDENTIFICATIONS FOR THE FOUR GROUPS OF KARHUNEN-LOÈVE COMPONENTS

Group	Divergence	Percentage of Correct Identifications
1	44.9	93
2	3.5	57
3	1.2	20
4	0.8	11

identify a speaker from an ensemble of 10 speakers. The percentage of correct identifications and the divergence for each of the four groups is shown in Table I. In each case, a decrease in divergence is accompanied by a decrease in the identification accuracy.

The divergence D takes account of the interparameter correlations and thus is a more appropriate measure of the effectiveness of a set of parameters than the F ratio. Of course, correlation is a measure of linear dependence between two parameters. A small correlation between two parameters does not imply lack of any relationship between them. Wolf [19] describes a technique for estimating pair-wise interparameter dependence (not necessarily linear) between a set of parameters. His method is based on estimating the degree of overlap between the parameter distributions of the individual speakers. Such a measure of interparameter dependence could prove to be useful for isolating cases of strong nonlinear dependence between the parameters.

C. Types of Acoustic Parameters

Acoustic parameters of speech can be broadly classified into two groups: steady-state (time-invariant) and time-varying. Ideally, one would like to think that the speaker-defining parameters should reflect fixed characteristics of the speech signal and therefore be time-invariant. Time-invariant parameters can be obtained either by averaging the time-varying behavior of the parameter or by performing measurements which reflect fixed anatomical properties of the vocal tract. The main advantage of the time-invariant parameters is that they are essentially independent of the spoken message and are therefore suitable for text-independent speaker recognition. However, they may suffer from important disadvantages. A large class of speaker-dependent properties of speech result from the idiosyncrasies in the speaking habits of individuals and these by nature vary from one sound to the other. Such useful speech characteristics cannot be represented by time-invariant parameters. Furthermore, time-averaged voice characteristics, such as average pitch or average spectrum, may be subject to easy mimicry since it would appear difficult that an impostor could easily mimic the entire time variation of a parameter.

Among the time-varying parameters, one should distinguish between parameters which are defined continuously as a function of time versus the parameters which are defined selectively for certain speech events. Both of these types have their merits. Continuously defined parameters are easier to measure but could result in a large set of data having a high degree of redundancy. Such redundancies can be identified by using the techniques based on the F ratio or the divergence as discussed earlier. Selective measurement of parameters at appropriately chosen locations in the utterance avoids this problem but requires the identification of the proper locations in the speech utterance—a task difficult to perform automatically. Most of the speaker-recognition work has been done with parameters measured at regularly spaced intervals in an utterance, and consequently, a large part of the results presented in this paper are related to such measurements. Both Wolf and Sambur [19], [20] have discussed selectively defined parameters exclusively and their published work is an important source of information for evaluating such measurements for speaker recognition.

D. Model of Speech Production

Speech sounds are produced as a result of acoustical excitation of the vocal tract which consists of the cavities in the pharynx and the mouth. The entire vocal tract can be thought of as an acoustic tube terminated by the lips at one end and by the vocal cords at the other end. The shape of the vocal tract is changed continually during speech production by the movement of different articulators such as the tongue, the jaw, and the lips. For most sounds, the sound is radiated at the lips. In the case of nasal consonants, the front part of the vocal tract is closed and the vocal tract is coupled through the velar opening to the nasal cavities, thereby producing sound radiation from the nostrils. During the production of nonnasal sounds, the velar opening is closed and no sound is radiated from the nostrils.

There are three important modes of exciting the vocal tract. It is customary to classify sounds according to their mode of excitation. Voiced sounds are produced as a result of excitation by a series of nearly periodic pulses generated by the vocal cords. Examples of voiced sounds are vowels, semivowels, voiced stops, and nasals. The fundamental frequency of the vocal-cord vibrations is determined by the mass and tension of the vocal cords, and by the subglottal air pressure. Both the tension of the vocal cords and the subglottal air pressure change during speech production. The range of fundamental frequency in speech from a normal adult is about 60 to 400 Hz. Unvoiced sounds are produced when the vocal tract is excited by a noise-like turbulent flow of air at a point of constriction. The resulting excitation has a relatively uniform spectrum. Examples of unvoiced sounds are various fricatives, such as f, s, sh, etc. and voiceless stop consonants like p, t, and k. Finally, the plosive sounds are produced by making a complete closure, building up pressure behind the closure, and abruptly releasing it. Examples of plosives are stop consonants like b, d, g, p, t, and k. In continuous speech, combinations of different types of excitation mentioned above are also used.

A simplified model of speech production can be made by representing the vocal tract as a linear time-varying filter excited either by a periodic pulse source or by a white-noise source as shown in Fig. 1. The output of the vocal tract is then expressed as a convolution between the excitation function and the impulse response of the time-varying linear filter representing the vocal tract. Thus, the speech signal $s(t)$ is given by

$$s(t) = \int_{-\infty}^{t} e(\tau)\, v(t, \tau)\, d\tau \qquad (11)$$

where $e(t)$ is the excitation function, and $v(t, \tau)$ is the response

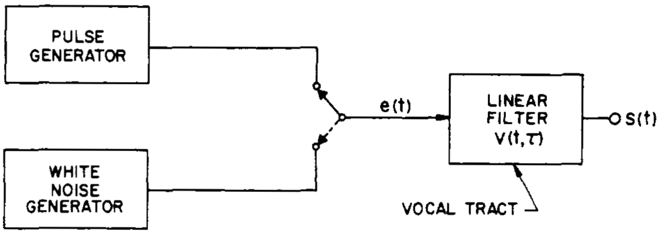

Fig. 1. Block diagram of a functional model of speech production.

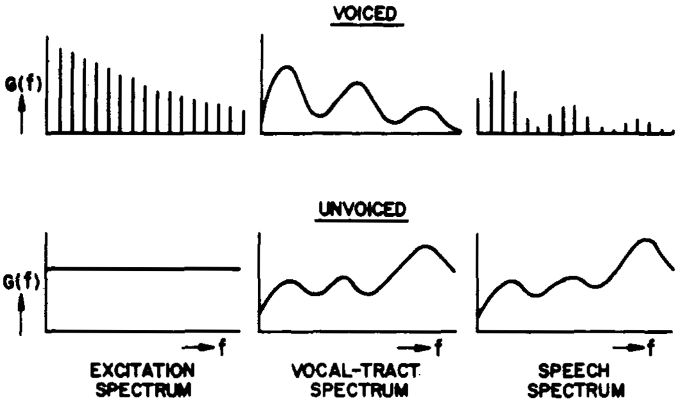

Fig. 2. Spectra of voiced and unvoiced sounds.

of the vocal tract at time t to a delta function input applied at time τ. For most sounds, the shape of the vocal tract changes slowly compared to the variations in the excitation waveform. Thus the speech production can be considered to be a quasi-stationary process. The quasi-stationarity of the model permits us to look at the speech characteristics in the frequency domain. For voiced speech, the vocal excitation function is nearly periodic with slowly varying periods. Its spectrum consists of a series of harmonics whose amplitude falls off at approximately 12 dB per octave. The spacing between adjacent harmonics is determined by the period of the vocal-cord vibrations. The spectrum of the vocal-tract response consists of a number of resonances whose locations depend upon the vocal-tract shape. The voiced-speech spectrum is thus composed of harmonically related frequencies whose amplitudes are determined by the vocal-tract response at these frequencies. For unvoiced speech, the excitation spectrum is uniformly distributed over a wide frequency range; the speech spectrum for these sounds reflects entirely the vocal-tract response. Such idealized representations of spectra for voiced and unvoiced speech are given in Fig. 2.

E. Short-Time Spectral Representation of Speech

A useful concept for describing properties of nonstationary signals such as speech is the time-varying or "short-time" spectrum [18], [27], [28]. In such a "short-time" representation, only a portion of the signal in the neighborhood of the present time is included in computing the spectrum. Mathematically speaking, for a signal $s(t)$, the short-time power spectrum is defined as a function of frequency f and time t as

$$G(f, t) = \left| \int_{-\infty}^{+\infty} s(\tau)\, w(t - \tau)\, e^{-j2\pi f\tau}\, d\tau \right|^2 \tag{12}$$

where $w(t)$ is a suitable window function. For most practical cases, the effective duration of $w(t)$ is in the vicinity of 20 to 30 ms. The frequency range of interest in speech usually extends from 0 to 10 kHz. The short-time spectrum of speech contains nearly all of the important information in speech and has formed the basis for most of the present methods of characterizing speech in a parametric form.

In the earlier days before the availability of fast digital computers, the short-time power spectrum was obtained by passing speech through a bank of band-pass filters. The energy at the output of each band-pass filter provided a good estimate of the short-time spectrum at the center frequency of the filter. The number of band-pass filters is governed by the bandwidth of each filter and the range of frequencies to be covered. An example of such a system is described in [19]. It consists of a 36-channel filter bank with center frequencies spaced linearly between 150 and 1650 Hz and logarithmically thereafter to 7025 Hz.

On a digital computer, the short-time power spectrum is obtained directly by using (12). A suitable window function $w(t)$ is the "Hamming" window [29] defined as

$$w(t) = \begin{cases} 0.54 + 0.46 \cos 2\pi t/T, & -T/2 \leqslant t \leqslant T/2 \\ 0, & |t| > T/2 \end{cases} \tag{13}$$

where the duration T is typically 25 ms. An example of the short-time power spectrum of speech for a voiced speech sound is shown in Fig. 3. The fine structure corresponding to the regularly spaced peaks in the spectrum is due to the periodicity of the voiced speech; the frequency spacing between the peaks is the pitch frequency. The amplitude of the peaks—often described as the spectral envelope—is a function of both the glottal-pulse shape and the vocal-tract response. Since the fine structure is determined completely by a single parameter, the pitch frequency, it is usually desirable to separate the fine structure of the spectrum from the envelope. Such a decomposition leads to a compact description of the spectrum.

F. Description of Acoustic Parameters

In the early work on speaker recognition, spectrographic data (time–frequency–energy pattern of speech) was almost exclusively used in speaker-recognition studies. Since then, a wide variety of additional measurements, based on both frequency and time-domain analyses, have been investigated for application to automatic speaker recognition. Many of these parameters are related to some property of the short-time power spectrum. We present here a brief description of the acoustic parameters which have been found useful for speaker recognition. The intention here is not to give an exhaustive list of parameters but to provide a fair sampling of different speech characteristics capable of distinguishing one speaker from another. The description of speech analysis methods for obtaining speech parameters is omitted here. An excellent survey on different speech analysis techniques and various parametric representations of speech appears in [18].

1) Intensity: One of the simplest characteristics of any signal is its intensity. For nonstationary signals such as speech, the intensity must be defined as a function of time. A suitable definition would be

$$E(t) = \int_{t-T/2}^{t+T/2} s^2(\tau)\, d\tau \tag{14}$$

where T is the averaging interval. The choice of T is somewhat

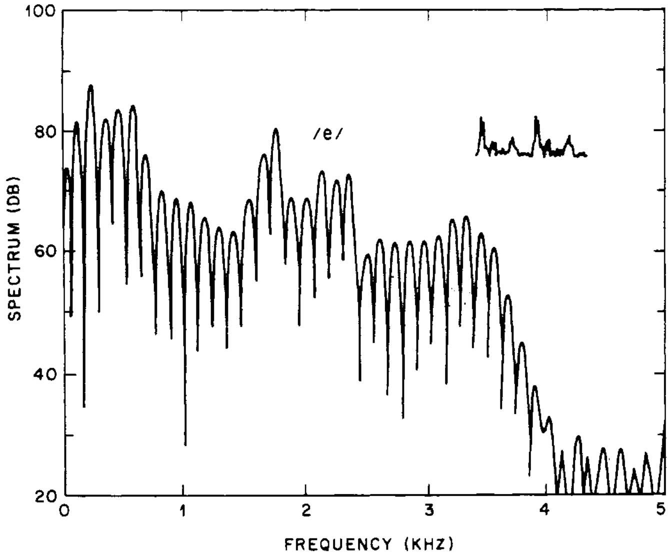

Fig. 3. An example of the short-time spectrum of voiced speech.

arbitrary; a value in the range 10 to 30 ms is adequate in most cases. The variations in intensity of speech are caused by the variations of both the subglottal pressure as well as the vocal-tract shape as a function of time and represent an important source of speaker-dependent information in speech [30]-[31].

2) Pitch: Pitch is the fundamental frequency of the vocal-cord vibrations. Accurate measurement of pitch has received considerable attention in speech research. Voice pitch can be determined either in the time domain by direct measurement of the period of the speech waveform or in the frequency domain by computing the frequency spacing of the spectral peaks [6], [32]-[35]. The temporal variation of pitch represents an important speech characteristic and has been found to be effective for automatic speaker recognition [6], [25].

3) Short-Time Spectrum: The definition of short-time spectrum given in (12) provides a three-dimensional representation of the speech signal, the coordinates being time, frequency, and energy. The short-time spectrum provides a complete (although not a very compact) description of the acoustical characteristics of speech. Both the exact representation of the short-time spectrum, defined in (12), and its approximation by filter-bank outputs have been found to be effective for automatic speaker recognition. [3], [4], [7], [8], [21].

4) Predictor Coefficients: Linear prediction analysis is an important method of characterizing the spectral properties of speech in the time domain [18], [36]-[39]. In this analysis method, each sample of the speech waveform is predicted as a linear weighted sum of the past p samples. The weights which minimize the mean-squared prediction error are called the predictor coefficients. Typically, 12 predictor coefficients are adequate for speech band-limited to 5 kHz. The predictor coefficients vary as a function of time and it is usually sufficient to compute them once every 20 ms.

5) Formant Frequencies and Bandwidths: Formant frequencies are defined as the resonance frequencies of the vocal tract. These too are speaker dependent. The chief difficulty with the formants lies in their measurement. Several methods have been described in the literature which provide a partial solution to the problem of determining formant frequencies [40]-[43]. Still, accurate and reliable determination of formants for both male and female speakers poses very difficult problems.

6) Nasal Coarticulation: In connected speech, due to slow movement of the articulators, the vocal-tract shape at any given time depends not only on the phoneme being spoken at that time but also on the neighboring phonemes. This phenomenon is known as coarticulation and it has been suggested that the nature of coarticulation in a given context is speaker dependent. One difficulty in using such information for speaker recognition is in obtaining a quantitative measure of such differences from speech. Coarticulation during the production of nasal consonants has been found useful for speaker recognition [45]. In this study, an acoustic measure of nasal coarticulation in a consonant-vowel context was obtained by measuring the spectral difference between the mean spectrum of a nasal consonant followed by a front vowel (such as /i/) and that of the same consonant followed by a back vowel (such as /a/).

7) Spectral Correlations: Significant degree of correlation exists between the short-time spectrum at different frequencies. These correlations have been found to vary consistently from one speaker to another [45]. Stable evaluation of such correlations, however, requires averaging over long utterances; about 30 s of speech is considered a minimum.

8) Timing and Speaking Rate: Relative timing of different speech events in spoken utterances differ from one speaker to another. Doddington has described a novel way of measuring such differences by determining the nonlinear deformation of the time axis of one utterance relative to that of another [10], [46].

III. Speaker-Recognition Procedures in Experimental Studies

Several computer-based studies have been conducted to test the feasibility of automatic speaker recognition. The objectives of these studies have been twofold: one to provide data about the comparative performance of different speech parameters in representing interspeaker differences and the other to determine the performance in terms of the error rates achieved in a complete speaker-recognition system using a selected set of parameters. A word of caution is in order here about interpreting the results of different speaker-recognition studies. These results depend on the task involved in the experiment—such as speaker identification or verification, text-dependent or text-independent speaker recognition. A meaningful comparison of results between these different cases cannot be made. Even in the absence of such differences, the comparisons may still not be straightforward. Every speaker-recognition study needs a data base consisting of several utterances spoken by many speakers. Unfortunately, there are no standard set of rules to be followed in collecting such a data base of speech utterances. The differences in data base can originate from several sources: number of speakers, type of speaker population, speech material, recording conditions, the time span over which the speech data is collected, and the elapsed time between the collection of test and training data. The decision rule employed for classification could also have a significant influence on the performance achieved in different studies. However, the results, in most cases, provide a rough measure of performance.

A. Collection of Speech Data

Ideally, the data base for speaker-recognition studies should include a large number of speakers. Collection of speech data from a large number of speakers poses many practical problems. Consequently, most studies use a relatively small population of speakers. Very few studies have tested speech data from one hundred or more speakers [7], [8], [47], [48]. In many cases, only male or female speakers have been used exclusively. With a small number of speakers, it is preferable not to include speakers with widely different speech characteristics. For example, male and female voices can often be distinguished from each other on the basis of average pitch alone. A small population study using speakers with widely different voice characteristics is not likely to provide a critical test of the suitability of a set of speech parameters for automatic speaker recognition. In some of the early work on speaker recognition, both the test and reference utterances were recorded in a single session. Later on, it became obvious that the results from such studies can be misleading. A speaker can often maintain a high degree of stability in his speaking habits over a short interval of time. However, the intraspeaker variability can often increase significantly with an increase in the time interval between recordings of test and reference utterances. As an example, Furui *et al.* have reported that percentage-recognition rate decreased from 96 percent to 52 percent when the interval between recordings of test and reference samples was increased from 3 days to 3 months [16]. Thus one must be careful about using the results from those speaker-recognition studies in which all the utterances from a speaker were collected over a short interval of time.

B. Formation of Training and Test Sets

In most speaker-recognition systems, a training set of speech utterances is used to obtain a statistical characterization of the acoustical parameters of the various speakers. In testing the performance of these systems, it is essential that the speech utterances used in the training process be kept separate from the utterances used to test the performance. In fact, clever pattern-recognition procedures can often be devised to work very well on the utterances in the training set. But there is no assurance that these procedures would work equally well on another set of utterances not used in the training process. The use of an independent test set would show if indeed the training procedure has learned the true characteristics of a speaker which are common to all of his utterances and not just to those represented in the training set.

C. Temporal Alignment of Utterances

Speech events in two utterances—even though they have the same text and are spoken by the same speaker—are seldom synchronized in time. This effect is attributable to the differences in the speaking rates. In comparing time-varying parameters, it is necessary that these be derived from similar speech events in the reference and the test utterances. Approximate time synchronization can be achieved by aligning the beginning and the end of the two utterances. More accurate synchronization can be obtained by pattern-matching techniques using nonlinear time deformation [30], [46], [49]. An alternative procedure would be to identify a number of "landmarks" in the utterances and to align the utterances by linear stretching or compression of the time scales between the landmarks [8].

D. Speaker-Classification Procedure

A speaker-recognition system involves two basic functions: measurement and classification. The measurement and selection of speaker-dependent parameters were discussed earlier in Section II. The classification is concerned with the decision-making process of determining if a particular person is the speaker of a given utterance.

The speaker-classification problem can be viewed as an exercise in statistical decision theory. A set of parameters derived from an utterance can be thought of as representing the utterance by a point in a multidimensional parameter space. In principle, the distribution of points corresponding to different utterances of a speaker can be characterized by a multivariate probability-density function. For speaker identification, the decision rule to assign a given point in the parameter space to a particular speaker would call for the selection of a speaker who has the highest probability of generating the given set of parameters. Similarly, the task of speaker verification (or authentication) would call for deciding if the probability of the parameters being generated by the claimed speaker is large enough for the particular application at hand. However, a straightforward application of this simple idea is not practically feasible. Even with relatively few parameters, the estimation of the multivariate probability-density function requires a very large number of utterances from a speaker. It is more practical to choose a parametric form of the probability density with some unknown parameters whose values could be estimated from the utterances in the training set. A convenient choice for this purpose is the well-known multivariate Gaussian density function. The Gaussian density function has an important property that is completely determined by its mean vector and a covariance matrix. The assumption of Gaussian density for the parameters is not as arbitrary as it might seem to be. For the decision rule to be correct, it is sufficient that the density be essentially unimodal and approximately Gaussian in the center of its range—a property often found to be true for physical measurements. Needless to say that such a decision rule would also be optimum for arbitrary monotonic functions of the Gaussian probability-density function.

As mentioned earlier, the Gaussian density is completely specified by its mean vector and a covariance matrix. An appropriate choice of the mean vector is the sample mean vector $\mu^{(i)}$ defined in (6), with the averaging being done over the various utterances of a speaker in the training set. The covariance matrix, strictly speaking, is also speaker dependent. However, in most practical situations, it is not possible to determine an accurate estimate of the matrix for each speaker. A more tractable choice for the covariance matrix is the pooled intra-speaker covariance matrix W defined in (5) representing an average matrix for all of the speakers. The N-dimensional Gaussian density function for the ith speaker is then written as

$$g_i(\boldsymbol{x}) = (2\pi)^{-N/2}\, |W|^{-1/2} \exp\left[-1/2(\boldsymbol{x} - \mu_i)^t W^{-1} (\boldsymbol{x} - \mu_i)\right] \tag{15}$$

where $|W|$ is the determinant of the matrix W. Let us consider speaker identification first. If we assume that the unknown utterance could have originated from any speaker in the population with equal apriori probability and that the cost of making an error is the same for all speakers, the optimum decision rule would assign the unknown utterance to the speaker j if [50]

$$g_j(\boldsymbol{x}) > g_i(\boldsymbol{x}), \qquad \text{for all } i \neq j. \tag{16}$$

Since logarithm is a monotonic function, the decision rule of

(16) can be rewritten as

$$d_j(\boldsymbol{x}) < d_i(\boldsymbol{x}), \qquad \text{for all } i \neq j \tag{17}$$

where

$$d_j(\boldsymbol{x}) = [(\boldsymbol{x} - \mu_j)^t W^{-1} (\boldsymbol{x} - \mu_j)]^{1/2}. \tag{18}$$

The above rule simply assigns the unknown utterance to a speaker on the basis of a minimum-distance classification rule. The quantity $d_j(\boldsymbol{x})$ of (18) would be called the distance between a vector representing the unknown utterance and the mean (reference) vector for the jth speaker. The commonly used Euclidean distance is a special case of this general definition and is obtained by equating the covariance matrix W by an Identity matrix in (18), thereby replacing the quantity under the square-root sign by the sum of the squared differences along each coordinate. However, the Euclidean distance does not provide the correct decision rule if the speech parameters show significant interparameter correlations or have widely different intraspeaker variances.

The distance metric $d_j(\boldsymbol{x})$ defined in (18) has several important properties: the distance $d_j(\boldsymbol{x})$ is invariant with respect to any arbitrary nonsingular transformation of the parameters. To prove this result, consider a representation of the utterances in the N-dimensional parameter space. Let T be a $N \times N$ matrix representing the linear transformation. The unknown utterance is represented in the transformed space by a vector $\hat{\boldsymbol{x}}$ given by

$$\hat{\boldsymbol{x}} = T\boldsymbol{x}. \tag{19}$$

The reference utterance of the jth speaker in the transformed space is given by

$$\hat{\mu}_j = T\mu_j. \tag{20}$$

The distance $\hat{d}_j$ between $\hat{\boldsymbol{x}}$ and $\hat{\mu}_j$ is then given by

$$\begin{aligned}\hat{d}_j &= [(\hat{\boldsymbol{x}} - \hat{\mu}_j)^t \hat{W}^{-1} (\hat{\boldsymbol{x}} - \hat{\mu}_j)]^{1/2} \\ &= [(\boldsymbol{x} - \mu_j)^t T^t \hat{W}^{-1} T(\boldsymbol{x} - \mu_j)]^{1/2}.\end{aligned} \tag{21}$$

Since $W^{-1} = T^t \hat{W}^{-1} T$, it follows that $\hat{d}_j = d_j$. This property of the distance metric of being invariant with respect to arbitrary linear transformations of the parameters is very useful. The distance metric of (18) would provide identical results with two parameter sets if they are linearly related to each other. For example, the Fourier transform is a linear transformation. Thus several frequency and time-domain representations of speech would be expected to provide similar results with this particular metric.

Consider the distribution of utterances in an N-dimensional parameter space. The non-Euclidean distance metric $d_j(\boldsymbol{x})$ is equivalent to the measurement of Euclidean distance after a suitable linear transformation of the parameters. This can be seen as follows: the covariance matrix W is a positive-definite and symmetric matrix and can thus be written as a product $W = CLC^t$, where C is a unitary matrix ($CC^t = I$), and L is a diagonal matrix with its ith element given by λ_i. The non-Euclidean distance measure of (18) is equivalent to the Euclidean distance if all the vectors are multiplied by a matrix $L^{-1/2} C^t$. This matrix transformation has the interesting property that the intraspeaker covariance matrix in the transformed space is the Identity matrix. Thus the transformed parameters have no interparameter correlations and have equal intraspeaker variances; the Euclidean distance is the proper distance measure under such conditions.

A distance metric can be regarded as a means of combining the contributions of a large number of measurements into a single measure of dissimilarity between two utterances. A good measure of distance must weight the contributions of different measurements in order of their importance in producing small intraspeaker variations. The distance metric of (18) does indeed possess this property while the Euclidean distance does not.

Both the Euclidean distance and the non-Euclidean distance metric obtained under the assumption of a Gaussian probability-density function of the parameters have been used in speaker-recognition studies with success. Some studies have used the degree of crosscorrelation between the two utterances—the unknown and the reference—as a measure of similarity between the two. In vector notations, the crosscorrelation is defined as

$$\rho_i = (\boldsymbol{x}^t \mu_j)/[(\boldsymbol{x}^t \boldsymbol{x})(\mu_j^t \mu_j)]^{1/2} \tag{22}$$

The crosscorrelation measure has an important property that it is unchanged even if either $\boldsymbol{x}$ or μ_j or both are multiplied by a constant. This property need not always be a desirable one but can be used with considerable advantage in some situations where parameter variations introduced by some random scale factor are to be ignored.

The distance metric of (18) and the crosscorrelation measure of (22) are also applicable to the speaker-verification problem. Here, the claimed identity of a speaker would be verified if the distance was less than a preselected threshold value or if the crosscorrelation exceeded the threshold value.

IV. Results of Experimental Studies

A fairly large number of speaker-recognition studies have been conducted during the past decade or so. It would not be possible to present in this paper an exhaustive survey covering all of them. Rather, we would attempt to use the results obtained in such studies to illustrate the effectiveness of various speech characteristics discussed earlier for automatic speaker recognition. Particular attention will be given to those results which enable one to compare the recognition performance of different sets of parameters.

Both speaker-recognition tasks, identification and verification, have been investigated in past experimental studies. Of the two, the identification task is more suited for comparing the performance of different parameters. In speaker identification, a single error rate can provide a measure of the performance, while in speaker verification, two kinds of errors, namely, the probabilities of false verification and false rejection as functions of a threshold parameter, determine the performance. Also, the identification accuracy is a more sensitive indicator of the ability of a parameter for discriminating speakers. Consequently, a large portion of the results in this section will relate to studies in speaker identification. Practical aspects of the speaker-verification problem are discussed in a companion paper by Rosenberg [48]. This paper also provides valuable details about the performance of a speaker-verification system being tested at Bell Laboratories under realistic practical environments [49]. To avoid duplication, this particular study will not be included in this paper.

A. Short-Time Power Spectrum

We will discuss here the results from two studies which have used the short-time power spectrum of speech for speaker identification. The first of these [3] is probably one of the earliest investigations on automatic speaker recognition. In this study, the spectrum analysis was accomplished by passing speech through a 17-channel filter bank covering the frequency range 100–7000 Hz. The first 16 channels were approximately equally spaced along a Koenig scale from 200–4000 Hz. The speech data consisted of 4 repetitions of 10 words excerpted from sentence-length utterances spoken by 10 speakers (7 male and 3 female). Each word was represented by a three-dimensional array describing spectrum as a function of frequency and time. The temporal alignment of the words was achieved by lining up the maximum of the energy-versus-time function of each word. For each speaker, three utterances of each of the ten words were used to form the reference pattern for that individual. The procedure used for recognition of speakers consisted of crosscorrelating the spectrographic pattern of the test utterance of each word with each of the ten reference patterns for that word using the product-moment definition of the correlation. The speaker corresponding to the reference pattern with the highest correlation was recognized as the speaker of the test utterance. Speakers were correctly recognized for 89 percent of the 393 cases tested. Results for individual speakers ranged from 77 percent to 98 percent. The recognition rate was not uniformly good for all of the words and ranged from 74 percent to 97 percent approximately.

In the same study, the three-dimensional time-frequency-intensity patterns were reduced to two dimensions by averaging over time for each of the 17 frequency bands. The averaging produced an intensity-versus-frequency array for each word. With the time dimension eliminated, speakers were still recognized correctly for 89 percent of the utterances—no change from the previous case. However, when the three-dimensional patterns were reduced to two dimensions by eliminating the frequency dimension, the recognition rate dropped to 47 percent thus implying that spectral information is more important than the energy-time information for speaker recognition. The low recognition rate may partly be due to the poor temporal alignment of speech events in the test and reference patterns. Such misalignment is not detrimental to the time-averaged spectral data but is highly critical in obtaining proper results from the energy-time data. Later studies by Doddington, using nonlinear time warping, show that intensity-versus-time variation in speech is very effective for speaker recognition [30], [31].

The same speech data was employed in another study by Bricker *et al.* [7]. Using the non-Euclidean distance metric of (18), they found a recognition rate of 97 percent for the two-dimensional spectral data. The non-Euclidean distance measure reduced the error rate by a factor of approximately 4 as compared to the simple crosscorrelation measure. The study was also extended to a much larger data base consisting of 172 speakers. The speech data in this study consisted of five digits spoken in isolation. The spectral data was obtained by using a 20-channel filter bank over the frequency range 20–2900 Hz. The speech data was recorded in an unattended booth in a relatively noisy concourse. The results of their study indicated that a fairly high recognition rate can be maintained even with such a large number of speakers. Based on the speech data from a single digit, the recognition rate was 84 percent; it increased to 94 percent by combining the information from two digits.

Both of the above studies succeeded in demonstrating the importance of spectral information in speech for speaker recognition. However, there is some difficulty in using time-averaged spectra for practical speaker-recognition systems. The spectral data as used in the above studies is strongly affected by the frequency characteristics of the recording and the transmission apparatus. Any variations in the frequency characteristics—not an uncommon occurrence—would be an additional source of randomness in the spectral data which could cause an adverse influence on the performance.

B. Parameters Derived by Linear Prediction of Speech Waveform

Speech analysis based on linear predictability of its waveform is carried out efficiently on digital computers. The predictor coefficients thus offer a convenient choice of parameters for speaker recognition. Recently, a number of studies have been carried out to determine the suitability of these parameters for speaker recognition [14], [51], [52].

The use of linear prediction parameters for automatic speaker recognition is desirable for several reasons. Most important of these is that it eliminates the necessity of deciding as to which of the speech characteristics—such as a particular formant frequency or its bandwidth or some property of the glottal wave—would be suitable for speaker recognition. The predictor coefficients represent the combined information about the formant frequencies, their bandwidth, and the glottal wave. Moreover, being independent of the pitch and intensity information, the predictor coefficients could be used to improve the reliability of another method of speaker recognition based on pitch and intensity information. The linear prediction analysis is also conveniently implemented by digital hardware.

In the linear prediction model [18], [36]–[39], one represents the combined filter transfer function of the vocal tract, the radiation, and the glottal-wave shape by a discrete all-pole linear filter with p poles as shown in Fig. 4. This filter is described completely by a set of p coefficients $a_1, a_2, \cdots, a_p$, defining the linear prediction characteristics of the speech waveform. The predictor coefficients are determined by minimizing the mean-squared prediction error between a speech sample and its linearly predicted value from the past p samples and are given as solutions of a set of linear simultaneous equations [18], [36]

$$\sum_{k=1}^{p} \phi_{jk} a_k = \phi_{jo}, \qquad j = 1, 2, \cdots, p \tag{23}$$

where

$$\phi_{jk} = \sum_n s_{n-j} s_{n-k} \tag{24}$$

$$s_n = n\text{th speech sample}$$

and

$$a_k = k\text{th predictor coefficient.}$$

The value of p is approximately determined by the number of poles of the vocal tract and the glottal wave within the frequency range of the speech signal. For speech band-limited to 5 kHz, a typical value of p is 12.

We present here a summary of the results reported in detail in [14]. In this study, the speech data base consisted of 60 utterances spoken by 10 speakers. All of the speakers were female and they spoke the sentence "May we all learn a yellow lion roar" with six repetitions. The speech recordings were made in an anechoic chamber using a high-quality microphone

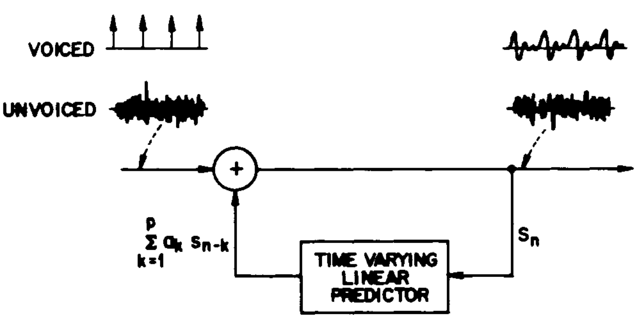

Fig. 4. All-pole model of speech production used in linear prediction analysis.

on two different days 27 days apart. On each day, the recordings were made in three separate sessions. The speech signal was band-limited to 5 kHz prior to analysis in the computer. In the analysis, each utterance was divided into 40 segments of equal duration. Approximate temporal alignment of the segments was achieved by making the segment durations in an utterance proportional to the duration of the utterance which ranged between 1.8 and 2.8 s. The silent portions and pauses in an utterance were eliminated automatically from the segments.

The basic parameter data used in this study consisted of 12 predictor coefficients obtained at 40 uniformly spaced time frames for each of the 60 utterances (6 repetitions X 10 speakers). Five utterances for each speaker were used to compute the reference vectors while the remaining sixth one was used as a test vector. Each of the six repetitions was used in turn as the test vector for each of the 10 speakers. The identification decision was based on the non-Euclidean distance measure defined in (18). The pooled intra-speaker covariance matrix was obtained by averaging over 50 utterances of the 10 speakers in the training set.

A number of different parametric representations of speech derived from the all-pole representation were examined to determine if any one of them is more effective than the other for speaker recognition. Amongst the parameters examined were the predictor coefficients, impulse response of the all-pole filter, autocorrelation function, area function, and cepstrum function of the impulse response—all of which have the interesting property that they can be derived from each other. The mathematical relationships between these different representations are discussed in detail in [14] and are omitted here except for the cepstrum function which proved to be the most effective of all for speaker recognition. The cepstrum by definition is the inverse Fourier transform of the logarithm of the transfer function [18], [53]. For the all-pole filter of Fig. 4, the logarithm of the transfer function is written as

$$\ln H(z) = C(z) = \sum_{n=1}^{\infty} c_n z^{-n} \tag{25}$$

where z is the usual z-transform variable [55]. The desired relationship between c_n's and a_n's is given by [14], [53]

$$\begin{aligned} c_1 &= a_1 \\ c_n &= \sum_{k=1}^{n-1} (1 - k/n) a_k c_{n-k} + a_n, \qquad 1 < n \leqslant p \\ c_n &= \sum_{k=1}^{n-1} (1 - k/n) a_k c_{n-k}, \qquad n > p. \end{aligned} \tag{26}$$

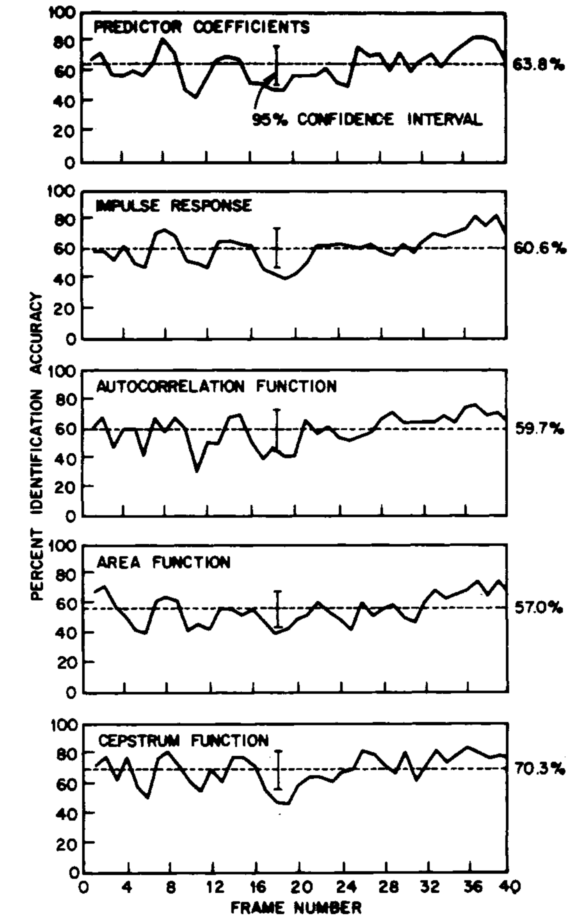

Fig. 5. Percent identification accuracy for different parametric representations of speech derived by linear prediction. Each result is based on a single frame of speech approximately 50 ms in duration.

The recognition accuracy for each of the 40 segments for the five sets of parameters is shown in Fig. 5. The number of parameters p in each case is 12 and each result is based on a total of 60 judgments. The recognition accuracy varied from one segment to another—the average standard deviation is approximately 6.0 percent. However, most of the time the variation is within the 95 percent confidence interval. The identification accuracy averaged over the 40 segments was found to be 63.8 percent for the predictor coefficients, 60.6 percent for the impulse response, 59.7 percent for the autocorrelation function, 57.0 percent for the area function, and 70.3 percent for the cepstrum function. These identification scores are quite high considering the fact that each segment has a duration of approximately 50 ms. The cepstrum function provided the highest score while the area function provided the lowest. The 95 percent confidence interval for the mean is approximately 2 percent. The mean identification score of cepstrum is thus significantly higher than the other parameters.

The identification accuracy was also determined as a function of the duration of speech included in computing the distances. An average distance was computed by averaging the distances from the individual frames. The results are shown in Fig. 6 for the 12-parameter cepstrum function. The solid curve was obtained when the duration was increased from zero onwards starting at frame 1 at the beginning of an utterance while the dashed curve was obtained when starting at the middle of the utterance. The differences between the two curves are obviously due to different texts of the two portions of the

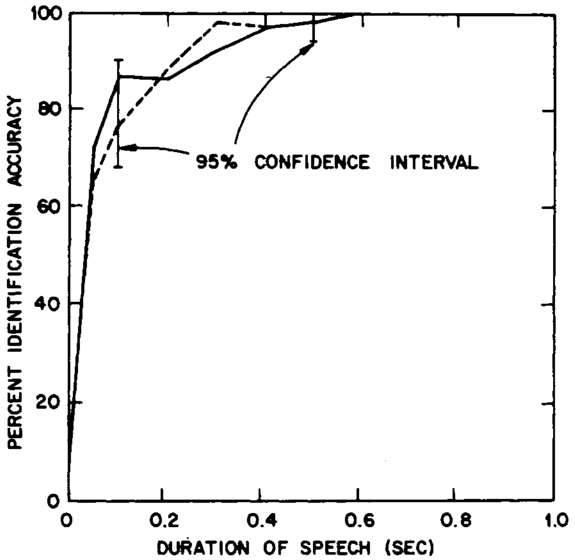

Fig. 6. Percent identification accuracy as a function of the duration of the speech sample based on 12 cepstrum parameters. The solid curve is for speech starting at the beginning of the utterance, while the dashed curve is for speech starting at the middle of the utterance.

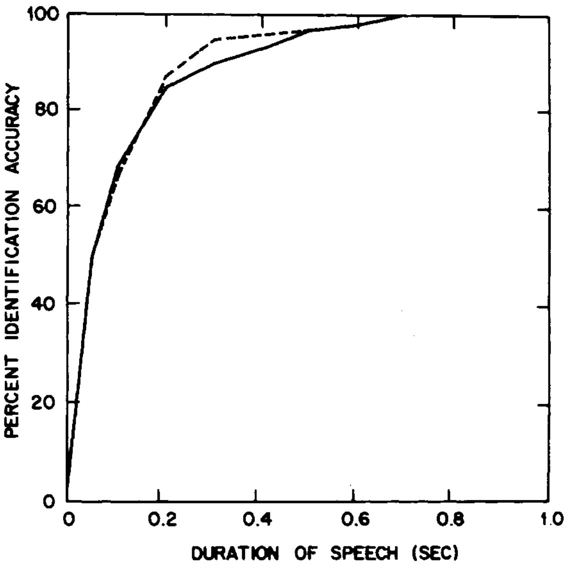

Fig. 7. Percent identification accuracy as a function of speech duration when the time average of the cepstrum parameters is eliminated from the data. The solid curve is for speech starting at the beginning of the utterance, while the dashed curve is for speech starting at the middle of the utterance.

utterance. The 95 percent confidence intervals are also marked at two places suggesting that the differences between the two curves are probably not significant. The recognition accuracy is approximately 80 percent for a duration of 0.1 s and is greater than 98 percent for durations of 0.5 s and higher. These results are very similar to those reported by Pruzansky for the spectral data [3].

All of the linear prediction parameters discussed so far are also affected by the frequency response of the recording apparatus as well as the transmission system. The cepstrum parameters, however, have the additional advantage that one can derive from them a set of parameters which are invariant to any fixed frequency-response distortions introduced by the recording apparatus or the transmission system. The new parameters are obtained simply by subtracting from the cepstrum parameters their time averages performed over the entire utterance. This result follows directly from the property of the cepstrum; namely, it is the logarithm of the overall transfer function. Since, the transfer function of the transmission system introduces a frequency-dependent multiplicative factor in the overall transfer function, the net result in the cepstrum is an additive frequency-dependent factor. If this factor does not vary in a short interval of a few seconds—a typical duration of an utterance—it is eliminated by subtracting the time-averaged cepstrum. The identification accuracy as a function of the duration of the spoken material based on the cepstrum parameters with their time averages removed is shown in Fig. 7. The significance of the solid and dashed curves is the same as in Fig. 6. There are no major differences between the two curves. The results show an identification accuracy of 68 percent for a duration of 0.1 s and an accuracy of greater than 98 percent for a duration of 0.6 s and higher. On comparing these results with the ones in Fig. 6—where the temporal averages were not removed—one finds the identification accuracy to be slightly lower (68 percent as compared to 80 percent for a duration of 0.1 s) but not significantly so. In fact, for durations larger than about 0.3 s, the differences are very minor. This result is not surprising. An increase in duration brings in more phonemes in the recognition process causing a decrease in the error rate. This decrease is slower in the case of Fig. 6 owing to higher correlation between the parameters of different phonemes. The higher correlations are introduced by the average frequency response of the utterance which is common to all phonemes in the utterance.

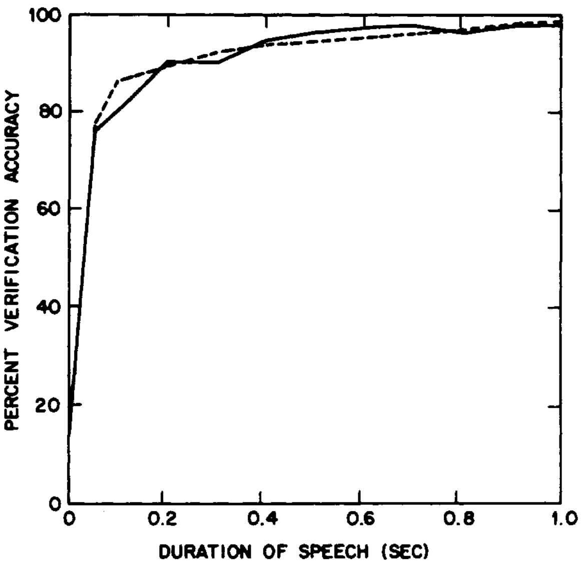

Fig. 8. Percent verification accuracy as a function of speech duration based on 12 cepstrum parameters with their time averages subtracted out. The solid and the dashed curves refer to the same conditions as in Fig. 7.

Tests were also conducted to determine the effectiveness of cepstrum parameters for automatic speaker verification. The speaker was verified if the distance between his utterance vector and the reference vector of the claimed speaker was smaller than a preselected threshold value. The results are shown in Fig. 8. A verification accuracy of 90 percent was achieved for a duration of 0.2 s and of 98 percent for a duration of 1.0 s.

The possibility of text-independent speaker recognition was also investigated in this study. Twelve cepstrum parameters

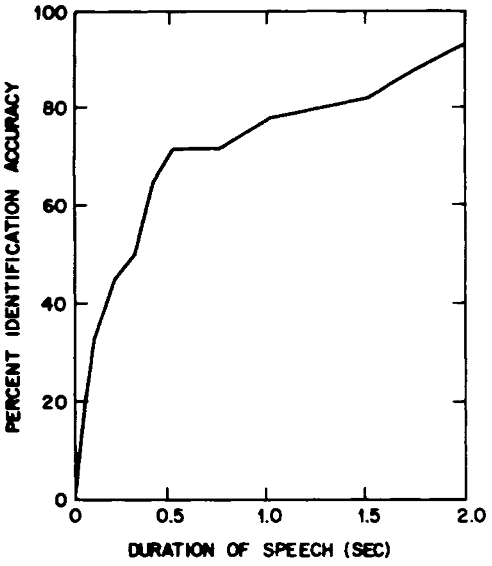

Fig. 9. Percent identification accuracy as a function of speech duration when the texts of the test and reference speech samples are different.

were again used with the distance measure of (18). The time synchronization between the reference and the test utterances was destroyed by cutting each utterance into 40 equal segments and later recombining them in random order to form a new utterance. The identification accuracy as a function of the duration of the spoken material is shown in Fig. 9. The identification accuracy is only 17 percent for a duration of 50 ms but increases to 72 percent for a duration of 0.5 s and to 93 percent for a duration of 2 s. The recognition accuracy, although lower than in the text-dependent case, is surprisingly high considering the difficult nature of the task involved.

A recent study by Furui and Itakura [55] using PARCOR coefficients [37, 39] reports similar results both for identification and verification tasks. The speech data base for the speaker identification study consisted of 4 words spoken by 9 speakers over a period of 3 months and yielded a recognition rate of 99.1 percent. The number of speakers were increased to 37 for the verification task and resulted in an accuracy of 99.2 percent.

C. *Pitch Contours*

The pitch or the fundamental frequency of vocal-cord vibrations is an important speaker-dependent speech characteristic. Several studies [6], [10], [19], [25], [30] have found pitch to be an effective parameter for automatic speaker recognition. The average pitch for a speaker varies considerably from one individual to another but by itself is not sufficient to distinguish between many speakers. A more interesting source of speaker-dependent information is the entire pitch contour describing the variation of pitch as a function of time in a sentence-length utterance. Pitch has an important advantage over the spectral information since it is not affected by frequency characteristics of the recording or the transmission system. A speaker-identification study based exclusively on the pitch information is described in [25]. The speech data base employed in the study is identical to the one used for the linear prediction characteristics [14] and thus it allows us to compare the relative performance of pitch and linear prediction information for speaker identification.

The pitch analysis in this study [6, 25] was carried out by performing a short-time correlation analysis on the cubed and low-pass filtered (3-dB attenuation at 1 kHz) speech waveform [6]. The time interval corresponding to the largest peak in the short-time correlation function determined the basic period of the waveform. Two examples of the pitch contours obtained by this method for each of the ten speakers included in the study are shown in Fig. 10. The speakers are identified in the figure by their initials. The ordinate is the duration of the pitch period in ms (pitch period is the reciprocal of the fundamental frequency) and the abcissa represents the time in seconds. The pitch contours are strongly speaker-dependent and yet are quite stable within the utterances of a single speaker.

For speaker identification, each pitch contour was represented by 40 samples spaced uniformly along the utterance. The pitch data was further compressed to a smaller set of parameters by the Karhunen-Loève (KL) transformation [26]. This compression was possible due to high correlations between neighbouring pitch samples. A 20-component KL representation accounted for all but 0.5 percent of the total variance. As before, five utterances of each speaker were used to form his reference pattern while the sixth one served as the test pattern. The identification decision was based on a non-Euclidean distance metric essentially similar to one described in (18). The overall percentage of correct identifications based on a total of 60 judgments was found to be 97 percent. A number of different recognition procedures were also tested on the same data to compare them with the non-Euclidean distance metric. The minimum-distance classification rule based on the Euclidean distance yielded only 68 percent correct identifications leading to the conclusion that the non-Euclidean distance metric reduced the misidentification rate from 32 percent to 3 percent—a reduction by a factor of over 10. The decision rule based on the crosscorrelation measure of (22) yielded a recognition rate of 70 percent—same as obtained by the Euclidean distance measure.

A somewhat different description of the pitch contours is obtained by describing each contour by the first-order probability distribution of the pitch samples in the contour. The first four central moments of the distribution yielded a recognition rate of 78 percent. Such a description retains information only about the variability of pitch in an utterance but not about the exact sequence of pitch variations as a function of time. For example, a particular sequence of pitch samples can be arranged in any arbitrary order without affecting the moments. These results suggest that the temporal information in the pitch contours is essential for reliable speaker recognition.

The results from the above study demonstrate the importance of pitch for automatic speaker recognition. These results can be directly compared with the results described earlier for the 12 cepstral coefficients [14]. The pitch-contour data provided a recognition accuracy of 97 percent for a speech utterance about 2 s in duration. On the other hand, a recognition accuracy of 98 percent was achieved for the 12 cepstral coefficients with a duration of only 0.5 s which is four times shorter than needed for pitch contours. Since the cepstral coefficients represent information about the spectral envelope of the speech signal, these results suggest that the spectral envelope is much more effective than the fine structure of the spectrum for automatic speaker recognition. Interestingly, a similar result was obtained by Miller in a study [56] involving human listeners to determine the relative importance of articulatory and source characteristics for speaker identifiability.

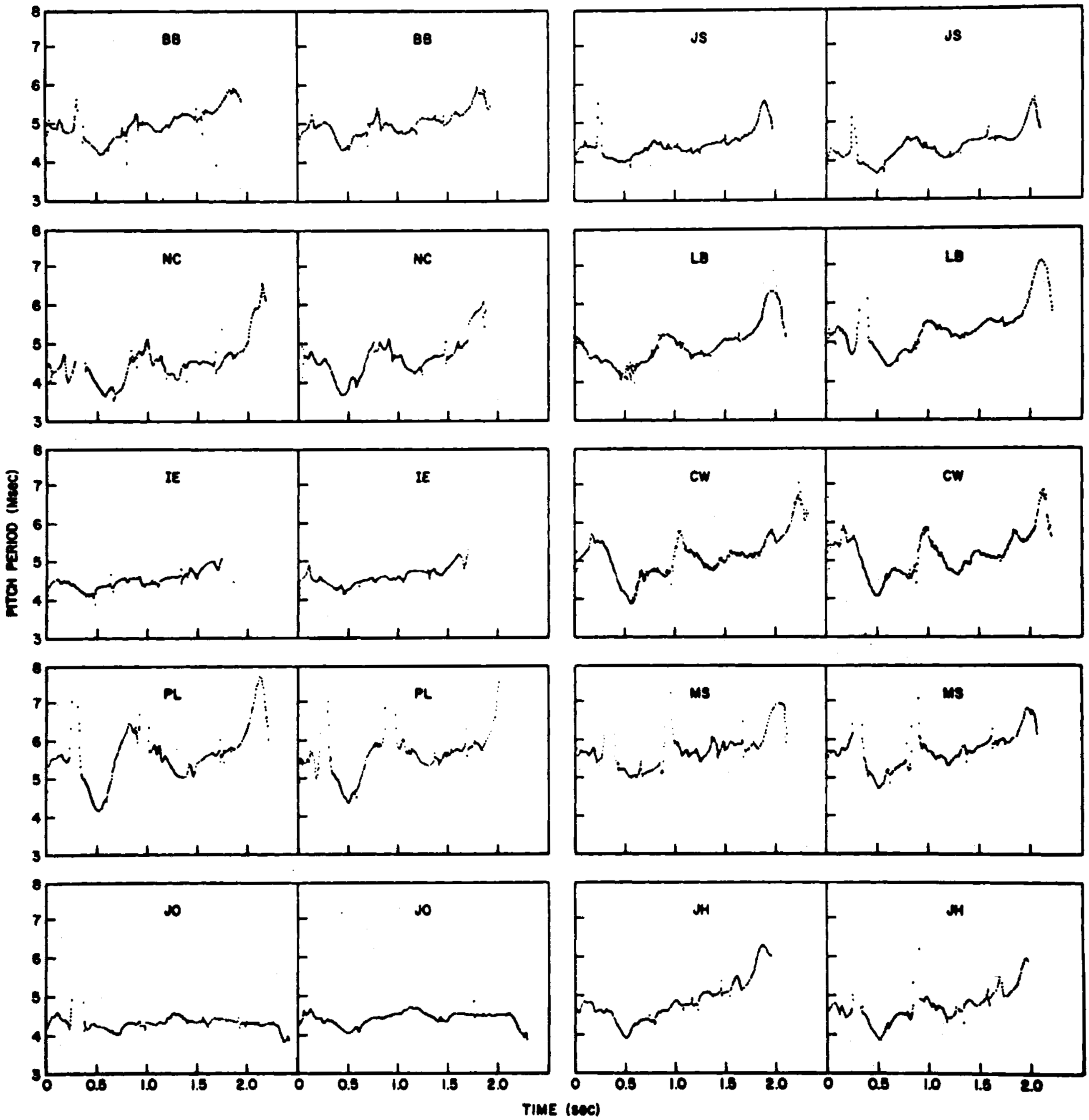

Fig. 10. Some examples of the pitch contours for the ten speakers. The sentence spoken was "May we all learn a yellow lion roar".

D. Spectral Correlations

The samples of short-time power spectrum of speech at different frequencies show significant correlations. Since these correlations are caused by the vocal-tract resonances, they are expected to be speaker dependent. These correlations are also a function of the text of the speech utterance used in computing the correlations. Such text dependence can be avoided by including a sufficiently long utterance containing a variety of speech events. Li *et al.* [45] have examined the correlational properties of speech spectra in considerable detail. Their main conclusions are: 1) the estimates of the correlation coefficients converge to a set of stable values after about 30 s of speech material is included; 2) the correlations exhibit strong speaker dependent behavior. Spectral correlations are usually characterized by a matrix of spectral correlations. In another study, Li and Hughes [15] measured the quantitative differences between the spectral correlation matrices of different speakers and of different utterances of the same speaker. The intraspeaker differences were found to be consistently less than the interspeaker differences. For a population of 30 speakers, the distribution of the two differences—interspeaker and intraspeaker—had about 1 percent overlap. These results suggest that the spectral correlations could be very useful for text-independent speaker recognition.

V. Speaker Recognition By Human Listeners

The results presented in the last section demonstrate clearly the feasibility of automatic speaker recognition—at least within the laboratory environment. The performance of automatic speaker recognition systems under "real world" conditions is discussed by Rosenberg in a companion paper [48]. In this section, we will present some results concerning the performance of human listeners in speaker recognition.

The question "How accurate are the automatic methods compared to the human listeners?" comes up quite often in any discussion on speaker recognition. It is also a question which is rather difficult to answer owing to the differences in the tasks performed by the automatic methods and the human listeners. For example, in order to perform satisfactorily, the present automatic methods require that the texts of the test and the reference utterances be identical. On the other hand, such a restriction is seldom necessary for human speaker recognition; the human listeners can recognize familiar voices with high accuracy irrespective of the text being spoken. A comparison between automatic methods and human listeners usually ends up in the selection of a speaker-recognition task that can be performed reasonably well by automatic methods, which is obviously unfair to human listeners. The results presented in this section should thus be interpreted with considerable caution.

A number of experimental studies have been conducted to investigate different aspects of human speaker recognition [57]-[63]. Several of these have been aimed at finding the perceptual factors which are important in human speaker recognition [59], [60]. The results of such studies however cannot be directly used for comparing the performance with automatic methods. At least two studies [61]-[62] have been reported in which the task before the human listeners was comparable to the one faced by the automatic methods—that is of identifying the speaker of an unknown utterance by direct listening to speech samples corresponding to the reference and the test utterances.

One of the earliest studies on speaker recognition by human listeners was conducted by McGehee [58]. The study included 5 speakers who read a paragraph of text to a group of listeners *unfamiliar* with the talkers. The identification accuracy was found to depend upon the elapsed time interval between the test and training sessions. For example, the identification accuracy decreased from 83 percent for a time interval of 1 day to 69 percent for a time interval of 2 weeks.

A more comprehensive study involving human listeners was carried out by Bricker and Pruzansky [61] in an attempt to determine the influence of the duration and phonetic content of the spoken utterance on the identifiability of speakers. Their study included 10 speakers and 16 listeners who were familiar with all of the speakers. The listeners were however not able to access the reference samples of different speakers for comparison during the test session. Each speaker recorded all of his material in a single session. The average identification score together with the durations of the spoken material are shown in Table II. The identification scores ranged from 56 percent for vowel excerpts to 98 percent for sentence-length utterances. The identification score was found to increase with the number of phonemes in the utterance. The scores presented in Table II are the average values for the 16 listeners. The individual scores varied considerably. As an example, for the vowel excerpts, the range of variation was between 39 and 75 percent.

Another study which allowed the listeners to hear the reference samples of the speakers is reported in [62]. This study included eight known speakers whose samples were available to the listeners for comparison in the tests. The speech material consisted of five repetitions of 11 words. A total of 10 listeners participated in the tests. The identification scores varied between different words, between different speakers, and between different listeners. The approximate identification score averaged over all the speakers and listeners was 89 percent for monosyllabic words and 91 percent for two-syllable words, phrases, and sentences. The identification score was not found to improve significantly once the number of syllables exceeded two. The average length of the two-syllable words was slightly under 1 s. The error rates for different words differed by a factor of 2.5 to 1. The error rates for different listeners differed by a factor of 3 to 1.

TABLE II
IDENTIFICATION SCORES OF HUMAN LISTENERS FOR FIVE TYPES OF SPEECH MATERIAL

Speech Material	Duration msec	Percent Correct
Vowel excerpts	117	56
CV excerpts	117	63
Monosyllables	498	81
Disyllables	446	87
Sentences	2400	98

A recent study by Rosenberg [11] reports the performance of human listeners in a speaker-verification task. The listeners were asked to respond whether a pair of test and reference utterances were spoken by the same or different speakers. The speech utterance was a 2-s long all-voiced sentence "We were away a year ago". The human listeners achieved an accuracy of 96 percent correct within a population of 40 speakers. An automatic method using formant, pitch, and intensity data from the same set of utterances provided a verification accuracy of 98 percent.

These studies provide a rough estimate of the performance of human listeners in identifying speakers. A strict comparison between the identification scores of automatic methods and human listeners cannot be made due to the differences in the speech data base as well as the differences in the test formats. However, it is interesting to compare these results with those presented in Fig. 6 based on the 12 cepstral parameters. The performance of the automatic method is at least comparable if not better. In this particular case, the results are quite intriguing if one considers the fact that human listeners could use the pitch, the intensity, and the timing information, as well as the spectral information, while the automatic method was constrained to use only the spectral envelope information contained in the 12 cepstral parameters.

VI. Concluding Remarks

The research on automatic methods for speaker recognition has reached a turning point. It is a proper time to look back and ask some basic questions. What have we learned so far? What are the remaining unresolved problems? Where do we proceed from here?

One of the original objectives of research on automatic speaker recognition was to find out if computers could be programmed to recognize speakers from their voices. It is obvious that automatic speaker recognition within small speaker populations is possible—at least, under laboratory environments—and, in this sense, the objective has been fulfilled.

A somewhat different motivation for speaker-recognition research came from a desire to isolate speaker-dependent parameters of speech from message-dependent ones. So far, little progress has been made in this respect. Most of the present evidence suggests that speaker-related and message-related information in speech are mixed in a fairly complicated fashion at the acoustic level. Almost every acoustic parameter derived from speech is speaker-dependent to some extent. Fortunately, the ability to define a set of simple and physically meaningful parameters for speech conveying information solely about the speaker is not crucial for the success of automatic methods. Modern statistical techniques enable us to combine the contributions of a large number of parameters—each of them individually might carry only a limited amount of information about the speaker—into a single measure which is highly selective in representing speaker-related information in speech. Our present knowledge of many of the transformations occurring during speech production at different levels—semantic, linguistic, articulatory, and acoustic—is very limited. Perhaps satisfactory answers to a number of basic questions relating to the manner in which speaker-related information is embedded in speech must wait till we gain a better understanding of the various physical processes involved in human speech production.

The motivation for future speaker-recognition research would come in part from a desire to find practical and economical applications of automatic speaker recognition. Past studies have left unanswered many questions of practical importance. For example, can the results based on small speaker populations be generalized to include large ones? Most practical applications are likely to involve considerably more speakers than have been used in experimental studies so far. The accuracy with which a particular speech parameter can be determined often depends upon the speaking environment. Results based on speech recorded under conditions which are free from noise and reverberation cannot be expected to provide a realistic estimate of the performance under practical conditions. One cannot always assume that the parameter-extraction method found to work well with a few speakers would continue to work satisfactorily for a large number of speakers. Very often, the speaker-recognition performance is found to depend upon the elapsed time between the recordings of the reference and the test utterances. It is inconvenient in practical systems to update the reference utterance of various speakers at frequent intervals. More research is needed to identify speech characteristics which show a high degree of stability over time. Finally, most speaker-recognition studies so far have used cooperative speakers who are not interested in fooling the machine, not a realistic situation for practical systems. Studies are needed to test the performance under the conditions where the impostors have some motivation to mimic.

REFERENCES

[1] F. McGehee, "The reliability of the identification of human voice," *J. Gen. Psychol.*, vol. 17, pp. 249–271, 1937.

[2] I. Pollack, J. M. Pickett, and W. H. Sumby, "On the identification of speakers by voice," *J. Acoust. Soc. Amer.*, vol. 26, pp. 403–406, May 1954.

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

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

[5] 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, Nov. 1966.

[6] B. S. Atal, "Automatic speaker recognition based on pitch contours," Ph.D. dissertation, Polytech. Inst. Brooklyn, Brooklyn, NY, June 1968.

[7] P. D. Bricker *et al.*, "Statistical techniques for talker identification," *Bell Syst. Tech. J.*, vol. 50, pp. 1427–1454, Apr. 1971.

[8] 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, Mar. 1971.

[9] J. E. Luck, "Automatic speaker verification using cepstral measurements," *J. Acoust. Soc. Amer.*, vol. 46, pp. 1026–1032, Oct. 1969.

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

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

[12] P. Garvin and P. Ladefoged, "Speaker identification and message identification in speech recognition," *Phonetica*, vol. 9, no. 4, pp. 193–199, 1963.

[13] M. H. L. Hecker, *Speaker Recognition: An Interpretive Survey of the Literature, ASHA Monogr.* 16 (American Speech and Hearing Association, Washington, DC), Jan. 1971.

[14] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, June 1974.

[15] 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, Apr. 1974.

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

[17] J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An algorithm for finding best matches in logarithmic time," Stanford Linear Accelerator Center Rep. SLAC-PUB-1549, Feb. 1975. This paper has been submitted for publication in the Communications of the ACM.

[18] R. W. Schafer and L. R. Rabiner, "Digital representations of speech signals," *Proc. IEEE*, vol. 63, pp. 662–677, Apr. 1975.

[19] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 51, pt. 2, pp. 2044–2055, June 1972.

[20] M. R. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 176–182, Apr. 1975.

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

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

[23] S. Kullback, *Information Theory and Statistics.* New York: Wiley, 1959, p. 6.

[24] T. Marill and D. M. Green, "On the effectiveness of receptors in recognition systems," *IEEE Trans. Inform. Theory*, vol. IT-9, pp. 11–17, Jan. 1963.

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

[26] S. Watanabe, "Karhunen-Loève expansion and factor analysis," in *Trans. IV Prague Conf. Information Theory, Statistical Decision, Functions, Random Processes.* Prague: Academia Publishing House, Czechoslovak Academy of Science, 1967, pp. 635–660.

[27] M. R. Schroeder and B. S. Atal, "Generalised short-time power spectra and autocorrelation functions." *J. Acoust. Soc. Amer.*, vol. 34, pp. 1679–1683, Nov. 1962.

[28] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd ed. New York: Springer-Verlag, 1972, pp. 141–161.

[29] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra.* New York: Dover, 1959.

[30] G. R. Doddington, "A method of speaker verification," *J. Acoust. Soc. Amer.*, vol. 49, pt. 1, p. 139(A), Jan. 1971.

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

[32] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, Feb. 1967.

[33] M. M. Sondhi, "New methods of pitch detection," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 262–266, June 1968.

[34] 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. 442–449, 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.