Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification*

B. S. Atal

Bell Laboratories, Murray Hill, New Jersey 07974
(Received 2 January 1974)

Several different parametric representations of speech derived from the linear prediction model are examined for their effectiveness for automatic recognition of speakers from their voices. Twelve predictor coefficients were determined approximately once every 50 msec from speech sampled at 10 kHz. The predictor coefficients and other speech parameters derived from them, such as the impulse response function, the autocorrelation function, the area function, and the cepstrum function were used as input to an automatic speaker-recognition system. The speech data consisted of 60 utterances, consisting of six repetitions of the same sentence spoken by 10 speakers. The identification decision was based on the distance of the test sample vector from the reference vector for different speakers in the population; the speaker corresponding to the reference vector with the smallest distance was judged to be the unknown speaker. In verification, the speaker was verified if the distance between the test sample vector and the reference vector for the claimed speaker was less than a fixed threshold. Among all the parameters investigated, the cepstrum was found to be the most effective, providing an identification accuracy of 70% for speech 50 msec in duration, which increased to more than 98% for a duration of 0.5 sec. Using the same speech data, the verification accuracy was found to be approximately 83% for a duration of 50 msec, increasing to 98% for a duration of 1 sec. In a separate study to determine the feasibility of text-independent speaker identification, an identification accuracy of 93% was achieved for speech 2 sec in duration even though the texts of the test and reference samples were different.

Subject Classification: 70.40, 70.55, 70.60.

INTRODUCTION

Parametric representation of speech derived by linear prediction of its waveform provides an accurate yet efficient representation of the speech signal for speech analysis–synthesis systems.[1,2] In linear prediction, the present speech sample is predicted as a linear combination of past speech samples; for a speech wave sampled at 10 kHz, 12 predictor coefficients defining the weights in the linear combination, together with additional parameters relating to the properties of the source, are sufficient to regenerate synthetic speech with little or no degradation in speech quality. The application of these techniques, of course, need not be limited to speech synthesis only. Speech analysis based on linear predictability of the speech waveform is carried out very efficiently on modern digital computers. The predictor coefficients thus offer a very convenient choice of parameters for representing speech efficiently in applications such as speech and speaker recognition. In this paper, some results on the effectiveness of the linear prediction characteristics of speech for automatic speaker identification and verification are presented.

Considerable research in the past few years has been directed towards finding speech characteristics which are effective for automatic speaker recognition.[3] Typical of the characteristics which have been investigated are the spectrographic data,[4–7] the pitch,[8–10] the intensity,[10] and the formants[10] of the speech signal. A summary of the results from several of these studies appears in Ref. 11. The linear prediction characteristics have many advantages for automatic speaker recognition: First of all, they are easily determined from the speech signal. Second, it is not necessary to make an *a priori* assumption as to which of the individual speech characteristics—such as a particular formant frequency or its bandwidth or some property of the glottal wave—would be effective. The 12 predictor coefficients represent the combined information about the formant frequencies, their bandwidth, and the glottal waveform.[1,2] Finally, being independent of the pitch and intensity information, the predictor coefficients could be used to improve the reliability of the existing methods of automatic speaker recognition based on pitch and intensity information.

I. LINEAR PREDICTION CHARACTERISTICS OF THE SPEECH WAVE

Consider the samples of the speech waveform band limited to 5 kHz and sampled at a frequency of 10 kHz. In the linear prediction model, samples $n-1$ to $n-p$ of the speech wave (p is an integer) are used to predict the nth sample. The predicted value of the nth speech sample is given by

$$\hat{s}_n = \sum_{k=1}^{p} s_{n-k} a_k, \tag{1}$$

where a_k's are the predictor coefficients and s_n is the nth speech sample. The value of p is determined by the total number of real and complex poles of the vocal tract and the glottal wave within a frequency range equal to half the sampling frequency. A typical value of p is 12, which is usually adequate for speech samples at 10 kHz.

Consider a speech segment having a duration of N samples. Under the assumption that neither the vocal tract shape nor the glottal waveform changes significantly over the analysis segment, the predictor coefficients are determined by minimizing the mean-squared prediction error between s_n and $\hat{s}_n$ defined as

$$\bar{\epsilon}_n^2=\frac{1}{N}\sum_n(s_n-\hat{s}_n)^2, \tag{2}$$

where the sum over n includes all the samples in the analysis segment. The coefficients a_k which minimize the mean-squared prediction error are obtained as solutions of the set of linear equations[1,2]

$$\sum_{k=1}^{p}\varphi_{jk}a_k=\varphi_{j0},\ j=1,2,\cdots,p, \tag{3}$$

where

$$\varphi_{jk}=\frac{1}{N}\sum_n s_{n-j}s_{n-k}. \tag{4}$$

A computationally efficient method of solving Eq. 3 is given in Ref. 2. In the linear prediction model, one represents the filtering action of the vocal tract, the radiation, and the glottal flow by a discrete linear filter with p poles. The transfer function $H(z)$ of this filter in the complex z domain is related to the predictor coefficients by the equation

$$H(z)=1\Big/\Big[1-\sum_{k=1}^{p}a_kz^{-k}\Big], \tag{5}$$

where z is the usual z-transform variable.[12] The transfer function $H(z)$ is related to the samples h_n of the impulse response of the filter by the equation

$$H(z)=\sum_{n=0}^{\infty}h_nz^{-n}. \tag{6}$$

On substituting Eq. 6 into Eq. 5, one obtains the following relationship between the predictor coefficients and the samples of the impulse response:

$$\begin{aligned}h_n&=\sum_{k=1}^{p}a_kh_{n-k}, && n>0,\\ &=1, && n=0,\\ &=0, && n<0.\end{aligned} \tag{7}$$

Furthermore, it can be shown that the first p samples $h_1, h_2, \cdots, h_p$ of the impulse response are sufficient to determine the p predictor coefficients uniquely. Thus, the linear prediction characteristics of the speech wave are also represented by the p numbers $h_1, h_2, \cdots h_p$. For mathematical purposes, the two representations are equivalent. It is interesting to enquire if they are also equivalent in their effectiveness for automatic speaker recognition. Later in this paper, we will compare the effectiveness of not only these two sets of parameters but also many other parameter sets which are related in a one-to-one manner to the predictor coefficients.

Consider next the autocorrelation function of the impulse response of the linear filter. Let r_k be the autocorrelation function at the kth sampling instant. By definition, r_k is given by

$$r_k=\sum_{n=0}^{\infty}h_nh_{n+|k|}. \tag{8}$$

The autocorrelation function samples r_n are related to the predictor coefficients a_k by the relationship[2]

$$r_n=\sum_{k=1}^{p}r_{|n-k|}a_k,\ n\geq 1, \tag{9}$$

and

$$r_0=\sum_{k=1}^{p}a_kr_k+1. \tag{10}$$

Furthermore, the relationship between the p autocorrelation function samples $r_1, r_2, \cdots, r_p$ and the p predictor coefficients $a_1, a_2, \cdots, a_p$ is unique. A knowledge of one is sufficient to determine the other.[13]

The transfer function $H(z)$ in Eq. 5 has exactly p poles. A transfer function with p poles is always realizable as the transfer function of a nonuniform acoustic tube formed by cascading p uniform cylindrical sections of equal length. The relationship between the cross-sectional areas of the p cylindrical sections and the predictor coefficients is also unique.[14] The p areas thus provide an alternative representation of the linear prediction characteristics of the speech wave.

Another important representation is obtained by considering the power series expansion of $\ln H(z)$, the logarithmic transfer function, in powers of z^{-1}. If all the poles of $H(z)$ are inside the unit circle, $\ln H(z)$ can be expressed by

$$\ln H(z)=C(z)=\sum_{n=1}^{\infty}c_nz^{-n}. \tag{11}$$

Since $z=\exp(j\omega t)$, where $\omega=$ frequency in radians and $T=$ sampling interval, c_n is the amplitude, at the nth sampling instant $t=nT$, of the inverse Fourier transform of $C(z)$ considered as a function of the frequency variable ω. A simple and unique relationship can be shown to exist between the parameters c_n's and a_n's.[15] To obtain this relationship, we substitute $H(z)$ from Eq. 5 into Eq. 11 and take derivatives on both sides of Eq. 11 with respect to z^{-1}; that is,

$$\frac{d}{dz^{-1}}\ln\Big[1\Big/\Big\{1-\sum_{k=1}^{p}a_kz^{-k}\Big\}\Big]=\frac{d}{dz^{-1}}\sum_{n=1}^{\infty}c_nz^{-n}, \tag{12}$$

which is simplified to

$$\Big\{\sum_{k=1}^{p}ka_kz^{-k+1}\Big\}\Big/\Big\{1-\sum_{k=1}^{p}a_kz^{-k}\Big\}=\sum_{n=1}^{\infty}nc_nz^{-n+1}, \tag{13}$$

and rewritten as

$$\sum_{k=1}^{p}ka_kz^{-k+1}=\Big(1-\sum_{k=1}^{p}a_kz^{-k}\Big)\sum_{n=1}^{\infty}nc_nz^{-n+1}. \tag{14}$$

If we now equate the constant term and the various powers of z^{-1} on the left and right sides of Eq. 14, we obtain the desired relationship between c_n's and a_n's namely,

$$c_1=a_1,$$

$$c_n=\sum_{k=1}^{n-1}(1-k/n)a_kc_{n-k}+a_n,\ 1<n<p,$$

and

$$c_n = \sum_{k=1}^{n-1} (1 - k/n)\, a_k c_{n-k}, \qquad n > p. \tag{15}$$

By defining a new variable $g_n = nc_n$, Eq. 15 can be written in a somewhat simpler form as

$$g_n = na_n + \sum_{k=1}^{n-1} a_k g_{n-k}, \tag{16}$$

where $a_n = 0$, $n > p$, and $g_1 = a_1$. It is interesting to note the similarity between Eqs. 7 and 16. Whereas h_n's, being the samples of the impulse response of the linear filter with the transfer function $H(z)$, represent the output of the filter in response to an input of unit impulse, g_n's are the samples at the output of the filter in response to input samples a_1, $2a_2$, $\cdots$, pa_p. It is well known that $H(z)$ can be expanded into a partial fraction expansion, each term of the expansion representing the contribution of a pole expressed by its residue at that pole. It can be shown that $C(z)$ too has a similar partial fraction expansion with all residues set to unity, implying that different poles have the same amplitude. The important difference between the two parameter sets h_n and g_n is thus in the amplitudes with which different poles or resonances are represented.

Equation 15 allows us to compute the coefficients c_n from the p predictor coefficients and the predictor coefficients from the p coefficients c_1, c_2, $\cdots$, c_p. To conform to the terminology used previously in the speech literature, we shall call c_n's the samples of the cepstrum function.[15] It may be pointed out that, traditionally, the cepstrum function is obtained by a double Fourier transform operation on the impulse response samples h_n. For a transfer function with poles only, the cepstrum can be obtained directly from the impulse response samples h_n by the equation

$$c_n = \sum_{k=1}^{n-1} (1 - k/n)\, h_k c_{n-k} + h_n, \quad 1 < n,$$

and

$$c_1 = h_1, \tag{17}$$

or from the predictor coefficients a_k by Eq. 15.[16]

The parameter sets discussed so far include the predictor coefficients, the impulse response, the autocorrelation function, the area function, and the cepstrum. Besides these, a large number of additional parameters can be generated by linear transformations of any one of the above sets of parameters. For example, power spectrum, being the Fourier transform of the autocorrelation function, represents simply a linear transformation of the autocorrelation function. Similarly, the logarithm of the power spectrum also represents a linear transformation of the cepstrum. It is not necessary to include such parameters in our discussion, since, by proper choice of a distance metric, the distance between any two points in a multidimensional space can be made invariant with respect to any arbitrary linear transformation. Thus, if the decision alogarithm is based on the distance between a test and a reference pattern, it is immaterial whether the sets of parameters are transformed linearly or not. Hence, as far as the recognition accuracy is concerned, we can regard a set of parameters, which can be obtained from each other by linear transformations, as completely equivalent.

II. SPEECH DATA COLLECTION

The speech data consisted of 60 utterances, consisting of six repetitions of the same sentence, spoken by 10 speakers. All of the speakers were female and they spoke the sentence "May we all learn a yellow lion roar." The recordings were made in an anechoic chamber on two different days with an interval of 27 days; on each day, the recordings were made in three separate sessions. These recordings were identical to ones used by the author in an earlier study on speaker recognition based on pitch.[8,9]

The speech signal was sampled after low-pass filtering[17] at a rate of 10 000 times/sec and quantized into 12-bit binary numbers in an analog-to-digital converter for further processing on a digital computer. In the analysis, each utterance was divided into 40 equally long segments; the duration of each segment was made proportional to the duration of the utterance to provide an approximate alignment of the time scales of the different utterances. Twelve predictor coefficients were determined using the procedure outlined in Sec. I. The sum over the index n in Eq. 4 included approximately 500 speech samples corresponding to an analysis interval of 50 msec; the actual analysis interval varied with the duration of the utterance. The silent portions and pauses in an utterance were eliminated from the analysis interval. These portions of the utterance were determined by comparing the energy in contiguous speech segments 10 msec in duration with a fixed threshold; a segment having energy less than the threshold was classified as either a silent portion or a pause and was removed from the analysis interval of Eq. 4.

The linear prediction analysis produced a set of 12 predictor coefficients for each of the 40 uniformly spaced time frames in an utterance. The analysis was carried out for all of the 60 utterances of the 10 speakers and the resulting data were used to test its effectiveness for automatic speaker recognition.

III. SPEAKER-RECOGNITION PROCEDURE

Speaker-recognition tests were carried out to determine the effectiveness of the various parameter sets (discussed in Sec. I) for automatic speaker recognition. The task in a speaker-recognition test could either be that of identifying an unknown speaker in a population of several speakers of known characteristics, or that of verifying whether the person is what he claims to be. Of the two, the speaker-identification task is more suited for comparing the effectiveness of the different parameters. In the speaker-identification task, a single error rate can provide a measure of the performance, while in the speaker-verification task two kinds of errors, namely, the probabilities of false verification and false rejection, as functions of a threshold parameter, determine the performance. Furthermore, the accuracy with which one can identify speakers correctly provides a

more sensitive indication of the ability of a parameter for discriminating speakers. Consequently, most of the results in this paper in evaluating the different parameters relate to speaker identification. The results relating to speaker verification are provided only for the parameter found most effective for speaker identification.

The basic parameter data used in this study consisted of 12 coefficients obtained at 40 uniformly spaced time frames for each of the 60 utterances (six repetitions for each of the 10 speakers). Each set of six repetitions of an utterance for a speaker was partitioned into design and test sets. The utterances in the design set were used to form the rules for speaker recognition; the utterances in the test set were used to test the effectiveness of the speaker-recognition procedure.

Let us first consider speaker recognition based on the parameters in a single time frame. These parameters can be represented as a vector in a 12-dimensional space. Let $\mathbf{x}_{\alpha j}$ be the 12-dimensional vector representing the αth utterance of the jth speaker in the design set. A reference vector is obtained for each speaker by averaging the vectors for that speaker. Let $\mathbf{r}_j$ be the reference vector for the jth speaker. Then,

$$\mathbf{r}_j = \langle \mathbf{x}_{\alpha j} \rangle_\alpha, \tag{18}$$

where $\langle\ \rangle_\alpha$ indicates averaging over the subscript α. Let $\mathbf{z}$ be a vector in the test set which is to be identified or verified. For determining how similar a given test vector $\mathbf{z}$ is to a given reference vector $\mathbf{r}_j$, we introduce a distance metric d_j, representing the distance between the vector $\mathbf{z}$ and the reference vector of the jth speaker and define it as

$$d_j \equiv [(\mathbf{z}-\mathbf{r}_j)^t\, W^{-1}(\mathbf{z}-\mathbf{r}_j)]^{1/2}, \tag{19}$$

where W is the pooled intraspeaker covariance matrix and $(\ )^t$ indicates the transpose of the vector inside the brackets. The matrix W is given by

$$W = \langle (\mathbf{x}_{\alpha j}-\mathbf{r}_j)(\mathbf{x}_{\alpha j}-\mathbf{r}_j)^t \rangle_{\alpha j}, \tag{20}$$

where the averaging is done over both subscripts α and j. For speaker identification, the distance d_j is computed for each speaker in the population and the test vector is associated with the speaker corresponding to the smallest distance. For speaker verification, the distance between the test vector and the reference vector of the claimed speaker, say j, is compared with a threshold. If d_j is found smaller than the threshold value, the speaker is verified; otherwise, he is rejected.

The choice of the distance metric d_j defined in Eq. 19 has several advantages: First, the distance d_j is invariant with respect to any arbitrary nonsingular linear transformation of the vector space. To prove this result, let $\hat{\mathbf{z}}$ and $\hat{\mathbf{r}}_j$ represent the test and reference vectors in the transformed space. Thus,

$$\hat{\mathbf{z}} = T\mathbf{z}$$

and

$$\hat{\mathbf{r}}_j = T\mathbf{r}_j, \tag{21}$$

where T is a 12×12 nonsingular matrix. The distance

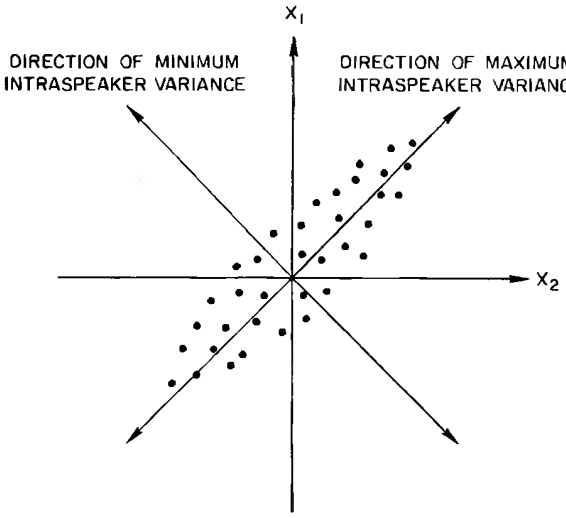

FIG. 1. Two-dimensional plot showing the directions of maximum and minimum intraspeaker variability.

$\hat{d}_j$ between the test and the reference vectors in the transformed space is given by

$$\hat{d}_j = [(\hat{\mathbf{z}}-\hat{\mathbf{r}}_j)^t\, \hat{W}^{-1}(\hat{\mathbf{z}}-\hat{\mathbf{r}}_j)]^{1/2}$$
$$= [(\mathbf{z}-\mathbf{r}_j)^t\, T^t \hat{W}^{-1}\, T(\mathbf{z}-\mathbf{r}_j)]^{1/2}. \tag{22}$$

Since $W^{-1} = T^t\hat{W}^{-1}T$ from Eq. 20, $\hat{d}_j = d_j$. This property of the distance metric of being invariant with respect to arbitrary linear transformations of the vector space is very important in allowing us to test a large number of parameter sets, related to each other by linear transformations, by testing only a single set in the speaker recognition test. As an example, the well-known Fourier transform is a linear transformation, and nothing is to be gained by considering those parameters which are Fourier transforms of an already existing set of parameters.

Second, an even more important property of the distance metric d_j is explained by means of Fig. 1, where a vector is represented by a point in a two-dimensional space. Consider now the vectors originating from a particular frame of a single speaker. It is obvious that, in general, the different coordinate directions are not equally good in producing a small variation within the different patterns of the same speaker. A good distance metric thus should weight the different coordinate directions in order of their importance in reducing the intraspeaker variations. It can be shown easily that the distance metric considered here does indeed possess this property.[18] The often-used Euclidean distance metric, on the other hand, does not have this property.

So far, this discussion was confined to distances for a single time frame. Actually, a total of 40 distances corresponding to 40 time frames are obtained in a single utterance, and the recognition performance can be improved by combining the distances from the individual frames. Let $d_j^{(k)}$ be the distance from the reference vector of the jth speaker in the kth time frame. Let us define an average distance D_j as

$$D_j = \langle \ln[1 + d_j^{(k)}] \rangle_k. \tag{23}$$

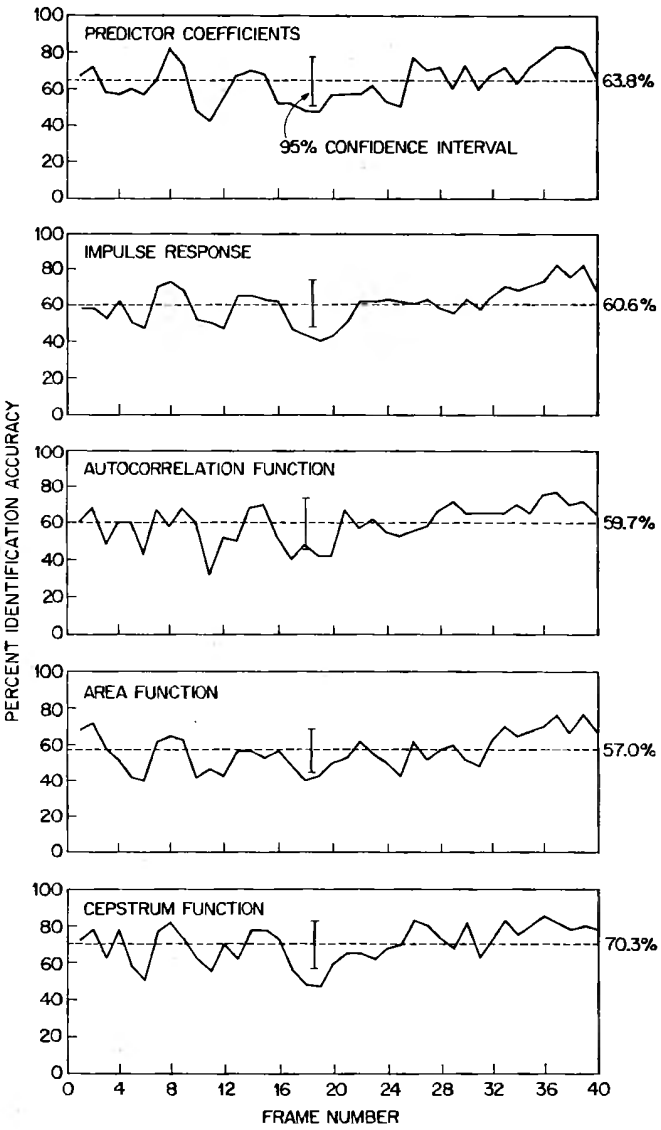

FIG. 2. Percent identification accuracy for different parametric representations of speech based on a 50-msec-long speech segment.

The reason for choosing the log operation prior to averaging is to avoid introduction of large errors in D_j due to a few large distances in some of the time frames. Similarly, the addition of 1 before the log operation is simply to avoid a disproportionate contribution to D_j by a zero distance in a single time frame. The "distance" D_j is used in the same manner as described earlier for d_j for each speaker-recognition task.

IV. RESULTS

A. Effectiveness of different parameters.

The identification accuracy for each of the 40 time frames for the five sets of parameters discussed in Sec. I is shown in Fig. 2. The number of parameters p in each case equals 12. 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 set for each of the 10 speakers to provide a total of 60 judgments for each time frame.

The identification accuracy averaged over the 40 time frames was found to be 63.8% for the predictor coefficients, 60.6% for the impulse response, 59.7% for the autocorrelation function, 57.0% for the area function, and 70.3% for the cepstrum. The cepstrum provides the highest identification score while the area function provides the lowest. The 95% confidence interval for the means is approximately ±2.0%; thus, the identification accuracy provided by the cepstrum is significantly higher than the rest of the parameters. The identification accuracy, of course, varies considerably from one timeframe to another—the average standard deviation equals 6.0%. However, as shown in Fig. 2, this variation, in most cases, is within the 95% confidence interval—the exceptions occur for frames 16–20 and 36–40. The lower identification accuracy for frames 16–20 is probably due to the pause which occurred in some of the utterances after the word "learn" and introduced considerable variability in the time frame just before and after the pause.

B. Identification accuracy as a function of duration of the spoken material

The results presented so far have been based on a single time frame with an average duration of 50 msec. By combining two or more time frames in the average distance metric defined in Eq. 23, the identification accuracy can be computed as a function of the duration of the spoken material. These results are shown in Fig. 3. The solid curve is obtained when the duration is increased from zero onwards by combining time frames starting at frame 1 at the beginning of the utterance. The dashed curve is obtained by combining frames starting at frame

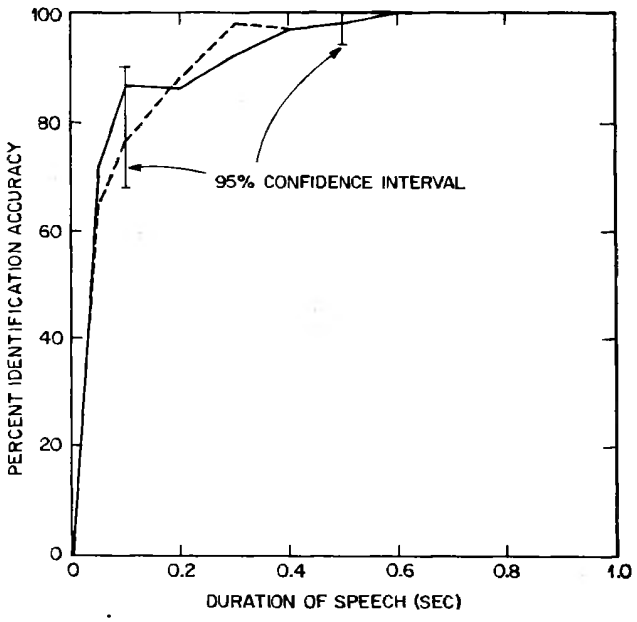

FIG. 3. Percent identification accuracy as a function of speech duration. The solid curve is for speech starting at the beginning of the utterance, while the dashed curve is for speech starting at the middle of the utterance.

21 at the middle of the utterance. The 95% confidence intervals at two places—one with maximum spread and the other with minimum spread between the solid and dashed curves—are also shown in Fig. 3, indicating that the differences between the two curves are not significant. The results show an identification accuracy of approximately 80% for a duration of 0.1 sec and an accuracy greater than 98% for durations of 0.5 sec and higher.

C. Comparison with results based on pitch contours

Speech characteristics such as pitch or the amplitude envelope function are not included in the parameter set used above. Thus, it is possible to compare the linear prediction characteristics with other speech characteristics for their effectiveness for speaker recognition. For example, for the same speech data set as used here, an identification accuracy of 97% for a duration of 2 sec was obtained using parameters derived from the pitch contours.[8,9] On the other hand, an identification accuracy of 98% is achieved for the 12 cepstral coefficients for a duration of only 0.5 sec, 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 information is more effective than the pitch contour information for automatic speaker recognition.

D. Comparison with human performance

It is interesting to compare the performance of an automatic method with the performance of a human listener for speaker identification. In a study discussing the effects of stimulus content and duration on talker identification by human listeners, Bricker and Pruzansky[19] give average identification scores of 56% for vowel excerpts (mean duration = 117 msec), 63% for CV excerpts (mean duration = 117 msec), 81% for monosyllables (mean duration = 498 msec), 87% for disyllables (mean duration = 446 msec), and 98% for sentences (mean duration = 2.4 sec). If these identification scores are compared with the results shown in Fig. 3 for different durations, the automatic method is found to perform much better than the average human listener. The result is even more intriguing if one considers the fact that the human listeners could use the pitch, the intensity, and the duration, as well as the spectral information, while the automatic method was constrained to use only the spectral envelope information contained in the 12 cepstral coefficients.

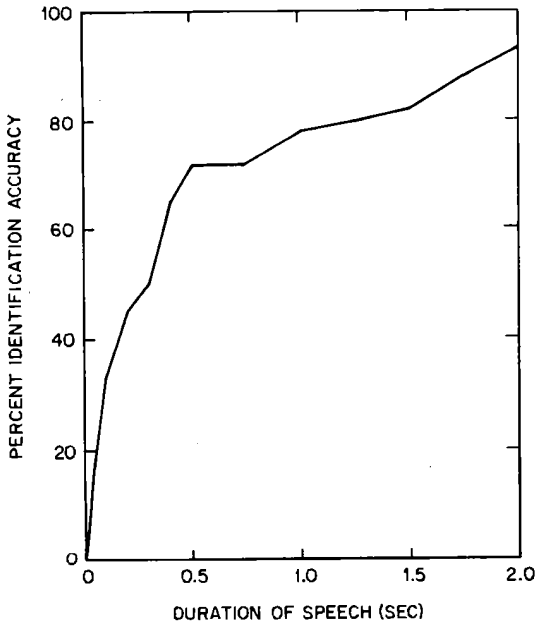

FIG. 4. Percent identification accuracy as a function of speech duration when the texts of the test and reference speech samples are different.

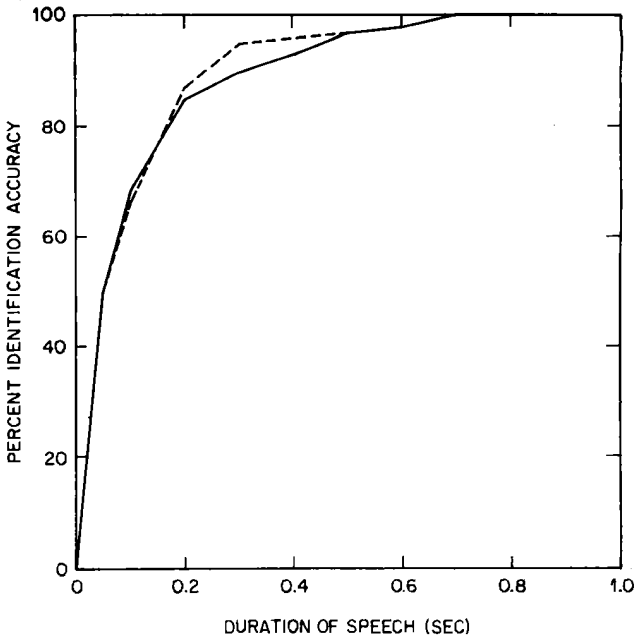

FIG. 5. Percent identification accuracy as a function of speech duration when the time average of the cepstral samples is eliminated from the data. The solid curve is for speech starting at the beginning of the utterance, while the dashed curve is for speech starting at the middle of the utterance.

E. Text-independent speaker recognition

In automatic speaker-recognition methods, it is generally required that the texts of the test and the reference utterances be identical. It is now interesting to enquire if such a restriction is really necessary for successful speaker recognition. After all, we, as human listeners, often recognize people from their voices, even though the spoken text is different from one occasion to another. This is not to say that the task in the text-independent case is not harder than when the text is the same. Indeed, text-independent speaker recognition is more difficult since one has now to cope with the additional variability due to the differences in the texts of the test and the reference utterances. However, this additional source of intraspeaker variability is in principle no different from others and the distance measure d_j defined in Eq. 19 is still a suitable distance measure in this case.

In order to use the existing speech recordings for the text-independent case, each of the 60 utterances was divided into 40 equal segments and the 40 segments were recombined in a random order to form a new utterance, the random order being different for each utterance. The randomization procedure destroyed the synchronization of the text present in the original set of utterances. Twelve cepstral coefficients were obtained for each segment of the randomized utterance and were used to identify a speaker using the procedure outlined in Sec. III, the identification procedure being identical to one used for the text-dependent case. The resulting identification accuracy, computed as a function of the duration of the spoken material, is shown in Fig. 4. The identification accuracy for a single segment 50 msec in duration is 17% and is increased to 72% and 93% for durations of 0.5 sec and 2.0 sec, respectively. For comparison, the results for the text-dependent case were 72% for 50 msec and 98% for 0.5 sec. The identification accuracy of 93% for the text-independent case, although much lower than the text-dependent case, is surprisingly high considering the fact that no pitch or intensity parameters have been used in the identification procedure.

F. Identification accuracy based on cepstral coefficients with time averages removed

The linear prediction parameters discussed so far suffer from an important limitation, namely, that the parameter data are influenced by the frequency response of the recording apparatus as well as the transmission system. The cepstral coefficients, 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 cepstral coefficients, a set of values representing their time averages over the duration of the entire utterance.

Let $\mathbf{c}_n$ be a 12-dimensional vector representing the 12 cepstral coefficients for the nth time frame in an utterance. The vector $\mathbf{c}_n$ can be written as the sum of two vectors, the first one representing the average value of the vector over the 40 time frames and the second representing the variation of the vector about its average value. Thus, let

$$\mathbf{c}_n = \bar{\mathbf{c}}_n + \hat{\mathbf{c}}_n, \tag{24}$$

where

$$\bar{\mathbf{c}}_n = \frac{1}{40}\sum_{n=1}^{40} \mathbf{c}_n. \tag{25}$$

The results presented so far have been based on the vector $\mathbf{c}_n$. It is now interesting to find out if the identification accuracy is lowered if the vector $\hat{\mathbf{c}}_n$ rather than $\mathbf{c}_n$ is used for representing a speaker. As pointed out earlier, an important advantage gained by removing the time averages from the data is that all time-invariant frequency distortions in the data are eliminated. Thus, the frequency response of the microphone and the transmission system has no influence on the data. The above 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 which is now eliminated due to averaging along the time axis. Furthermore, removal of time averages is likely to make an automatic speaker identification or verification system more reliable from the point of being fooled by mimics. An impostor may find it much easier to mimic those voice characteristics of a speaker which remain fixed in time as compared to those which vary in time.

The identification accuracy as a function of the duration of the spoken material using the vectors $\hat{\mathbf{c}}_n$, that is, the cepstral coefficients with the time averages removed, is shown in Fig. 5. The solid curve of Fig. 5 is obtained when the duration is increased from zero onwards by combining frames starting at frame 1 at the beginning of the utterance. The dashed curve in the figure is obtained by starting at frame 21 at the middle of the utterance. There are differences between the two curves but, mostly, these differences are not significant. The results show an identification accuracy of 68% for a duration of 0.1 sec and an accuracy of greater than 98% for duration of 0.6 sec and higher. On comparing these results with the ones presented in Fig. 3, where the temporal averages were not removed from the data, one finds that the identification accuracy is slightly lower (for example, 68% as compared to 80% for a duration of 0.1 sec) but not significantly so. In fact, for durations larger than about 0.3 sec, the differences are very small. To some extent, this result is to be expected. As the duration is increased, more phonemes are added, causing a decrease in the error rate. This decrease is going to be slower when time averages are not removed owing to higher correlation between the frames corresponding to different phonemes.

G. Speaker-verification results

For the speaker verification tests, one of the 10 speakers was designated as the speaker to be verified and the remaining nine were considered impostors. As before, five of the utterances of a speaker were used to form the reference vectors and the remaining sixth was used as a test utterance. For verification, the distance d_j (see Eq. 19) between the test vector and the reference vector of the claimed speaker, namely j, is computed and averaged over time to form the average distance D_j. If the distance D_j is smaller than a preselected threshold value, the speaker is verified; otherwise, he is rejected. Two kinds of errors are possible in a speaker-verification task. The error of the first kind, namely, that of false verification, occurs when an impostor is verified as the claimed speaker. The error of a second kind, namely, that of false rejection, occurs when an honest speaker is rejected. The overall error rate depends upon the *a priori* probabilities that a random customer in the population is an honest person or an impostor. For purposes of the present discussion, these *a priori* probabilities are assumed to be equal. The total error rate is then the average of the error rates

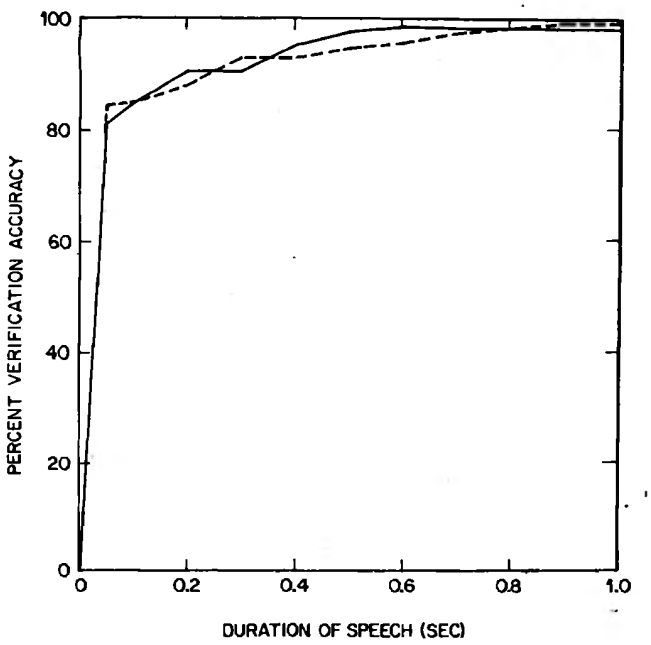

FIG. 6. Percent verification accuracy as a function of speech duration. The solid curve is for speech starting at the beginning of the utterance, while the dashed curve is for speech starting at the middle of the utterance.

for the two kinds of errors. The verification accuracy is defined as one minus the average error rate expressed as a percentage. The verification accuracy, of course, depends upon the threshold parameter used in comparing the distance D_j. A very small value of the threshold parameter increases the errors of false verification while decreasing the errors of false rejection. The reverse is true for very large values of the threshold parameter. An optimum value for the threshold parameter is selected to be one which produces the highest verification accuracy. A single threshold value is assumed for all of the speakers in the population, although, in principle, one could select a different threshold for each speaker.

The results of the speaker-verification tests are shown in Fig. 6. The parameter set consists of 12 cepstral coefficients. As before, the solid curve is obtained when the duration of speech is increased from zero onwards by combining frames starting from frame 1 at the beginning of the utterance, while the dashed curve is obtained by starting at frame 21 at the middle of the utterance. These results show a verification accuracy of approximately 90% for a duration of 0.2 sec, 95% for a duration of 0.5 sec, and 98% for a duration of 1 sec. The results of the speaker-verification tests after the time averages have been removed from the data are shown in Fig. 7. For durations of 0.2 sec and greater, there are no significant differences between the results of Figs. 6 and 7. Thus, removal of time averages does not adversely affect the performance of the speaker verification procedure.

V. CONCLUSIONS

The effectiveness of several different parametric representations of speech derived from the linear predictor coefficients was determined for automatic recognition of speakers from their voices. The different parameter sets investigated were, in addition to the predictor coefficients, the impulse response function corresponding to the transfer function based on the predictor coefficients, the autocorrelation function of the impulse response, the area function of an acoustic tube with an identical transfer function, and the cepstrum representing the logarithmic transfer function. The various parameters differed somewhat in the realized accuracy of identification of speakers, but not by a wide margin; the cepstrum produced the highest average identification accuracy of 70.3% for 50 msec of speech, while the area function produced the lowest value of 57.0% under the same conditions. The identification accuracy increased with the duration of the spoken material, yielding a value of 98% for a duration of 0.5 sec for the cepstrum. Furthermore, the identification accuracy was not affected significantly by removal of the time averages from the cepstrum samples. This result can be used to a great advantage for eliminating the influence of any frequency-response distortions introduced by the recording or the transmission system from the cepstrum data. A non-Euclidean distance metric was used to determine the distance between a test vector and a reference vector. This particular distance metric has the important property of weighting down the components in those directions in the original coordinate space along which the intra-speaker variance is large. The distance metric was also found to be very effective for text-independent speaker recognition. Speaker identification with an accuracy of 93% was achieved for a 2-sec-long speech sample even though the texts of the test and reference speech samples were different. Tests were also conducted to determine the efficiency of the cepstrum parameters for automatic speaker verification. A verification accuracy

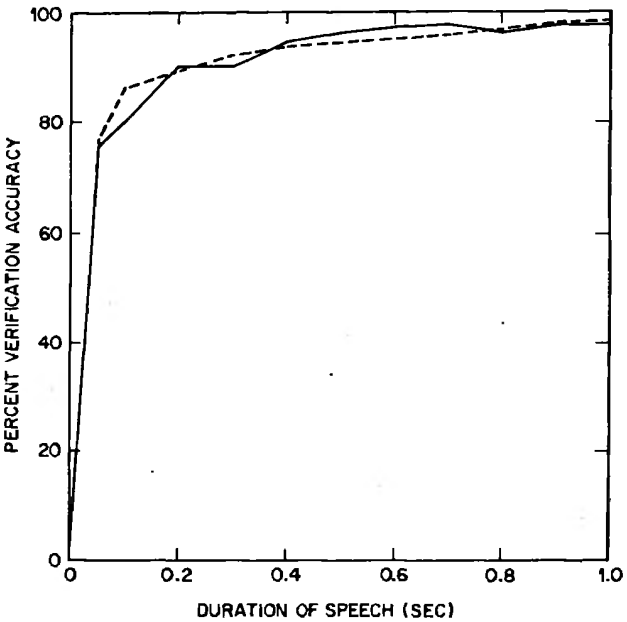

FIG. 7. Percent verification accuracy as a function of speech duration when the time average of the cepstral samples is substracted out from the data. The solid and the dashed curves refer to the same conditions as in Fig. 5.

of 90% was achieved for speech with a duration of 0.2 sec and of 98% for a duration of 1 sec.

The results of this paper show that the predictor coefficients, or an equivalent set of parameters derived from them, provide a very effective representation of speech for automatic speaker recognition. The linear prediction characteristics provide a representation of the spectral envelope of the speech signal. The recognition accuracy achieved was found to be significantly higher than achieved by pitch or intensity function of the speech signal. Of course, in any practical implementation, all of these speech characteristics could be used to provide reliable speaker recognition by machines.

*Presented in part at the 83rd Meeting of the Acoustical Society of America, Buffalo, N.Y. 18–21 April 1972 [J. Acoust. Soc. Am. 52, 181(A) (1972)].

[1]B. S. Atal and M. R. Schroeder, "Adaptive Predictive Coding of Speech Signals," Bell System Tech. J. 49, 1973–1986 (1970).

[2]B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Preduction," J. Acoust. Soc. Am. 50, 637–655 (1971).

[3]Throughout this paper, the word, "recognition" is used to include both identification and vertification.

[4]S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition," J. Acoust. Soc. Am. 35, 354–358 (1963).

[5]K. P. Li, J. E. Dammann, and W. D. Chapman, "Experimental Studies in Speaker Verification Using an Adaptive System," J. Acoust. Soc. Am. 40, 966–978 (1966).

[6]P. D. Bricker, R. Gnanadesikan, M. V. Mathews, S. Pruzansky, P. A. Tukey, K. W. Wachter, and J. L. Warner, "Statistical Techniques for Talker Identification," Bell System Tech. J. 50, 1427–1454 (1971).

[7]S. K. Das and W. S. Mohn, "Pattern Recognition in Speaker Verification," AFIPS Conf. Proc., Fall Joint Computer Conf. 35, 721–732 (1969).

[8]B. S. Atal, "Automatic Speaker Recognition Based on Pitch Contours," Ph.D. Thesis, Polytechnic Inst. Brooklyn (June 1968).

[9]B. S. Atal, "Automatic Speaker Recognition Based on Pitch Contours," J. Acoust. Soc. Am. 52, 1687–1697 (1972).

[10]G. R. Doddington, "A New Method of Speaker Verification," J. Acoust. Soc. Am. 49, 139(A) (1971).

[11]J. L. Flanagan, *Speech Analysis and Perception*, (Springer–Verlag, New York, 1972), pp. 199–204.

[12]J. R. Ragazzini and G. F. Franklin, *Sampled-Data Control Systems* (McGraw–Hill, New York, 1958).

[13]Ref. 2, pp. 642–643.

[14]Ref. 2, pp. 653–655.

[15]A. V. Oppenheim and R. W. Schafer, "Homomorphic Analysis of Speech," IEEE Trans. Audio Electroacoust. AU-16, 221–226 (1968).

[16]It may be pointed out here that the cepstrum obtained from Eqs. 15 or 17 is not identical with the cepstrum obtained by taking the logarithm of the short-time Fourier transform of the signal even when the transfer function has no zeros. The difference between the two is due to the truncation of the signal to a finite duration prior to Fourier analysis which produces a time series with zeros only irrespective of whether the transfer function has poles or zeros. The cepstrum obtained from the predictor coefficients, on the other hand, is based on the explicit assumption that the transfer function has only poles.

[17]The cutoff frequency of the low-pass filter was 5.0 kHz.

[18]G. S. Sebestyen, *Decision-Making Processes in Pattern Recognition* (The Macmillan Company, New York, 1962), pp. 24–30.

[19]P. D. Bricker and S. Pruzansky, "Effect of Stimulus Content and Duration on Talker Identification," J. Acoust. Soc. Am. 40, 1441–1449 (1966).