Speech Communication 1 (1982) 257–267
North-Holland Publishing Company

TEXT-INDEPENDENT SPEAKER RECOGNITION: A REVIEW AND SOME NEW RESULTS

M. SHRIDHAR and N. MOHANKRISHNAN

Electrical Engineering Department, University of Windsor, Windsor, Ont., Canada N9B 3P4

Received 19 February 1982
Revised 17 September 1982

Abstract. The development of a high accuracy (about 99%) text-independent speaker recognition system is discussed in this paper in two stages. The first stage deals with the evaluation of the speaker selectivity characteristics of the various parameter sets that characterize human speech. The second stage utilizes any two parameter sets of the first stage tests and combines these logically to obtain a significantly higher recognition accuracy than is possible with any single speaker-sensitive parameter set. The algorithm utilizes additional utterances to resolve contradictory decisions from the two parameter sets resulting from the first test utterance. For completeness of discussion a brief review of relevant literature in the field is also presented.

Zusammenfassung. Der vorliegende Beitrag befasst sich in zwei Abschnitten mit der Entwicklung eines textunabhängigen Sprechererkennungssystems hoher Genauigkeit (Erkennungsrate ungefähr 99%). Der erste Abschnitt untersucht die Eigenschaften der verschiedenen parametrischen Darstellungsmöglichkeiten des Sprachsignals hinsichtlich ihrer Sprecherselektivität. Im zweiten Abschnitt werden jeweils zwei im ersten Abschnitt behandelte parametrische Darstellungen getestet und logisch verknüpft. Hiermit lässt sich eine erheblich höhere Erkennungswarhscheinlichkeit erreichen, als wenn nur ein einziger sprechersensitiver Parametersatz verwendet wird. Der Algorithmus untersucht zusätzliche sprachliche Äusserungen (des gleichen Sprechers), wenn sich aus der ersten Testäusserung für zwei verschiedene parametrische Darstellungen widersprüchliche Entscheidungen ergeben. Ein kurzer Überblick über die einschlägige Literatur rundet die Diskussion ab.

Résumé. Le développement d'un système de reconnaissance du locuteur à haute précision (≈ 99%) et indépendant du texte est envisagé dans ce travail en deux étapes. La première phase traite de l'évaluation des caractéristiques sélectives du locuteur pour les différents ensembles de paramètres qui caractérisent le langage humain. Dans la seconde étape, on apparie deux à deux tous les ensembles de paramètres et on les combine logiquement pour obtenir une précision de reconnaissance plus élevée que celle qui pourrait être atteinte avec chaque ensemble isolé. Cet algorithme permet d'utiliser des phrases supplémentaires pour résoudre les décisions contradictoires résultant de l'application à la première phrase-test de la procédure à deux ensembles de paramètres. Pour compléter la discussion, une brève revue de la littérature pertinente est également présentée.

Keywords. Speech, speaker recognition, text-independent.

1. Introduction

The development of a reliable computer-based text-independent speaker recognition system is expected to result in significant advances in such diverse fields as forensic sciences, commercial banking systems for telephone-based transactions and military applications involving release of classified information over communication channels. Most speaker recognition systems that have been proposed or developed utilize the following procedure:

(a) Derivation of a set of acoustic features for each speaker in the participating population, to be stored in the computer library.

(b) Derivation of the acoustic features of the test speaker from an analysis of the test speaker's utterance.

(c) Derivation of a quantitative measure of similarity or dissimilarity between the test speaker's acoustic features and the reference acoustic features stored in the computer library.

(d) A recognition decision based on the similarity/dissimilarity measure.

The recognition system is classified as text-dependent or text-independent on the basis of whether the chosen utterances for derivation of the acoustic features are prespecified or not. A further classification into identification and verification is made on the basis of whether the recognition system is required to identify or verify the identity claim of the test speaker.

In this paper, a brief review of the relevant literature in this field is presented in Section 2, with a view to bringing into focus the unique problems and the unique approaches to speaker recognition. Section 3 describes the procedure for speech data acquisition and the preprocessing of data prior to implementing the recognition algorithms. The parameter sets investigated for their speaker discrimination potential are listed in Section 4. The process of orthogonalization of the parameter sets is described in Section 5. Section 6 discusses the results obtained from an investigation of the convergence properties of the covariance matrix of the speech parameters, a critical step in the orthogonalization process. The results or the recognition tests conducted to evaluate the performance of the speech parameter sets is presented in Section 7. This is then followed by a discussion of the author's new approach towards the development of a high accuracy text-independent identification/verification system in Section 8. In this new work the authors discuss the feasibility of logically combining the results of two independent tests to increase the recognition accuracies up to about 99%. The proposed scheme utilizes additional utterances to resolve conflicting decisions by the two independent tests. It is shown that these high accuracies may be achieved with small duration test utterances (5 sec). Finally, the conclusions of this work are summarized in Section 9.

2. Literature review

A survey of the literature reveals the existence of quite a few text-dependent speaker recognition systems with accuracies better than 99%. However, there are very few high accuracy text-independent speaker recognition systems which do not have to resort to extensive averaging of feature vectors over long segments of speech.

A wide range of voice characteristics have been used in the past for speaker recognition. These include such features as pitch, intensity, linear prediction coefficients, vocal tract resonances and area functions, formants, spectral parameters, cepstral coefficients, etc. Atal [1] and Rosenberg [2] have conducted an extensive survey of the parameters used and their performances in speaker recognition tasks. We shall list only a few of the more conspicuous research efforts in this area.

Atal [3] investigated the effectiveness of pitch contours in speaker recognition and obtained an identification accuracy of 97% for a speech utterance about 2 sec in duration. He also investigated the effectiveness of the linear prediction characteristics of the speech wave in speaker recognition [4] using the same data base. The parameter sets investigated included the predictor coefficients, the impulse response corresponding to the transfer function $H(Z)$ of the vocal tract, the autocorrelation function of the impulse response, the area function and the cepstrum coefficients. The highest identification accuracy of 98% was obtained when the cepstral coefficients were used with an utterance duration of only 0.5 sec.

Sambur [5] proposed the use of orthogonalised speech parameters for speaker recognition. This technique uses the eigenvectors and eigenvalues of the covariance matrix of the speech parameters, estimated over the reference utterances of a speaker, to orthogonally transform both the reference and test parameters of the speakers under consideration. The distance between the test and reference parameters are then evaluated in the orthogonal domain. Sambur surmised that the least significant orthogonal parameters, which exhibit small variances across the analyzed utterance, were indicative of the talker's identity, whereas the most significant parameters with the largest variances, were reflective of the linguistic content of the utterance. The identification and verification scores obtained were better than 99% when the Parcor and log area ratio coefficients were used as the feature set.

Several recognition schemes have used the spectral speech characteristics of a speaker to formulate recognition schemes. The work of Pruzansky and Mathews [6] and Das and Kohn [7] were some of the earliest efforts in this area. In the former, 17 frequency channels covering the range from 100 Hz to 10 KHz were sampled every 10 msec to form the spectrogram. Elemental energies in one frequency channel and one time interval of the spectrogram of single words excerpted from sentences of 10 talkers were used as features. The effect of averaging groups of elemental energies over a rectangular area of the spectrogram, formed from several adjacent frequency channels and time intervals, was also investigated. A maximum recognition accuracy of 88% was obtained when all features were used, an accuracy that was maintained when only 25% of the features selected on the basis of their *F*-ratios (ratio of inter-speaker to intra-speaker feature variance), were used. It was also shown that while averaging of elemental energies over time improved performance, averaging over frequency worsened it.

Das and Mohn [7] used an extensive data base of phrases recorded from 50 real speakers and 68 impostors over a period of more than 5 weeks. The spectrum of the speech was obtained by passing it through a bank of 20 bandpass filters ranging from 188–8203 Hz and sampling every 20 msec after full-wave rectification and smoothing. An elaborate segmentation procedure was used to isolate various phonetic speech events in the utterances, to enable identical sections of different utterances to be compared. Based on the segmentation points, 405 features were calculated for each utterance, such as filter averages around segmentation points, filter averages of linear time-normalized data, formant information, time difference between adjacent pairs of segmentation points, fundamental frequency, etc. Based on the *F*-ratio criterion a subset of 200 of the 405 features was selected. A mean error rate of misclassification of 1% was obtained. Furui et al. [8] formulated a recognition scheme based on the long-time average spectrum of a short sentence as speaker-characterizing parameters. A suitable time-spaced reference measure was used to account for the long-term variation of the spectrum pattern. Identification scores of 91% and verification scores of between 93 and 95% were obtained with test samples recorded 3 months after the last reference sample.

Another technique based on the location and isolation of acoustic events with a powerful potential for speaker discrimination is the nasal coarticulation scheme proposed by Su et al. [9]. It is based on the premise that in connected speech, as a result of the inertia of the articulators, the vocal tract shape is not only a function of the current phoneme but also of neighbouring phonemes (coarticulation). The difference between the mean spectrum of a nasal consonant followed by a front vowel, and that of the same consonant followed by a back vowel, was used as an acoustic measure of nasal coarticulation in a consonant-vowel context. Identification scores of 100% were obtained with a group of trained speakers, but the authors did suggest that with a larger untrained population, this was not to be expected. They also concluded that the nasal coarticulation measure is superior to measures bases directly on nasal spectra. It is appropriate to mention here that most schemes involving location of nasal events rely on visual observation to perform the segmentation, an impractical proposition.

Bunge [10] developed a system called AUROS as a research tool for investigating the effect of various processing aspects on speaker recognition performance. Tests conducted using a large speech data base yielded identification and verification error rates of less than 1% for both text-dependent and text-independent speaker recognition. For the latter, a text length of at least 11 sec was found to be necessary to obtain high recognition rates.

One of the earliest investigations of text-independent speaker recognition was conducted by Atal [4]. The text-independent data base for this study was created in an artificial manner from the utterances for the text-dependent investigation mentioned earlier. The 40 segments into which each utterance was divided, were recombined in a random fashion to destroy the synchronization in the text of the utterances. The features used were 12 cepstral coefficients for each of the 40 segments. The distance metric used combined the contributions from each one of the segments into a composite measure. An identification accuracy of 93% was obtained for speech 2 sec in duration.

Sambur [5] conducted a preliminary investigation, using a rather limited text, to investigate the suitability of using the technique of orthogonal linear prediction for text-dependent speaker recognition. He obtained an identification accuracy of 94%. Follow-up efforts in this area include the work of Shridhar et al. [11] and R.E. Wohlford et al. [12]. The former conducted a more exhaustive study of the use of long-term orthogonal linear prediction parameters for text-independent speaker recognition and obtained an identification accuracy of 93% and an equal-error speaker verification rate of 96.7%. The latter tested and compared four techniques of automatic speaker recognition based on four different feature sets, such as correlation of short and long-term spectral averages, cepstral measurements of long-term spectral averages, orthogonal linear prediction and long-term average LPC reflection coefficients, pitch and overall gain. He found that the two techniques based on cepstral and spectral data did not perform as well as the LPC-based systems, which yielded recognition accuracies of about 95% with 10 minutes of reference speech and 13 sec of unknown test speech.

Several workers have investigated the potential of spectral parameters of speech for text-independent speaker recognition. Li and Hughes [13] attempted to quantify inter-speaker and intra-speaker differences based on correlation matrices derived from continuous speech spectra. They found that a minimum of about 30 sec of text was needed to ensure stability of the correlation matrices. Identification and verification tests yielded a minimum error rate of between 1% and 3%.

Markel and Davis [14] have made a significant contribution to the area of text-independent speaker recognition. They used an extensive data base of over 36 hours of conversational speech recorded from 17 speakers over a period of over 3 months. The reference parameters for each speaker were obtained by extensive averaging of the feature vectors over approximately an hour's time-spaced speech for each speaker. Using a 22-feature reference parameter set composed of the mean and standard deviations of the fundamental frequency and 10 reflection coefficients, an identification accuracy of 98% and an equal-error verification rate of 4.25% was obtained.

Two large-scale functional text-dependent speaker recognition systems are in existence. These are the Texas Instruments Entry Control System [2] and the Bell Labs automatic speaker verification system [2]. The Texas Instruments system makes use of spectral amplitude information on precisely located sections of the utterance to obtain a speaker characterization. Using a 4-phrase sequential decision strategy, a false rejection rate of 0.3% and a false acceptance rate of 1% was achieved. The features used in the Bell Labs system were dynamic contour measurements of pitch and intensity as they varied over a sentence-long utterance recorded over dialed-up telephone lines. The false acceptance and false rejection error rates stabilized to about 4% after adaptation.

It is quite evident that in general, the accuracy obtainable in text-independent speaker recognition systems, without having to resort to extensive averaging (an impractical proposition) is rather limited with the existing speaker-characterizing parameters. The authors suggest that higher accuracies could be obtained by logically combining the decisions made by two different systems based on different, speaker-characterizing parameter sets. A composite scheme utilizing this technique is presented in this work.

3. Data collection and preprocessing

Twelve male speakers, all Canadians (20 to 35 years old) from the Windsor area, with no noticeable differences in their accents or speaking styles, participated in this study. Each participant was asked to read a magazine or newspaper or a passage from a text that he was familiar with. There were about six sessions per speaker and the sessions were spaced a week to two weeks apart. The speakers spoke into a high-fidelity microphone and speech was recorded on a high quality tape recorder. All the recordings took place in normal office room environments. Each session consisted of about 10 minutes of recorded speech.

Recorded speech was passed through a bandpass filter with cutoff frequencies at 150 Hz and 4.5 KHz. The filtered speech was then digitized at 10 KHz sampling rate by a 14 bit resolution A/D converter and stored in the disk storage medium of a Data General NOVA 840 minicomputer. The minicomputer is equipped with an Array Processor (AP120B) for high speed data processing.

A further preprocessing step prior to parameter extraction involved the elimination of silence, pauses and unvoiced segments from the digitized speech data through the use of a simple energy threshold criterion [15].

4. Speech parameters

Two classes of speech parameters were used in this study:

(a) Parameters obtained through time domain analysis.

(b) Parameters obtained through frequency domain analysis.

4.1. Time domain parameters

The time domain parameters consisting of the linear prediction coefficients and parameters obtained through nonlinear transformations of the linear prediction coefficients were:

(i) the Linear Prediction Coefficients,
(ii) the Reflection Coefficients,
(iii) the Log Area Ratio Coefficients,
(iv) the Cepstrum Coefficients.

The derivation of these parameter sets from analysis of the digitized speech samples is presented in the Appendix. In this study digitized speech of about 60 sec duration from each of the six sessions of every speaker was utilized for evaluation of the parameters. The entire speech was divided into 20 msec (200 samples) segments called frames. The frames corresponding to the silent portions and pauses in the utterances were removed using an energy threshold criterion [15]. The speech samples in the remaining frames were digitally pre-emphasized by a first order filter with transfer function (1-0.94z^{-1}) and subsequently multiplied by a Hamming window function. A 12th order linear prediction analysis was applied to the remaining frames which amounted to a minimum of 1000 frames (20 sec) per session for each speaker.

4.2. Frequency domain parameters

The frequency domain parameters consisted of the magnitude spectra of the linear prediction inverse filter and spectral data obtained by passing continuous speech through a bank of 16 bandpass filters spanning the frequency range from 200 Hz to 5 KHz.

(i) The inverse filter is characterized by

$$A(z) = 1 + \sum_{i=1}^{P} \alpha_i z^{-i} \qquad (1)$$

where $z = e^{j\omega T_s}$, and T_s is the sampling interval.

The Inverse Filter Spectral Coefficients were obtained as follows.

(a) Compute the FFT of the sequence

$$\{1, \alpha_1, \alpha_2, \ldots, \alpha_P, 0, 0, \ldots, 0\}$$

where 51 zeroes were appended to the sequence to obtain the required frequency resolution [16] by increasing its length to 64.

(b) Compute the spectral magnitudes of the 33 Fourier coefficients.

(c) Evaluate the logarithm of the magnitudes.

The above procedure essentially evaluates the spectral characteristics of the LPC model.

(ii) The Speech Spectrum Parameters were obtained through a direct spectral analysis of speech as illustrated in Fig. 1. The same speech utterances on which linear prediction analysis was performed,

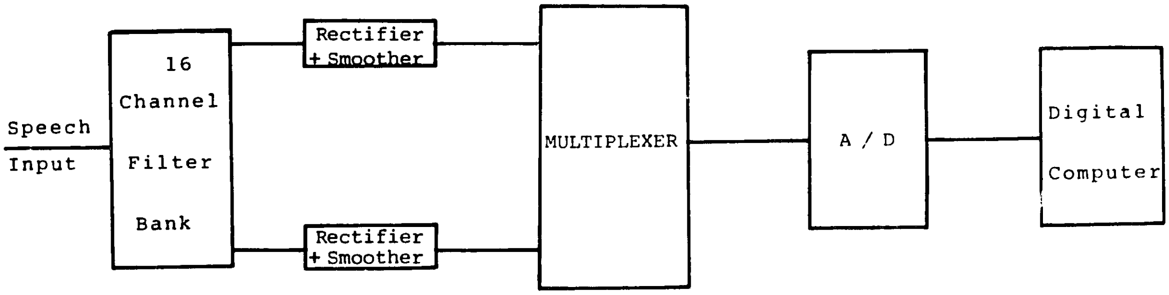

Fig. 1. Generation of speech spectrum parameters.

were passed through a bank of 16 bandpass filters. The filters were centred at frequencies ranging from 160 Hz to 5 KHz and were spaced a third of an octave apart. The output of the bandpass filters were rectified, smoothed and sampled every 10 msec. The sampled values represent the speech spectrum parameters.

The parameter sets obtained by the above procedures were orthogonalized through the Karhunen–Loeve transformation [17] as described in the next section.

5. Orthogonalization of parameter sets

The parameter sets were orthogonalized through the Karhunen–Loeve transformation as follows [5,17]:

(a) Let $(x_{ij}: i = 1, 2, \ldots, M; j = 1, 2, \ldots, \mathrm{NF})$ be the parameter set where x_{ij} is the ith parameter of the jth frame, M is the number of parameters in the set for each frame and NF is the total number of frames in the speech.

(b) Compute the covariance matrix $[C]$ of the parameter set, where $\{[(C)]: (c_{lm}: l = 1, 2, \ldots, M; m = 1, 2, \ldots, M)\}$ is given by

$$c_{lm} = \frac{1}{(\mathrm{NF}-1)} \sum_{j=1}^{\mathrm{NF}} (x_{lj} - \bar{x}_l)(x_{mj} - \bar{x}_m) \tag{2}$$

and

$$\bar{x}_l = \frac{1}{NF} \sum_{j=1}^{NF} x_{lj} \tag{3}$$

is the average value of the lth parameter.

(c) Compute the eigenvalues $\{\lambda_l: l = 1, 2, \ldots, M\}$ and the eigenvectors T_l of the matrix $[C]$ by solving $|C - \lambda I|$ for λ_l's and by solving $CT_l = \lambda_l T_l$ for T_l.

(d) Normalize T_l to unit length.

(e) Evaluate the orthogonal parameters $(\phi_{ij}: i = 1, 2, \ldots, M; j = 1, 2, \ldots, \mathrm{NF})$ as follows:

$$\phi_{ij} = \sum_{l=1}^{M} t_{il} x_{lj} \tag{4}$$

where, ϕ_{ij} is the ith orthogonal parameter in the jth frame, t_{il} is the lth element of the ith eigenvector T_i.

(f) The average value $\bar{\phi}_i$ of the ith orthogonal parameter is defined as

$$\bar{\phi}_i = \frac{1}{\mathrm{NF}} \sum_{j=1}^{\mathrm{NF}} \phi_{ij}. \tag{5}$$

(g) It is easy to show [17] that the variance of ϕ_{ij} is

$$\mathrm{Var}(\phi_{ij}) = \lambda_i. \tag{6}$$

(h) A quantitative measure of dissimilarity between two feature sets, termed 'distance', is given by

$$d_{sq} = \sum_{i=1}^{M} \frac{[(\bar{\phi}_i)_s - (\bar{\phi}_i)_q]^2}{(\lambda_i)_m} \tag{7}$$

where d_{sq} is the 'distance' between speaker 's' and speaker 'q', $(\cdot)_s$ and $(\cdot)_q$ refer to arguments for speaker 's' and speaker 'q' respectively, $(\bar{\phi}_i)_s$ is the mean value of the ith orthogonal parameter for speaker 's', $(\bar{\phi}_i)_q$ is the mean value of the ith oithogonal parameter for speaker 'q' and is defined by

$$(\bar{\phi}_i)_q = \frac{1}{\mathrm{NF}} \sum_{j=1}^{\mathrm{NF}} \left[\sum_{l=1}^{M} (t_{il})_s (x_{lj})_q \right]. \tag{8}$$

6. Convergence properties of the covariance matrix

An accurate estimate of the covariance matrix of the speech parameters is required for the computation of the orthogonal parameters. Since *a priori* statistics of the speech parameters are unknown, the covariance matrix is evaluated through eq. (2) under the assumption of ergodicity of the speech parameters. The first step in the estimation process is the choice of the duration of the speech utterance across which the covariance matrix is to be evaluated. The choice is governed by the convergence properties of the covariance matrix as computed using eq. (2) with the number of frames being the independent variable. The convergence properties were evaluated by computing the change in the computed covariance matrix through the use of the 'Average Absolute Difference' measure termed AAD [13]. The AAD is defined as

$$\mathrm{AAD}(n) = \frac{2}{M(M+1)} \sum_{i=1}^{M} \sum_{j=1}^{M} |\gamma_{ij}(n+1) - \gamma_{ij}(n)|$$
$$n = 1, 2, \ldots, L-1 \qquad (9)$$

where $\gamma_{ij}(n)$ is the nth estimate of the ijth element or the normalized covariance matrix $[\Gamma]$.

Our studies revealed that the AAD measure was effective only if a separation of 200 frames (4 sec) was maintained between the two consecutive estimates of $[\Gamma]$. Thus $\gamma_{ij}(4)$ was computed over 4×200 ($= 800$) frames of speech while $\gamma_{ij}(5)$ was computed over 5×200 ($= 1000$) frames of speech. In this study a total of 6000 frames were utilized and this resulted in L being 30 ($= 6000/200$). It is further noted that one frame of speech consists of 200 speech samples (20 msec).

Fig. 2 illustrates the behaviour of AAD as a function of 'n' when the covariance matrix was computed using linear prediction coefficients. It is observed that the AAD converges to 'mud level' after about 2000 frames of speech are utilized, indicating that about 40 sec of speech are required to accurately estimate the covariance matrix.

In this study the covariance matrix was normalized by the transformation [18]

$$\Gamma = \Lambda^{-1} C \Lambda^{-1} \qquad (10)$$

where Λ is a diagonal matrix whose elements are the standard deviations of the input parameters. This normalization causes the elements of the matrix to be less than 1 in magnitude, since any element γ_{ij} of $[\Gamma]$ represents the correlation coefficient between the ith and jth parameters.

7. Performance evaluation of speech parameter sets

The speech parameter sets defined in Section 4 were evaluated for their speaker discrimination characteristics through an exhaustive speaker identification/verification study, using the procedure described in Section 5. In this study the reference parameters (consisting of the average values of the

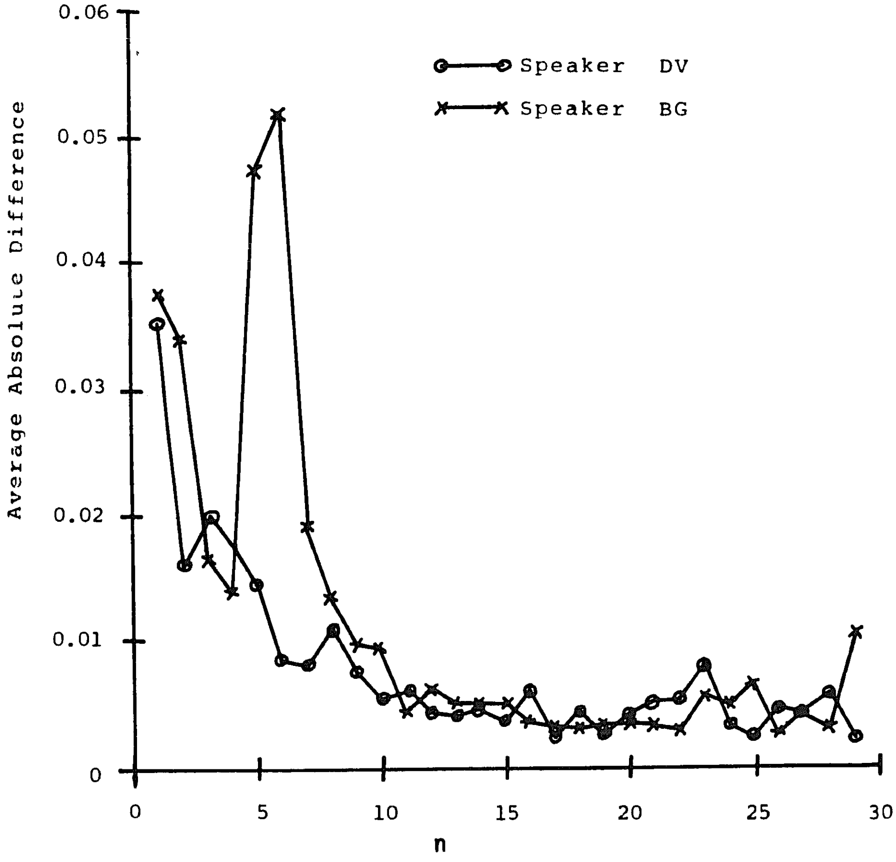

Fig. 2. Convergence properties of covariance matrix of speech parameters.

orthogonal parameters) were evaluated from about 100 sec of speech (20 sec from five of the six sessions), while the test parameters were obtained from about 5 sec of speech from the sixth session. Each of the six sessions was in turn used to compute the test parameters. With twelve speakers participating in this study the above procedure yielded four sets of test data per speaker for every reference set and six reference sets per speaker. There were in all 288 identification tests and 3456 verification tests.

The results of this study are illustrated in Table 1. It is interesting to observe that the inverse filter spectral coefficients obtained from the LPC model yielded the highest recognition accuracy (97.4%) while the lowest accuracy (88.54%) was obtained when the same linear prediction coefficients were directly used. The remaining parameter sets yielded comparably similar results with accuracies in the range of 91–93%. The above study also revealed that a direct implementation of the recognition algorithm could yield at the most an accuracy of about 97%. Improvements, if any, that could be realized through the use of more than one parameter set are discussed in the next section.

8. A composite scheme for speaker recognition

Several techniques have been proposed for improving recognition scores by incorporating a sequential or deferred decision strategy [2]. In such schemes an additional test utterance is called for when the recognition algorithm does not yield an unambiguous decision. In the case of speaker identification, such a situation would arise when the identity of the test speaker cannot be established with confidence. In the case of verification, any uncertainty in the distance measure would cause the algorithm to reject the identity claim of a genuine test speaker or accept the identity claim of a test speaker who is an impostor. The authors decided to investigate the effects of logically combining decisions from two independent tests involving two different parameter sets.

The two parameter sets initially chosen for the composite scheme were the hardware spectral parameters and the parameters of the frequency response $A(z)$ of the inverse filter. These two parameter sets also happen to be obtained from two distinct techniques of speech analysis, which is an added advantage of their choice. The logical

Table 1
Performance evaluation of speech parameter sets

Parameter Set	Number of coeffs./ samples	IDENTIFICATION		VERIFICATION			
		Accuracy (%)	Error (%)	Accuracy (%)	Error		
					False Acceptance (%)	False Rejection (%)	Total (%)
LPC Coeffs.	12	88.54	11.46	94.14	2.93	2.93	5.86
Reflection Coeffs.	12	92.19	7.81	94.54	2.73	2.73	5.46
Log Area Ratio Coeffs.	12	93.23	6.77	92.02	2.99	2.99	5.98
Cepstrum Coeffs.	12	91.67	8.33	94.14	2.93	2.93	5.86
Inv. Filt. Spec. Samp.	33	97.4	2.6	95.06	2.47	2.47	4.94
Speech Spec. Samples	16	93.23	6.77	94.02	2.99	2.99	5.98

decision procedure is described below:

(a) If the decisions from the two independent tests are identical, accept this decision.

(b) If the two decisions are different, then repeat the recognition using a different test utterance. If the two decisions are now identical, accept the decision.

(c) If the two decisions are different after two such repeats, then reject the identity claim of the test speaker. In the case of identification, classify the decision as 'unknown identity'.

The results of the above scheme are presented in Table 2. An identification accuracy of 98.32% was realized with the repeat feature having to be invoked in 4.47% of the tests. In the case of speaker verification, the false acceptance error rate was 0.56% and the false rejection error rate was 1.48% giving a total error rate of 2.04%. The repeat feature was utilized in 4.01% of the verification tests.

The above results point to only a 1% improvement in the identification accuracy over the 97.4% obtained with the inverse filter spectral coefficients alone. However the significance of the proposed composite scheme becomes evident if one observes that the error rate has decreased from 2.6% to about 1.7%, a decrease of more than a third. In the verification case the false acceptance rate fell from 2.4% to 0.56%, while the false rejection rate decreased to 1.48% from the 2.47% obtained with inverse filter spectral coefficients alone. It should be observed that this new scheme is biased in favour of false rejection. In order to fully assess the potential of the proposed composite scheme, a second test, utilizing the log area ratio coefficients and the speech spectral coefficients was conducted. The choice of these two parameter sets was based on comparable (in fact, equal) identification (about 93%) and verification (about 94%) accuracies when used separately, as seen from Table 1. With the composite scheme, the identification accuracy increased to nearly 98% while the total verification error fell to 2.8% as seen from Table 2. The percentage of tests that had to be repeated were 6.97% and 3.81% for identification and verification respectively. This is a significant improvement (about 4.5%) in the performance and clearly illustrates the power of the proposed scheme. Further tests utilizing two sets of parameters with comparable recognition accuracies confirmed that significant improvement is always obtained with the composite scheme.

Table 2
Recognition accuracies with proposed composite schemes

Parameter Sets in Composite Scheme	IDENTIFICATION				VERIFICATION				
	Accuracy (%)	Error (%)	Unknown Identity (%)	Percentage of tests repeated	Accuracy (%)	Error			Percentage of tests repeated
						False Acceptance (%)	False Rejection (%)	Total (%)	
Inverse Filter Spectrum + Speech Spectrum	98.32	1.12	.56	4.47	97.96	.56	1.48	2.04	4.01
Log Area Ratio + Speech Spectrum	97.67	1.16	1.16	6.97	97.20	.91	1.89	2.8	3.85

9. Conclusions

(a) The linear prediction coefficients and the parameters obtained through nonlinear transformations of the linear prediction coefficients are shown to possess good speaker discriminating characteristics.

(b) The orthogonal recognition algorithm is shown to be feasible for text-independent speaker recognition.

(c) Small duration (5 sec) test utterances are shown to be adequate for effective recognition.

(d) The proposed composite scheme utilizing a logical combination of decisions from tests based on different parameter sets is shown to be effective in improving the recognition scores significantly.

Appendix

Time Domain Speech Parameters

In this section the basic principles of linear prediction and speech parameterization are presented.

(a) *Linear Prediction Coefficients.* A very popular and effective characterization of speech [19,20] is realized through the use of a linear discrete model defined by the transfer function

$$H(z) = \frac{G}{1 - \sum_{i=1}^{P} \alpha_i z^{-i}} = \frac{G}{A(z)} \tag{A1}$$

where G is the gain of the model, α_i's are termed the linear prediction coefficients and $A(z)$ is the inverse filter. An equivalent time domain description is obtained as

$$\hat{s}_n = \sum_{i=1}^{P} \alpha_i s_{n-i} + G e_n \tag{A2}$$

where $\hat{s}_n$ is the estimated value of speech sample at the nth instant and e_n is the value of the excitation signal(input). In speech analysis the α_i's and G are updated about every 20 msec (assuming speech is sampled at 10 KHz). The excitation signal e_n is either a sequence of pulses spaced by the pitch period for voiced speech or pseudorandom wideband signal for unvoiced speech.

The coefficients α_i are obtained by minimizing the mean squared error between the original speech samples s_n and their estimates $\hat{s}_n$ as defined in eq. (A2). This procedure results in a set of linear equations for the α_i's. These are

$$\sum_{n} \sum_{i=1}^{P} \alpha_i s_{n-i} s_{n-j} = \sum_{n} s_n s_{n-j},$$
$$j = 1, 2, \ldots, P. \tag{A3}$$

An efficient solution to the above equation is realized through the use of the autocorrelation method [19,20] for computing the coefficients α_i in the above equation (A3). This results in the following set of equations:

$$\sum_{i=1} \alpha_i r_{|i-j|} = r_j, \quad j = 1, 2, \ldots, P \tag{A4}$$

where

$$r_j = \sum_{n=1}^{N-j} s_n s_{n-j} \tag{A5}$$

and N is the number of samples in the analysis interval.

(b) *Reflection Coefficients.* The modelling of the vocal tract as a nonuniform acoustic tube formed by concatenating P uniform cylindrical sections of equal length [21,22] leads to a set of parameters called the reflection coefficients. The reflection coefficients (K_i, $i = 1, 2, \ldots, P$) are related to the linear prediction coefficients (α_i: $i = 1, 2, \ldots, P$) by the following relations [20]:

$$K_i = a_i^{(i)},$$

$$a_j^{(i-1)} = \frac{a_j^{(i)} + a_i^{(i)} a_{i-1}^{(i)}}{1 - K_i^2},$$

$$1 \leqslant j \leqslant i - 1,$$

$$i = P, P-1, \ldots, 1. \tag{A6}$$

The recursion starts with the ordinary predictor coefficients obtained through eq. (A4).

(c) *Log Area Ratio Coefficients.* The acoustic tube model yields another set of parameters that are related to the ratios of the areas of the cylindrical sections. These are called the Log Area Ratio

Coefficients (G_i; $i = 1, 2, \ldots, P$) and these may be derived from the reflection coefficients as follows [20]:

$$G_i = \log\left[\frac{A_{i+1}}{A_i}\right] = \log\left[\frac{1-K_i}{1+K_i}\right], \quad 1 \leqslant i \leqslant P \tag{A7}$$

where (A_i, $i = 1, 2, \ldots, P$) are the areas of the P cylindrical sections of the model.

(d) *Cepstrum Coefficients.* These coefficients (C_i, $i = 1, 2, \ldots, P$) are obtained by computing the cepstrum of the inverse filter $A(z)$ and hence are related to the linear prediction coefficients α_i's as follows [20]:

$$C_i = \alpha_i + \sum_{j=1}^{i-1} \frac{j}{i} C_j \alpha_{i-j}$$

$$i = 1, 2, \ldots, P. \tag{A8}$$

The cepstrum coefficients may also be derived nonrecursively from the linear prediction coefficients. This procedure developed by Schroeder is discussed in [23].

References

[1] B.S. Atal, 'Automatic recognition of speakers from their voices', *Proc. IEEE*, Vol. 64, April 1976, pp. 460–475.

[2] A.E. Rosenberg, 'Automatic speaker verification: A review', *Proc. IEEE*, Vol. 64, April 1976, pp. 475–487.

[3] B.S. Atal 'Automatic speaker recognition based on pitch contours', *J. Acoust. Soc. Am.*, Vol. 52, Dec. 1972, pp. 1687–1697.

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

[5] M.R. Sambur, 'Speaker recognition using orthogonal linear prediction', *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-24, August 1976, pp. 283–289.

[6] S. Pruzansky and M.V. Mathews, 'Talker recognition procedure based on analysis of variance', *J. Acoust. Soc. Am.*, Vol. 36, Nov. 1964, pp. 2041–2047.

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

[8] S. Furui et al., 'Talker recognition by long-time averaged speech spectrum', *Electron. Commun. Jap.*, Vol. 55A, Oct. 1972, pp. 54–61.

[9] L.S. Su et al., 'Identification of speakers by use of nasal coarticulation', *J. Acoust. Soc. Am.*, Vol. 56, Dec. 1974, pp. 1876–1882.

[10] E. Bunge, 'Automatic speaker recognition system AUROS for security systems and forensic voice identification', *Proc. 1977 Intern. Conf. Crime Countermeasures-Science and Engineering*, Lexington, KY, July 25–29, 1977, pp. 1–7.

[11] M. Shridhar et al., 'Text-independent speaker recognition using orthogonal linear prediction', *Proc. IEEE Intern. Conf. Acoust. Speech Signal Process.*, Atlanta, GA, March 30–April 1, 1981, pp. 197–200.

[12] R.E. Wohlford et al., 'A comparison of four techniques for automatic speaker recognition', *Proc. IEEE Intern. Conf. Acoust. Speech Signal Process.*, Denver, CO, April 9–11, 1980, pp. 908–911.

[13] K.P. Li and G.W. Hughes, 'Talker differences as they appear in correlation matrices of continuous speech spectra', *J. Acoust. Soc. Am.*, Vol. 55, April 1974, pp. 833–837.

[14] J.D. Markel and S.B. Davis, 'Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base', *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-27, February 1979, pp. 74–82.

[15] L.R. Rabiner and R.W. Schaefer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978, Ch. 4.2, pp. 122.

[16] J.D. Markel and A.H. Gray, *Linear Prediction of Speech*, Springer-Verlag, 1976, Ch. 6.6, pp. 160.

[17] N. Ahmed and K.R. Rao, *Orthogonal Transforms for Digital Signal Processing*, Springer-Verlag, Berlin, 1975, Ch. 9.1, pp. 200.

[18] N. Ahmed and K.R. Rao, *Orthogonal Transforms for Digital Signal Processing*, Springer-Verlag, Berlin, 1975, Appendix 8.1, pp. 195.

[19] J.D. Markel and A.H. Gray, *Linear Prediction of Speech*, Springer-Verlag, Berlin, 1976, Ch. 1, pp. 1.

[20] L.R. Rabiner and R.W. Schaefer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978, Ch. 8, pp. 396.

[21] J.D. Markel and A.H. Gray, *Linear Prediction of Speech*, Springer-Verlag, Berlin, 1976, Ch. 4, pp. 60.

[22] L.R. Rabiner and R.W. Schaefer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978, Ch. 3, pp. 38.

[23] M.R. Schroeder, 'Direct (nonrecursive) relations between cepstrum and predictor coefficients', *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-29, April 1981, pp. 297–301.