On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition

FRANK K. SOONG, MEMBER, IEEE, AND AARON E. ROSENBERG, FELLOW, IEEE

Abstract—**The use of instantaneous and transitional spectral representations of spoken utterances for speaker recognition is investigated. LPC-derived cepstral coefficients are used to represent instantaneous spectral information, and best linear fits of each cepstral coefficient over a specified time window are used to represent transitional information. An evaluation has been carried out using a database of isolated digit utterances over dialed-up telephone lines by 10 talkers. Two vector quantization (VQ) codebooks, instantaneous and transitional, are constructed from each speaker's training utterances. The experimental results show that the instantaneous and transitional representations are relatively uncorrelated, thus providing complementary information for speaker recognition. A rectangular window of approximately 100 ms duration provides an effective estimate of the transitional spectral features for speaker recognition. Also, simple transmission channel variations are shown to affect the instantaneous spectral representations and the corresponding recognition performance significantly, while the transitional representations and performance are relatively resistant.**

I. INTRODUCTION

SHORT-TIME spectral features of speech signals have long been used successfully in both speech and speaker recognition applications. In addition to conveying phonetic information, the spectral features carry speaker related information correlated with talking behavior as well as physiological structure of the vocal apparatus.

There are several important issues involved in automatic speech and speaker recognition, including:

1) how to extract short-time spectral information from raw speech signals;
2) how to efficiently represent instantaneous spectral information at any time instant;
3) how to reliably characterize transitional spectral information associated with the time-varying properties of speech signals in a compact form;
4) how to use instantaneous and transitional spectral features to measure the similarity (or dissimilarity) between two given running spectra; and
5) how to make use of instantaneous and transitional spectral features in a complementary way.

Short-time spectral information of speech signals is usually extracted through a filter bank, an FFT, or an LPC (linear predictive coding) spectral analysis. In this study, we use LPC as our spectral analysis. However, the results obtained in this study are believed to be equally applicable to both filter bank and FFT spectral analysis front ends.

Atal [1] compared several different spectral representations of speech spectra including LPC predictor coefficients, autocorrelation coefficients, LPC-derived cepstral coefficients, etc., and found that the LPC cepstral coefficients, when used with a Mahalanobis distance measure, gave the best speaker recognition performance. In the same paper, he also investigated the feasibility of removing the effects of varying transmission channels by subtracting the long term average of the cepstral coefficients. Both text-dependent and text-independent experiments were conducted, and the performance of the text-dependent speaker recognition system was found to be better than that of the text-independent one.

Furui [2] used both instantaneous and transitional spectral information in his LPC cepstrum-based speaker verification experiments to characterize a sentence-long utterance. High verification performance was achieved with this system. It was also shown, in the same paper, that a log area ratio representation is less effective for speaker verification than a cepstral representation.

Recently, a speaker-based vector quantization (VQ) approach to speaker recognition was proposed [3]. The approach, although intrinsically text-independent, can easily be extended to text-dependent speaker recognition. The LPC likelihood ratio distortion measure is used for both VQ codebook generation and for recognition tests. This particular VQ-based speaker recognition system, when evaluated over a 100 speaker database, achieved a 98 percent speaker identification accuracy when the input speech consisted of 10 distinct, randomly ordered digits. Related VQ approaches have been reported by Li and Wrench [4], Helms [5], Shikano [6] and Buck *et al.* [7].

In this paper, we discuss the relative contributions of instantaneous (static) spectral information and transitional (dynamic) spectral information to our VQ-based speaker recognition performance. We use LPC-based cepstral coefficients to characterize instantaneous spectral information and orthogonal polynomials to characterize the time trajectories of cepstral coefficients over a finite length time window (i.e., the transitional spectral information). Each speaker's short-time spectral features are efficiently represented by two VQ codebooks—an instantaneous codebook and a transitional codebook. Statistically

Manuscript received January 15, 1986; revised December 4, 1987.

The authors are with AT&T Bell Laboratories, Murray Hill, NJ 07974-2070.

IEEE Log Number 8820418.

weighted spectral distances are used in order to equalize the contributions from individual components of the feature vectors. The corresponding spectral distances from test vectors to the two VQ codebooks are optimally combined to make a final recognition decision.

The paper is organized as follows. In Section II, we review the cepstral representation of instantaneous and transitional information of a speaker's short-time spectrum. The ideas of inverse variance weighting of cepstral distance and its resemblance to the quefrency weighting of the spectral slope metric are discussed. In Section III, the speaker-based VQ recognizer is reviewed and a description of the telephone database is given. In Section IV, the set of experiments used to evaluate the performance of the recognizer and the results are presented. Finally, in Section V, we discuss the experimental results and give some conclusions.

II. A Two-Dimensional Spectral Representation

A. Cepstral Representation, Discrete Cosine Transform, and Karhunen-Loève Transform

The logarithm of the time varying short-time spectrum can be represented as

$$\log |S(\omega, t)| = \sum_{m=-\infty}^{\infty} c_m(t)\, e^{-j\omega m} \tag{1}$$

where $c_m(t)$ is the well-known mth cepstral coefficient at time t.

Due to the symmetry of the power spectrum, i.e., $c_m(t) = c_{-m}(t)$, (1) can be rewritten as a pure cosine series expansion

$$\log |S(\omega, t)| = 2 \sum_{m=1}^{\infty} c_m(t) \cos(\omega m t) + c_0(t). \tag{2}$$

The cosine series expansion can be further approximated by a finite-term summation, called the discrete cosine transform (DCT). It was shown by Zelinsky and Noll [8] that the DCT gives a very efficient representation of speech signals in the frequency domain, and good quantization performance when used in an adaptive transform speech coder. In the same paper, they also showed that the intrinsic basis functions of the DCT have a close resemblance to the eigenvectors of the optimal Karhunen-Loève transform (KLT). Thus, DCT representation of a short-time spectrum approximates well the optimal orthogonal KLT decomposition. Experimentally we have confirmed this conclusion by observing that the covariance matrix of the cepstral coefficients is diagonal dominant.

B. Euclidean Cepstral Distance, Weighted Euclidean Distance, and Spectral Slope Distance

Given c_m and c'_m, the cepstral representations of two speech spectra, $\log |S(\omega)|$ and $\log |S'(\omega)|$, respectively, the L_2 norm or the Euclidean cepstral distance is defined as

$$d_{L_2} = 2 \sum_{m=1}^{\infty} (c_m - c'_m)^2 + (c_0 - c'_0)^2. \tag{3}$$

Using only the first p cepstral coefficients without the gain term in comparing the spectral shapes between two cepstrally smoothed log spectra, we obtain

$$d_{\text{CEP}} = \sum_{m=1}^{p} (c_m - c'_m)^2. \tag{4}$$

Note that without affecting the relative magnitude of the distance, we have dropped the factor 2 in front of the summation sign.

It can be shown, due to the minimum phase nature of all-pole modeling, that LPC-derived cepstral coefficients decay at least as fast as $1/m$, and the cepstral coefficients tend to be concentrated around $m = 0$ [10]. As a consequence, when a cepstral distance, d_{CEP}, is computed in (4), usually, the main contributions to the finite sum are likely to be from the first few cepstral coefficients which have much larger magnitudes than the higher order ones. If higher order cepstral coefficients do carry useful information for speaker recognition, they are not effectively used in the Euclidean cepstral distance as given in (4). To illustrate this point further, we use Fig. 1 to show histograms of the inter- and intraspeaker distance components $(c_m - c'_m)^2$ in (4). From the corresponding histograms, it is clear that the higher order cepstral coefficients are as important as the lower order cepstral coefficients in their ability to separate one speaker from the others. In order to equalize the contributions from individual cepstral coefficients and, hopefully, maximize the recognition performance, a weighted cepstral distance seems desirable. The Mahalanobis distance, which uses inverse covariance weighting, has shown to be very effective in speaker recognition experiments [1]. The distance has the following form:

$$d_{\text{mah}} = (\boldsymbol{c} - \boldsymbol{c}')^T \boldsymbol{R}^{-1} (\boldsymbol{c} - \boldsymbol{c}') \tag{5}$$

where $\boldsymbol{R}$ is the pooled intraspeaker covariance matrix and $\boldsymbol{c}$ and $\boldsymbol{c}'$ are the cepstral coefficient vectors of the log spectra $\log |S(\omega)|$ and $\log |S'(\omega)|$, respectively. The inverse covariance matrix is used here to decorrelate as well as to normalize the cepstral coefficients.

Since the estimated covariance matrix $\boldsymbol{R}$ is essentially diagonal, we simplify it to a diagonal matrix, and the corresponding weighted cepstral distance is then

$$d_{\text{WCEP}} = \sum_{m=1}^{p} (c_m - c'_m)^2 w_m \tag{6}$$

where the weighting coefficient, w_m, is the reciprocal of the variance of the corresponding pooled intraspeaker variance of the mth cepstral coefficient.

Using an isolated digit database of 100 talkers, we estimated the pooled intraspeaker variances of the 8 LPC-derived cepstral coefficients. The sampled intraspeaker

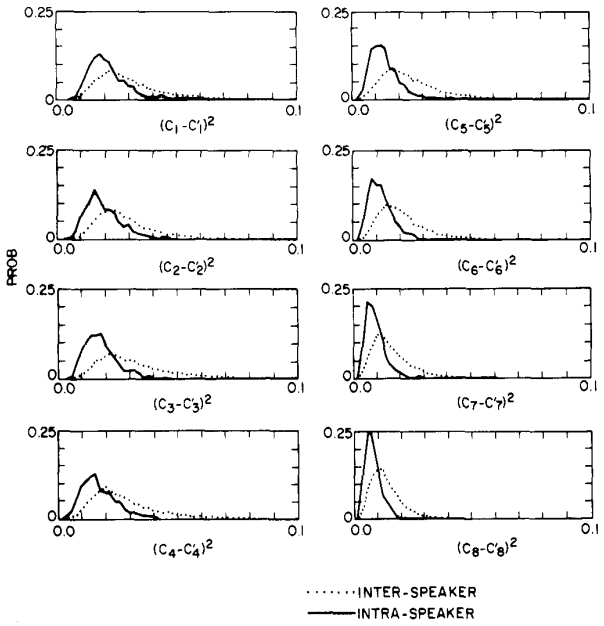

Fig. 1. Histograms of inter- and intraspeaker distances of each individual component $(c_i - c_i')^2$, $i = 1, \cdots, 8$.

variances and the corresponding weighting coefficients are depicted in Fig. 2. It is observed that the lower order cepstral coefficients are weighted less than the higher order ones in the distance computation. The weighting function shows a rate of increase faster than a linear ramp, m, although not as fast as m^2. It is also interesting to see the similarity between this weighted cepstral distance and the Euclidean distance between the spectral slopes (in frequency) of two given cepstrally smoothed spectra. The spectral slope of a cepstrally smoothed spectra is given by

$$\frac{\partial \log |S(\omega)|}{\partial \omega} = \sum_{m=-p}^{p} (-jm)\, c_m e^{-j\omega m}. \tag{7}$$

The Euclidean spectral slope distance between two spectra S and S' is then

$$d_{\text{slope}} = 2 \sum_{m=1}^{p} m^2 (c_m - c_m')^2. \tag{8}$$

This distance, although different in form, is closely related to the weighted slope metric proposed by Klatt [11] and later investigated by Nocerino *et al.* [12]. The spectral slope also has an interpretation as the power sum of LPC polynomial roots as noted by Schroeder [13]. The spectral slope distance was used by Paliwal [9] in a vowel recognition experiment. Recently, both the spectral slope distance and the weighted cepstral distance were investigated by Tohkura [14] in a speech recognition experiment. Significant improvement in recognition performance over the log likelihood ratio distance measure was obtained.

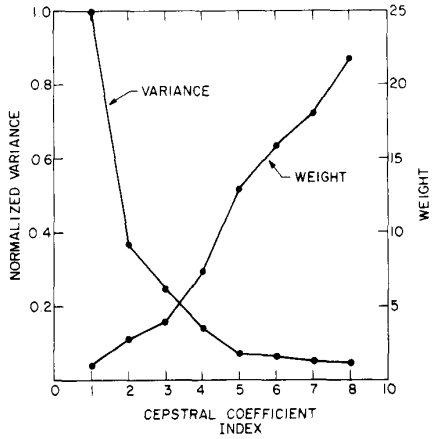

Fig. 2. Normalized statistical cepstral coefficient variances and the corresponding weights.

C. Orthogonal Polynomial and Transitional Spectral Information Characterization

In the previous section, we have dealt only with a sampled spectrum at a specific time instant. In this section, we use an orthogonal polynomial to characterize spectral transitional information. Spectral variation in time (i.e., spectral transitional information) is represented by

$$\frac{\partial \log |S(\omega, t)|}{\partial t} = \sum_{m=-\infty}^{\infty} \frac{\partial c_m(t)}{\partial t} e^{-jm\omega}. \tag{9}$$

The sampled time series, $c_m(t)$, usually does not have an analytic form and its time derivative, $\partial c_m(t)/\partial t$, can only be approximated by a finite difference. The 1st-order finite difference, as it is usually used, is intrinsically noisy. To alleviate this difficulty, Furui [2] suggested the use of an orthogonal polynomial fit of each cepstral coefficient trajectory over a finite length window. The zeroth-order coefficient, or the constant term of the orthogonal polynomial, is

$$\bar{c}_m(t) = \frac{\sum_{k=-K}^{K} h_k c_m(t+k)}{\sum_{k=-K}^{K} h_k} \tag{10}$$

where h_k is a window (usually symmetric) of length $2K + 1$ frames. The 1st-order orthogonal polynomial coefficient, or the generalized spectral slope (in time), denoted as $\Delta c_m(t)$, has the following form:

$$\frac{\partial c_m(t)}{\partial t} \approx \Delta c_m(t) = \frac{\sum_{k=-K}^{K} k h_k c_m(t+k)}{\sum_{k=-K}^{K} h_k k^2}. \tag{11}$$

Higher order orthogonal polynomial coefficients can be similarly derived [15]. However, for efficient representation of spectral dynamics over time, Furui [2] has shown that a 1st-order polynomial characterization of spectral

change is usually adequate. A window of reasonable length has to be used to ensure a smooth fit to the data points from one frame to the next. A rectangular window is usually adequate. If further smoothing is needed, a more effective smoothing window, such as a Hamming window, can be used instead.

D. Weighted Transitional Spectral Distance

Similar to the inverse variance weighted cepstral distance, a weighted Euclidean distance between two given sets of Δ-cepstrum coefficients is defined as

$$d_{W_{\Delta\text{CEP}}} = \sum_{m=1}^{P} u_m(\Delta c_m - \Delta c'_m)^2 \qquad (12)$$

where the weighting coefficient, u_m, is the reciprocal of the pooled intraspeaker variance of Δc_m. Just like the covariance matrix of cepstral coefficients, the pooled intraspeaker covariance matrix of the Δ-cepstrum coefficients is also diagonal dominant. The variances of the covariance matrix decrease with the quefrency index, and the ratio of variances between the largest (1st) term and the smallest (8th) term is about 15.

III. VQ-Based Speaker Recognition System and Database

A. The Speaker-Based VQ Codebook Generation and Recognition System

The speaker recognition system used in this study is similar to the one proposed by Soong *et al.* [3], with the following differences. In addition to the instantaneous spectrum VQ codebook for each speaker, a transitional spectrum codebook is also used to characterize the time variation of a speaker's voice. The distortion measures used in the present system, namely, the weighted and unweighted Euclidean distances, are also different from the LPC likelihood ratio distortion measure used previously. The VQ codebook generation algorithm is the well-known generalized Lloyd algorithm [16], as used previously, and can be summarized as follows. Given a set of I training vectors, $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_I\}$, from a specific speaker, we want to find a nonoverlapping partitioning of the feature vector space into Q distinct regions, $\{S_1, S_2, \cdots, S_Q\}$, where the whole feature space is $S = S_1 \cup S_2 \cup \cdots \cup S_Q$. Each partition, S_q, forms a convex region which is represented by its centroid vector, $\boldsymbol{b}_q$, in such a way that the average distance of all training vectors to the corresponding centroids yields the minimum average distortion, i.e.,

$$D = \frac{1}{I}\sum_{i=1}^{I} \min_{1 \le q \le Q} d(\boldsymbol{r}_i, \boldsymbol{b}_q). \qquad (13)$$

D is the average distortion over the whole training set, and $d(\boldsymbol{r}_i, \boldsymbol{b}_q)$ is the distortion between the training vector $\boldsymbol{r}_i$ and codebook entry $\boldsymbol{b}_q$.

A block diagram of the text-independent speaker identification system is shown in Fig. 3. The input analog signals are first bandpassed filtered from 200 to 3200 Hz and digitized at a 6.67 kHz sampling rate. The digitized signals are then appropriately endpointed, preemphasized by a 1st-order digital network with a transfer function, $H(z) = 1 - 0.95z^{-1}$, and blocked into 45 ms frames every 15 ms. A Hamming window is used to compute the first 9 autocorrelation coefficients, and a standard LPC analysis is performed. Durbin's recursion is used to solve for the LPC coefficients. Cepstral coefficients are calculated recursively from the LPC coefficients. Each cepstral coefficient time trajectory is then modeled by a 1st-order orthogonal polynomial over a finite length time window.

Two VQ codebooks (one instantaneous, one transitional), each with 64 entries, are generated and used as a model for that particular speaker. In the recognition (test) phase, the LPC cepstral coefficient vector, $\boldsymbol{c}(l)$, and the corresponding spectral slope (in time) vector, $\Delta\boldsymbol{c}(l)$, for the lth input frame are compared to the instantaneous and transitional codebooks of each speaker, respectively. The codebook entries which are closest to the input vectors are found using an exhaustive search, and the corresponding distortions are recorded. The transitional and instantaneous distances are then equalized by their corresponding averages of pooled intraspeaker distances, $\bar{d}_{\Delta\text{CEP}}$ and $\bar{d}_{\Delta\text{CEP}}$, and linearly combined as

$$d^n = \alpha \frac{d^n_{\Delta\text{CEP}}}{\bar{d}_{\Delta\text{CEP}}} + (1 - \alpha)\frac{d^n_{\text{CEP}}}{\bar{d}_{\text{CEP}}} \qquad (14)$$

where n refers to the nth speaker (codebook) and α is a combination factor between 0 and 1.

The final decision block of the text-independent identification system is a minimum distance classifier. The unknown speaker's identity is chosen as the one whose resultant accumulated distance is the minimum among N different speakers.

The system, as described above, operates intrinsically in a text-independent mode. In certain applications where a speaker can be identified using a prescribed text, the system can be modified to operate in a text-dependent mode as shown in Fig. 4. In this case, a time registration and comparison block is inserted between the VQ codebook search block and the minimum distance classification decision block. The reference templates of the specified text are stored in terms of VQ index sequences, and are time aligned in the recognition phase with the input utterance using a dynamic time warping (DTW) procedure. A normalized and combined distance, similar to the one in (14), is used in the optimal path search in the DTW procedure.

B. Database

We used a 10-speaker (5 male and 5 female), isolated digit database in this study. Each speaker recorded 200 digits, 20 utterances/digit, over local dialed-up telephone lines. The 200 digits were equally divided and recorded in 5 different sessions over about a 2 month period. The first 100 utterances were used as training data to generate VQ codebooks. The first 50 training utterances were also

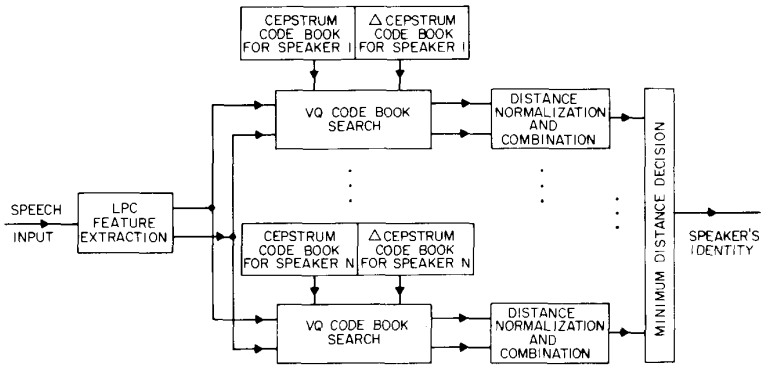

Fig. 3. VQ-based speaker identification system block diagram (text-independent mode).

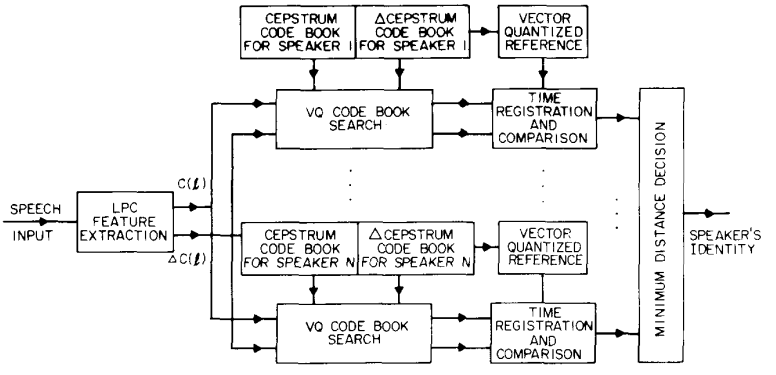

Fig. 4. VQ-based speaker identification system block diagram (text-dependent mode).

used to generate VQ indexed word reference templates for each digit, which were used in our text-dependent speaker recognition experiments. The second 100 utterances were used for testing. The 10-speaker database used in this pilot parameter study is rather small for a full-fledged speaker recognition experiment. However, since the goal of this study is set more on studying the properties of the instantaneous and transitional spectral representations than to carry out a full-scaled speaker recognition task, we feel that results obtained in this study should be relatively correct when tested in a much larger database. The validity of our results was confirmed in a companion paper where a superset of this database consisting of 100 speakers (50 male and 50 female) was used for full scale evaluations [17].

IV. Experiments and Results

In this section, various experimental effects are discussed, including: the effects of window length used for computing the orthogonal polynomial coefficients, the correlation between the cepstral distance and the Δ-cepstral distance, combination of instantaneous and transitional spectral information, transmission channel effects, the effects of distance weighting, cross-sex and within-sex speaker confusions, and a text-dependent recognition experiment.

A. Effects of Window Length in Computing Spectral Variation

In this section, the effects, on recognition performance, of fitting a 1st-order orthogonal polynomial to the cepstral coefficient trajectory are studied. Euclidean distances are used for VQ codebook construction and testing. Text-independent speaker identification results using zeroth- and 1st-order orthogonal polynomial coefficients over different length rectangular time windows are shown in Figs. 5 and 6, respectively. It should be noted that the two figures are plotted in different scales. As seen in Fig. 5, there exists some appreciable, although relatively small, performance advantage by introducing some averaging of cepstral coefficients. There exists a relatively small difference in recognition performance between one frame "instantaneous" cepstral coefficients and cepstral coefficients averaged over 3, 5, and 7 frames. The result is true for test tokens of different length, including: 1, 2, 4, and 10 random digits. Based upon the above observation, we decided to use only the original cepstral coefficients to characterize "instantaneous" spectral information in all later experiments.

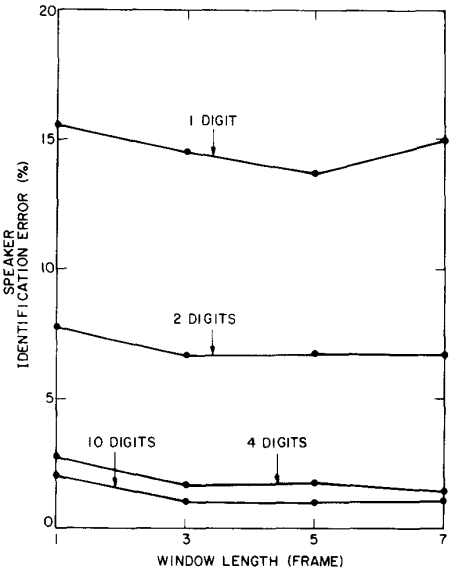

Fig. 5. Speaker identification error rate versus different window lengths when zeroth-order orthogonal polynomial coefficients (spectral average over time) are used.

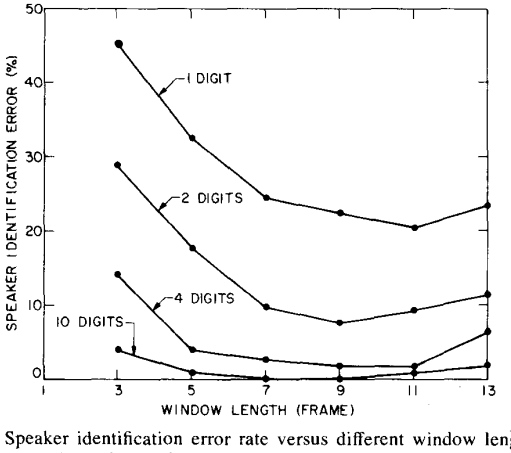

Fig. 6. Speaker identification error rate versus different window lengths when 1st-order orthogonal polynomial coefficients (spectral derivative) are used.

Speaker identification performance versus different window lengths is rather different, however, when transitional spectral information, or Δ-cepstrum, was used. As shown in Fig. 6, when a relatively short window size, say 3 frames, was used, the 1st-order orthogonal polynomial coefficients are rather noisy, and, as a consequence, the identification error is high. At the other extreme, when a very long window is used, say 13 frames, the performance degrades also. The degradation due to long window lengths can be partially attributed to the edge effects at the isolated word boundaries. The optimal choice of window length seems to fall in the range between 7 and 11 frames. We chose this 7-frame rectangular window for estimating spectral transitional information in later experiments. Results using Hamming windows, although not given here, are comparable.

It may look surprising, or even counterintuitive, that a window of such a long duration of 7–11 frames can give good estimates of spectral transitions. The spectral transitional features used in our experiments should be viewed as smoothed and, hopefully, reliable estimates of the trend of spectral change rather than a salient characterization of any spectral transients. On the other hand, a fast spectral transient, in any case, cannot be reliably estimated with an LPC analysis window of 45 ms used in our experiments. From our experience in both speech and speaker recognition, a relatively smoothed but more consistent spectral estimate usually yields better recognition performance than a higher resolution but less consistent estimate.

B. Correlation Between Instantaneous and Transitional Spectral Information

In this subsection, we will demonstrate that the instantaneous and transitional spectral features, as measured here, are complementary and can be used together to improve the overall system performance. The intraspeaker distances, denoted as d_{CEP} and $d_{\Delta\mathrm{CEP}}$, normalized by their corresponding standard deviations, are shown in Fig. 7 in a scatter diagram. The normalized correlation coefficient of 0.6 between these two intraspeaker distances is not high. Since both features carry speaker related information, this relatively low correlation between the two distances indicates that they are not redundant and should be complementary to each other for improving speaker recognition performance.

C. Combination of the Two Spectral Distances

As shown in the previous subsection, since the two spectral features are relatively uncorrelated, they can be used jointly to improve speaker recognition performance. As described in (14), the instantaneous and transitional spectral distances are first equalized by the corresponding averages of these intraspeaker distances, and then combined with a combination factor α. The speaker identification error rate using single-digit test tokens over the 10-speaker population is shown in Fig. 8. It can be seen that the performance obtained by using either the transitional ($\alpha = 0$) or the instantaneous ($\alpha = 1$) information alone is worse than the performance obtained by combining them. For both the unweighted and the inverse variance weighted distances, rather broad regions of α near the optimal performance points (around $\alpha = 0.5$) are observed. The performance improvement for combined spectral features is rather consistent throughout the whole speaker population as shown in Fig. 9. It can be seen that the combined performance is better than either the transitional feature or the instantaneous feature for every speaker. The identification error rates are averaged over the whole testing database which consists of 100 randomly chosen, single-digit test tokens. Similarly, the combined features outperform the transitional or the instantaneous spectral features regardless which digit is used as a test token as shown in Fig. 10.

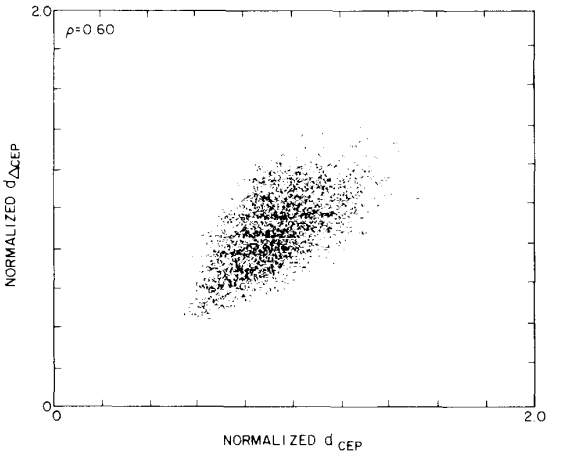

Fig. 7. Scatter diagram showing the correlation between the instantaneous and transitional cepstral distances.

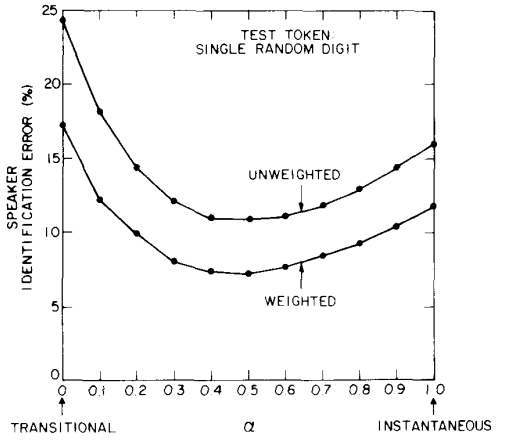

Fig. 8. Speaker identification error rate versus the combination factor α.

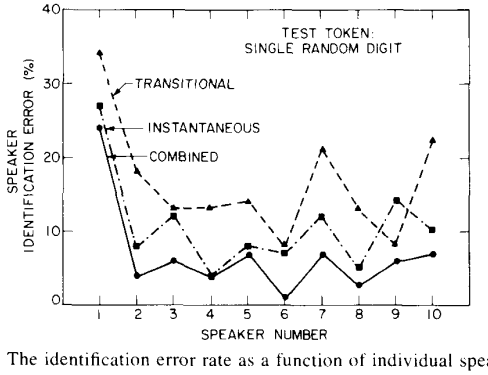

Fig. 9. The identification error rate as a function of individual speakers.

D. Resistance of the Two Spectral Features to Channel Variations

To evaluate resistance of the instantaneous and transitional spectral features to transmission channel variations, we used the original training data to generate instantaneous and transitional VQ codebooks for each speaker.

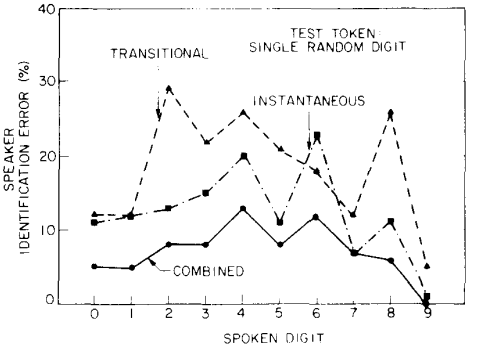

Fig. 10. The identification error rate as a function of individual spoken digits.

However, the test data are passed through a fairly mild 1st-order preemphasis filter with a transfer function, $H(z) = 1 - 0.3z^{-1}$, to introduce an artificial transmission channel mismatch between training and test data. The spectral tilt caused by this mild preemphasis dramatically increases the instantaneous spectral distance between the testing data and the VQ codebooks. The identification results are depicted in Fig. 11. (A broken line designates a channel mismatch and a solid line designates no channel mismatch.) With statistical weighting, which weights the lower order cepstral coefficients less than the higher order ones, the system performance is more resistant to this channel mismatch. As shown in the figure, a 6.7 percent degradation of performance is observed for the weighted cepstral distance and a more dramatic degradation, 21.2 percent, is observed when no statistical weighting is applied.

The transitional spectral features, due to their differential nature (in time) in the log spectral domain, are more resistant to channel variation than the instantaneous features. For both weighted and unweighted transitional cepstral features, the degradation of performance introduced by the artificial channel mismatch is minimal (i.e., < 1 percent). The weighted transitional cepstral distance still outperforms the unweighted distance by about 7 percent in both channel conditions.

E. Cross-Sex and Same-Sex Speaker Confusions

It has long been observed by speech researchers that speech spectra of speakers tend to form two distinctive clusters according to the sex of the speaker. A similar phenomenon is observed in our experiment, as shown in Fig. 12. When weighted instantaneous cepstral features are used, only 9.4 percent of the total speaker confusions are cross-sex confusions, the remaining 90.6 percent confusions are either male-to-male or female-to-female confusions. However, when the transitional cepstral features are used, the confusion pattern is shifted toward more cross-sex confusions. The rate of cross-sex confusions, although still less than that of the same-sex confusions, is now at 26.4 percent of the total errors. This result indicates that transitional cepstral features are less effective

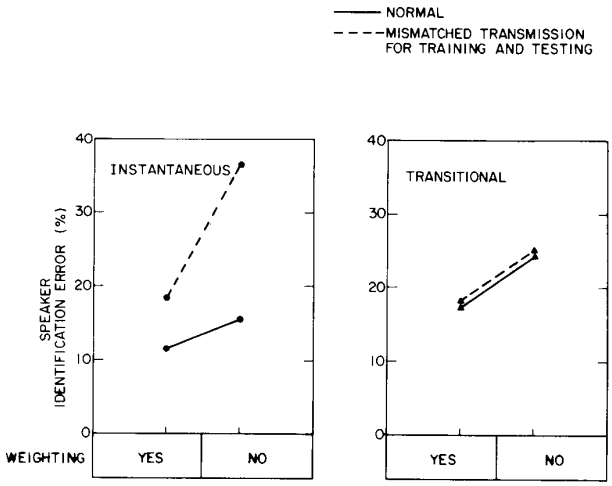

Fig. 11. Identification error rate when a mismatch of transmission channel conditions between the VQ training and testing is introduced.

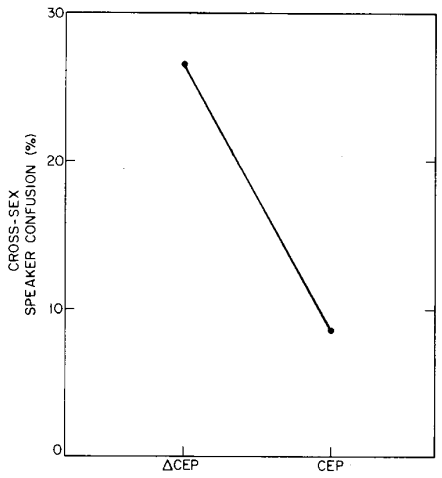

Fig. 12. Cross-sex speaker confusion percentage of the total speaker confusions as a function of the spectral representation.

in differentiating male speaker from female speaker than instantaneous spectral features.

F. Text-Dependent Speaker Recognition

It is reasonable to conclude that transitional spectral features, as suggested in this study, are useful in capturing the temporal, contextual information of a speaker's voice. However, the most direct and precise approach for exploiting contextual information in spoken utterances for speaker recognition applications is to stipulate that the test and reference utterances be tokens of the same word or phrase. This is the so-called text-dependent approach to speaker recognition. In this approach, temporal and contextual information is accounted for directly. Comparisons, using dynamic time warping alignment techniques, are more precise since related portions of a given word or phrase are compared sequentially. Each test digit is compared to 5 prototypes for that digit, encoded both in instantaneous and transitional forms. The minimum distance over the 5 prototypes is taken as the distance for that input digit.

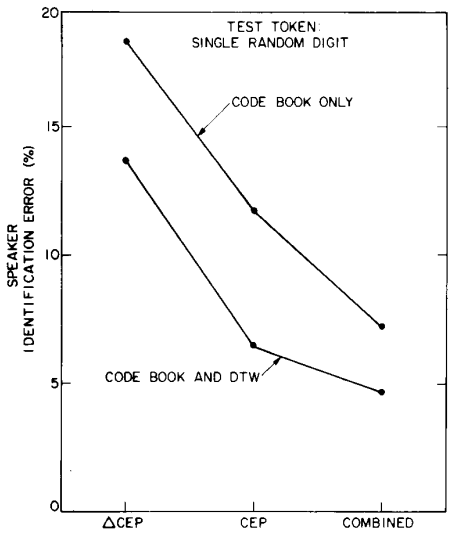

Fig. 13. Speaker identification error rate as a function of different spectral representations (text-dependent and text-independent).

Speaker identification performance for the text-dependent mode ("codebook and DTW") is shown in Fig. 13 and compared to results for the text-independent mode ("codebook only"). It can be seen that there is a significant and consistent performance advantage for the text-dependent mode over the text-independent mode. In the same figure it can be seen that, in a text-dependent mode, the rate of performance improvement obtained using combined spectral features is less than that in a text-independent mode. This suggests that DTW alignment offsets improvements attributable to transitional spectral features.

It should be noted that the temporal structure of a speaker's utterances can also be captured by using a hidden Markov model (HMM) as demonstrated by Poritz [18].

V. Summary and Concluding Remarks

In this study, we investigated several issues which are relevant to characterizing the spectral features of a speaker's voice, including: instantaneous and transitional spectral representations of spectral information, weighted and unweighted spectral distances, and different ways to exploit the dynamic temporal structure of speech sounds in both text-independent and text-dependent speaker recognition systems.

The results are summarized in Table I, where the speaker identification performance is tabulated for various different experimental conditions. For all the results in Table I, single-digit test tokens were used in the experiments. It is concluded that the instantaneous spectral features carry more speaker relevant information (hence are more useful) than transitional spectral features in automatic speaker recognition. We have also shown that the instantaneous and transitional spectral features are fairly

TABLE I
SPEAKER IDENTIFICATION RATES FOR VARIOUS EXPERIMENTAL CONDITIONS (SINGLE-DIGIT TEST TOKENS WERE USED)

Transitional	•	•			•	•	•
Instantaneous			•	•	•	•	•
Weighting		•		•		•	•
DTW							•
Speaker Identification Rate (%)	75.5	82.6	84.3	88.3	89.1	92.8	95.3

uncorrelated so that they can be used jointly to improve speaker recognition performance. We have also found that Euclidean distances, when inversely weighted by corresponding variances of individual components, give better recognition performance than unweighted Euclidean distances. In terms of their resistance to transmission channel variations, the instantaneous cepstral features have been shown to be more susceptible to a transmission channel mismatch between a speaker model (i.e., his VQ codebook) and the test speech material, than the transitional cepstral representations. The transitional spectral features are also associated with more cross-sex confusions than the instantaneous features. Finally, performance in a text-dependent mode, whether spectral features are combined or not, is uniformly better than performance in a text-independent mode.

REFERENCES

[1] B. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, pp. 460–475, Apr. 1976.

[2] S. Furui, "Cepstrum analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 254–272, Apr. 1981.

[3] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, "A vector quantization approach to speaker recognition," *AT&T Tech. J.*, vol. 66, pp. 14–26, 1987.

[4] K. P. Li and E. H. Wrench, Jr., "An approach to text-independent speaker recognition with short utterances," in *Proc. ICASSP*, vol. 2, 1983, pp. 555–558.

[5] R. E. Helms, "Speaker recognition using linear prediction vector codebooks," Ph.D. dissertation, Southern Methodist Univ., 1981.

[6] K. Shikano, "Text-independent speaker recognition using vector quantization," *J. Acoust. Soc. Amer.*, suppl. V.77, p. 511, 1985.

[7] J. T. Buck, D. K. Burton, and J. E. Shore, "Text-dependent speaker recognition using vector quantization," in *Proc. ICASSP*, vol. 1, 1985, pp. 391–394.

[8] R. Zelinsky and P. Noll, "Adaptive transform coding of speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 299–309, Aug. 1977.

[9] K. K. Paliwal, "On the performance of the quefrency-weighted cepstral coefficients in vowel recognition," *Speech Commun.*, vol. 1, pp. 151–154, 1982.

[10] A. V. Oppenheim, Ed., *Applications of Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[11] D. Klatt, "Prediction of perceived phonetic distance from critical band spectra: A first step," in *Proc. ICASSP*, vol. 2, 1982, pp. 1278–1281.

[12] N. S. Nocerino, F. K. Soong, L. R. Rabiner, and D. H. Klatt, "Comparative study of several distortion measures for speech recognition," in *Proc. ICASSP*, vol. 1, 1985, pp. 25–28.

[13] M. R. Schroeder, "Direct (nonrecursive) relations between cepstrum and predictor coefficients," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 297–301, Apr. 1981.

[14] Y. Tohkura, "A weighted cepstral distance measure for speech recognition," in *Proc. ICASSP*, vol. 1, Apr. 1986, pp. 761–764.

[15] P. R. Bevington, *Data Reduction and Error Analysis for the Physical Sciences*. New York: McGraw-Hill, 1969.

[16] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantization," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.

[17] A. E. Rosenberg and F. K. Soong, "Evaluation of a vector quantization talker recognition system in text indepenent and text dependent modes," in *Proc. ICASSP-86*, vol. 2, 1986 pp. 873–876.

[18] A. Poritz, "Linear predictive hidden Markov models," in *Proc. ICASSP*, vol. 2, 1982, pp. 1291–1294.

Frank K. Soong (S'76–M'82), for a photograph and biography, see p. 48 of the January 1988 issue of this TRANSACTIONS.

Aaron E. Rosenberg (S'57–M'63–SM'83–F'84) received the S.B. and S.M. degrees in electrical engineering from M.I.T. in 1960 and the Ph.D. degree in electrical engineering from the University of Pennsylvania in 1964.

He is a member of the Technical Staff in the Speech Research Department at AT&T Bell Laboratories, Murray Hill, NJ. He has been at Bell Labs since 1964 where his research interests have included auditory psychophysics, speech perception, and currently, speech and speaker recognition. He has authored or co-authored over 45 papers in these fields.

Dr. Rosenberg is a Fellow of the Acoustical Society of America and a member of Sigma Xi. He is a member of the IEEE Acoustics, Speech, and Signal Processing Society's Conference Board. He has served as a member of the Society's Administrative Committee, as Chairman of the Society's Technical Committee on Speech Communication, and as Associate Editor for the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING.