Effectiveness of Orthogonal Instantaneous and Transitional Feature Parameters for Speaker Verification

A M Ariyaeeinia and P Sivakumaran

Division of Electrical and Electronic Engineering,
University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, UK

Abstract: The effectiveness, for text-dependent speaker verification, of orthogonal instantaneous and transitional feature parameters of speech is investigated. Instantaneous spectral features are represented by cepstral coefficients obtained through a linear prediction analysis of speech. Transitional spectral information is characterised using differential cepstral coefficients. Sets of orthogonal parameters are obtained by applying an eigenvector analysis to instantaneous and transitional feature coefficients. The experimental work is based on the use of a subset of the BT Millar speech database, consisting of repetitions of isolated digit utterances 1 to 9 and zero spoken by twenty male speakers. The investigation includes an examination of the relative speaker discrimination abilities of the above two types of orthogonal feature parameters. It is shown experimentally that the equal error rate in speaker verification can be reduced significantly by forming a spectral distance based on a combination of orthogonal instantaneous and transitional feature parameters. It is further demonstrated that, when the input utterance consists of a sequence of five digits, an equal error rate of less than 0.5% can be achieved.

I. INTRODUCTION

Speaker verification is defined as the automatic authentication of the identity claimed by an unknown speaker, given only the samples of the speaker's voice. This has been an area of active research for over two decades [1-5]. The main impetus for research into this field has been the development of new and revolutionary applications in many diverse areas. Examples of these are protection of confidential computer files, access control for telephone banking, control of entry to restricted areas in secure buildings, and automatic telephone transactions such as credit card verification. In general, speaker verification can be either text-dependent or text-independent. The former mode of operation provides additional security due to the requirement for a password, and is believed to be the one closer to a practical implementation.

The classical approach to text-dependent speaker verification is through the use of short-term spectral analysis [6]. An effective and widely used analysis of this type is that based on the linear predictive coding (LPC). The results of some earlier experimental studies [7] have shown that, through an eigenvector analysis of LPC-derived speech feature coefficients, a set of orthogonal parameters can be obtained which may then be effectively used for the purpose of speaker discrimination. It has been demonstrated that [8] amongst various types of LPC-based orthogonal feature parameters, cepstral coefficients possess the highest speaker discrimination ability. An important advantage of the orthogonalisation technique has been found to be due to the fact that it can be used for text-dependent speaker verification without the need for a separate time normalisation procedure.

It has been demonstrated that [9] the transitional spectral features, associated with the time varying properties of speech, can be effectively represented through the use of an orthogonal polynomial fit of each cepstral coefficient trajectory over a finite time window. The resultant so called Δ-cepstral coefficients have been shown to possess excellent speaker discrimination properties [9,10].

In the following sections a description of the methods used for extracting orthogonal cepstral and Δ-cepstral coefficients is presented, and the spectral distances employed in the experiments are discussed. The dependence of the performance of the transitional feature parameters on the length of the time window used for computing these is then experimentally analysed. Finally an investigation into the relative effectiveness, for speaker verification, of cepstral coefficients, Δ-cepstral coefficients, and a combination of the two is presented.

II. PROCEDURE FOR SPEAKER MODELLING AND VERIFICATION

In the present work, the instantaneous and transitional spectral features are represented using orthogonal cepstral

0-7803-2627-X/95 $4.00

and Δ-cepstral parameters respectively. Cepstral parameters are derived recursively from the LPC coefficients, and Δ-cepstral coefficients are obtained as a first-order polynomial characterisation of the spectral change over a finite length window [9,10], i.e.:-

$$\Delta c_m(t) = \frac{\sum_{k=-K}^{K} k h_k c_m(t+k)}{\sum_{k=-K}^{K} h_k k^2} \quad (1)$$

where $\Delta c_m(t)$ is the m^{th} Δ-cepstral coefficient in the t^{th} frame, h_k is a symmetric time window (usually, and also in this study, of the rectangular type) with a length of $2K + 1$ frames, and $c_m(t+k)$ is the m^{th} cepstral coefficient in the $(t+k)^{th}$ frame. Although higher order Δ-cepstral coefficients may similarly be derived, it has been indicated that [9,10] for efficient representation of spectral dynamics over time, a first order polynomial characterisation is usually adequate.

In order to orthogonalise cepstral feature parameters a conventional linear transformation can be used [7]. Representing the matrix of cepstral parameters by $\mathbf{c}$, this transformation may be expressed as:-

$$\boldsymbol{\phi} = \mathbf{b}^T \mathbf{c} \quad (2)$$

where $\boldsymbol{\phi}$ is the matrix of orthogonal cepstral parameters, and $\mathbf{b}^T$ is the transpose of the matrix of the eigenvectors of the covariance matrix, $\mathbf{R}_c$, of cepstral feature parameters. These eigenvectors are obtained by:-

$$\lambda_i \mathbf{b}_i = \mathbf{R}_c \mathbf{b}_i, \qquad 1 \le i \le p \quad (3)$$

where λ_i represent the eigenvalues of the covariance matrix $\mathbf{R}_c$, and p is the order of cepstral feature vectors. The eigenvalues λ_i, i=1,2,...,p are also the diagonal terms of the covariance matrix, $\mathbf{R}_\phi$, of the orthogonal parameters obtained using (2). Due to the orthogonality of these parameters, the off-diagonal terms of $\mathbf{R}_\phi$ are all zero.

For the purpose of generating reference models, each enrolling speaker can be characterised by the mean values of the orthogonal parameters (i.e. $\bar{\phi}_i$, i=1,2,...,p), the eigenvalues, and the matrix of the eigenvectors [8].

The degree of dissimilarity between a test utterance and the reference model of the proposed speaker can be measured using a combination of the weighted Euclidean distances associated with the mean values and eigenvalues of the orthogonal parameters [8]. This may be expressed as:-

$$d_c = \sum_{i=1}^{p} \frac{(\lambda_i - \gamma_i)^2}{w_i} + \sum_{i=1}^{p} \frac{(\bar{\phi}_i - \bar{\psi}_i)^2}{w_i'} \quad (4)$$

where γ_i and $\bar{\psi}_i$ are the measured variances and the mean values of the set of parameters obtained by linearly transforming the test feature coefficients through the use of the eigenvectors of the proposed speaker. w_i in the above equation represent the variances in the estimation of λ_i. Similarly, w_i' are the variances in the estimation of $\bar{\phi}_i$.

The technique described above can also be used for orthogonalising Δ-cepstral parameters. In this case, however, the mean values of the orthogonal parameters will depend disproportionally on the few starting and ending frames [11]. As a result the distance metric for Δ-cepstral parameters is based only on the variance information. This metric is given by:-

$$d_{\Delta c} = \sum_{i=1}^{p} \frac{(\tilde{\lambda}_i - \tilde{\gamma}_i)^2}{\tilde{w}_i} \quad (5)$$

where $\tilde{\lambda}_i$, $\tilde{\gamma}_i$ and $\tilde{w}_i$ have the same meanings as their corresponding parameters in (4).

III. EXPERIMENTAL WORK AND RESULTS

This section describes a set of speaker verification experiments conducted using orthogonal cepstral and Δ-cepstral feature parameters. The first part of this study is concerned with the effects, on the verification accuracy, of the length of the time window used in computing transitional spectral information. In the second part, a method for combining cepstral and Δ-cepstral spectral distances is investigated, and the relative effectiveness of each of these and their combination is examined.

A. Speech Database and Feature Extraction

For the purpose of these experiments a subset of the BT Millar Speech database is adopted. This subset which was collected in five recording sessions over a period of about three months, has a bandwidth of 3.1 kHz and a sample rate of 8.0 kHz. The subset consists of 25 repetitions of isolated digit utterances 1 to 9 and zero spoken by 20 male English speakers of about the same age. The first 10 versions of each utterance (obtained over the first two recording sessions) are

reserved for training. The last 15 repetitions of the utterances (recorded over the last three sessions) form the standard test set of the data.

In order to extract the required feature parameters, utterances are pre-emphasised using a first-order digital filter with the transfer function $H(z) = 1 - 0.95z^{-1}$. Each utterance is then segmented into 25 ms frames at intervals of 12.5 ms using a Hamming window, and subjected to a 12^{th}-order LPC analysis based on the autocorrelation method. Finally, the methods described earlier are used for the extraction of orthogonal cepstral and Δ-cepstral parameters.

B. Dependence of Δ-Cepstrum Performance on the Duration of the Utilised Window

The estimation of the spectral variation, as indicated in equation (1), depends on the length of the window employed for extracting Δ-cepstral parameters. In order to investigate the effects of the window size on the verification accuracy, experiments were conducted with sets of orthogonal Δ-cepstral parameters derived using an increasing length of the time window. In the first set of these experiments, verification tests were based on the use of the individual digit utterances 1 to 9 and zero. The equal error rates (EER) calculated for the individual digits were then averaged to obtain an overall equal error rate for each given length of the time window. A plot of this is presented in Figure 1. As seen in this figure, a consistent and sharp decrease in verification error rate is achieved (from around 27% to about 10%) when the window size is increased continuously from 3 frames to 13 frames. For windows longer than 13 frames, the EER appears to oscillate around 10%. In fact further investigations have revealed that the window length must exceed 31 frames, before the verification error rises significantly again.

The utterances used in the next stage of experiments consisted of digit sequences formed by cascading up to five different randomly selected individual digits. Each of these utterances was used in turn to conduct a set of verification tests based on orthogonal Δ-cepstral parameters and an increasing length of the time window. The results of this study (Figure 2) indicate that, as expected, the equal error rate decreases considerably as the digit contents of the utterance is increased. It is also observed in Figure 2 that the verification error in all cases reaches a minimum when the window length is about 13 frames, and tends to gradually increase with further increases in the time window size. The error rate for a window of 13 frames (corresponding to 175 ms) has been found to range from about 9% to under 0.8%.

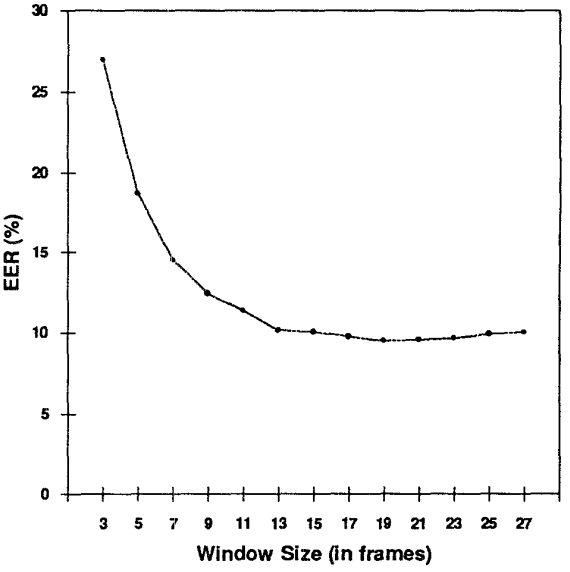

Figure 1: Average equal error rate for single digit utterances in experiments based on Δ-cepstral parameters and an increasing size of the time window.

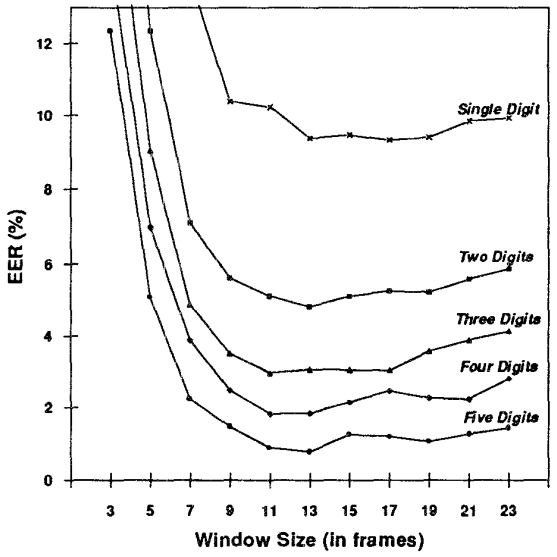

Figure 2: Dependence of orthogonal Δ-cepstrum performance on the size of the utilised window and the utterance duration.

Based on these results, and those obtained for single digit utterances, it was decided to use a time window of 13 frames for the following parts of the investigation.

C. Relative Effectiveness of Instantaneous, Transitional and Mixed Spectral Information

The main aim of this part of the investigation was to experimentally examine the relative speaker discrimination abilities of cepstrum, Δ-cepstrum and a combination of these. The distance for the combined information was obtained as:-

$$d = \alpha d_{nc} + (1-\alpha) d_{n\Delta c} \quad (6)$$

where α is the combination factor, and a subscript *n* indicates that the distance has been normalised by the average of its corresponding intraspeaker distances.

In order to determine the most effective value of α, a set of speaker verification tests were conducted using single digit utterances. For each digit utterance the value of α was varied from 0 to 1 in steps of 0.1 and, in each case, the equal error rate in verification was computed. An overall equal error rate for each value of α was then obtained by averaging the equal error rates associated with individual digits.

The results of this experimental study are given in Figure 3. By comparing the error rates for two values of α of 0 and 1 it becomes evident that for single digit utterances, the distance based on cepstral parameters is considerably more effective than that based on Δ-cepstral information.

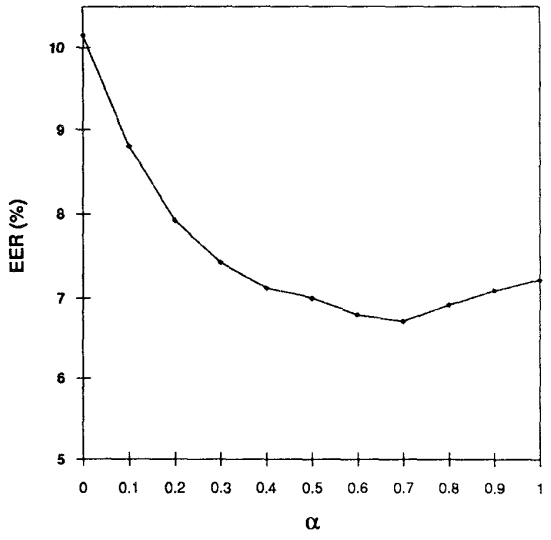

Figure 3: Average equal error rate associated with the combined spectral distance as a function of the combination factor α.

It is also observed in this figure that the error rate reaches its minimum (about 6.7%) for a combination factor of 0.7. This confirms that a higher accuracy in verification can be achieved by using the combined distance than using any of the two cepstral and Δ-cepstral distances alone. In order to investigate the relative verification abilities of these distances more thoroughly, two other sets of experiments were conducted using single digit utterances and sequences of digits. The combination factor used in these experiments was 0.7, and the digit sequences were the same as those used for the experiments discussed in part *B* of this section.

Figure 4 illustrates the experimental results obtained for single digit utterances. It can be observed in this figure that the error rates for cepstral parameters are consistently, and in some cases significantly, less than those associated with Δ-cepstral parameters. This figure also shows that for most digit utterances, the error rates obtained using the combined distance is less than those resulted by using cepstral parameters. The slightly better performance of the cepstral distance in the cases of digit utterances four and six is evidently due to the large error rates associated with Δ-cepstral distance in these cases.

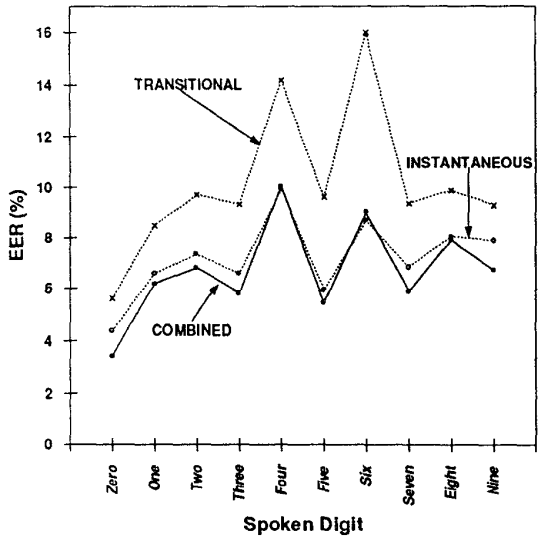

Figure 4: Equal error rates in verification experiments based on single digit utterances.

The results of experiments with digit sequences are presented in Figure 5. It is interesting to note from these results that the effectiveness of Δ-cepstral parameters relative to instantaneous feature parameters improves significantly as the duration of the spoken material increases. Figure 5 shows

that in fact, for four-digit and five-digit sequences, Δ-cepstral parameters perform better than cepstral parameters. The error rate obtained for the combined distance, ranging from just under 6% for a single-digit utterance to under 0.5% for a sequence of five digits, is consistently less than those for cepstral and Δ-cepstral distances.

The above results further suggest that since the relative effectiveness of cepstral and Δ-cepstral distances vary with the utterance duration, the value of α leading to the best result for the combined distance may not be the same for utterances of different lengths.

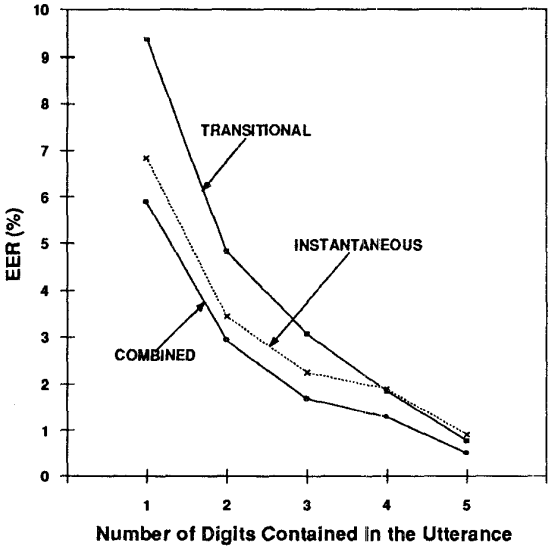

Figure 5: Equal error rates for different spectral distances as a function of the utterance duration.

IV. CONCLUSIONS

An investigation into text-dependent speaker verification has been conducted. In this study, the instantaneous and transitional features of speech have been represented using orthogonal cepstral and Δ-cepstral parameters respectively.

The experiments carried out using orthogonal Δ-cepstral parameters have shown that the duration of the time window used in computing these parameters highly affects their speaker discrimination ability. The experimental results have also indicated that for single-digit utterances, orthogonal cepstral coefficients perform significantly better than orthogonal Δ-cepstral parameters. It has, however, been demonstrated that the relative speaker discrimination ability of the orthogonal Δ-cepstral parameters improves considerably as the duration of the utterance is increased. The results of verification tests with digit sequences of different lengths have further indicated that when the utterance contains a sequence of four or more digits, the error rates for these parameters are less than those obtained using orthogonal cepstral parameters.

Further experimental studies have shown that by combining the distances associated with the instantaneous and transitional spectral information a significant improvement in verification accuracy can be achieved. The equal error rate obtained using the combined distance has been found to range from about 6% for a randomly selected single-digit utterance, to under 0.5% for a sequence of five digits.

V. ACKNOWLEDGEMENTS

The authors wish to express their thanks to Mr M Pawlewski of BT Laboratories for his support and stimulating discussions.

The investigation presented in this paper is part of the research work supported by funding from BT Laboratories.

VI. REFERENCES

[1] A. E. Rosenberg and M. R. Sambur, "New Techniques for Automatic Speaker Verification", *IEEE Trans. Acoust, Speech, and Signal Processing*, vol. 23, pp. 169-176, April 1975.

[2] A. E. Rosenberg, "Automatic Speaker Verification: A Review", *Proc. IEEE*, vol. 64, pp. 475-487, April 1976.

[3] R. E. Bogner, "On the talker Verification Via Orthogonal Parameters", *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 29, pp. 283-289, Aug. 1981.

[4] D. V. Burton, "Text-Dependent Speaker verification Using Vector Quantization Source Coding", *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 35, pp. 133-143, Feb. 1987.

[5] T. Matsui and S. Furui, "Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's", *IEEE*

Trans. Speech and Audio Processing, vol. 2, pp. 456-458, July 1994.

[6] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification", *J. Acoust. Soc. Am.*, vol. 55, pp. 1304-1312, June 1974.

[7] M. R. Sambur, "Speaker Recognition Using Orthogonal Linear Prediction", *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 24, pp. 283-289, Aug. 1976.

[8] A. M. Ariyaeeinia and P. Sivakumaran, "Speaker Verification Based on the Orthogonalisation Technique", *in Proceedings of the IEE European Convention on Security and Detection (ECOS'95)*, No. 408, May 1995, pp. 101-105.

[9] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 29, pp. 254-272, April 1981.

[10] F. K. Soong and A. E. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 36, June 1988.

[11] H. Gish and M. Schmidt, "Text-Independent Speaker Identification", *IEEE Signal Processing Magazine*, vol. 11, pp. 18-32, Oct. 1994.