Proceedings of the Institute of Acoustics

SPEAKER VERIFICATION USING ORTHOGONAL LINEAR PREDICTION

E Abadjieva

The MicroCentre, Dept. of Mathematics and Computer Science, †
The University, Dundee, DD1 4HN, Scotland

1. INTRODUCTION

The main task in speaker verification is to accept or reject the claimed identity of a tested speaker. Its successful solution can have many practical applications. For instance it can be used to provide an additional level of security for privileged access, based on the unique biometric features of the speaker. The speaker verification problem has been under consideration in many major speech laboratories. It is difficult to compare the reported results because they depend strongly on the size and type of the test population and the experimental conditions. There is still a need for research work aimed at the design and development of a real time speaker verification system reliable enough to be included in security applications.

Traditionally there is a general tendency to use techniques that have proven to be successful for speech recognition such as Dynamic Time Warping (DTW), Vector Quantisation (VQ) and Hidden Markov Modelling (HMM) to solve the speaker verification task. When adopting this approach the benefits introduced for the speech recognition task, such as time normalisation when using DTW, the data compression provided by VQ, or the compensation for time and frequency variability as a result of HMM, have to be disabled by the use of extra processing cepstral weighting for the VQ technique [4], or segmentation and labeling for the HMM approach, in order to reflect the specific verification problem. It is more logical to use a technique that is initially designed to emphasize the difference in the vocal tract parameters and thus separate different speakers. Such a method was proposed in [1] and different versions of it were investigated in [2,3]. It is based on the idea that a linear transformation of the linear prediction parameters using the eigenvectors and values of the covariance matrix can provide a high speaker discrimination potential. The reported results were encouraging despite the lack of a time normalisation procedure. An extended version of this method combined with cepstral analysis and its prototype implementation are discussed in this paper.

The objectives of this research are to explore the possibility of implementation of an orthogonal linear predictive cepstral analysis and an extended Mahalanobis distance measure, including variance information, in a real-time prototype speaker verification system based on a standard PC configuration with a DSP processing board.

2. ORTHOGONAL LINEAR PREDICTION CEPSTRAL ANALYSIS

Identifying speech characteristics that are effective for automatic speaker verification has been the subject of a great deal of research. An effective speaker verification feature should measure some aspect of the acoustic signal that reflects the unique properties of the speaker vocal apparatus, and should contain little or no information about the linguistic content of the speech. If the selected

† The work reported here was carried out while the author was employed at Centre for Speech Technology Research, Edinburgh University.

feature is indicative of not only the speaker but also of what is being said, then the full speaker discrimination potential of the feature can only be realized when the recognition process is confined to a comparison of speech samples with exactly the same speech content. The non-trivial and non - error free operations of segmentation and time normalisation are usually required to guarantee the success of systems that depend on such features.

The most effective features which have been reported for application in automatic speaker verification systems are pitch, gain and the linear prediction parameters of the speech waveform. The pitch and gain parameters are related to the properties of the speaker's glottal source and the linear prediction coefficients are indicative of the speaker's vocal tract. Since these three features are by themselves sufficient to produce a high quality synthesis, they must necessarily contain a high degree of information about the speaker's identity. Unfortunately these parameters are also quite obviously influenced by the exact linguistic content of the speech signal and can, therefore, not be regarded as "ideal" recognition measurements.

In the field of speech processing, the term linear prediction refers to a highly successful representation of the speech signal as the output of an all-pole filter that is excited by a sequence of pulses separated by the pitch period for voiced sounds, or pseudo-random noise for unvoiced sounds. The technique of orthogonal linear prediction was introduced to exploit the experimental observation that the linear prediction parameters exhibit significant redundancy for speaker verification purposes. This redundancy implies that a conventional eigenvector analysis can be used to reduce the dimensionality of the linear prediction space. The eigenvector analysis involves the generation of a set of statistically uncorrelated parameters that are formed by a linear combination of the given linear prediction parameters. The redundancy in the linear prediction parameters is reflected in the fact that only a small subset of the orthogonal parameters will demonstrate any significant variation across a speech utterance. The remaining orthogonal parameters can be effectively considered constant and only a knowledge of their specific mean values is required. The method of orthogonal parameters uses the eigenvalues and eigenvectors of the covariance matrix of measurements made on a set of utterances from an enrolled speaker to transform the measurements made on an utterance from a speaker claiming to be that authorised speaker.

The advantages of the proposed theoretical approach are derived from the fact that eigenvector analysis can be used to provide a reliable estimation of speaker specific vocal tract features, according to the verification results presented in the literature [1,2]. The method can be used in both text-dependent or text independent mode and is language independent. In terms of engineering realisation the orthogonal linear prediction algorithm, has a structure which is well suitable to real time implementation on a DSP processing board. In addition the stored references for each user of the system are compact.

3. THE SPEAKER VERIFICATION SYSTEM

The developed speaker verification system is a real-time prototype system based on a standard PC configuration with a DSP processing board. The front end processing algorithm includes standard LPC - cepstral analysis software, which was already developed at CSTR (Edinburgh University), as a part of a speech recognition project. Some modifications and adaptations have been added according to the requirements of the speaker verification task. The transformation of the speech waveform includes a standard set of speech processing techniques and results in a compressed description as a sequence of cepstral parameters. The real-time realisation is achieved by making

maximal use of the DSP32C processing card internal architecture at assembler level. The resulting description of the speech signal is transferred and stored on the PC. It is further processed by a principal components analysis procedure to form reference patterns for each user in training mode. Finely, at the verification stage, a Mahalanobis distance measure extended to include variance information is calculated.

The system operates in three independent modes:

1. Data Collection;
2. Training;
3. Verification.

A new user can be included at any time and tested as an impostor against any of the stored references of previous users or verified using his own characteristics. He is introduced and verified by a 7 character identification name. In the testing experiments the speakers preferred to use their e-mail name.

DATA COLLECTION MODE - For speaker verification it is important to record users' vocal tract characteristics through different time intervals. The Data Collection mode organizes and maintains a file structure with the training data for each user, keeping track of the date of the recording, the text and the training parameters. A list of phonetically balanced sentences is displayed and the user has the choice to combine them when collecting data for text independent experiments. The chosen sentence is displayed together with instructions on how to operate for the recording. The processing time for each sentence is about 30 seconds depending on its length.

TRAINING MODE - The speaker verification system requires training for each user. It is done using the material from the user's data file. Just one or up to ten sentences can be used for training. In this mode a file containing the reference parameters for each user is created and updated every time a new user is introduced to the system. The stored information for each user within the file is 686 bytes. The training time depends on the number of sentences and is about 20 seconds per sentence.

VERIFICATION MODE - A list with the identification names of all the users the system is trained with is displayed. The speaker can claim an identity by choosing one of them. He is invited to say a test sentence and is verified against the chosen person. His verification score is printed. There is an option to store all results in a history file for later statistical analysis.

4. PROCESSING ALGORITHM

The flow charts of the processing algorithm for the three modes of the verification system are shown in Figures 1, 2 and 3. Details about its implementation are as follows:

- Microphone input of the speech signal with sampling rate 10 KHz; maximal duration of the input utterance 6 sec; no end point detection implemented. The processed part of the signal included portions with background noise before and after the utterance.

- Standard segment duration 256 msec; overlap between segments 10 msec. Each segment is processed by standard LPC analysis of order p=12 and includes the following procedures:

  • Preemphasis filtering:

$S'(n) = S(n) - 0.9375\, S(n-1)$ where S(n) is the input signal;

- Hamming windowing with $w(n) = 0.54 - 0.46 \cos(2\pi n / N)$, $0 \le n \le N-1$, N=256;

$S''(n) = S'(n)\, w(n)$;

- Autocorrelation analysis:

$$r(i) = \sum_{n=0}^{N-i-1} S(n)\, S(n+1), \qquad i = 0,...,p;$$

- Linear Predictive analysis:

Using the Durbin-Levinson recursive procedure the LPC coefficients are computed recursively from the following relations:

$H(0) = r(0)$;

$$k_i = \left(r(i) + a_1^{(i-1)} r(i-1) + ... + a_{i-1}^{(i-1)} r(1) \right) / H(i-1);$$

$$a_i^{(i)} = k_i ;$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)}, \qquad j = 1,...,i-1;$$

$$H(i) = (1 - k_i^2) H(i-1);$$

The coefficients $a_j^{(i)}$, j=1,...,i are the LPC coefficients of an i-th order model.

- Cepstral analysis:

$$c_1 = -a_1, \qquad c_k = a_n - \sum_{i=1}^{k-1} c_i\, a_{k-i}, \qquad 1 \le k \le p;$$

- Calculation of the covariance matrix E of the CEP coefficients across the input utterance:

$$e_{ik} = (1/(J-1)) \sum_{j=1}^{J} (c_{ij} - \bar{c}_i)(c_{kj} - \bar{c}_k), \qquad i, k = 1, 2, ..., p;$$

where: J is the total number of segments in the utterance,
c_{ij} is the i-th CEP coefficient in the j-th segment.

$$\bar{c}_i = (1/J) \sum_{j=1}^{J} c_{ij} .$$

For each user a separate file with training data is created. It contains the CEP coefficients and their covariance matrices for each utterance of the training set (Figure 1).

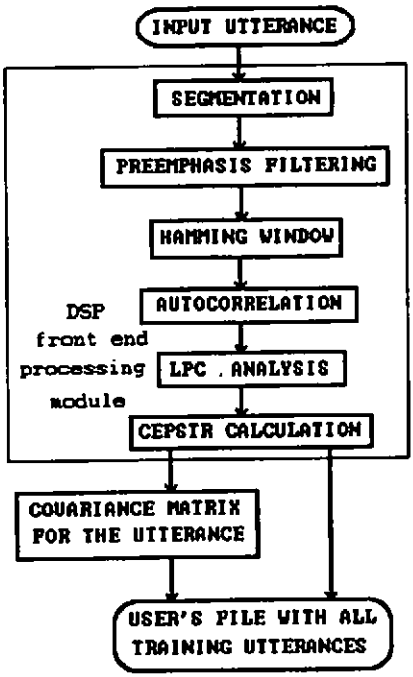

Fig.1
Data collection

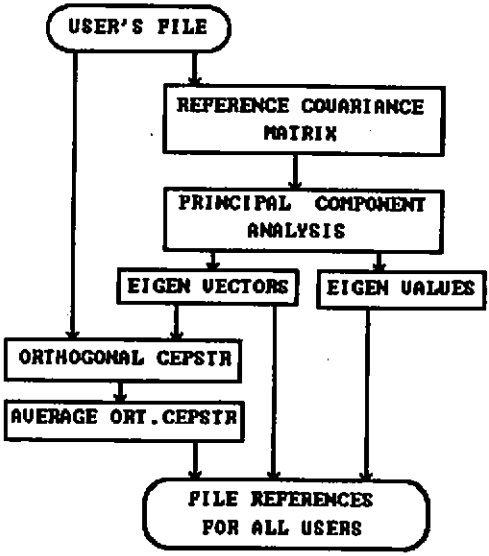

Fig.2
Training

In the training mode (Figure 2) the reference covariance matrix for the *m*-th talker is defined as the weighted average of the calculated covariance matrices of the selected utterances from the training set.

$$R = \left(1 / \sum_{l=1}^{L} J_l \right) \sum_{l=1}^{L} J_l E_l \; ;$$

where: L is the number of the training utterances;

J_l is the number of segments in the l-th utterance.

Principal components analysis is performed to obtain the eigenvalues (the statistical variances) and their corresponding mutually orthogonal eigenvectors.

$$|R - \lambda I| = 0;$$
$$\lambda_i b_i = R b_i; \qquad i = 1,2,...,p;$$

where: I is the identity matrix;
λ_i is the i-th eigenvalue;
b_i is an eigenvector with p elements.

The orthogonal CEP coefficients are calculated by:

$$\phi_{ij} = \sum_{k=1}^{p} b_{ik} c_{kj}; \qquad i = 1,2,...,p; \quad j = 1,2,...J.$$

The average value of the i-th orthogonal CEP coefficient for the *m*-th speaker is:

$$\bar{\phi}_i = \left(1/\sum_{l=1}^{L} J_l\right) \sum_{l=1}^{L} \sum_{j=1}^{J_l} \phi_{ij};$$

The eigenvectors, the eigenvalues and the mean orthogonal CEP coefficients form the reference parameters for each user. They are stored in a common file.

In verification mode the tested speaker claims an identity *v* from the reference set and speaks an input utterance *t* /fig.3/. The spoken utterance *t* is processed in real time by the DSP front end processing module. The resulting CEP coefficients are othogonalised by the eigenvectors of the claimed identity *v* and their mean values $\bar{\phi}_{it}$ are calculated.

The dissimilarity between the tested speaker *t* and the verified identity *v* is calculated by:

$$D_{tv} = d_1 + d_2;$$
$$d_1 = J_v \sum_{i=1}^{p} \left(\left(\bar{\phi}_{iv} - \bar{\phi}_{it} \right) / \sqrt{\lambda_{iv}} \right)^2;$$

where: λ_{iv} is the reference eigenvalue for the i-th orthogonal paramter for the *v* speaker;
J_v the average number of segments in the utterances of the *v* speaker's training set

The second part of the distance measure is an extension including variance information of the form:

$$d_2 = \frac{1}{2} \bar{J}v \sum_{i=1}^{p} \left(\left(v_{if} - \lambda_{iv} \right) / \lambda_{iv} \right)^2 ;$$

where : v_{if} is the measured variance of the i-th orthogonal parameter of the tested speaker .

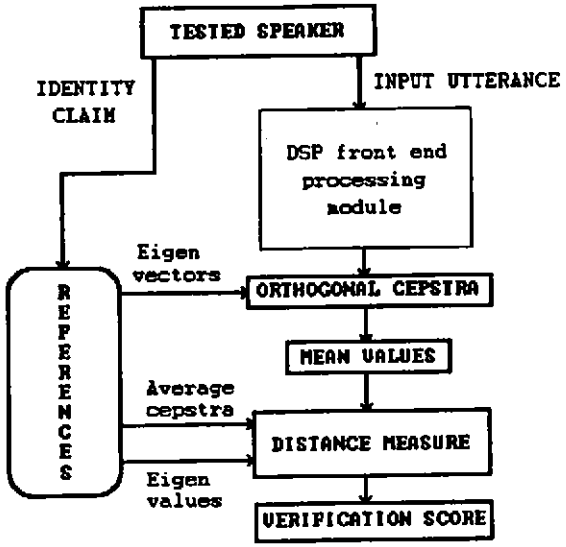

Fig.3
Verification

5. EXPERIMENTAL RESULTS

Text-dependent and text-independent experiments were carried out using a set of phonetically balanced sentences (each one shorter than 6 sec.). The accuracy obtained is 99.7% for text dependent mode, tested over 32 speakers (training and testing done with the same sentence during different sessions), and 95.1% for text independent mode (training for each speaker done with 5 different sentences recorded at different times within 3 weeks, and testing with 5 utterances of a sentence not included in the training set). The verification trials were done on-line; the speakers knew each other's way of speaking and in most cases were trying to imitate it. The results presented were obtained in conditions close to real life situations and prove that the approach forms a good basis for a reliable real time speaker verification system.

6. CONCLUSION

One original contribution of this work is the real time implementation of the orthogonal linear prediction technique in a prototype speaker verification system. The basic approach is enhanced by the use of a new feature set of linear predictive cepstral coefficients and by adding speaker adaptive threshold evaluation. The on-line experiments provide a more severe test for the system. In summary, the positive aspects of the presented work are:

1. The proposed approach is computationally cheap and suitable for real time implementation on a DSP board, as the main part of the processing algorithm consists of matrix multiplications. The stored references for each user are compact (680 bytes) and do not impose any significant limitation on the number of the users.

2. The method operates on the long-term statistics of the cepstra hence it can be used in totally text-independent mode with no linguistic constraints on the testing text. After the system is trained, any sufficiently long sensible utterance (approximately 6 sec) can be used for testing. This feature provides an additional flexibility compared to the commonly used approaches, where training is done on strings of digits and another digit combination is used for testing, which is not a totally text-independent mode [4,5].

3. The accuracy obtained is in the same range as the results presented by other authors [4,5,7], is better for text-independent mode, and is also achieved in real time with less training material (1 to 5 utterances) and computational effort.

7. ACKNOWLEDGEMENT

The author would like to thank CSTR for providing the facilities and support for the reported work, and all the colleagues who took part in the experiments.

8. REFERENCES

[1] M R SAMBUR, 'Speaker Recognition Using Orthogonal Prediction', IEEE Trans. Acoustics,Speech and Signal Processing, vol.ASSP-24, NO.4 (1976);
[2] R E BOGNER, 'On Talker Verification Via Orthogonal Parameters' IEEE Trans. Acoustics,Speech and Signal Processing, vol.ASSP-29, NO.1 (1981);
[3] E ABADJIEVA, 'A Method for Speaker Identification using Orthogonal Linear Prediction', Proceedings of the Institute of Forensic Research,Sofia (1986);
[4] WANG REN-HUA, HE LIN-SHEN, HIROY FUJISAKI, 'A weighted distance measure based on the fine structure of feature space: application to speaker recognition', ICASSP 1990;
[5] A E ROSENBERG, C H LEE,F K SOONG, A McGEE, 'Experiments in automatic Talker Verification Using Sub-word Unit Hidden Markov Models', ICSLP 1990;
[6] E ABADJIEVA-LOGAN, A A WRENCH, A M SUTHERLAND & M A JACK, 'A Real Time Speaker Verification System Using Hidden Markov Models', Colloquium on "Systems and Applications of Man-Machine Interaction Using Speech I/O", IEE, March 1991.
[7] W FEIX, M DeGeorge, ' A Speaker Verification System for Access-Control', ICASSP 1985.