Speaker Verification With Combined Time And Spectral Domain Methods

PETRE G. POP, EUGEN LUPU
Comm. Dept.
Technical University of Cluj-Napoca
26-28, Baritiu str., 400020
ROMANIA

Abstract: - This paper presents two approaches to speaker text dependent verification using a combined VQ-DTW method and a combined TESPAR-DTW method. In first case VQ is used in the training stage to generate a model for each speaker from training utterances as well as in the test stage to compute a model for the test utterance. Before applying DTW for final distance, the test sequences of centroids is rearranged such as in the first position is put that test centroid which is closest to first centroid in the reference sequence and so on. The second approach aim to avoid the complicated process associated with TESPAR alphabet generation. We used DTW to align sequences of epochs in order to generate a verification decision. TESPAR coding is used to generate speech features (sequences of epochs, each with specific shape and duration) for each speaker utterance, in training and testing stage. DTW is used for successive alignments to compute a speaker model in training stage and also in testing stage in order to evaluate a distance between the speaker's models and the test utterance model.

Key - Words: Speaker Verification, DTW, VQ, TESPAR.

1. Introduction

The speaker verification consists of automatically authenticating the identity claimed by a speaker, given only some samples of his voice. There are three categories of approaches in speaker verification. In the first one, the verification system is trained on a particular utterance and the same utterance is latter spoken by the speaker who claims that identity, making up *text-dependent speaker verification*. Within the second approach, verification decisions are based on utterances selected by the speaker and not previously known by the verification system, making up the *text-independent speaker verification.* In *text-prompted* approach, the verification system generate the text that each speaker has to utter in both training and testing stage.

A typical approach to the text dependent speaker verification is DTW (Dynamic Time Warping) [6] in which the unknown speaker's utterances are time aligned to the reference stored for the speaker whose identity is claimed, and the decision to accept or reject is based on a measure of similarity between two time series of parameters corresponding to the utterances.

Vector quantization (VQ) is a data compression technique, with several successful applications in speech and image coding or speech/speaker recognition [1] [2]. In VQ each source vector is coded as one of a prestored set of codewords that minimises the distortion between itself and the source vector. For speech, a VQ codebook is obtained from a training sequence containing typical speech.

TESPAR is a method based on the approximations to the locations of real and complex zeros, derived from an analysis of a band-limited signal [7]. The key features of the TESPAR coding in the speech processing field are the following:

- the capability to separate and classify many signals that are indistinguishable in the frequency domain;
- an ability to code the time varying speech waveforms into optimum configurations for processing with Neural Networks.

2. TESPAR

2.1. TESPAR Basics

TESPAR method is based on the approximations to the locations of the real and complex zeros, derived from the analysis of a band limited signal. The real zeros correspond to the zero crossings of the signal while the complex zeros are associated with local maxima. Numerical descriptors of the signal waveform may be obtained via the classical Shanon numbers resulted from the analysis [8].

The Shanon model involves detecting the ordinates of a waveform at a series of points equally spaced at *1/2W*. A variety of mainly linear transforms (e.g. Fourier, LPC, Wavelet or Walsh) has been developed for describing and classifying key features of the sampled data set. This coding strategies involve the following requirements:

- the use of amplitude descriptors;

- the use of regular sampling;
- an approximation domain dependent upon the numbers of bits per sample.

TESPAR coding is based on the zero-crossings of the signal analysed.

2.2. TESPAR Coding

The key in the interpretation of the TESPAR coding possibilities consists in the complex zeros concept. The band-limited signals generated by natural information sources include complex zeros that are not physically detectable. The real zeros of a function (represented the zero crossing) and some complex zeros can be detected by visual inspection (Fig. 1), but the detection of all zeros (real and complex) is a complex task.

Locating all complex zeros involves the numerical factorization of a $2TW^{th}$-order polynomial [9]. A signal waveform of bandwidth W and duration *T*, contains *2TW* zeros, a number which usually exceeds several thousand [9]. The numerical factorization of a $2TW^{th}$-order polynomial is computationally infeasible for real time.

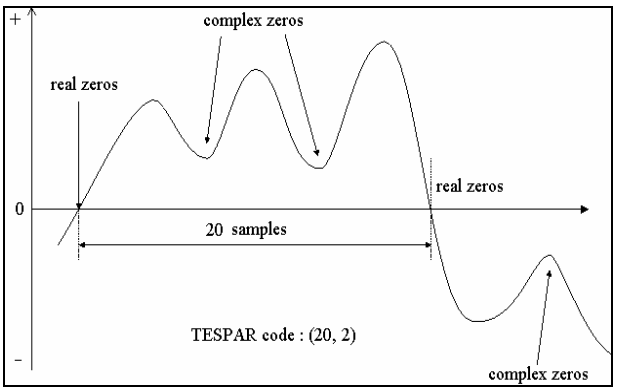

Fig.1. TESPAR waveform analysis

The key to overcome this drawback is to introduce an approximation in the complex zeros location.
Instead of detecting all zeros of the function the following procedure may be used:

- the waveform is segmented between successive real zeros;
- this duration information is combined with simple approximations of the wave shape between these two locations.

Only some of the complex zeros that can be identified directly from the waveform by these approximations detect. In this transformation of signals, from time-domain in the zero-domain, the real zeros, in the time-domain, are identical to the locations of the real zeros in the zero-domain, and the complex zeros occur in conjugate pairs associated with features (minima, maxima, points of inflexion etc.) in the wave shape that appear between the real zeros. In this way an important subset of complex zeros may be identified by examining the features of the wave shape between its successive real zeros.

In the simplest implementation of the TESPAR method [9], two descriptors are associated with every segment or *epoch* of the waveform, in order to generate the TESPAR symbol alphabet :

- the duration between successive real zeros (in number of samples);
- the shape between two successive real zeros.

In this simple TESPAR model implementation, not all complex zeros can be identified from the wave shape, so the approximation is limited to those zeros that can be so identified.

The band-limitation of the signal imposes significant restrictions upon the maximum and minimum duration of any epoch, and also upon the maximum number of significant waveform extreme points that each epoch may contain. The longest epoch may have a duration approximately equal to half the period of the lowest frequency component allowed by band-limiting; the shortest epoch may have a duration approximately equal to half the period of the highest frequency component allowed within the band of signal. Also, short epochs have no or few features, whilst long epochs may contain few or many features. For the simplest implementation, each epoch may be classified in terms of its *duration (D) - number of samples* and the *number of minima (S),* that it contains.

The TESPAR coding process is presented in Fig.2, using an alphabet (symbol table) to map the duration/shape (*D/S*) attributes of each epoch to a single descriptor or symbol.

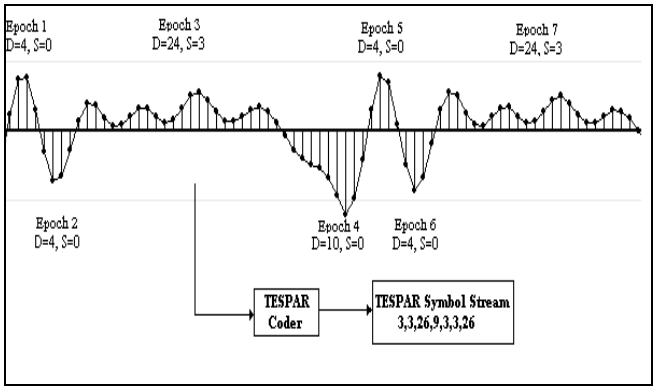

Fig.2. TESPAR coding process

The TESPAR symbols string may be converted into a variety of fixed-dimension matrices [9]. For example, the S-matrix is a single dimension *1xN* (N - number of symbols of the alphabet) vector, which contains the histogram of symbols that appear in the data stream. Another option is the A-matrix, which is

a two dimensional *NxN* matrix that contains the number of times each pair of symbols appears with a possible lag of *n* symbols. The matrices obtained in the training phase are compared to that obtained in the testing phase allowing tasks like verification or recognition. The TESPAR alphabet may be generated in vector quantization process or using neural networks.

3. Decision Threshold

The main problem in applying this source coding approach to speaker verification consists in formulating a criterion for rejecting the speaker. To decide whether to reject a speaker or not, for a particular utterance, a threshold is associated to each speaker codebook. An unknown speaker is rejected if its distortion exceeds the threshold.

One way to compute the threshold for a given speaker is to estimate the parameters for two Gaussians distributions: the in-class distribution of the distortion obtained by encoding utterances from that speaker in his codebook and an out-of-class distribution of the distortion obtained by encoding utterances spoken by other speakers [4][5]. Equalising the overlapping areas of the two distributions, thus equalising the expected numbers of false acceptances and false rejections, chooses the threshold. The threshold computation involves the following steps:

- compute the mean distortion μ_i^{in} resulted from encoding the training set of the speaker "i" in his codebook and the corresponding standard deviation σ_i^{in} ;
- compute μ_i^{out}, the mean distortion obtained by encoding utterances not spoken by the speaker "i", using the "i" speaker's codebook and the corresponding standard deviation σ_i^{out} ;
- to equalize the numbers of false rejection and false acceptances, the threshold T_i, is chosen to be at an equal number of standard deviations away of each mean (Fig. 3):

$$T_i = \frac{\mu_i^{in}\sigma_i^{out} + \mu_i^{out}\sigma_i^{in}}{\sigma_i^{in} + \sigma_i^{out}} \quad (1)$$

This method for threshold computation assumes Gaussians distributions. As distortion metric, the Euclidean distance was employed.

Another way to compute the decision threshold for each speaker enrolled in the experiment implies the clustering of the reference and test utterances in the parameters space (Fig. 4).

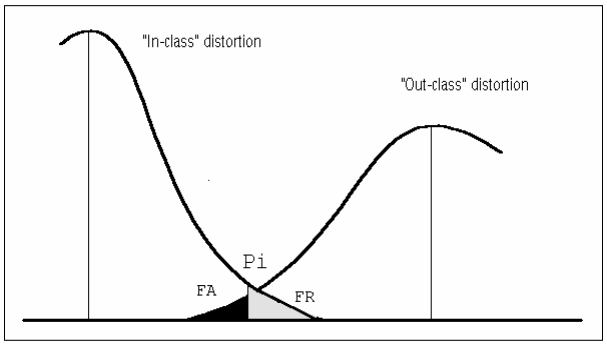

Fig.3. Threshold computing based on mean distortions and standard deviations

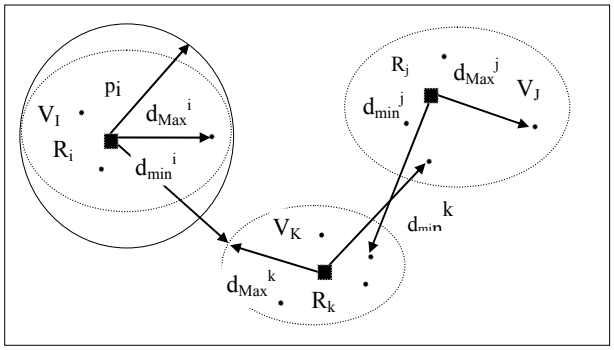

Fig.4. The decision threshold estimation based on averaged distances

First, the maximum distance between the optimal reference and the other references for the "i" speaker is computed:

$$d_{MAX}{}^i = d(R_i, r_{im}), \ m=1,...,M, \quad (2)$$

where *d* is Euclidean distance and M the number of references for each speaker.
Second, the minimum distance between the optimal reference of speaker "i" and the other references of the speakers different from "i" is also computed:

$$d_{min}{}^i = d(R_i, r_{km}), \ m=1,...M \ ; \ k=1,...K, \quad (3)$$

where K is the number of the enrolled speakers.
Finally, the threshold p_i, for the speaker "i", is computed as the average of the two distances previously defined:

$$p_i = \frac{d_{Max}^i + d_{min}^i}{2} \quad (4)$$

4. Combining VQ and DTW

In DTW approach, both reference and test patterns are represented as sequences of spectral parameters and then are time-aligned resulting a score. This score and the decision threshold are used for verification decision.

In VQ approach, reference patterns are represented as sequences of codes but test patterns are represented as sequences of spectral parameters. For each test utterance, an average quantization distortion is evaluated over reference patterns, which is then used together with decision threshold in verification decision process.

If both reference and test patterns are represented as sequences of VQ codes sequences instead of spectral parameters sequences, the amount of computation and storage can be greatly reduced [2]. Then, DTW may be used in order to obtain a score between reference and test VQ codes, score that will be used to take the verification decision.
This approach presents some advantages:

- it is possible to use different number of centroids for training and for testing;
- if enough training utterances are available, one can use a small number of centroids because we expect to obtain reliable sequences of codes; if the number of speakers is high the data storage is greatly decreased; then a higher number of centroids can be used in the test stage to increase verification accuracy;
- if a small number of training utterances an available, is better to use a higher number of centroids for training and is possible to use a smaller number of centroids for testing.

Because one cannot predict the order in which the centroids will appear for test utterance, the direct time-alignment with DTW lead to fluctuant results. To overcome this situation we rearranged the sequences of centroids before applying DTW: we put on the first position in test sequence that test centroid which is closest to first centroid in the reference sequence and so on.

Both types of thresholds (see eq. 1 and eq. 4) can be used together with the final score for speaker verification decision.

5. Combining TESPAR and DTW

The key problem in TESPAR is to define the alphabet used for the coding process, alphabet usually generated by a quantization process. This process is influenced by the maximum number for shape and duration that strongly depend on sampling rate. To avoid this complicated process, in our approach we used DTW (Dynamic Time Warping) to align sequences of epochs in order to generate a verification decision. We used TESPAR to generate speech features (sequences of epochs, each with specific shape and duration) for each speaker utterance, in training and testing stage. DTW is used for successive alignments to compute a speaker model in training stage and in testing stage in order to evaluate a distance between the speaker's models and the test utterance model.

Our speaker verification system works as follows [10]:

- all training utterances are processes with TESPAR method and the resulting epochs (D/S) are saved; before TESPAR coding an endpoint detection process is applied to each utterance;
- DTW is used to build a model for each speaker by successive alignments between epochs files corresponding to each training utterance; each model consist of a sequence of epochs;
- speaker's models and training epochs files are used to compute a threshold for each speaker which will be used in testing stage;
- each test utterance is processed with TESPAR method;
- the resulting epochs string and speaker models are time-aligned with DTW and a distance is computed for each speaker;
- this distance is compared with each speaker threshold and a verification decision is emerged.

We used a weighted Euclidean distance in DTW alignment to discriminate between the contribution of shape and duration.

6. Results

Two particular Romanian utterances, *"Lămâia ia anemia"* (U_1) and *"Aoleu lâna are molii"* (U_2) were used for speaker verification experiments. These test phrases were chosen because they are composed mainly by voiced sounds, which exhibit important energy. The experiments involved 25 speakers (15 males and 10 females) and 5 utterances for each test phrase were collected from each speaker, 3 utterances were used for training and 2 utterances for testing.

We used Error Recognition Rate (ERR) as verification criteria:

$$\mathrm{ERR} = \sqrt{\mathrm{FAR} \cdot \mathrm{FRR}}\ , \tag{5}$$

where FAR is the False Acceptance Rate and FRR is the False Rejection Rate.
In case of combining DTW and VQ we used LPC derived cepstral parameters (LPCC) and MFCC coefficients as features.
We studied the variation of ERR with:

- the number of reference centroids N_R(16, 32, 64, 128);
- the number of test centroids N_T (8, 16, 32, 64, 128);
- the type of decision threshold (Th_1-eq.1, Th_2-eq.4).

The experiments were carried out for each utterance (U_1, U_2), using 16 coefficients for LPCC and 13 coefficients for MFCC [3].
The speaker verification results are presented in the following figures (Fig.5,…, Fig.8).

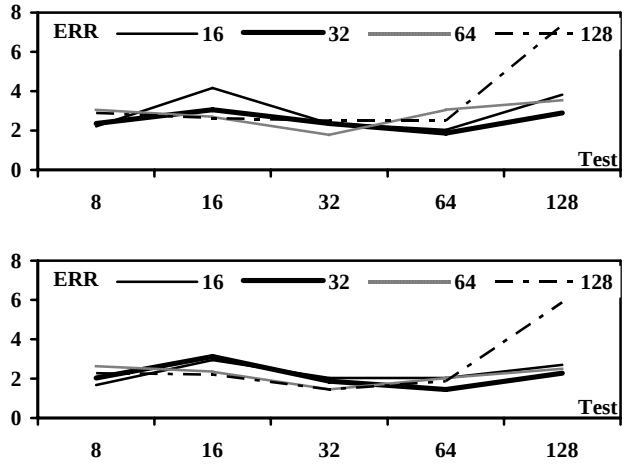

Fig.5. ERR for LPCC, U_1, Th_1/ Th_2

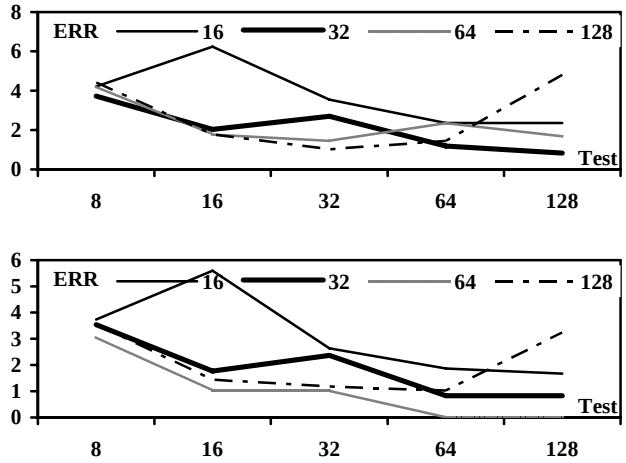

Fig.6. ERR for LPCC, U_2, Th_1/ Th_2

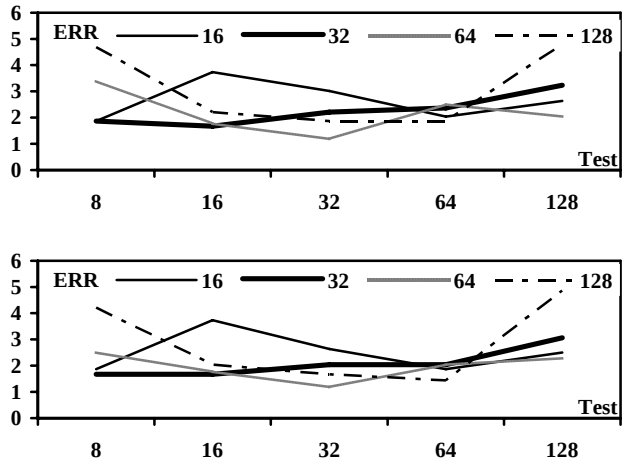

Fig.7. ERR for MFCC, U_1, Th_1/ Th_2

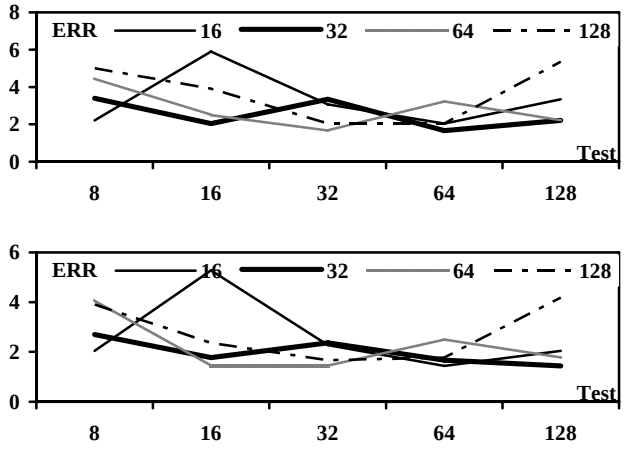

Fig.8. ERR for MFCC, U_2, Th_1/ Th_2

Certain conclusions can be outline from these experiments:
- the best results are obtained for $N_R = 32$ and $N_R = 64$;
- verification performances are better if $N_T = 2*N_R$ or $N_T = N_R/2$ than other values for N_T;
- if $N_T = N_R$ the verification performances go down;
- the Th_2 decision threshold lead to better results than Th_1;
- if $N_R = 16$, the verification performances are still good enough for $N_T >= 32$.

The speaker verification results for DTW-TESPAR combined method are presented in the following table.

Table 1

	Th_1	Th_2
ERR[%]	11.4	9.7

Certain conclusions can be outline from these experiments:
- the Th_2 decision threshold lead to better results than Th_1;
- the verification performances are worse than other methods (where ERR<=2%);
- data reduction is very high, because a set of speech sample corresponding two successive zero-crossing is replace by two TESPAR values, S(shape) and D(duration);
- all calculations are made in time domain.

But, if we added a new feature, an average energy for each epoch, we obtained the following results:

Table 2

	Th_1	Th_2
ERR[%]	6.9	5.1

This new feature, the epoch energy averaged by the number of samples from epoch, add more speaker

dependent information which lead to better results, more closed to those obtained with traditionally methods.

7. Conclusions

In this paper, we presented two approaches to speaker text dependent verification using a combined VQ-DTW method and a combined TESPAR-DTW method

In first approach, we used a combined VQ-DTW method in which VQ is used in the training stage to generate a model for each speaker from training utterances as well as in the test stage to compute a model for the test utterance. Before applying DTW for final distance, the test sequences of centroids is rearranged such as in, the first position is put that test centroid which is closest to first centroid in the reference sequence and so on. The verification experimental results shows very good performances for N_R=32 or 64, and for $N_T = 2*N_R$ or $N_T=N_R/2$. The decision threshold based on averaged distances generates better results than threshold based on mean distortions and standard deviations. In case there are many training utterances available, a small number of reference centroids (i.e. 32) and a medium number of test centroids (i.e. 64) lead to very good verification performances.

In second approach, we used a combined TESPAR-DTW method in which TESPAR coding is used to generate speech features and DTW is used to generate a model for each speaker from training utterances as well as in the test stage to compute a distance between the test utterance and speaker's models. The verification experimental results show medium performances. The decision threshold based on averaged distances generates better results than threshold based on mean distortions and standard deviations. This approach seems to be promising because all calculations are made in time domain and data reduction is about 15-20 times.

References:

[1] BURTON, D.K., *Text dependent speaker verification using vector quantization source coding,* IEEE Transactions on Acoustics, Speech and Signal Processing, Vol ASSP-35 Feb. 1987.

[2] FURUI, S., *Vector-Quantization-Based Speech Recognition and Speaker Recognition Techniques*, IEEE 1058-6393/91.

[3] PICONE, J.W., *Signal modeling techniques in speech recognition,* Proc. IEEE, Vol. 81, Sept. 1993.

[4] RABINER, L.R., JUANG, B.H., *Fundamentals of speech recognition,* Prentice-Hall International , Inc., 1993.

[5] FURUI, S., *Digital Speech Processing, Synthesis and Recognition,* Marcel Dekker Publications, 1989, 2001.

[6] HUANG, X., ACERO, A., HON, H., *Spoken Language Processing*, Prentice Hall, 2001.

[7] R. A. KING , T. C. PHIPPS, *Shannon, TESPAR And Approximation Strategies*, ICSPAT 98, Vol. 2, pp. 1204-1212, Toronto, Canada, September 1998.

[8] A. A. G. REQUICHA, *The zeros of entire functions, theory and engineering applications,* Proceedings of the IEEE, Vol. 68, no. 3, pp. 308-328, March 1980.

[9] J. HOLBECHE, R. D. HUGHES AND R. A. KING, *Time Encoded Speech (TES) descriptors as a symbol feature set for voice recognition systems*, IEE Int. Conf. Speech Input/Output; Techniques and Applications, pp. 310-315, March 1996.

[10] POP, G.P., LUPU, E., *Speaker Verification Employing TESPAR Coding and DTW*, Proceedings of the Symposium on Electronics and Telecommunications "Etc. 2004", Timisoara, Oct. 22-23, 2004, Buletinul ştiinţific al Universităţii POLITEHNICA Timişoara, Tomul 49(63), Fascicula 1, pp. 275-278.