COMBINING VQ AND DTW IN TEXT DEPENDENT SPEAKER VERIFICATION

Petre G. Pop
Electronics and Telecommunications Faculty,
Communications DEPT.,
Cluj-Napoca Technical University
26-28, str. BARIŢIU, 3400 CLUJ-NAPOCA
ROMANIA ,
Petre.Pop@com.utcluj.ro

Eugen Lupu
Electronics and Telecommunications Faculty,
Communications DEPT.,
Cluj-Napoca Technical University
26-28, str. BARIŢIU, 3400 CLUJ-NAPOCA
ROMANIA ,
Eugen.Lupu@com.utcluj.ro

Abstract – **This paper presents an approach of text dependent speaker verification using both VQ and DTW methods. We used VQ in the training stage, to generate a model for each speaker based on training utterances. In the test stage, we use also VQ to compute a model for the test utterance then we use DTW algorithm to evaluate a distance between the speaker's models and the test utterance model. Before applying DTW, the sequences involved must be rearranged in order to generate better results.**

Keywords: speaker verification, VQ, DTW, LPCC, MFCC.

I. INTRODUCTION

The speaker verification consists of automatically authenticating the identity claimed by a speaker, given only some samples of his voice. There are two 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.*

Vector quantization (VQ) is a data compression technique, with several successful applications in speech and image coding or speech/speaker recognition [3]. 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. A typical approach to the text dependent speaker verification is DTW (Dynamic Time Warping) 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.

In addition to the template matching method described above, statistical methods are also used. These methods require large amounts of training data to estimate the probability densities of the parameters chosen to represent the speaker.

II. SPEECH PROCESSING

The speaker specific information can be found within the short time spectrum of a speech segment. First, the speech signal delivered by the microphone is amplified and passed through an antialiasing filter (0-5kHz). Then, the speech signal is proccesed by:

- *preemphasis:* the speech samples are filtered by a pass-high digital filter in order to offset the natural slope (attenuation of 20dB/decade) due to physiological characteristics of the speech production system, thereby improving the efficiency of analysis;
- *end-point detection:* intended to remove silence and noise zones from the speech signal;
- *frame blocking*: the speech signal is blocked into frames of N samples, with a frame-overlap factor of frames M (M<= N); (a 50% value was used for the overlap factor);
- *windowing:* favors the samples towards the center of the window; this fact coupled with the overlapping analysis performs an important function in obtaining smoothly varying parametric estimates; the Hamming window is the typical window in speech processing.

Finally, the feature vectors can be derived. In case of LPC derived cepstrum, the following recursive relations were used [4]:

$$\begin{aligned} c_1 &= -a_1 \\ c_m &= -a_m - \sum_{k=1}^{m-1}(1-\frac{k}{m})c_{m-k}a_k \quad ,1 < m \le p \\ c_m &= -\sum_{k=1}^{m-1}(1-\frac{k}{m})c_{m-k}a_k \quad ,m > p \end{aligned} \tag{1}$$

where a_i denote the LPC coefficients.

The computation of MFCC implies the following steps [7]:

- first, the short-term spectrum of the vocal segment is evaluated;
- then, this spectrum is integrated over gradually widening frequency intervals on the Mel scale;
- the resulting Mel-warped spectrum is projected on a cosine basis to generate the Mel frequency cepstral coefficients (MFCC).

III. DYNAMIC TIME WARPING

DTW is a recursive recognition algorithm, which is usually used to evaluate a distance between a previously stored reference set, and a test set of speech parameters. The main advantage of this method consists of temporal alignment of the two compared sets of data. In fact, DTW seeks a way between reference and test data so that the cumulated distance to be minimal (fig. 1).

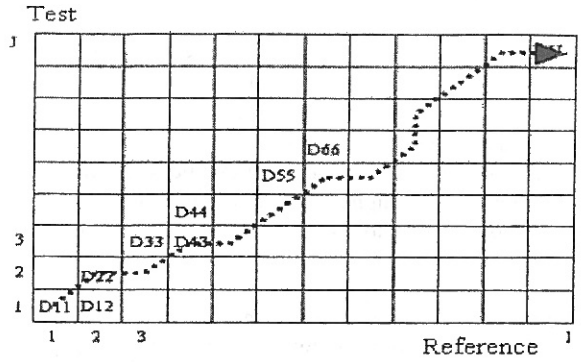

Fig.1. The optimal way in DTW algorithm

The cumulated distance between the two utterances A and B is given by [6]:

$$D(F) = D(A,B) = \frac{g(I,J)}{(I+J)}, \tag{2}$$

where:

$$g(i,j) = \min \begin{pmatrix} g(i,j-1) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i-1,j) + d(i,j) \end{pmatrix}, \tag{3}$$

and:

$$g(1,1) = 2 \cdot d(1,1) \tag{4}$$

In previous equations $d(i,j)$ is the Euclidean distance between two parameters vectors of the reference and test utterances.

IV. VECTOR QUANTIZATION

In speaker verification we represent each speaker by a VQ codebook designed from a training sequence composed of repetitions of a particular utterance [1]. The same utterance is used later by an unknown speaker, which claims that identity. This test utterance is coded using the speaker codebook, and the resulting quantization distortion is compared to a threshold. If the distortion lies below the threshold, the speaker is accepted.

Each speaker is represented by a time series of parameters (reference template) extracted from a particular utterance. The parameters are chosen to reflect speaker specific organic differences in the structure of vocal apparatus or to reflect specific learned differences in the use of vocal apparatus. The main parameters employed by this technique are: the pitch, the short time energy, the short time spectrum and LPC coefficients or parameters derived from these coefficients (such as MFCC – Mel Frequency Cepstral Coefficients).

The algorithm used for vector parameters quantization is the LBG (Linde-Buzo-Gray) algorithm [7]. A codebook may be small in the beginning and may be gradually expanded to the final size. One method is to split an existing cluster into two smaller clusters and assign a codebook entry to each. The following steps describe this method of designing the codebook:

- create an initial cluster consisting of the entire training set; this initial codebook contains a single centroid for the entire set;
- split this cluster into two subclusters, getting a codebook of twice the size;
- repeat this cluster-splitting process until the codebook reaches the desired size; ideally, each cluster should be divided by a hyperplan normal to the direction of maximum distortion [6].

V. 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]. 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. 2).

$$T_i = \frac{\mu_i^{in}\sigma_i^{out} + \mu_i^{out}\sigma_i^{in}}{\sigma_i^{in} + \sigma_i^{out}} \tag{5}$$

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

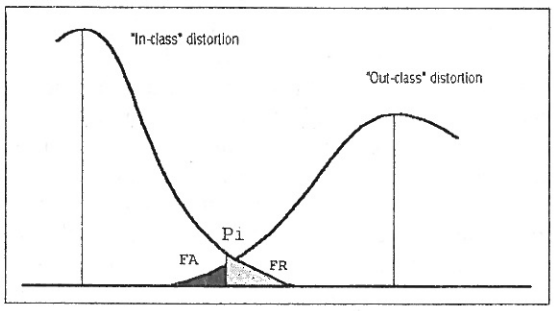

Fig.2. Threshold computing based on mean distortions and standard deviations

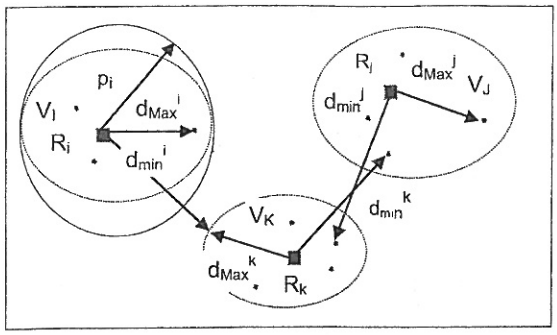

Fig.3. The decision threshold estimation based on averaged distances

Another way to compute the decision threshold for each speaker enrolled in the experiment was proposed in [5] and implies the clustering of the reference and test utterances in the parameters space (fig. 3).

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

$$d_{MAX}{}^{i} = d(R_i, r_{im}), \quad m=1,...,M, \tag{6}$$

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:

$$d_{min}{}^{i} = d(R_i, r_{km}), \quad m=1,...M; \; k=1,...K, \tag{7}$$

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 [5]:

$$p_i = \frac{d^i_{Max} + d^i_{min}}{2} \tag{8}$$

VI. 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 utterrances 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. 5 and eq. 8) can be used together with the final score for speaker verification decision.

VII. EXPERIMENTAL RESULTS

Two particular Romanian utterances, *"Lămâia ia anemia"* (U1) and *"Aoleu lâna are molii"* (U2) 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. As features, we used LPC derived cepstral parameters (LPCC) and MFCC coefficients.

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

$$ERR = \sqrt{FAR \cdot FRR} \tag{9}$$

where FAR is the False Acceptance Rate and FRR is the False Rejection Rate.

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 (Th1-eq.5, Th2-eq.8).

The experiments were carried out for each utterance (U1, U2), using 16 coefficients for LPCC and 13 coefficients for MFCC.

The speaker verification results are presented in the following figures (fig.4,..., fig.11). 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 Th2 decision threshold lead to better results than Th1;
- if $N_R = 16$, the verification performances are still good enough for $N_T >= 32$.

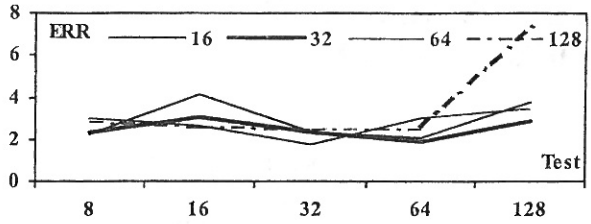

Fig.4. ERR for LPCC, U1, Th1

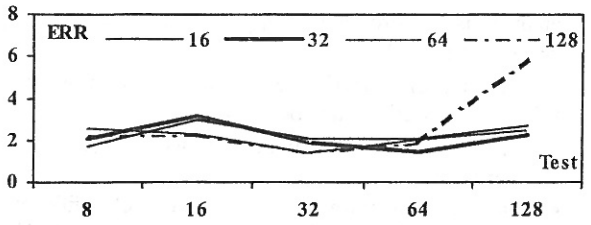

Fig.5. ERR for LPCC, U1, Th2

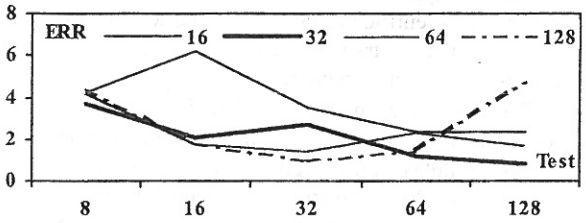

Fig.6. ERR for LPCC, U2, Th1

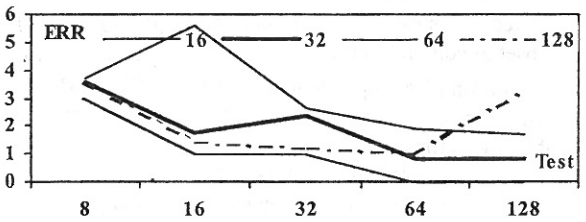

Fig.7. ERR for LPCC, U2, Th2

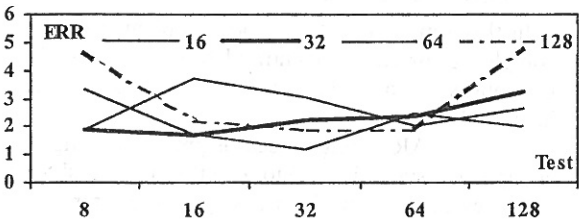

Fig.8. ERR for MFCC, U1, Th1

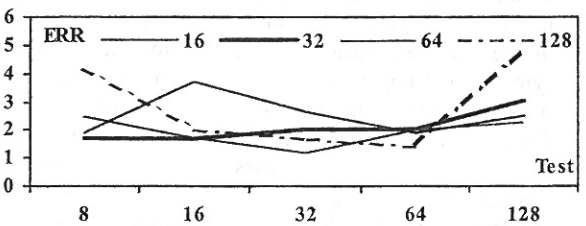

Fig.9. ERR for MFCC, U1, Th2

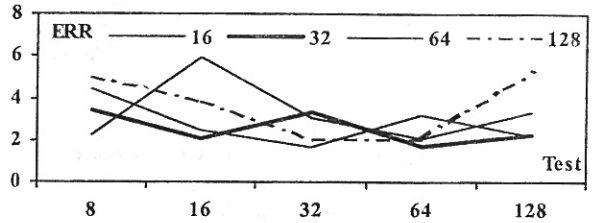

Fig.10. ERR for MFCC, U2, Th1

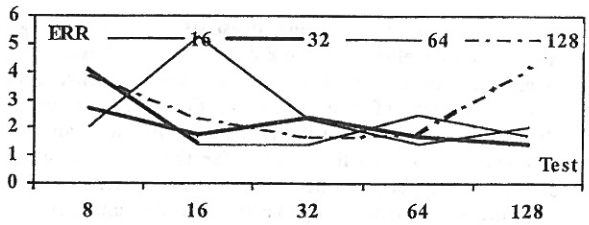

Fig.11. ERR for MFCC, U2, Th2

VIII.CONCLUSIONS

In this paper, we presented an approach to speaker text dependent verification using 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.

IX. 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] Lupu E., Pop G.P., Toderean G., "Speaker Verification Using Vector Quantization", *Proceedings of the First Workshop on Text, Speech, Dialog (TDS '98) Brno, Czech Republic 23-26 sept. 1998, p. 275-380.*

[6] Furui, S. *Digital Speech Processing, syntesis and recognition,* Marcel Dekker Publications, 1989, 2001.

[7] Huang, X., Acero, A., Hon, H., *Spoken Language Processing*, Prentice Hall, 2001.