MULTISTAGE VECTOR QUANTIZATION WITH ACOUSTIC CONSTRAINTS FOR SPEAKER VERIFICATION

P.Pierucci, A.Paladin
Alcatel-Face Research Centre
Pomezia, Italy

ABSTRACT - In this paper a new method to build multisection codebooks for the speaker recognition task is investigated. Different methods of threshold evaluation are then discussed for the proposed approach, and a comparison with single section VQ and previously reported Multisection VQ is discussed, in a fixed text speaker verification experiment.

INTRODUCTION

The Vector Quantization is a data compression principle which has been recently applied to the speaker recognition task(Buck,1985). In this application good results have been obtained, either in a text-dependent or text-independent mode, although other techniques based on fixed-text approaches and temporal alignement did offer better performances (Naik,1987). In the Vector Quantization coding method, after dividing the speech into frames and performing a parametric analysis (LPC cepstral coefficients in this paper), the parameter space generated from a speaker is quantized into several partitions using a clustering algorithm[3]. For each partition a representative vector (centroid) is chosen, according to some distance criteria. The set of centroid vectors form the codebook. In speaker verification schemes, a codebook is generated for each speaker. The recognition of a speaker identity is carried out by computing the average quantization distortion of the parametric vector sequence of the unknown input phrase over the reference codebook, and comparing it with a threshold. No temporal information is contained in the average quantization distortion. A generalization of the VQ approach, in order to include the temporal information, is the Multisection Vector Quantization technique, in which a temporal ordered sequence of VQr are used. The speech model is composed of a number of sub-sources (sections), each described with a single section VQ (section codebook). The general method is as follows : each word is divided in a fixed number of equal length partitions; several different procedures can be adopted to do this, here we mention the one proposed in (Burton,1985). The speech signal is divided in equal length frames and parametric analysis is performed; the number of partitions is fixed to m and the length of each section is set to n=F/m, where F = number of frames of the input phrase. For each partition a VQ quantizer is designed with several repetitions of the same training utterance. During classification the same segmentation procedure is applied to the unknown

utterance, and the speech data corresponding to a partition is encoded with the corresponding codebook, obtaining a set of average distortion values, one for each partition. The classification decision can be based, for example, on the minimum average distortion value, but many other strategies can be adopted. The major problem in the MSVQ is the segmentation procedure, which does not take into account variations in the speech rate of the utterance; moreover the number and length of each section must be adjusted according to the input utterance length.

THE PROPOSED APPROACH

Acoustic constraints for Multisection Codebook

In this paper a new method to build multisection codebooks for the speaker recognition task is investigated. The method builds each codebook section using speech data corresponding to transitions between two steady-state spectral configurations, in order to characterize, with each section, acoustical correlates of articulatory movements. The number of sections in the Multisection codebook is not fixed, but is tied to the number of steady-state spectral configurations that can be detected in the fixed input utterance. In this way each section will represent the speech signal transition between two spectral steady-state configurations, which has been shown to exhibit speaker dependent characteristics, at least for certain classes of sounds (Sambur,1975). Futhermore the approach does not require any temporal normalization. The method is as follows : the speech signal corresponding to an utterance is broken into equal length frames and represented as a sequence of parametric vectors, using LPC-Cepstral analysis; the utterance is divided in sections using a spectral distance measure between adjacent frames; the chosen distance is based on the log-area ratio coefficients, because it allows to compare speech frames spectra keeping apart the contribution of energy and pitch(Barnwell,1982). The expression of the distance measure is :

$$D(l,m)= \sum_{i=1}^{12} \left[\left(Ln \frac{1-K(i,l)}{1+K(i,l)} \right) - \left(Ln \frac{1-K(i,m)}{1+K(i,m)} \right) \right]^2$$

K(i,l/m)= i-th LPC coefficient, frame l/m

The section borders are located on the minima of the log-area ratio distance measure; a supervision stage is provided so far to have a better estimate of the performances of the method. Given the phonetic content in the selected vocabulary, as described in next sections of the paper, we obtained a 3-section decomposition for most of the words.

For each section we designed a single section VQ using several repetitions of each word of the vocabulary, from the users of the system. During the classification phase, the

same length-normalization procedure is applied to the unknown utterance, and speech data corresponding to a section is encoded with the corresponding section codebook, obtaining a number of quantization distortions equal to the number of sections for the word. The classification decision is based on the set of average quantization distortion of the different sections, and a nuber of different strategies will be discussed later.

Threshold Evaluation

The classification decision for the speaker verification task, when usin VQ techniques, is based on a match of the average quantization distortion of the unknown utterance, over the reference codebook, with a pre-defined threshold value. The evaluation of the optimum threshold value is based, in general, on the estimate of two quantization distortions distributions : the in-class distribution, obtained encoding the speech signal of the user with his own codebook, and the out-class distribution, obtained encoding speakers other than the user with the user's codebook. If an equal error rate is requested for the speaker verification system, the threshold must be evaluated in order to equalize the overlap area of the two distribution : in this case the expected number of imposter acceptances (false acceptances) would equal the expected number of rejections of admissible speakers (flase rejections). Usually gaussian in-class and out-class distributions are assumed, and the equal error rate criteria gives, for the threshold, the value :

$$T = \frac{\mu^{in}\,\sigma^{out} + \mu^{out}\,\sigma^{in}}{\sigma^{in} + \sigma^{out}}$$

$\mu^{in/out}$: average in/out-clas quantiz. distortion

$\sigma^{in/out}$: in/out-class quanti distortion variance

The quantization distortion for the multisection method is:

$$D = \sum_{i=1}^{nsect} w(i)*d(i);$$

w(i) = i-th section weight

d(i) = i-th section quantiz. distortion

nsect= number of sections

For each multisection codebook we estimate the in-class and out-class quantization distortion distributions of D, and the threshold evaluation can be made using still the gaussian assumption for the two distributions.

SPEECH DATABASE

A dedicated speech database, with a population of 5 users and 15 impostors, is used to score the system. The vocabulary contains 10 different words (italian language) : vaglio, metto, monto, la, scocca, penna, rampa, bianca, buia, vuota. The speech signal has been recorded in a quiet office environment over a time interval of 45 days for the users of

the verification system. 50 repetitions of each word of the vocabulary have been recorded from the users of the verification system, and each impostor recorded 10 repetitions. The analysis conditions are : LPC-Cepstral analysis on a 16 ms frame with 16 ms displacement; pre-emphasis=0.95 and filter order=12. Each utterance is extracted from silence using energy criteria.

RESULTS

The preliminary results we present here are relative to a subset of the speech material : the three words vaglio, penna, vuota have been used. The results are relative to 5 different users and 15 different impostors (in each verification experiment of one user the other 4 are used as impostors). For each codebook under test, we had 10 attempts/user and 5 attempts/impostor for a total impostor attempts of 95. Two different distance criteria have been used : 1)Euclidean distance ; 2)Weighted Euclidean distance, with weights obtained from the training material.
The first set of experiments are relative to single section VQ with rate 4 and 5. The second set of experiments are relative to Length Normalized Multisection VQ(Burton,1985), 3 sections, rate 3. The third set are relative to the proposed method, with 3 sections and rate 3. For the Multisection VQ experiments, we first investigated the evaluation of the weights w(i) in (3); we checked the following hypothesis :

$$w(i)=1 \ , \ i=1,3 \ ; \qquad w(i) = \mu_i^{out} - \mu_i^{in} \ ; \qquad w(i) = 1 / \mu_i^{in} \ .$$

Although preliminar experiments showed us that different sections within the word had different discrimination potentiality in terms of overall verification rate, the best result for the multisection experiment was obtained using the uniform weighting. In the table the results concerning the two multisection approaches are presented. The length normalized MSVQ of [4] is indicated as **MSVQ**; the MSVQ with acoustic constraints is indicated as **MSVQ1**. Each row indicates the results for each speaker used as reference in the verification experiments. The results for each speaker are expressed in terms of false acceptance errors (FA)(i.e. an impostor has been accepted as the user) and false rejection errors (FR)(i.e. a user has not been authenticated as the user). The two final rows give the overall performance measure. In the figures 1,2,3 we present the comparison of the two MSVQ approaches with the single stage VQ, with rate 4 and 5, referred as **VQ16** and **VQ32**.
In general we can see that no better performances have been obtained using MSVQ1 in comparison with VQ32, although the results of MSVQ1 are more biased toward the false rejection errors, usually considered as the "best" of the two errors. We think that this is caused by the use of very short utterances in the test experiments, and we expect that in the next experiments, where we will use phrases composed by

sequences of three words of the selected vocabulary, better performances of MSVQ1 will be observed. We can observe a slight better overall performance of MSVQ1 against MSVQ, and a similar behaviour of the performances of the two multisection approaches is evident when compared with the single section VQ; in particular the speaker A is "in trouble" with MSVQ and MSVQ1. Finally the use of Weighted Euclidean Distance Measure seems to be very appropriate for the MSVQ1 method, where it contributes to a strong decrease of the FR errors.

CONCLUSIONS

In this paper we evaluated the performances of a new method of Multisection Vector Quantization based on Acoustic Constraints for the Speaker Verificatin task, and compared it with the performances of previously reported multisection and single section approaches. The results showed a slight improvement of performances that we think would be increased if longer utterances are used as test material. This will be verified in a future work, where we will address the problem of making the segmentation procedure, based on the location of steady-state spectral configuration, completely automatic.

REFERENCES

Barnwell,T.P and Quackenbush,S.R,(1982),"An Analysis of Objectively Computable Measures for Speech Quality Testing", ICASSP 82

Buck,J.T.,Burton,D.K.,Shore,J.E.,(1985),Speaker Recognition Using Vector Quantization",IEEE Int.Conf.on Acoust.Speech, Signal Processing (ICASSP 1985), pp 11.5.1-11.5.4

Burton,D.K.,Shore,J.E.,Buck,J.T.,(1985),"Isolated-Word Speech Recognition Using Multisection Vector Quantization Codebooks", IEEE Trans.Acoust.,Speech,Signal Processing,ASSP-33,n.4,August 1985, pp.837-849

Linde,Y.,Buzo,A.,Gray,R.M.,(1980),"An Algorithm for Vector Quantizer Design",IEEE Transactions on Communication, COM-28, 1980,pp.84-95

Naik,J.M.,Doddington,G.R.,(1987),"Evaluation of a high performance Speaker Verification System for Access Control", IEEE Int.Conf. Acoust.,Speech,Signal Processing (ICASSP 1987), pp. 34.15.1-34.15.4

Sambur,M.R.,(1975),"Selection of Acoustic Features for Speaker Identification",IEEE Trans.Acoust.,Speech,Signal Processing, ASSP-23,n.2, April 1975, pp.176-182

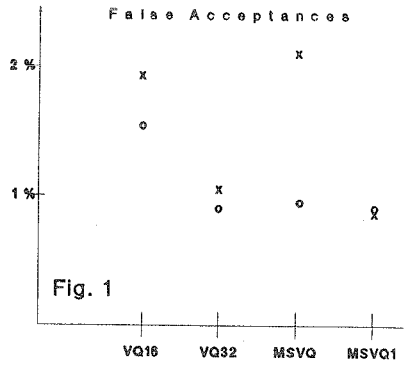

Fig. 1

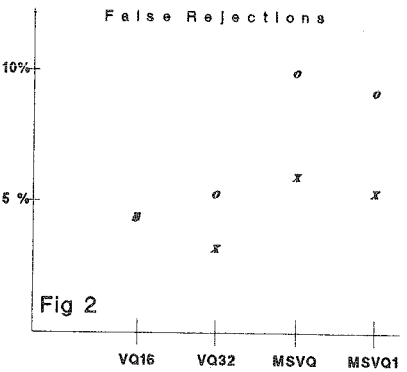

Fig 2

o : euclidean distance x : weigthed euclidean dist.

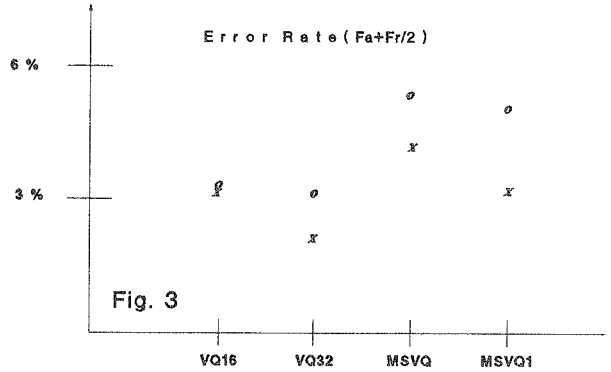

Fig. 3

	MSQV				MSQV1			
	euclidean		W. euclidean		euclidean		W. euclidean	
	FR	FA	FR	FA	FR	FA	FR	FA
User A	9	4	5	4	9	4	5	2
User B	3	2	1	5	2	3	1	5
User C	2	6	2	9	2	3	1	3
User D	1	1	1	8	1	3	1	3
User E	0	1	0	5	0	0	0	0
Total	15	14	9	31	14	13	8	13
%	10.0	0.98	6.0	2.17	9.33	0.91	5.33	0.91