A VQ BASED SPEAKER RECOGNITION SYSTEM BASED IN HISTOGRAM DISTANCES. TEXT INDEPENDENT AND FOR NOISY ENVIRONEMENTS.

Enric Monte, Ramón Arqué, Xavi Miró.

Dpt.TSC.Universitat Politécnica de Catalunya Barcelona.Spain

E-mail:enric@gps.tsc.upc.es

ABSTRACT

In speaker recognition systems based on VQ, normally each speaker is assigned a codebook, and the classification is done by means of the a distortion distance of the utterance computed by means of each codebook. In [1] we proposed a system which instead of having a codebook for each speaker, had only one codebook for all the speakers, and for each speaker one histogram. This histogram was the occupancy rate of each codeword for a given speaker. This means that the information of the histogram of a given speaker is the probability that the speaker utters the information related to the codeword. So we approximated the pdf of each speaker by the normalized histogram.

In this paper we present an exhaustive study of different measures for comparing histograms: Kullbach-Leiber, log-difference of each probability, geometrical distance, and the Euclidean distance.

We have done also an exhaustive study of the properties of the system for each distance in the presence of noise (white and colored), and for different parameterizations:

LPC, MFCC, LPC-Cepstrum-OSA (One sided autocorrelation sequence), LCP-Cepstrum. (Cepstrum with/without liftering).

As the combination of experiments was high, the conclusions were drawn after an analisis of variance (ANOVA), and T-tests. Thus the conclusions, with significance levels, can be drawn about the differences and interactions between kind of. distance, parametrizacion, kind of noise and level of noise.

1. INTRODUCTION

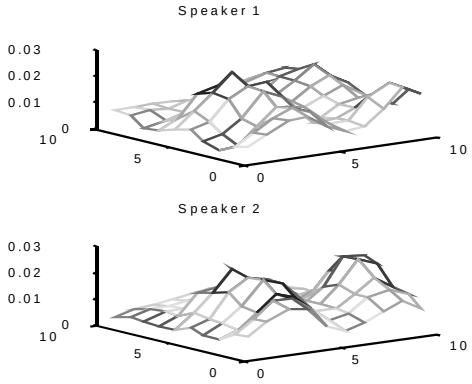
There are several approaches to automatic speaker identification, the main strategies that have been proposed are:

- The use of a codebook per speaker [2]
- The use of probabilistic models [3]
- The use of DTW [4]
- The use of metrics like the arithmetic-harmonic spherity measure [5]

This work has been supported by the TIC95-1022-C05-03

The system that we propose is based on one codebook for all the speakers, and an occupation historgram of each codeword for each speaker [1]. Thus the system that we propose is a non parametric classifier, and no hypothesis are done about the distribution of the parameters. In order to show that the occupancy histogram is an adecuate tool for classifying speakers, in figure 1 we show two histograms that correspond to two different speakes, it can be seen that the occupancy rate is different for each of them.

Figure 1: Occupancy histogram of two different speakers for a



codebook of 10x10.

In this case the codebook consisted of codewords that represented the cepstral coefficients. We used as a quantification tool the self organizing map in order to make use of the fact that neighbouring codewords in the feature space are neighbours in the topological map [6], this explains the fact that the occupancy histogram forms a surface. In figure 2 we also show the contents of the codebook, it can be seen that neighbouring codewords correspond to similar codewords.

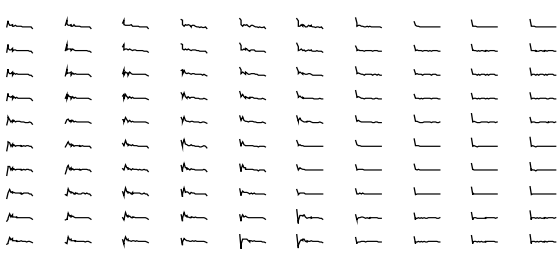


Figure 2: The Contents of a codebook of cepstral coeficients.

The reason for chossing this architecture was to make the system as much independent of the text as possible, and flexible. The advantatges of using only one codebook for all the speakers are:

- There is more material for training the codebook, than in the case of one codebook per speaker.
- The number of parameters of the system is reduced.
- The probabilistic model of each speaker (i.e. histogram) is separeted from the explicit representation of the speech signal.
- New speakers can be incorporated, without changing the codebook, because only a new histogram has to be computed.

2. SYSTEM DESCRIPTION

The classification system consists of a library of histograms (one per speaker), a codebook, a module that computes the histogram of an input that corresponts to the speaker to be classified and a distance module. The distance module compares the histograms in the library with the histogram of the input signal. A summary of the training of the system is shown in figure 3.

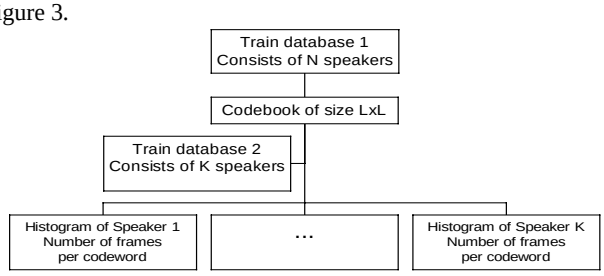


Figure 3: Summary of the training of the system.

Remark that the training of the VQ, does not necessary need to be done with the same database that the one used for computing the histograms, nor the same number of speakers are needed. The only restriccion is that the material used for training should be representative enough. In our system we have used for training the VQ, the same speakers to be recognized, but different utterances that the ones used for computing the histograms, i.e. we used two different dabases for training the system. Once the histograms of each speaker are computed, the recognition phase takes place as it is summarized in figure 4.

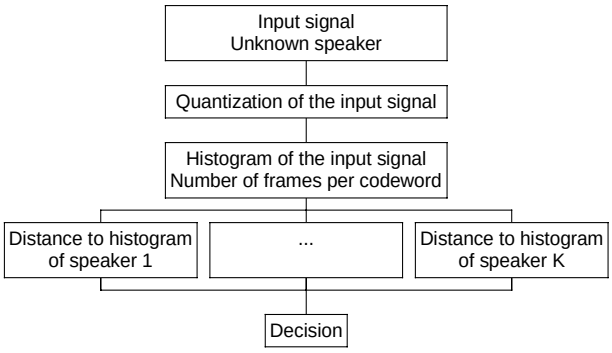


Figure 4: Recognition phase of the system.

The crucial part of the whole recognition system is the distance module. As the objets to be compared we histograms that can be interpreted as aproximations of probability density functions, we decided to use as distance measure the Kulbach-Leiber mesure. Also other measures were used in order to take into account the fact that we are not working with real pdf.

- The selected distances were:

  Kulbach Leiber, Log difference, Geometrical mean between probabilities, and the Euclidean distance between probabilities

Also we evaluated the dependency of the system in relation to the parametrization in the presence of noise. It is well known that different parametrizations yield different results for different levels of noise.

- The parametrizations that were studied were:

  LPC, MFCC, LPC-Cepstrum (with/without liftering), LPC-Cepstrum-OSA (one sided autocorrelation sequence)[7]

It is also known that the influence of the parametrization in recognition results depends not only on the level of noise, but also on the spectrum of the noise. In order to acess this factor we decided to test the system with white noise, and with low pass noise. The reason for chosing low pass noise (cut-off frequency of 500 Hz), is due to the fact that most of the environmental noise has this characteristic (i.e., office noise, car noise).

- Thus the design will take into account:

  The sizes of codebooks and histograms, distances between test histogram and reference histograms, parametrizacions, and influence of the kind of noise.

3. DATA BASE AND PREPROCESSING

The database used was the TIMIT. Although this database was recorded in one session, and in the speaker recognition problem the variability between sessions is important, we used it for the purporses of testing the system. In the future we will test the system and analize its performance with a database specific for speaker recognition.

The window size used for all experiments was of 32 ms., preemfasis of 0.985, a Hamming window, an overlap of 2/3 and the analysis order for the LPC parameters was 24.

The database used consisted of 100 speakers, with 30 files per speaker for the train databse and 30 files per speaker for the test database.

4. TRAINING OF THE CODEBOOK

The codebook consisted of a Self Organizing Feature Map, trainned by means of the Kohonen algorithm. The size of the codebook was fixed to 20x20 after some preliminary experiments. The neighborhood function had an initial radius of

10, and the topological neighborhood was taken hexagonal. The coarse training was done for 500 epochs, and the fine tuning for 100000. The initialization of the codevectors was done with random values in the interval [-1,1].

5. EXPERIMENTAL FRAMEWORK

The performance of the system was evaluated for different combinations of distance meaures, parametrizations, kind of noise and levels of noise. The number of combinations is so high that we decided to use a methology for drawing conclusions. The chosen methology was the analysis of variance ANOVA, and the comparison of means by means of T of student tests. This methodology gives confidence values that allows the posibility of selecting a combination of factors with a certainty about the decision.

6. EXPERIMENTAL RESULTS

The analysis of the system with respect to the factors distance, noise kind, noise level, and parametrization, was done.

6.1. Dependency of the kind of noise and level with the distance measure.

The first test was done in order to acess if there was a distance that behaved differently with respect to a given signal noise ratio and kind of noise. In the following table we present the probability level that all the distances are equal (F-test), when the noise is white and when it is low pass. The variances were computed using the results obtained for five different parametrizations and the three different distances.

	Distance	
SNR (dB)	**white noise**	**low pass noise**
∞	0.152	0.167
30	0.622	0.005
20	0.583	0.02
10	0.339	2e-11

Table 1: Study of the dependency of the distance factor with respect to the kind of noise and level. The numbers presented are the probability that the distance coefficient is zero.

The conclusion that can be drawn from table 1, is that for white noise all the distances differently, i.e. there is a distance that is significantly different than another. Nevertheless, when the noise is low pass the distances behave similarly for all levels of noise.

6.2. Dependency of the kind of noise and level with the parametrization measure.

This test was done in order to acess if there were parametrizations that behaved differently with respect to the kind of noise and the level. As in the other table, the variances were computed using the results obtained for five different parametrizations and the three different distances.

	Parametrization	
SNR (dB)	**white noise**	**low pass noise**
∞	0.013	0.002
30	1e-11	0.042
20	2e-9	0.186
10	8e-9	0.303

Table 2: Study of the dependency of the parametrization factor with respect to the kind of noise and level. The numbers presented are the probability that the parametrization coefficient is zero.

The conclusion that can be drawn from table 2 is that for white noise is that the probability remains low for all SNR, while for low pass noise, the probability increases. This means that the use of robust parametrizations is important in the presence of white noise, while in the low pass noise there is not much difference in the use of one parametrization or another.

6.3. Interaction between parametrization and distance

We will study if there is an interaction between parametrization and distance, that is, if there is a significant dependency between factors. In order to have enough experiments for computing the interaction, two replications were done.

	Interaction Parametrization-Distance	
SNR (dB)	**white noise**	**low pass noise**
∞	0.065	0.484
30	0.021	0.51e-9
20	3.6e-7	3.0e-11
10	0.009	9.3e-8

Table 3: Study of the interaction between the parametrization factor and distance with respect to the kind of noise and level. The numbers presented are the probability that the interaction coefficient is zero.

In the case of low pass noise there is a significant interaction between parametrization and the distance measure in the presence of noise (for snr=∞, most of the recognition results were 100% or near, so in this case the interaction is indiferent). The detailed analysis of the results showed that euclid, Kullbach and log-diff, distances behaved differently depending on the parametrization. The pair which yielded the best results for all the signal noise ratios and low pass noise was the LPC-Cesptrum, with geometrical distance. On the other hand for the experiments with white noise the interaction between distance and parametrization is lower and as in the other case, the interaction increases as the SNR decreases. In this case the pair that yielded the best results was the Kullbach distance with the LPC-Cepstrum-OSA. Which could be expected, due to the fact that the LPCC-OSA, is known to work well in the presence of white noise. The second best pair was the LPC-Cesptrum with geometrical distance.

6.3. Final results of the recognition system

Finally we present the final results of the system for the pairs that had the best performace for the two kinds of noise that were used. The results presented in figure 5 and 6 are the mean of the recognition rate and the confidence margin for the 95% of cases.

We found that for most of the experiments the variance of the results was higher in the case of low pass noise than for the case of white noise, but the recognition results were in general better.

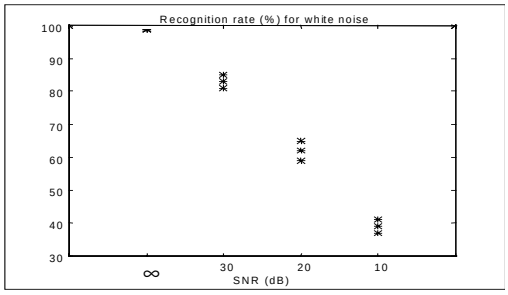


Figure5 Recognition rate for white noise

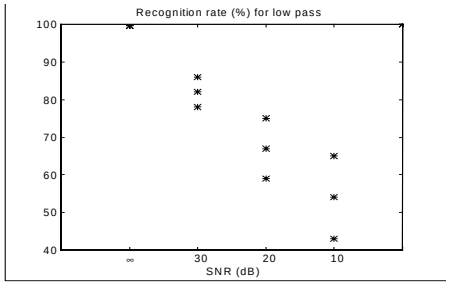


Figure 6 Recogniton rate for low pass noise

In the figures 7 and 8 we compare the results obtained with the best parametrizations with the best distance in each case (ie. Kullbach for white noise, geometrical for low pass noise). The most important conclusion that can be drawn from the figures is that the conficende margins do not overlap, when the difference in the recognition rate is important. That is for the case of high noise Another point that was noticed is that the other parametrizations usually behaved similarly to the second best and the confidence margins overlapped, and that for low pass noise the OSA yielded poor results, comparable to the ones obtained by the LPC alone.

7. CONCLUSIONS

We have presented a speaker recognition system, independent of the text and robust to noise. We have also studied the behaviour of the system with respect to the distance measure in the comparison block and the parametrization. In order to draw conclusions of which system behaves better we have used tha ANOVA methodology.

8. REFERENCES

1. E.Monte, A. Adolf, X. Miró, J. Hernando ‘Text Independent Speaker Identification on Noisy Environments by Means of Self Organising Maps’. *ICSLP* , 1996.
2. F.K. Soong, A.Rosenberg, L.Rabiner, B.Juang, ‘A vector quantization approach to speaker recognition’, *ICASSP*. 1985.
3. H. Gish y M. Schmidt, ‘Text independent speaker identification’, *IEEE signal processing magazine*, 1994.
4. Naik, J. “Speaker Verification: A Tutorial”, IEEE Communication Magazine, Jan.90.
5. Bimbot F., Mathan L., “Text-Free Speaker Recognition Using an Arithmetic-Harmonic Sphericity Measure”, Proc.EUROSPEECH´93, Belin, Sept.1993.
6. Kohonen,T., "Self Organisation and Associative Memory". Springer-Verlag, 1984.
7. Hernando J., Nadeu C., Villagrasa C. and Monte E., "Speaker Identification in Noisy Conditions using Linear Prediction of the One-Sided Autocorrelation Sequence",ICSLP 94. Sept 94.Yokohama,Japan.

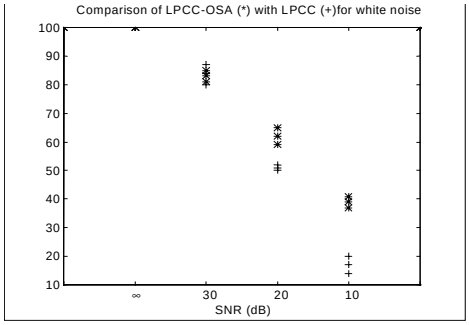


Figure 7. Comparison of the recogniton rate of the LPCC-OSA and LPCC for white noise

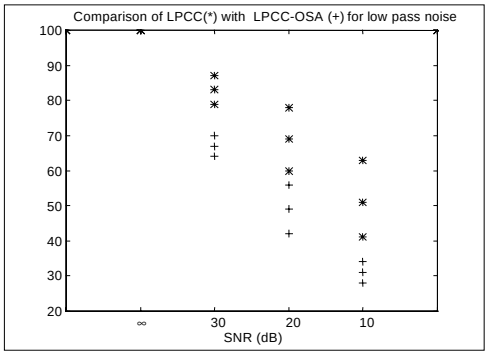
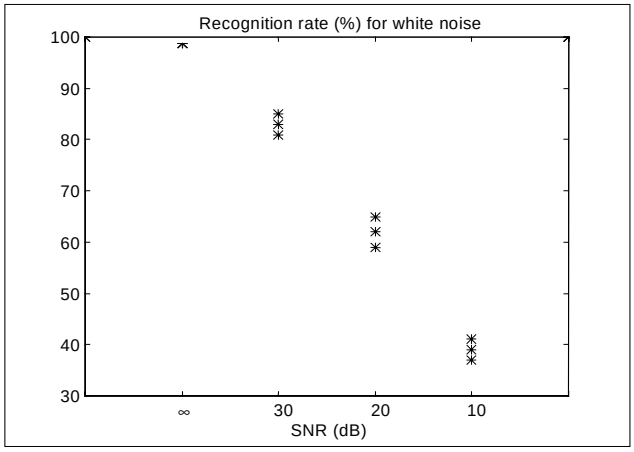
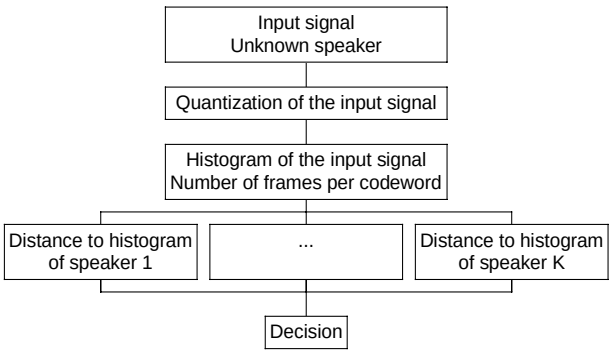


Figure 8. Comparison of the recogniton rate of the LPCC with the LPCC-OSA for low pass noise.


Recognition rate (%) for white noise
100
90
80
70
60
50
40
30
∞
30
20
10
SNR (dB)


Input signal
Unknown speaker
Quantization of the input signal
Histogram of the input signal
Number of frames per codeword
Distance to histogram of speaker 1
...
Distance to histogram of speaker K
Decision