VQ speaker verification with sentence codebook

Filipe Moreira, Carlos Espain
CEFAT / DEEC / FEUP / Universidade do Porto

ABSTRACT

Experiments on text independent speaker verification (with speakers of the same sex) using vector quantisation are reported in this paper. Several situations were considered using different features sets, training times and testing utterances durations, both with speaker independent and speaker dependent thresholds. The amount and location of silence in the testing sentences is usually a problem in text-independent speaker verification systems. We proposed and tested the use of string codebooks to attenuate the problem. A codebook is extracted from the utterance and it is this codebook that is quantised again into the claimer's codebook. It is also reported an additional experiment combining different techniques of speaker verification.

INTRODUCTION

Many systems of text-independent speaker verification use VQ methods [1]. The problem of the silence/noise which might be present, to a variable extension and localisation, is usually adressed by using a speech detector. Here we present a different method of dealing with it.

We designed a regular system which performs a quantisation of the string frames into the claimer's codebook normalising the total quantisation error in accordance with the number of frames of the utterance. A voice detector can be used to eliminate silence between words and therefore eliminate the disturbance caused by different noise amplitudes. Instead a second system was designed producing first a codebook from each utterance and then quantising this codebook into the claimer's codebook. This way silence is reduced to a few points of the sentence codebook whatever its duration and/or location might be [3]. In both cases noise has not been filtered.

We studied several situations using different features sets, training times and testing utterances durations, both with speaker independent and speaker dependent thresholds and tried different techniques of combining results from distinct verifiers.

TRAINING CONDITIONS

The experiments were done using a subset of the TIDIGITS database, consisting of 110 speakers randomly chosen (55 men and 55 women), each one uttering 77 strings of digits, of which only 39 were used, 17 for training the codebooks and 22 for testing. We only used the strings with 2, 3, 5 and 7 digits.

The speech signal is sampled at 20 kHz, pre-emphasised with a 0.97 coefficient filter and Hamming windowed into 20 ms frames starting every 9 ms. From each frame the normalised energy and D-energy and 8 cepstral coefficients (CCs) along with their first-derivatives were extracted.

A LBG type algorithm was used to generate a 64 point codebook. Clustering and centroid calculation were done using the euclidean distance. Two types of codebooks were produced. The first type (Cbk 1) was generated from 6 3-digits strings from each speaker to be verified and the second one (Cbk 2) was generated from 11 5-digits strings. Average training time was 9.6 s for Cbk 1 and 25.2 s for Cbk 2.

Within each type, two kinds of codebooks were produced according to the features chosen as components of the vector representing each frame, CCs plus D-CCs (F1) and CCs (F2). Normalised energy and D-energy were used in both situations.

TESTING CONDITIONS

Two subsets of the database were chosen for testing conditions. The first one consisted of 11 2-digits strings from the claimed speaker and one 2-digits string from each of 54 impostors of the same sex. The second one consisted of 11 7-digits strings from the claimed speaker and one 7-digits string from each of 54 impostors of the same sex. The average duration of each 2-digits string is 1.3 s and each 7-digits string has an average duration of 3.2 s.

All speakers were in turn considered as claimers which gives a total of 1210 sessions for testing the false rejection rate (FR) and 5940 sessions for testing the false acceptance rate (FA) for all claimers.
We, therefore, have two training situations differing on the amount of training time, two sets of features and two testing subsets corresponding to two types of digits strings, a short one (1.3 s) and a long one (3.2 s).

For each test subset, two different methods were tried, as referred before.
The first one (S1) consisted of a normalised quantisation of the frames into the claimer's codebook, taking the Euclidean distance of the frame to the nearest codebook point as its quantisation error. These errors for the entire utterance were added and the sum was divided by the total number of frames of the utterance.
The resulting quantity was then compared to a threshold: the claimer speaker was rejected if the sum was higher then the threshold and accepted if otherwise.
The second method (S2) consisted in generating a sentence 64 points codebook from the spoken string itself. This codebook was then quantised into the claimer's codebook. The total quantisation error was again compared to a threshold.

In both methods the Equal Error Rate (EER) was used as the assessment parameter in all situations. Two kinds of results will be presented. They were obtained according to two methods of determining the EER thresholds. In the first kind a single threshold for all the speakers, i. e., a speaker independent threshold (SIT), was determined; in the second kind a speaker dependent threshold, i. e., a threshold for each speaker (SDT), was determined.

RESULTS

For the subset consisting of the 2-digits strings, the results are shown in Table 1 and Table 2.

	Cbk 1		Cbk 2	
	CCs+DCCs	CCs	CCs+DCCs	CCs
SIT	**23.4**	**21.5**	**22.5**	**19.4**
SDT	**22.5**	**20.7**	**20.2**	**18.1**

Table 1 EER (%) for normalised distance quantisation using different features sets and thresholds

	Cbk 1		Cbk 2	
	CCs+DCCs	CCs	CCs+DCCs	CCs
SIT	**23.8**	**20.2**	**20.4**	**16.6**
SDT	**21.3**	**17.8**	**18.1**	**14.5**

Table 2 EER (%) for the string codebook quantisation using different features sets and thresholds

The results for the 7-digits strings are shown in Table 3 and Table 4.

	Cbk 1		Cbk 2	
	CCs+DCCs	CCs	CCs+DCCs	CCs
SIT	**23.1**	**23.1**	**21.6**	**21.6**
SDT	**17.5**	**19.3**	**17.3**	**18.2**

Table 3 EER (%) for normalised distance quantisation using different features sets and thresholds

	Cbk 1		Cbk 2	
	CCs+DCCs	CCs	CCs+DCCs	CCs
SIT	**10.1**	**10.8**	**6.6**	**7.8**
SDT	**6.3**	**6.3**	**3.7**	**4.7**

Table 4 EER (%) for the string codebook quantisation using different features sets and thresholds

Results showed, that particularly in the case of the long sentences (7-digits), as expected, the use of sentence codebooks decreased the E.R.R. in an average of 66%.

COMBINATION OF TECHNIQUES

Several techniques of combining the results of different statistically independent classifiers are known [2, 4]. We applied four combination techniques to the two methods described above: the quantisation of the frames into the claimer's codebook (S1) and the quantisation of the string codebook into the claimer's codebook (S2).

For all of the techniques the total quantisation error for each testing utterance in both methods is computed.

In the first technique, called Maximum (MAX), the maximum of the quantisation errors given by methods S1 and S2, for each sentence, was taken as the final output.
In the second technique, called Minimum (MIN), the minimum of the quantisation errors given by methods S1 and, for each sentence, was instead considered as the final output.
In the third technique, named Product (PROD), the final output is the product of the quantisation errors.
Finally, in the fourth technique, named Sum (SUM), the final output is the average of the errors.
In all the combinations, the final output was compared to a speaker independent EER threshold: the claimer speaker was rejected if the ouput was higher then the threshold and accepted if otherwise.

For the subset consisting of the 2-digits strings, the results are shown in Fig. 1 and Fig.2.

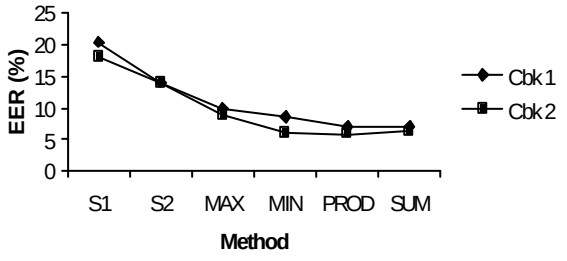

Fig. 1. EER(%) for the several methods using the CCs plus ΔCCs features set

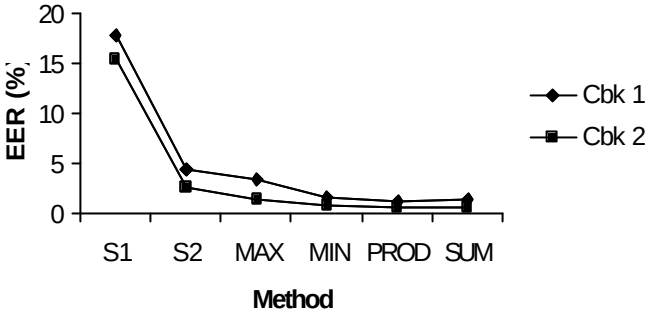

Fig. 2. EER(%) for the several methods using the CCs features set

The results for the 7-digits strings are shown in Fig.3 and Fig.4.

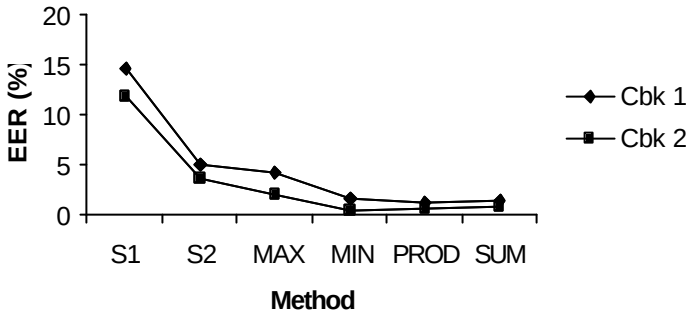

Fig. 3. EER(%) for the several methods using the CCs plus ΔCCs features set

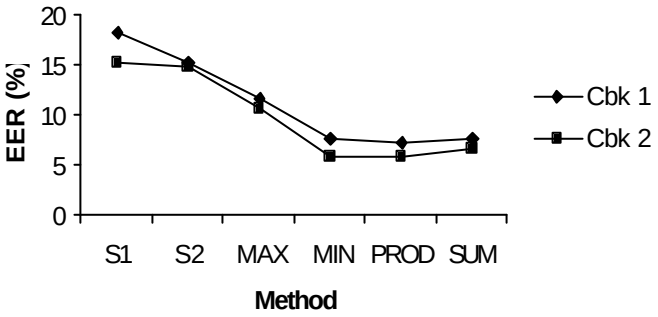

Fig. 4. EER(%) for the several methods using the CCs features set

CONCLUSIONS

The use of codebooks generated from the strings to be verified proved to be a much better technique than the simple use of a direct quantisation of the utterance.

Concerning the combination of techniques the results are according to the expectations, giving a significant improvement in the overall performance of the system. The system with Cbk 2, 25.2 s. of training time, CCs plus DCCs, using the combinations PROD and SUM between S1 and S2, gave an EER of 0.6% with testing utterances of 3.2 s.

ACKNOWLEDGMENTS

This work was granted by the project PRAXIS XXI 2/2.1/tit/1558/95 and by Portuguese-Spanish Action CRUP E-12/98.

REFERENCES

[1] - Matsui, T., Furui, S., *Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMMs*, Proc. IEEE Int. Conf. ASSP, San Francisco, 1992.

[2] – Kittler, J., Hatef, M., Duin, R. P. W., Matas, J., *On Combinig Classifiers*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, n.o 3, Março 1998.

[3] - Moreira, F., Espain, C., *Text-independent speaker verification using string codebooks*,COST 250 MCM, Porto, 1999.

[4] - Rodríguez-Linares, L., García-Mateo, C., *A novel technique for the combination of utterance and speaker verification systems in a text-dependent speaker verification task*, Proc. ICSLP'98, Sidney, 1998.