Speaker Identification Based on Vector Quantization*

Vlasta Radová and Zdeněk Švenda

University of West Bohemia, Department of Cybernetics,
Univerzitní 22, 306 14 Plzeň, Czech Republic
{radova, svendaz}@kky.zcu.cz

Abstract. In this paper a method of text-independent speaker recognition using discrete vector quantization is presented. The identification experiments were performed in a closed set of 599 speakers and two various types of features were tested: cepstral mean subtraction coefficients and mel-frequency cepstral coefficients. The effect of the various codebook size on the speaker identification performance was investigated.

1 Introduction

There are many various methods that can be used for the design of speaker recognition systems. At present, a majority of systems is based on vector quantization, hidden Markov models or artificial neural networks [1], [3], [4]. The method described in this paper is based on discrete vector quantization and can be used for text-independent speaker identification. The principle of the method is explained in Section 2. In Section 3 the structure of the speech database and the feature analysis techniques used in our experiments are described. Achieved experimental results are presented in Section 4.

2 Principle of the Method

The principle of the speaker identification using vector quantization consists of 2 phases – training and identification.

During the *training phase* a codebook for each reference speaker is formed from the training data of that speaker. To do it, we used so-called uneven binary decision and the modified MacQueen (k-means) algorithm described in [5]. Achieved codebooks have a tree structure, where each leaf node represents a codebook vector. Due to this structure, a very simple algorithm can be used to find the nearest codebook vector to an input vector.

During the *identification phase*, the speech signal of an unknown speaker is first analyzed and converted to a sequence of N feature vectors $\boldsymbol{x}(1), \ldots, \boldsymbol{x}(N)$. Then each feature vector of the unknown speaker is quantized using codebooks

* This work was supported by the Ministry of Education of the Czech Republic, project no. VS 97159.

of all reference speakers and the average quantization distortion Q_k is computed for each of K reference speakers according to the formula

$$Q_k = \frac{1}{N} \sum_{n=1}^{N} d(\boldsymbol{x}(n), \boldsymbol{v}_{*k}), \tag{1}$$

where $d(\boldsymbol{x}(n), \boldsymbol{v}_{*k})$ is a distortion measure between the input vector $\boldsymbol{x}(n)$ and its nearest codebook vector $\boldsymbol{v}_{*k}$ of the k-th reference speaker. The unknown speaker is then identified as the reference speaker with the minimum average distortion, i.e.

$$k^* = \underset{k=1,\ldots,K}{\operatorname{argmin}} Q_k, \tag{2}$$

where k^* represents the identified speaker.

3 Speech Database and Feature Analysis

The method described in Section 2 was used to identify "unknown" speakers in a group of 599 speakers (317 male, 282 female). Every speaker spoke a different set of several short Czech sentences and isolated words in the total duration of about 70s. Approximately 40s of speech of each speaker were regarded as training data and were used to form the codebook of that speaker. The remaining about 30s of speech were used for identification tests.

The speech signal of each speaker was recorded during one session through a telephone channel, sampled at the rate of 8kHz and stored in a μ-law 8bit format. Before further processing, however, the 8bit μ-law samples were converted to linear 16bit PCM samples. Finally, the digitalized utterances were converted to sequences of segments using a 16ms rectangular window.

In our experiments various types of features were tested – cepstral mean subtraction coefficients (CMSCs) and mel-frequency cepstral coefficients (MFCCs). To obtain the vectors of CMSCs the utterances were first converted to sequences of vectors of 12 LPC coefficients. These vectors were then converted to vectors of 12 LPC cepstral coefficients (LPCCs) and finally to the vectors of 12 CMSCs according to the formula [2]

$$\boldsymbol{c}_{cmsc}(n) = \boldsymbol{c}_{lpcc}(n) - \overline{\boldsymbol{c}}_{lpcc}, \tag{3}$$

where $\boldsymbol{c}_{lpcc}(n)$ is the n-th vector of LPC cepstral coefficients and $\overline{\boldsymbol{c}}_{lpcc}$ is the mean vector of LPC cepstral coefficients over the whole speech data of a speaker, i.e

$$\overline{\boldsymbol{c}}_{lpcc} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{c}_{lpcc}(n), \tag{4}$$

where N is the number of segments in the speech of the speaker.

In the experiments with MFCCs the speech signal was processed first by a set of 26 triangual filters spaced uniformly on a mel-scale. The output energy of

the filters were then transformed into vectors of 16 MFCCs using the discrete cosine transform [2].

The feature vectors were formed in several various ways. First, either vectors of only 12 CMSCs or vectors of only 16 MFFCs were used as feature vectors. Next the feature vectors were extended and to the CMSCs or MFCCs their delta coefficients or delta and acceleration coefficients were added. The delta coefficients were computed using the formula [6]

$$\boldsymbol{d}(n) = \frac{\sum_{i=1}^{I} i[\boldsymbol{c}(n+i) - \boldsymbol{c}(n-i)]}{2\sum_{i=1}^{I} i^2}, \tag{5}$$

where $\boldsymbol{d}(n)$ is a vector of delta coefficients for the n-th segment of speech and $\boldsymbol{c}(n)$ is a vector of 12 CMSCs or 16 MFCCs for the n-th segment of speech. The value of I was set equal to 2 in our experiments. Similarly the vectors of acceleration coefficients $\boldsymbol{a}(n)$ were computed from the delta coefficients $\boldsymbol{d}(n)$ according to the formula

$$\boldsymbol{a}(n) = \frac{\sum_{i=1}^{I} i[\boldsymbol{d}(n+i) - \boldsymbol{d}(n-i)]}{2\sum_{i=1}^{I} i^2}. \tag{6}$$

Since equations (5) and (6) rely on past and future speech segments some modification is needed at the beginning and end of the speech. One of the possible modifications is simply to replicate the first or last vector as needed [6].

4 Experimental Results

All types of feature vectors described in Section 3 were used to form a codebook of either 80 or 320 codebook vectors. To compute the distortion measure $d(.,.)$ in (1) the cepstral measure [5] was used. Achieved results are presented in Table 1 and Table 2. In these Tables, CMSCΔs means feature vectors composed of CMSCs and delta coefficients, and CMSC$\Delta\Delta$s means feature vectors composed of CMSCs, delta coefficients and acceleration coefficients. Similar denotation is valid for the MFCCs as well.

Table 1. Identification results using the codebook of 80 vectors

Coefficients	# correct	Correct [%]
CMSCs	564	94.16
CMSCΔs	569	94.99
CMSC$\Delta\Delta$s	567	94.66

Coefficients	# correct	Correct [%]
MFCCs	595	99.33
MFCCΔs	595	99.33
MFCC$\Delta\Delta$s	586	97.83

Table 2. Identification results using the codebook of 320 vectors

Coefficients	# correct	Correct [%]
CMSCs	570	95.16
CMSCΔs	578	96.49
CMSC$\Delta\Delta$s	572	95.49

Coefficients	# correct	Correct [%]
MFCCs	597	99.67
MFCCΔs	597	99.67
MFCC$\Delta\Delta$s	594	99.17

As the Tables show, using the feature vectors with the MFCCs better results were achieved than using feature vectors with the CMSCs. The size of the codebooks has a certain effect on the speaker identification performance as well – when the codebook with 320 vectors was used more speakers were identified correctly than with the codebook of 80 vectors.

5 Conclusion

A method of speaker identification based on vector quantization was presented in this paper. Various types of feature vectors were tested and as many as 99.67% of speakers in a group of 599 speakers were identified correctly. Such a result may be regarded as a promising way to a high-performance speaker identification system. However, it has to be taken into account that the speech data used in the experiments were recorded during one session. When there is a time interval between the recording of training and test data the achieved results may not be so excellent. Our experiments on data where the time interval between the recording of test and training data are from several days to several weeks are now in progress.

References

1. Cheng, Y., Leung, H. C.: Speaker Verification Using Fundamental Frequency. In: Proc. of the ICSLP (1998) 161–164
2. Mammone, R. J., Zhang, X., Ramachandran, R. P.: Robust Speaker Recognition. A Feature-based Approach. IEEE Signal Processing Magazine **5** (1996) 58–71
3. Monte, E., Adolf, A., Miró, X., Hernando, J.: Text Independent Speaker Identification on Noisy Environments by Means of Self Organising Maps. In: Proc. of the ICSLP (1996) 1804–1087
4. Monte, E., Arqué, R., Miró, X.: A VQ Based Speaker Recognition System Based in Histogram Distances. Text Independent and for Noisy Environments. In: Proc. of the ICSLP (1998) 185–188
5. Psutka, J.: Communication with Computer by Speech. Academia, Praha (1995) (in Czech)
6. Young, S., Jansen, J., Odell, J., Ollason, D., Woodland, P.: The HTK Book (for HTK V2.0). Cambridge University (1996)