Text-Independent Speaker Identification Using Vector Quantization

Noha Korany

E. E. Dept., Faculty of engineering, Alexandria University, Egypt. E-Mail: nokorany@hotmail.com

Abstract

This paper investigates an effective method for text-independent speaker recognition. Vector quantization (VQ) is used as the recognition engine. It is necessary to reduce the operator's load. The paper aims to determine the dimension of the codebook vectors within the model and to find out the best feature that fits with this model.

Various Kinds of features are extracted, VQ classifiers are applied independently to each feature and the identification rate is calculated for each case. Furthermore, the performance of the recognition process is investigated while varying the length of the feature vector employed by the classifier. Finally, recorded speech database is used to evaluate the system for text-independent speaker identification in both clean and noisy environments.

Introduction

Speaker recognition process aims to automatically establish the identity of an individual based on his voice.

Vector quantization (VQ) is used as recognition engine. The recognition task is achieved using two phases, the training and the test phases. Within the training phase, speaker-specific feature vectors are extracted, then they are clustered to generate a speaker-specific VQ codebook for each known speaker. In the test phase, feature vectors are extracted from speech samples from unknown speaker, then they are compared to the speakers codebooks. The identified speaker is the one whose codebook is the closest to the observed feature vectors. Mel-frequency cepstral coefficients, *MFCC* is frequently employed for the classification problem, whereas relative spectral perceptual linear predictive cepstral coefficients RASTA-PLPCC is used for classification in noisy environments.

It is necessary to reduce the computational complexity of text-independent speaker identification process. Many parameters affects the performance of the process, such as the best feature for the classification, the dimension of the feature vector used within the model. Other factors e.g. acoustical noise and variations in recording environments present a challenge to speaker recognition technology.

The aim of this paper is to evaluate the performance of the recognition process for text-independent speaker in clean and noisy environments.

Feature Extraction

Feature extraction is the process that transforms the raw data into data that can be used by a classifier. This paper uses *MFCC*, *and RASTA-PLPCC* to describe the speech signal.

The MFCC is defined as the real cepstral of a windowed short-time signal derived from the FFT of that signal. The speech signal is first subjected to non-uniformly spaced filters (Mel-filters) [1], and then a modified Discrete Cosine Transform is applied. The principle of RASTA processing is that, linearly filtering the time trajectories of each spectral component of the auditory spectrum of the signal, then the all-pole model is estimated to obtain RASTA-PLPCC[2].

Recognition Engine

VQ is a lossy data compression method based on the principle of block coding. The VQ algorithm is based on a training sequence[3]. In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his training feature vectors into a predefined number of cluster centroids. Thus, each speaker is modeled by a set of K clusters of feature vectors. For a well-known speaker, if the training template $X = \{x_1, x_2, \dots, x_T\}$ consists of a set of T vectors, then the reference template $\theta = \{\theta_1, \theta_2, \dots, \theta_k\}$ is obtained by dividing X into k clusters represented by their mean vectors. The objective of the K-means algorithm is to minimize the total distortion (quantization error), D, that is given by equation (1).

$$D = \sum_{i=1}^{T} \sum_{j=1}^{k} \|x_i - \theta_j\|^2 \qquad (1)$$

Within the test phase, theory of VQ [4] can be applied in template matching. If the test template is denoted as $Y = \{y_1, y_2, \dots, y_T\}$, then the average quantization distortion of Y using θ, $D_Q(Y, \theta)$, is computed, and the speaker whose codebook is the closest to the observed feature vectors is identified. $D_Q(Y, \theta)$ is defined in equation (2), where d (¢; ¢) is Euclidean distance.

$$D_Q(Y, \theta) = \frac{1}{T} \sum_{t=1}^{T} \min_{1<j<k} d(y_t, r_j) \qquad (2)$$

Database

The Database consists of 10 English words that are spoken by 15 speakers and are used for the classification problem. The words are one, two, three, four, five, six, seven, eight, nine, ten. Now the data set contains (10×15= 150) audio files from 15 different speakers. Each file was recorded using mono-format with same microphone, same sound card. Each file was sampled at 44100 Hz, and 16-bit quantization level was used. Next the data were segmented into approximately 20 ms frame length, overlapped by 50% of this frame. A Hamming window was then applied to each frame.

Simulations & Results

Two recorded words per speaker are employed within the training phase, whereas the remaining ones are used within the test phase. Three experiments are conducted. For the

first and the second experiments, clean data is employed within the train and the test phases, whereas for the third experiment, clean data is used for training, whereas additive white Gaussian noise is added to the remaining words. The noisy words are used within the test phase.

Experiments description

There is no theoretical solution to find the optimal number of clusters for a given data set. The first experiment is conducted to determine the best number of cluster centroids for the classification problem. MFCC are employed for the classification. Various number of Mel-filter bank are used, and the identification rate is calculated for different numbers of cluster centroids, as shown in figure 1.

The second experiment aims to improve the identification rate by determining the most important elements within a feature vector. The experiment runs using 64 Mel-filter bank, and hence 63 MFCC discarding 0^{th} order cepstral coefficient, the dimension of the feature vector is first reduced by eliminating a number of coefficients within the vector, and finally 8 cluster centroids are employed. Let the extracted feature vector is V = {$MFCC_1$, $MFCC_2$, ..., $MFCC_n$}, then after dimension reduction the feature vector that is used within the training phase is M ={$MFCC_m$, $MFCC_{m+1}$,..., $MFCC_l$}, where 1< m < n, 1 < l < n. This vector is denoted by M[m, l]. While varying the dimension of the vector M[m, l], the identification rate is calculated. Moreover, Twenty four RASTA-PLPCC, R[1,24], are extracted, whereas the feature vector is first reduced, and then the reduced vector R[m, l] is employed by the classifier. The identification rate is obtained for various number of coefficients. The results are shown on table 1, and table 2 using MFCC and RASTA-PLPCC respectively.

The aim of the third experiment is to evaluate the classifier performance in noisy environments. MFCC are employed by the classifier, the identification rate is calculated for different signal-to-noise ratio (SNR). The identification rate is recalculated while using RASTA-PLPCC. The results are shown in figure 2.

Results and Discussion

Figure 1 shows that the identification rate is not affected significantly by the number of cluster centroids, especially when the number of Mel-filter bank is increased.

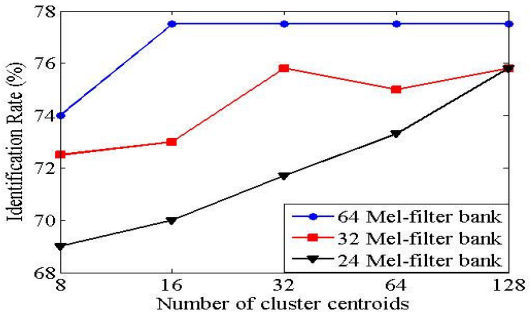

Figure 1: Identification Rate (%) versus number of cluster centroids

Table 1, and table 2 show that the dimension of the RASTA-PLPCC feature vector does not affect significantly the identification rate, whereas the identification rate increases when the low-order elements of the MFCC feature vector are discarded. Maximum identification rate is reached for MFCC[8,50], i.e. the first seven elements of the feature vector are discarded.

Figure 2 shows that for high SNR (SNR >25 dB), the feature vector MFCC[8, 40] yields to the highest identification rate, whereas for lower SNR, RASTA-PLPCC provides the highest identification rate.

Table 1: Identification Rate, IR (%) in clean environment using various dimension of the MFCC feature vector, MFCC[m, l]

[m, l]	[1, 63]	[8, 40]	[10,40]	[4, 40]	[8, 50]	[8, 63]
IR(%)	74	83.3	74.2	79.2	85	82.5

Table 2: IR (%) in clean environment using various dimension of the RASTA-PLPCC feature vector, RASTA[m, l]

[m, l]	[1, 24]	[4, 24]	[7, 24]	[1, 12]	[4, 12]	[1, 8]
IR(%)	42.5	39.2	36.7	43.3	34.2	36.7

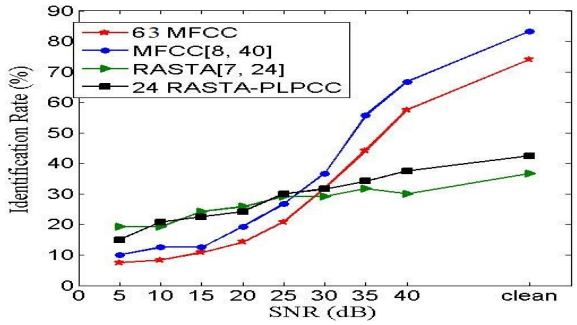

Figure 2: Identification Rate (%) versus SNR using various dimension of the feature vector

Conclusions

This paper employs MFCC and RASTA-PLPCC, to describe the speech signal. Vector Quantization method uses each feature type to investigate the performance of the recognition process. The paper specifies 64 Mel-filter bank, 16 cluster centroids for the classification problem. Increasing the number of centroids, the identification rate remains the same. In addition, discarding the seven low-order elements of the MFCC feature vector improves the identification rate for high SNR, whereas for low SNR RASTA-PLPCC provides the highest identification rate.

References

[1] S. Davis, and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. Acoust., Speech, Signal Processing 28 - 4 (1980), 357–366.

[2] H. Hermansky, and N. Morgan, "RASTA Processing of Speech," IEEE Transactions on Speech & Audio Processing 2 (1994), 587–589.

[3] H. B. Kekre, and V. Kulkarni ,Speaker Identification by using Vector Quantization, International Journal of Engineering Science and Technology 2 - 5 (2010), 1325-1331.

[4] A. Gersho, and R. Gray, Vector Quantization and Signal Compression. Kluwer, Academic Publishers, Boston, 1991.