Speaker Recognition using Support Vector Machine

Geeta Nijhawan
Faculty of Engineering and Technology,
Manav Rachna International University, Faridabad

M. K. Soni, Ph.D
Faculty of Engineering and Technology,
Manav Rachna International University, Faridabad

ABSTRACT

Speaker recognition is the process of recognizing the speaker based on characteristics such as pitch ,tone in the speech wave.Background noise influences the overall efficiency of speaker recognition system and is still considered as one of the most challenging issue in Speaker Recognition System (SRS). In this paper mel-frequency cepstral coefficients (MFCC) feature is used along with Vector Quantisation(VQ)-LBG [Linde, Buzo and Gray, 1980] algorithm for designing SRS. MFCC feature is extracted from the input speech and then vector quantization of the extracted MFCC features is done using VQLBG algorithm. It reduces the dimensionality of the input vector .These MFCCs are used as the speaker features for matching via Support Vector Machine (SVM) method. The experimental results show that the proposed text-dependent speaker identification system gives an accuracy rate of 95.0%.

Keywords

Feature extraction, vector quantization, MFCC, SVM

1. INTRODUCTION

Speaker identification is the process of automatically identifying a speaker by machine using some characteristics of speaker's voice [1].

Speaker recognition can be categorized into identification and verification [2]. Speaker identification is the process of identifying the speaker from the database whereas Speaker verification is the process of accepting or rejecting the identity of a speaker. For more than fifty years, development on speaker recognition techniques has been an active area of research. Many methods [2] like simple template matching, dynamic time-warping approaches, and statistical pattern recognition approaches, such as neural networks and Hidden Markov Models (HMMs) [Siohan, 1998] have been used in the past. Gaussian Mixture Modeling (GMM) [Reynolds, 1995], multi-layer perceptrons [Altosaar and Meister, 1995], Radial Basis Functions [Finan et al., 1996] and genetic algorithms [Hannah et al., 1993] have also been used for speaker recognition.

Speaker recognition methods can be text-independent and text-dependent. In a text-independent system, speaker recognition is done without taking into consideration what one is saying whereas in text-dependent system, the speakers are identified based on them speaking some specific phrases, like passwords, card numbers, PIN codes, etc. [2].Figure 1 shows the block diagram of a Speaker recognition system.

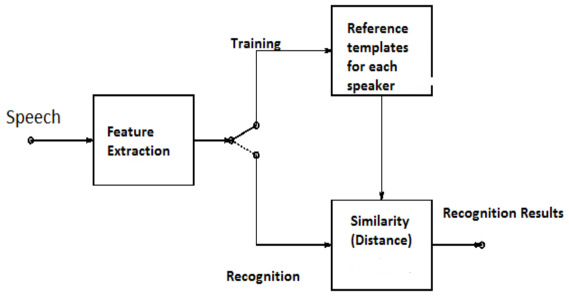

Fig.1 Block Diagram of Speaker Recognition System

Speaker recognition is a 1: N match where one unknown speaker's extracted features are matched to all the templates in the reference model for finding the closest match. The speaker feature with maximum similarity is selected. The selection of speaker features is important in the designing of a speaker identification system. The speaker features set should have high inter-speaker variability and lower intra-speaker variance. Also the selected speaker features should be independent of each other so as to reduce redundancy. The aim of this paper is to develop a more efficient system for text-dependent speaker identification using MFCCs and SVM. Previous researches [3-5] have shown that MFCCs represents detail characteristics of individual speakers and are robust. SVM is a two-class classifier based on the principles of structural risk minimization .It has well generalization ability when compared to hidden Markov model and neural network based classifier [6]. The following steps are being followed for designing Speaker Recognition System:

i. Data Acquisition through Microphone

ii.Feature Extraction

iii.Data Compression Using VQ LBG algorithm

iv.Feature Matching

The remainder of this paper is organized as follows. In section II, MFCC used for feature extraction from the input voice is presented in detail. Section III deals with the details about the support vector machine method used for feature matching. Section IV gives the experimental results and finally in section V conclusions are drawn.

2. FEATURE EXTRACTION

The purpose of Feature Extraction module is to extract the acoustic feature vectors which are used to characterize the spectral properties of the time varying speech signal .These feature vectors are used for recognition of speaker. There are a number of techniques available for parametrically representing the speech signal for the speaker recognition.Linear prediction coding (LPC), mel-frequency cepstrum coefficients (MFCC) are the most commonly used technique. MFCCs are based on the known variation of the human ear's critical bandwidths with frequency. The MFCC technique makes use of two types of filter, i.e.linearly spaced filters and logarithmically spaced filters. Mel frequency scale is used. MFCCs are less prone to the variations in speech waveform due to physical condition of speakers vocal cord [1].

2.1. The MFCC processor

Mel frequency cepstral coefficients (MFCC) is probably the best known and most widely used for both speech and speaker recognition [8]. A mel is a unit of measure based on human ear's perceived frequency. The mel scale has approximately linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. The approximation of mel from frequency can be expressed as

mel(f) = 2595*log(1+f/700) (1)

where f denotes the real frequency and mel(f) denotes the perceived frequency. The block diagram showing the computation of MFCC is shown in Figure 2.

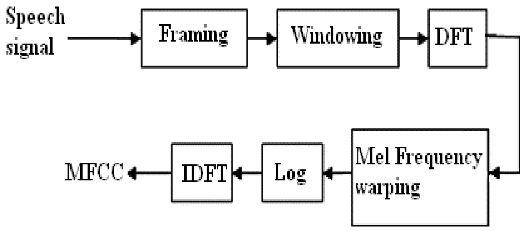

Fig.2 MFCC Extraction

In the first stage speech signal is divided into frames with the length of 20 to 40 ms and an overlap of 50% to 75%. In the second stage windowing of each frame with some window function is done to minimize the discontinuities of the signal by tapering the beginning and end of each frame to zero. In time domain window is point wise multiplication of the framed signal and the window function. A good window function has a narrow main lobe and low side lobe levels in their transfer function. In our work hamming window is used to perform windowing function. In third stage DFT block converts each frame from time domain to frequency domain [10]. Hamming window is given by

$$w(n)=0.54-0.46\cos\frac{2\Pi n}{N-1} \quad \text{------ (2)}$$

where N represents the width, in samples, of a discrete-time, symmetrical window function w[n],0 $\leq$n $\leq$N-1.

In the next stage mel frequency warping is done to transfer the real frequency scale to human perceived frequency scale called the mel-frequency scale. The mel frequency warping is normally realized by triangular filter banks with the center frequency of the filter normally evenly spaced on the frequency axis. Figure 3 shows the mel space filter bank.The warped axis is implemented according to Equation 1 so as to mimic the human ears perception. The output of the ith filter is given by-

$$y(i)=\sum_{j=1}^{N}s(j)\Omega_i(j) \quad \text{------ (3)}$$

S(j) is the N-point magnitude spectrum (j =1:N) and Ω_i(j) is the sampled magnitude response of an M-channel filter bank (i =1:M). In the fifth stage log of the filter bank output is computed and finally DCT (Discrete Cosine Transform) is computed. The MFCC may be calculated using the equation-

$$C_s(n,m)=\sum_{i=1}^{M}(\log Y(i))\cos[i\frac{2\pi}{N'}n] \quad \text{--- (4)}$$

where N' is the number of points used to compute standard DFT.

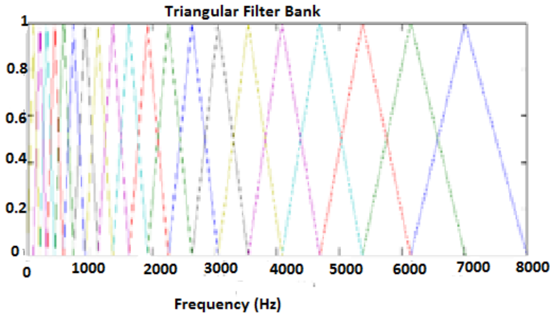

Fig.3 Triangular filter bank

Figure 4 shows screen shot of GUI developed using MATLAB of the input speech, MFCC, pitch and power plots.

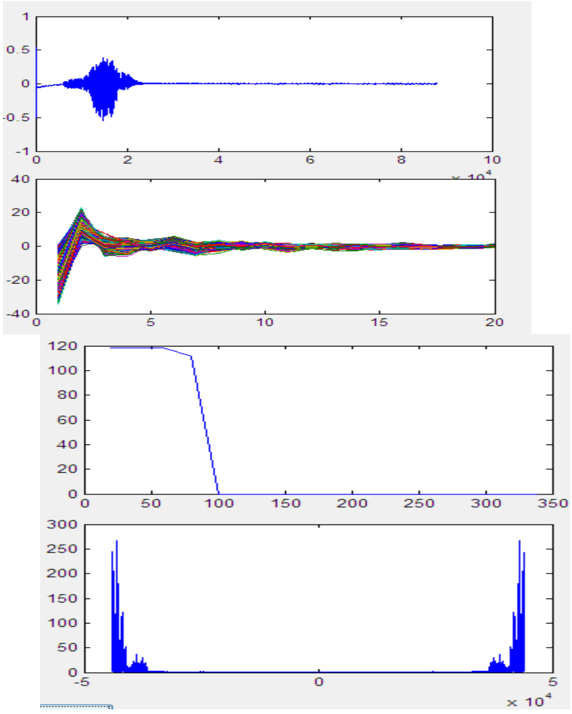

Fig. 4 GUI waveforms showing input speech, MFCC, pitch and power plots.

The aim of VQ is to compress data .We select the more effective features instead of using the whole feature vectors. By clustering the speaker's feature vector into a known cluster numbers the speaker models are formed. Each cluster is called centroid and is represented by a code vector .The code vectors constitute a codebook [3]. Each feature vector of the input is then compared with all the other codebooks. The codebook which gives the minimum distance is selected as the best [5].The next section describes how these feature vectors can be used for feature matching.

3. FEATURE MATCHING

The state-of-the-art feature matching techniques used in speaker recognition include Hidden Markov Models (HMM), Dynamic time warping (DTW), Vector Quantization (VQ) and neural network techniques. We have used Support Vector Machine due to its high accuracy.

3.1. Support vector machine

Support vector machine (SVM) was developed by Vapinik (1998).It is one of the most important developments in pattern recognition in the last 10 years. Other techniques like Hidden Markov models (HMM) and Gaussian mixture models (GMM) which are used for feature matching are prone to over fitting and they do not directly optimize discrimination [14].

SVM is a linear classifier. For a set of training examples, an SVM training algorithm builds a model that assigns new examples into one category or the other. It is important to choose the right demarcation line to remove the misclassification. It is required to maximize the margin between two classes [15]. The optimal hyperplane is calculated using kernel functions. The samples closest to the separating hyperplane are called support vectors. Optimal hyperplane is completely defined by support vectors.

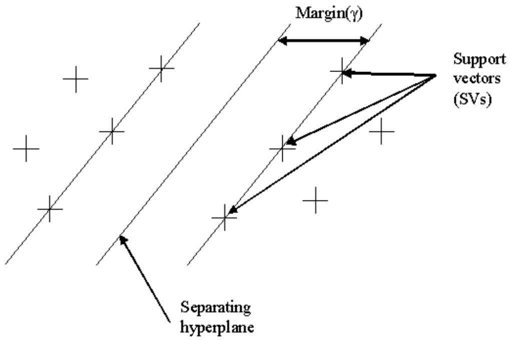

Fig. 5: A linear support vector machine.

To find the optimal hyperplane, we have to solve the following optimization problem:

Minimize $\frac{1}{2}\ ||w||^2$ ------------- (5)

Subject to (x_i .w+b) y ≥1

When the data are not linearly separable then no hyperplane exists for which all points satisfy the inequality. Therefore slack variables ξ_i are included into the inequalities. This relaxes the inequality and some points are allowed to be misclassified. The objective function becomes:

$$\frac{1}{2}||w||^2 + C\sum_i L(\zeta_i) \quad \text{----------- (6)}$$

subject to (x_i .w+ b) $y_i \geq 1 - \zeta_i$ for all i.

The second term of Equation 6 is the empirical risk associated with those points that are misclassified, L is the loss function (cost function) and C is a hyper parameter .It trades off the effects of minimizing the empirical risk against maximizing the margin.

Kernels are used to non-linearly map the input data to a high-dimensional space (feature space). The new mapping is then linearly separable.

The polynomial and radial basis function (RBF) kernels are commonly used, and are given by the following:

$$K(x_i, x_j) = (x_i^T, x_j) \quad \text{--------- (7)}$$

$$K(x_i, x_j) = (x_i.x_j + 1)^n \quad \text{------------------ (8)}$$

$$K(x_i, x_j) = \exp[-\frac{1}{2}(\frac{||x_i - x_j||^2}{\sigma})^2] \quad \text{--------- (9)}$$

where n is the order of the polynomial and σ is the width of the radial basis function.

RBF is the main kernel function because the RBF kernel nonlinearly maps samples into a higher dimensional space unlike to linear kernel.The RBF kernel has less hyperparameters and less numerical difficulties than polynomial kernel.

4. RESULT

A database of eight speakers is created .The feature extraction was done by using MFCC (Mel Frequency Cepstral Coefficients). The speakers were modeled using Vector Quantization (VQ). A VQ codebook is generated by clustering the training feature vectors of each speaker and then stored in the speaker database. The LBG algorithm was used for clustering purpose. All of the input speech signals are sampled at 8000 Hz with 16-bit resolution. A speaker identification system comprises of a training phase and a test phase. In the training phase the SVM models are created for each speaker. In testing phase the stored data are compared with the claimed SVM model and a decision is made. The Equal Error Rate (EER) is used to measure the system performance.

We made the comparison against the two types of kernel functions used in SVM implementation. As a kernel functions RBF and Polynomial (degree 2 and 3) are used. The results are presented in Table 1. The best results are obtained with RBF kernel function.

Table 1: Mean EER for different kernels and coefficients

Method	EER
MFCC+Polynomial(order 2)	10.25
MFCC+Polynomial (order 3)	9.85
MFCC+RBF	9.45

Table 2 shows the effect of changing the number of centroids on the identification rate of the system. It can be seen that an accuracy rate of 95.0% is obtained for MFCC of order 24.

Table 2. Comparison of SVM based text-dependent speaker identification system with different MFCC orders

MFCC Order	***Accuracy rate***
8	***85.7%***
12	***90.5%***
16	***92.0%***
20	***93.7%***
24	***95.0%***

5. CONCLUSIONS

This paper successfully presents an approach based on SVM for speaker identification. The MFCC technique has been applied for feature extraction. VQ is used to minimize the data of the extracted feature. The result shows that as number of centroids increases, identification rate of the system increases but it comes at the expense of increasing computational time. Also the combination of Mel frequency and Hamming window gives the best performance. It can be concluded that the outcome of the work clearly indicates that the proposed model can be used as an attractive and effective means for the recognition.

In future, GMM techniques or Neural Network technique can be used to improve the performance and to increase the accuracy. Also Voice Activity Detection can be employed to distinguish between silence and speech which would definitely improve the identification rate further.

To design a robust speaker identification system different feature extraction techniques like MFCC,LPC etc can be combined . We can add other soft computing techniques like Genetic Algorithm etc. to it and can make an improved hybrid system also. We can use dynamic speaker recognition methods also.

6. REFERENCES

[1] Ch.Srinivasa Kumar, Dr. P. Mallikarjuna Rao, 2011, "Design of an Automatic Speaker Recognition System using MFCC, Vector Quantization and LBG Algorithm'', International Journal on Computer Science and Engineering,Vol. 3 No. 8 ,pp:2942-2954.

[2] Amruta Anantrao Malode,Shashikant Sahare,2012 , "Advanced Speaker Recognition", International Journal of Advances in Engineering & Technology ,Vol. 4, Issue 1, pp. 443-455.

[3] A.Srinivasan, "Speaker Identification and verification using Vector Quantization and Mel frequency Cepstral Coefficients",Research Journal of Applied Sciences,Engineering and Technology 4(I):33-40,2012.

[4] Vibha Tiwari, "MFCC and its applications in speaker recognition",International Journal on Emerging Technologies1(I):19-22(2010)

[5] Md. Rashidul Hasan,Mustafa Jamil,Md. Golam Rabbani Md Saifur Rahman, "Speaker Identification using Mel Frequency Cepstral coefficients",3rd International Conference on Electrical & Computer Engineering,ICECE 2004,28-30 December 2004,Dhaka ,Bangladesh

[6] Fu Zhonghua; Zhao Rongchun; "An overview of modeling technology of speaker recognition", IEEE Proceedings of the International Conference on Neural Networks and Signal Processing Volume 2, Page(s):887 – 891, Dec. 2003.

[7] Seddik, H.; Rahmouni, A.; Sayadi, M.; "Text independent speaker recognition using the Mel frequency cepstral coefficients and a neural network classifier"First International Symposium on Control, Communications and Signal Processing, Proceedings of IEEE 2004 Page(s):631 – 634.

[8] John G. Proakis and Dimitris G. Manolakis, "Digital Signal Processing", New Delhi: Prentice Hall of India. 2002.

[9]Rudra Pratap. Getting Started with MATLAB 7. New Delhi: Oxford University Press, 2006

[10] D.A. Reynolds, "Experimental evaluation of features for robust speaker identification," IEEE Trans. Speech Audio Process., vol. 2(4), pp. 639-43, Oct. 1994.

[11] L. Rabiner, and B.H. Juang"Fundamentals of Speech Recognition", Singapore: Pearson Education, 1993.

[12] B. Yegnanarayana, K. Sharat Reddy, and S.P. Kishore, "Source and system features for speaker recognition using AANN models," in proc. Int. Conf. Acoust., Speech, Signal Process., Utah, USA, Apr. 2001.

[13] C.S. Gupta, "Significance of source features for speaker recognition," Master's Thesis, Indian Institute of Technology Madras, Dept. of Computer Science and Engg., Chennai, India, 2003.

[14] Shi-Huang Chen and Yu-Ren Luo, Speaker Verification Using MFCC and Support Vector Machine, Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I,IMECS 2009, March 18 - 20, 2009, Hong Kong

[15] S. M. Kamruzzaman, A. N. M. Rezaul Karim, Md. Saiful Islam and Md. Emdadul Haque, Speaker Identification using MFCC-Domain Support Vector Machine.