International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 5 Issue 5|| May 2016 || PP.62-65

Speaker Identification based on GFCC using GMM-UBM

Jayanth M[1], B Roja Reddy[2]
[1]*(Student, TCE Dept., R V College of Engineering, Bengaluru, INDIA)*
[2]*(Asst. Professor, TCE Dept., R V College of Engineering, Bengaluru, INDIA)*

Abstract : *The accuracy of the speaker identification system reduces severely in the presence of noise. The auditory based features called gammatone frequency ceptral coefficient (GFCC) which resembles to the ability of human ear to identify the speaker identity's in noisy environment. The GFCC is based on gammatone filter bank which models the basilar membrane as a sequence of overlapping band pass filters. The system uses GFCC features and Gaussian Mixture Model - Universal Background Model (GMM - UBM) for modeling of features of the speaker. The experiment are conducted on the English Language Speech Database for Speaker Recognition (ELSDR) databases. In order to find out the performance of the system, the test utterances are mixed with noises at various SNR levels to simulate the channel change. The results shows that the GFCC features has a good identification accuracy not only in clean sample, but also for noisy condition.*

Keywords: *Auditory based feature, GFCC features, GMM-UBM, Noise.*

I. Introduction

Communication between humans is carried out by means of speech. It contain unique characteristics of the speaker which are useful in speaker recognition. Speaker recognition can be broadly classified into two categories : speaker identification and speaker verification. Speaker identification is a process of automatically recognize who is speaking whereas speaker verification is process of claiming identity of a speaker [1].

Speaker identification consists of two phase namely : training and testing. In training phase, feature vectors are extracted which are speaker dependent and speaker models are developed for each speaker's feature set. In testing phase, for unknown speaker feature vectors are extracted and are compared against all speaker models and the most likely speaker identity is decided.

In Bell Labs researches proposed frame based features combining the cepstral coefficients to increase robustness [2]. In the mid-1980's a Speech Group was developed to study advanced speech processing techniques by the National Institute of Standards and Technology (NIST). Clustering technique for effective compressing of feature vectors using Vector Quantization [3]. Since 1996 Speaker Recognition Evaluation (SRE) to derive the technology for finding promising approaches [4]. Different modeling technique have been investigated such as Gaussian Mixture Models (GMM) and Support Vector Machine (SVM) classifier [5].

In this study, we will be using gammatone frequency cepstral coefficients for feature extraction, speaker modeling using gaussian mixture model - universal background model (GMM-UBM) and to calculate score maximum likelihood classifier being used.

The rest of the paper is organized as follows. Section II deals with overview of the system. Section III deals with feature extraction. Section IV deals with speaker modeling, followed by experimental setup and results in section V. We conclude this paper in section VI.

II. The System Model

Fig. 1 shows the schematic diagram of the proposed system. In training phase the speech signal is passed through auditory feature extraction to extract gammatone features (GF) and gammatone frequency cepstral coefficients (GFCC). After feature extraction, the cepstral coefficients are passed to GMM-UBM to obtain the statistical models and they are stored in database. During testing, for the unknown speech utterance features are extracted and log likelihood ratio are calculated based on the models and decides the most likely utterance.

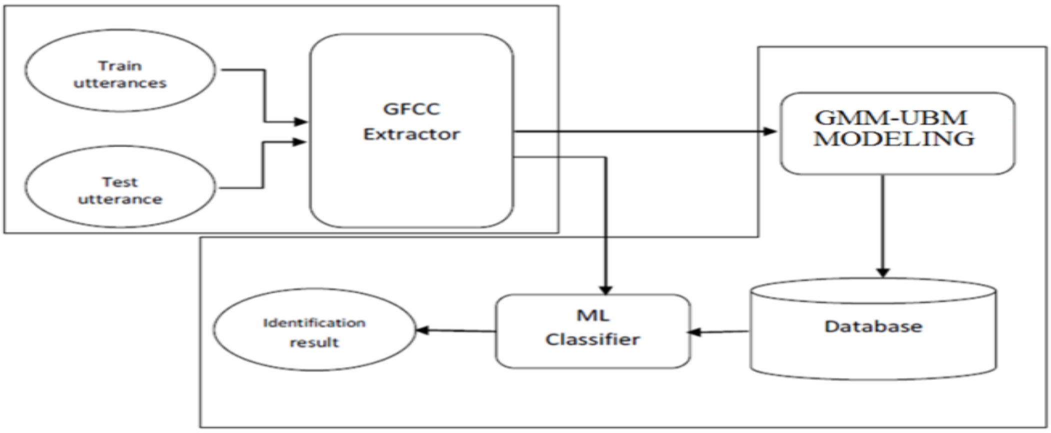

Fig. 1. Schematic diagram of the proposed speaker identification system.

III. Auditory Feature Extraction

Gammatone filters are derived from psychophysical and physiological observations of the auditory periphery and this filter bank is a standard model of cochlear filtering [6]. The system first performs auditory filtering by decomposing an input signal into the T-F domain using a bank of gammatone filters. The impulse response of a gammatone filter centered at frequency f is :

$$g(f,t)=\begin{cases} t^{a-1}e^{-2\pi bt}\cos(2\pi ft), & t\geq 0 \\ 0, & else \end{cases} \quad (1)$$

t refers to time, filter order a=4,b is the rectangular bandwidth which increases with the center frequency f.

We use a bank of 64 filters whose center frequencies range from 50 Hz to 4000 Hz or 8000 Hz depending on the sampling frequency of speech data. Since the filter output retains the original sampling frequency, we decimate fully rectified 64-channel filter responses to 100Hz along the time dimension. This yields a corresponding frame rate of 10 ms, which is used in many short-time speech feature extraction methods. The magnitudes of the decimated outputs are then loudness-compressed by a cubic root operation.

$$G_m[i]=||g|_{decimate}\{i,m\}|^{\frac{1}{3}} \quad (2)$$

Here, N=4 refers to the number of frequency (filter) channels. M is the number of time frames obtained after decimation. The resulting responses form a matrix, representing the T-F decomposition of the input. This T-F representation is a variant of cochleagram.

We call a time slice of the above matrix gammatone feature (GF), and use G[i] to denote its ith channel. Time index m is dropped for simplicity. Here, a GF vector comprises 64 frequency components. Additionally, because of the frequency overlap among neighboring filter channels, GF components are correlated with each other. In order to reduce dimensionality and de-correlate the components, we apply a DCT to a GF. We call the resulting coefficients gammatone frequency cepstral coefficients (GFCCs). Cepstral coefficients, C[j], j=0,........N-1 , are obtained from a GF as follows:

$$C[j]=\sqrt{\frac{2}{N}}\sum_{i=0}^{N-1}G[i]\cos\left(\frac{j\pi}{2N}(2i+1)\right)=0....N-1 \quad (3)$$

Note that the zeroth-order coefficient summates all the GF components. Thus, it relates to the energy of a GF vector. Rigorously speaking, the newly derived coefficients are not cepstral coefficients because a cepstral analysis requires a log operation between the first and the second frequency analysis for the deconvolution purpose [7]. Here, we call them ceptral coefficients because of the functional similarities between the above change and that of a ceptral analysis in the extraction of MFCC.

IV. Speaker Modeling

The GMM has been the predominant approach for speaker modeling for past decade [5]. In this work the GMM framework along with the universal background model(UBM) is adopted for speaker modeling [8].

Universal Background Model (UBM)
Two hypothesis tests are used for identification of speaker. The first test is the one in which the speech signal *Z* does come from the hypothesized speaker and the second one where it does not come from the hypothesized speaker.
The *likelihood* of the hypothesis H_i given the
speech signal can be defined as the probability density function *p* (*Z* | H_i*)*. Then likelihood ratio test given by the two hypotheses to determine the decision. For text independent speaker recognition the most successful model for the creation of likelihood ratio is the Gaussian mixture models (GMM). A GMM could be thought of as a Gaussian distribution describing a one dimensional random variable *X*. The variable *X* is defined as a vector described by the mean, variance and weight. The mixture density for a feature vector, *X* can be defined as:

$$P\left({}^{X}\!/_{\lambda}\right) = \sum_{i=1}^{M} W_i P_i(X) \tag{4}$$

The density is a weighted linear combination of M uni-modal Gaussian densities, P_i(X), each parameterized by a mean D*1 vector, □$_i$, and a D*D covariance matrix, Σ_i ;

$$P_i(X) = \frac{1}{2\pi^{D}/_{2}|\Sigma_i|^{1}/_{2}} exp\left\{\frac{-1}{2}(X-\mu)'(\Sigma_i)^{-1}(X-\mu_i)\right\} \tag{5}$$

The UBM is trained using the Expected- Maximization (EM) algorithm. The EM algorithm refines the parameters of the GMM iteratively to increase the likelihood of the estimated model for the feature vectors being observed.

Adaption speaker model
The speaker-specific model is adapted from the UBM using the maximum a posteriori (MAP) estimation. The adaptation increases the performance and provides a tighter coupling between the two models. According to the alignment of the training vectors to the UBM can be computed as follows

$$\mathrm{P_r}(i\ |x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_j)} \tag{6}$$

Finally, the new adapted mean, variance and wieght of the mixture i of training data is as follows :

$$\widehat{w_i} = [\alpha_i^{w}\, {}^{n_i}\!/_{T} + (1-\alpha_i^{w})w_i]\gamma \tag{7}$$

$$\widehat{\mu_i} = \alpha_i^{m} E_i(x) + (1-\alpha_i^{m})\mu_i \tag{8}$$

$$\hat{\sigma}_i^2 = \alpha_i^{v} E_i(x^2) + (1-\alpha_i^{v})(\sigma_i^2+\mu_i^2) - \hat{\mu}_i^2 \tag{9}$$

V. Experimental Setup And Results

The experiment are conducted on the English Language Speech Database for Speaker Recognition (ELSDR) databases, ELSDR database consists of 22 speaker samples out of which 12 are male and 10 are females samples. Length of training sample is taken to be 24 seconds whereas testing sample is taken to be of 6 seconds. Each training samples are added with Additive White Gaussian Noise (AWGN) at different signal to noise ratio (SNR) values in order to now the robustness of the speaker identification system.
We are using 32 gaussian components to train the input samples. The performance of the system is measured using speaker identification accuracy. Table.1 gives the performances of the system at clean test sample along with different SNR values. Fig.2 represents the graphs of identification accuracy v/s different SNR values.

Table.1 Accuracy At Different SNR Values
And Clean Sample

Testing Sample at Different SNR	Accuracy %
Clean sample	95.4545
0 dB	31.8182
5 dB	50
10 dB	77.2727
15 dB	86.3636
20 dB	86.3636
25dB	90.9091

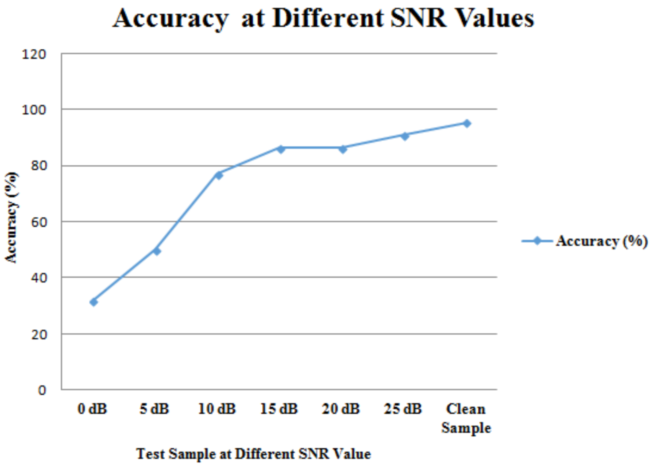

Fig. 2. Graph Of Identification Accuracy (%) V/S Testing Samples At Different SNR Values

VI. Conclusion

The results shows that the gammatone frequency cepstral coefficient (GFCC) along with gaussian mixture model - universal background model (GMM-UBM) has a good identification performance not only in clean speech environment but also in noisy condition. The performance of Speaker identification systems has improved due to recent advances in speech processing techniques but there is still need of improvement in noisy condition.

References

[1] J. P. Campbell, ``Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437--1462, September 1997.

[2] S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition," *Journal of the Acoustical Society of America*, pp. 354-358, 1963.

[3] E. Rosenberg and F. K. Soong, .Evaluation of a vector quantization talker recognition system in text independent and text dependent models,. Computer Speech and Lamguage, pp. 143-157,1987.

[4] D. A. Reynolds, “Speaker identification and verification using gaussian mixture speaker models,” *Speech Commun.*, vol. 17, pp. 91–108, 1995.

[5] M. Mclaren, R. Vogt, B. Baker and S. Sridharan, "A Comparison of Session Variability Compensation Techniques for SVM-Based Speaker Recognition," in *8th Annual Conference of the International Speech Communication Association*, Antwerp, Belgium, 2007.

[6] R. D. Patterson, J. Holdsworth, and M. Allerhand, “Auditory models as preprocessors for speech recognition,” in *The Auditory Processingof Speech: From Sounds to Words*, M. E. H. Schouten, Ed. Berlin, Germany: Mouton de Gruyter, 1992, pp. 67–83.

[7] V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

[8] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.