ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(19), DOI: 10.17485/ijst/2016/v9i19/93870, May 2016* ISSN (Online) : 0974-5645

Speaker Identification System using Gaussian Mixture Model and Support Vector Machines (GMM-SVM) under Noisy Conditions

R. Dhineshkumar[1*], A. Balaji Ganesh[2] and S. Sasikala[3]

[1]School of Computing Science and Engineering, Vellore Institute of Technology University, Chennai Campus, Vellore – 632014, Tamil Nadu, India; r.dhineshkumar2014@vit.ac.in
[2]TIFAC-CORE, Velammal Engineering College, Chennai - 600066, Tamil Nadu, India; abganesh@velammal.edu.in
[3]Department of Computer Science, Institute of Distance Education, University of Madras, Chennai - 600005, Tamil Nadu, India; sasikalarams@gmail.com

Abstract

Background: Automatic Speaker Identification (SID) systems has been a major breakthrough and crucial in many real-world applications. **Methods:** This work addresses the SID task based on GMM-SVM in a three stage process. Firstly, the Gammatone Frequency Cepstral Coefficients (GFCC) and Mean Hilbert Envelope Coefficients (MHEC) of the speakers are extracted. Secondly, these features are modeled using Gaussian Mixture Model (GMM), on adapting the extracted acoustic features by mean, the corresponding super vectors are found and these vectors are trained using Support Vector Machine (SVM). Finally, the actual recognition is done by feeding the super vectors of them asked noisy test utterance by Ideal Binary Mask (IBM) into SVM model and their accuracy of recognition is compared for GFCC, MHEC and RASTA-MFCC in different noisy conditions. **Findings:** Evaluation results show that SID performance carried out with MHEC is extensively better than the performance of other two features. **Applications:** Major areas that implements automatic SIDs are forensics, surveillance and audio biometrics etc.

Keywords: GMM-SVM, Gamma tone Frequency Cepstral Coefficients, Ideal Binary Mask, Mean Hilbert Envelope Coefficients, Robust Speaker Identification

1. Introduction

Given an identity and a speaker model, the goal of SID is to determine whether the claim is true or false. Significant research has been carried out in improving the robustness of text independent automatic SID's performance in noisy and reverberant conditions over past few decades.

The underlying factor for SID task is to extract the right feature which captures the complete phonetically important characteristics of the speech utterance to the fullest. MFCCs[1] and Perceptual Linear Predictive (PLP) coefficients are the extensively adopted features for SID applications, since spectral features are more precise than temporal features. However, the MFCC based SID systems are prone to recognition mismatch in noisy and reverberant conditions. Recently, GFCC[2] and MHEC[3] features have shown that the spectrum estimation carried out in these features are robust to background noise or reverberation.

Typically the speaker verification task is done based on Gaussian Mixture Model[4] (GMM). This approach of speaker recognition is carried out by training the GMM and creating a Universal Background Model (UBM) from the training speech signals via Expectation Maximization (EM) algorithm. The hypothesized speaker-specific model is then framed by updating the well-trained parameters in the UBM via Bayesian learning or Maximum A Posteriori (MAP) adaptation[5]. The verification is done based on log-

* *Author for correspondence*

likelihood ratios of observed features of given speaker model. This method of verification has proven extremely successful in SID systems in clean conditions. However, if noise is taken into considerations, they yield poor identification performance.

Considering the daily acoustic environments, room reverberation, additive noise and channel or handset variations that combine to pose considerable challenges to SID systems, several methods like cepstral mean normalization[6], RASTA filtering[7], speech segregation and speech enhancement method such as spectral subtraction have been examined, motivated by auditory masking phenomena[8] the concept of ideal binary mask IBM is introduced. IBM typically isolates speech from background noise by producing a binary Time-Frequency (T-F) mask that determines whether a specific T-F unit is speech dominant or noise dominant. We employed this mask in our work to segregate speech signal from noise signal which are mixed artificially.

Recently Support Vector Machines (SVM) has proven to be an effective and novel method for speaker recognition[9]. SVM perform a non-linear mapping from an input space to a support vector feature space then linear classification techniques are applied in potentially high-dimensional space. The basic working of this method is by using latent factors, the MAP adapted means of a GMM are modeled to describe variation. A key feature of this approach is to use the GMM super vector consisting of the stacked means of the mixture components. Super vectors can be used to characterize the speaker and channel using Eigen voices and Eigen channels methods respectively. This representation of a speech utterance using a single vector effectively replaces the conventional computationally demanding data to model type of identification with the training speaker model.

The rest of the paper is organized as follows. The components of the GMM-SVM based SID system is described in Section 2. Section 3 describes the auditory feature extraction methods and mask estimation. Section 4 discusses the GMM super vectors. SID experiments and evaluations are presented in Section 5.

2. System Overview

The block diagram of the GMM-SVM based speaker recognition is shown in Figure 1. Feature space of speech utterances are trained using GMM and UBM is constructed for all speakers in the training data. SVM model is constructed for the super vectors obtained from adapting the features of the enrollment data with the UBM. The super vectors of the test utterances are then classified using the SVM model created earlier.

3. Feature Extraction Methods

3.1 GFCC Feature Extraction

The Gammatone Frequency Cepstral Coefficients[9,10] are extracted by using a 64 channel Gammatone filter bank with central frequencies ranging from 50 Hz to 8000 Hz equally spaced on Equivalent Rectangular Bandwidth (ERB) scale, the Gammatone filter bank basically is a series of band pass filters described by an impulse response. The rectified outputs are decimated into time frames of 10 ms (100 Hz along the time dimensions). The loudness of the decimated outputs is then compressed by a cubic root operation. The resultant matrix represents T-F decomposition of the input and referred as GF feature

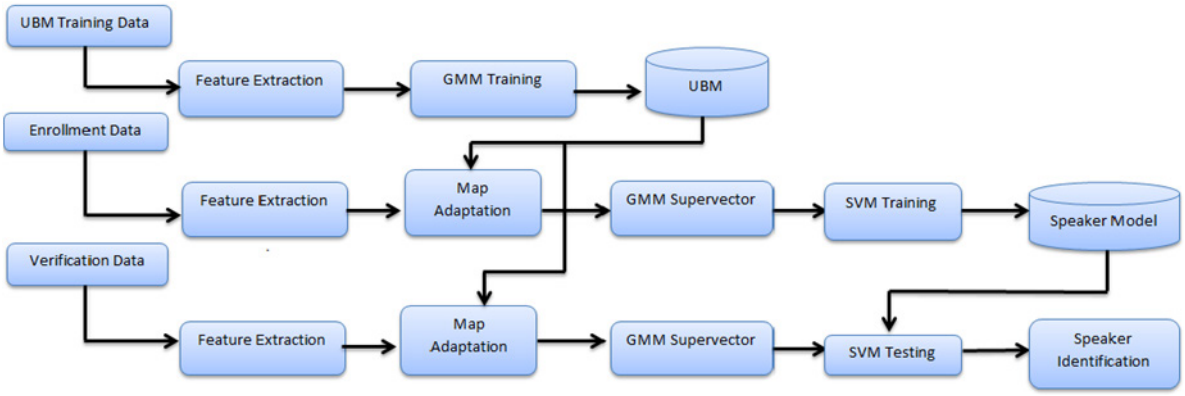

Figure 1. GMM-SVM based speaker Identification system.

components which are correlated with each other. In order to de-correlate and reduce the dimensionality, we apply Discrete Cosine Transform (DCT) to get GFCC.

3.2 MHEC Feature Extraction

Mean Hilbert Envelope Coefficients extraction[3] is done by pre-emphasizing the speech of sample frequency F_s(8000 Hz) signal with a high-pass filter and passed through a 24 channel Gammatone filter banks with central frequencies ranging from 300 Hz to 3400 Hz are equally spaced in ERB scale. Next, to find the temporal envelope, we apply Hilbert transform to each channel of the filter responses. Consider *S(t, i)* being the response from the filter bank and the analytical signal formed $S_a(t, i)$ from transform is:

$$S_a(t,i) = S(t,i) + \hat{i} * S'(t,i) \quad (1)$$

Where *S′(t, i)* the Hilbert is transform of *S(t, i)* and $\hat{i}$ is the imaginary unit. Now the envelope can be obtained as:

$$E(t, i) = (S(t, i))^2 + (S'(t, i))^2 \quad (2)$$

Where *E(t, i)* is the Hilbert Envelope of *S(t, i)*. The enveloped signal is then smoothed to suppress the redundant high-frequency components using a low-pass filter with a cut-off frequency f_c = 20 Hz.

$$E_s(t, i) = (1-\mu)\, E(t, i)) + \mu * E_s(t-1, i) \quad (3)$$

Where μ is the smoothing factor and it is related to cut-off frequency as:

$$\mu = \exp\left(-\frac{2\pi f_c}{F_s}\right) \quad (4)$$

A Hamming Window is applied to each decomposed frames of smoothed Hilbert envelope having 25 ms duration with a skip rate of 10 ms in order to minimize the discontinuity around the edges.

The mean envelope amplitude for a frame *p* can be obtained from the following equation:

$$S(p,i) = \frac{1}{N}\sum_{t=1}^{N-1} w(t) * E_s(t,i) \quad (5)$$

Here *w(t)* is the Hamming window and *N* is the frame size in samples. *S(p, i)* represents short term spectral nature of the speech signal. To suppress the dynamic range of the spectral attributes, natural logarithm is applied and DCT is then applied to convert spectral parameters to cepstral feature and the overlapping vectors are decorrelated. Thus formed feature is called Mean Hilbert Envelope Coefficients. To capture the dynamic pattern of the speech we append the first and second temporal cepstral derivatives of the first 12 coefficients of the static feature resulting in 36-dimensional feature.

3.3 Mask Estimation

The important goal of Computational Auditory Scene Analysis (CASA) is Ideal Binary Mask[11] (IBM), where each element corresponds to a T-F unit in the Cochleagram[12], a cochleagram is a T-F representation of a signal as show in Figure 2. With such a representation, a binary T-F mask delivers the vital information about whether a specific T-F unit is speech-dominant or noise. IBM defined as follows:

$$IBM(t, f) = \begin{cases} 1, if\ SNR(t, f) > LC \\ 0, \quad otherwise \end{cases} \quad (6)$$

IBM(t, f) is indexed by time *t* and frequency *f*. *SNR(t, f)*. refers to the local *SNR* value of the IBM and *LC* denotes the threshold for *SNR(t, f)* called local criterion.

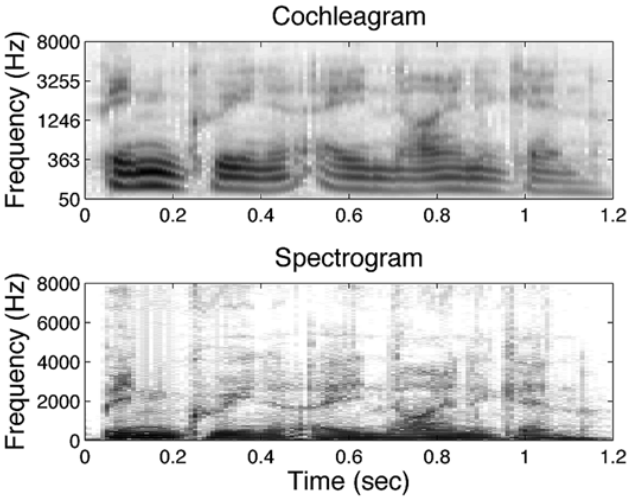

Figure 2. T-F representation of cochleagram (top) and spectrogram (bottom) on a clean speech signal[12].

4. GMM Super vectors

Consider a Gaussian Mixture Model-Universal Background Model (GMM-UBM).

$$g(x)=\sum_{i=1}^{N}\lambda_i N(x,m_i,\Sigma_i) \quad (7)$$

Where $N()$ is a Gaussian, λ_i, m_i, Σ_i are the mixture weights, mean and covariance of the Gaussians respectively.

Given a sample utterance of a speaker, and GMM UBM[4], the GMM supervector[13] is generated by adapting the extracted acoustic feature from the input sample with GMM UBM model by mean using MAP adaption algorithm. The adapted mean vectors are stacked into one another to produce the super vector and it is a mapping between an utterance and a high and fixed dimensional vector.

5. Experimental Setup

The speaker recognition experiments are conducted with the TIMIT phone labelled database corpus[14], which consists of 6300 sentences, ten from each of 630 speakers from eight dialectical regions of the United States. Each speaker has 10 utterances and we chose 9 for training and 1 for testing. The implementation is carried out with Matlab R2015a.

The GFCC feature vectors are extracted for the training speech utterances ,we use only the first 22 GFCC vectors from the 64- dimensional extracted output, as these lower order coefficients depicts nearly all the GF components found before applying DCT in the extraction process. These coefficients are found for each of the speakers in the corpus for training and GMM-UBM model[4]is constructed with 32 Gaussian mixtures. We adapt the extracted features of each speaker with GMM-UBM by mean using MAP adaptation with relevance factor of 16 and corresponding super vectors are found. Now these super vectors along with the labels for each speaker in the data set are fed into multi-class SVM to create a model. Lib-SVM is used for this purpose[15].

The test utterances of randomly selected 10 speakers from the database are mixed with babble noise, street noise, car noise, airport noise and exhibition noise as interfering signals from the Noizeusdatabase[16] to study the performance of the SID system under different types of noisy conditions. Individual noise is mixed with test utterances at various SNR levels from -10 dB to 5 dB at 5 dB intervals and from 10 dB to 30 dB at 10 dB intervals.

Given the mixed target and interference signals, the IBM is constructed with LC set at 0 dB to indicate the source is stronger, and them asked signal extracted from IBM contains the properties of the near-original speech signal. GFCC feature is then extracted for this signal and adapted with the GMM UBM by mean; super vector for this test signal is found then classified with the SVM model generated earlier.

For a meaningful comparison, we extract RASTA-MFCCs as acoustic features from speech signals for recognition purpose. Unlike conventional MFCC extraction[1,17] processes, the speech signal is first pre-emphasized using RASTA filtering[7], to suppress any constant or slowly varying components in the sample and windowed using Hamming window. Followed by windowing, Fast Fourier Transform (FFT) and logarithmic 26-channel Mel-Scale filter bank is applied for each windowed frame correspondingly. Finally, DCT is applied to the outputs and first 13-dimensional cepstral features excluding the 0^{th} coefficient is extracted.

The detailed analysis of the differences between RASTA-MFCC, GFCC and MHEC can be seen in Table 1. The noticeable difference is that the frequency scale employed in these individual features. MHECs and GFCCs are equally spaced on ERB scale while the RASTA-MFCCs on a Mel-Scale. In addition, the Non-linear rectification methods engaged in MHECs and RASTA-MFCCs is logarithmic operation, which transforms convolution between exciting source and filter as an additive term to the spectral domain unlike in GFCCs that uses Cubic root operation. Besides these other differences like Pre-emphasis and Number of frequency bands used are described in Table 1.

Table 1. Differences between RASTA-MFCC, GFCC and MHEC

Category	GFCC	MHEC	RASTA-MFCC
Pre-emphasis	No	Yes	Yes
No. of Frequency bands	64	24	26
Frequency Scale	ERB	ERB	Mel-Scale
Non-Linear Rectification	Cubic root	Logarithmic	Logarithmic

The same setup is carried out for 36-dimensional MHEC and 13-dimensional RASTA-MFCC features for recognition and their accuracy is compared.

Table 2. SID performance (%) based on RASTA-MFCC, GFCC and MHEC features for various noisy conditions

Babble Noise	-10 dB	-5 dB	0 dB	5 dB	10 dB	20 dB	30 dB	Average
RASTA-MFCC	23.06	25.65	72.34	69.14	89.53	94.20	96.50	67.20
GFCC	23.50	28.87	70.12	73.52	92.61	98.00	98.16	69.25
MHEC	25.73	27.52	75.04	86.53	94.34	98.38	99.00	72.36

Street Noise	-10 dB	-5 dB	0 dB	5 dB	10 dB	20 dB	30 dB	Average
RASTA-MFCC	23.51	35.07	70.92	81.53	91.53	93.95	95.76	70.32
GFCC	26.12	32.26	64.95	73.03	90.62	95.73	96.72	68.49
MHEC	26.34	45.37	71.45	83.74	90.87	96.59	98.57	73.27

Car Noise	-10 dB	-5 dB	0 dB	5 dB	10 dB	20 dB	30 dB	Average
RASTA-MFCC	26.86	50.05	68.07	70.45	75.62	81.52	97.13	67.10
GFCC	30.35	50.23	68.15	74.03	85.87	90.39	96.13	70.73
MHEC	35.02	56.35	70.52	79.10	88.64	91.32	94.27	73.60

Airport Noise	-10 dB	-5 dB	0 dB	5 dB	10 dB	20 dB	30 dB	Average
RASTA-MFCC	11.15	37.52	50.42	70.75	80.06	90.13	96.94	62.42
GFCC	11.31	42.03	58.62	74.05	82.31	92.15	98.06	65.50
MHEC	27.92	45.54	70.36	80.47	84.45	95.56	97.90	71.74

Exhibition Noise	-10 dB	-5 dB	0 dB	5 dB	10dB	20dB	30dB	Average
RASTA-MFCC	22.36	37.06	72.63	81.53	82.33	91.17	95.35	68.91
GFCC	24.32	37.12	70.16	79.14	84.12	85.41	96.83	68.16
MHEC	26.64	40.64	71.36	86.26	90.33	94.39	97.02	72.37

As it can be seen from Table 2 it is observed that on an average MHEC and GFCC feature based SID performance under various noisy conditions outperforms the SID based on MFCC feature. The performance gain over MFCC by GFCC is smaller as compared to performance gain achieved by MHECs, reflecting its robustness under various noisy conditions. From the Figure 3 it is evident that the performance of MHEC yields better performance results as compared to RASTA-MFCC and GFCC.

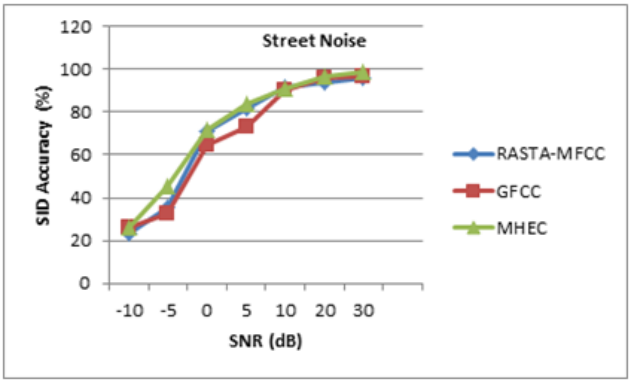

Figure 3. SID accuracy (%) comparison for different features under street noise.

6. Conclusion

In this work, the GMM-SVM based SID performances for three acoustic features namely RASTA-MFCC, GFCC, and MHEC in different noisy conditions at various SNR levels are examined and compared with each other. Segregation of added noise from the target speech through IBM is investigated and from the acquired results it is apparent that SID performance carried out with MHECs betters the performance of other two competitive acoustic features.

7. Acknowledgement

I would like to dedicate this work to my High-School teachers Mr. Sasanka Sekhar Dash, Mr. Gurucharan Singh Makhija, Mrs. Sundari. I would like to thank Mr. R. Venkatesan for his valuable ideas and support.

8. References

1. Dave N. Feature extraction methods LPC, PLP and MFCC in speech recognition. International Journal for Advance Research in Engineering and Technology. 2013; 1(6):1-4.
2. Shao Y, Wang DL. Robust speaker identification using auditory features and computational auditory scene analysis. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2008. p. 1589-92.
3. Sadjadi SO, Hansen JHL. Mean Hilbert Envelope Coefficients (MHEC) for robust speaker and language identification. Speech Communication. 2015; 72:138-48.
4. Reynolds DA, Rose RC. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing. 1995; 3(1):72-83.
5. Gauvain J-L, Lee C-H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing. 1994; 2(2):291-8.
6. Sadaoki F. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech and Signal Processing. 1981; 29(2):254-72.
7. Hynek H, Morgan N. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing. 1994; 2(4):578-89.
8. Moore BCJ. An introduction to the psychology of hearing. Academic San Diego. 1997; 313:159-67.
9. Shao Y, Jin Z, Wang DL, Srinivasan S. An auditory-based feature for robust speech recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2009. p. 4625-8.
10. Zhao X, Wang Y, Wang DL. Robust speaker identification in noisy and reverberant conditions. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014; 22(4):836-45.
11. Wang, DL. On ideal binary mask as the computational goal of auditory scene analysis. Speech Separation by Humans and Machines. US: Springer; 2005. p. 181-97.
12. Zhao X, Shao Y, Wang DL. CASA-based robust speaker identification. IEEE Transactions on Audio, Speech and Language Processing. 2012; 20(5):1608-16.
13. Campbell WM, Sturim DE, Reynolds DA, Solomonoff A. SVM based speaker verification using a GMM super vector kernel and NAP variability compensation. IEEE Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2006. p. I-I.
14. Garofolo J, et al. TIMIT acoustic-phonetic continuous speech corpus LDC93S1. Philadelphia: Linguistic Data Consortium; 1993.
15. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM TIST. 2011; 2(3):27.
16. Loizou PC. Subjective evaluation and comparison of speech enhancement algorithms. Speech Commun. 2007; 49:588-601.
17. Kari B, Muthulakshmi S. Real time implementation of speaker recognition system with MFCC and neural networks on FPGA. Indian Journal of Science and Technology. 2015; 8(19):1.