Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 30 (2012) 319 – 326

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Communication Technology and System Design 2011

Fused Mel Feature sets based Text-Independent Speaker Identification using Gaussian Mixture Model

R.Shantha Selva Kumari [a], S. Selva Nidhyananthan [ba*], Anand.G [c]

[a,b]*Department of ECE, Mepco Schlenk Engineering College , Sivakasi -626005, INDIA*
[c]*Department of ECE, PSRR College of Engineering for Eomen, Sevalpatti- 626140 , INDIA*

Abstract

This paper provides an efficient approach for text-independent speaker identification using a fused Mel feature sets and Gaussian Mixture Modeling (GMM). The individual Gaussian components of a GMM are shown to represent some speaker specific spectral shapes that are effective for modeling speaker identity. Two different set of features which are complement to each, other, Mel Frequency Cepstral Coefficient (MFCC) and Inverted Mel Frequency Cepstral Coefficient (MFCC) features are obtained for each speaker and are trained using Expectation Maximization algorithm. Two GMM models; one for MFCC feature sets, other one for IMFCC feature sets are created. During testing phase, the likelihood of unknown speaker's features with each of the GMM models is determined. By using a weighted sum of these likelihood values, a fused score is created for each speaker and speaker with maximum score is the identified speaker. The performance of this fusion GMM system is evaluated using a part of the TIMIT database consisting of 120 speakers and a maximum identification efficiency of 93.88% is achieved.

Keywords: Speaker Identification; Filter Bank design; Mel Frequency Cepstral Coefficients (MFCC); Inverted Mel Frequency Cepstral Coefficients (IMFCC); Gaussian Mixture Modeling (GMM); Fusion. Fusion.

1. Introduction

In the modern world technology, speech based techniques are being widely used in many applications, ranging from voice recognition in ordinary PCs to biometric and forensic applications. The major speech based techniques that rule these applications are speech recognition and speaker recognition. Speaker recognition uses the acoustic features of speech that have been found to differ between individuals. The vocal tract shape of each person is generally is an important physical distinguishing factor of speech. The

* S. Selva Nidhyananthan. Tel.: +91-04562-235409; fax: +91-04562-235111.
E-mail address: nidhyan@mepcoeng.ac.in.

1877-7058 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2012.01.867

vocal tract shape can be estimated from the spectral shape of the voice signal. These acoustic patterns reflect both anatomy (e.g., size and shape of the throat and mouth) and learned behavioral patterns (e.g., voice pitch, speaking style). Speaker recognition is classified into two major categories namely speaker identification and speaker verification. A speaker verification system must confirm a declaration of identity claim whereas a speaker identification system must identify an unknown speaker from a finite set of registered speakers i.e. the system must answer the question: "Who has spoken the given utterance?" The text-independent Speaker Identification systems are most commonly used for speaker identification because they require very little cooperation by the speaker. Text-independent speaker identification system can be implemented using any of the following techniques: Hidden Markov Model (HMM) [1], Gaussian Mixture Model (GMM) [2], Vector Quantization (VQ) [3], Neural Networks (NN) [4] and Discrete Wavelet Transform (DWT) [5]. In this paper Gaussian Mixture Model is used for speaker identification. Gaussian mixture density provides smooth approximation to the sample distribution of observations obtained from utterances of a given speaker.

Several speech corpuses such as TIMIT, YOHO are readily available for speaker identification process. Pre-processing of speech signals involve noise removal, silence removal, pre-emphasize, framing and windowing to capture the dynamic characteristics of vocal tract system [6], [7]. The speaker identification performance improves significantly when complementary information are fused with either by simple concatenation or by combining models' scores. The examples of such complementary information are pitch [8], residual phase [9], prosody [10], dialectical features [11] etc. In [12], it has been shown that complementary information can be captured easily from the high frequency part of the energy spectrum of a speech frame via reversed filter bank methodology. In this project, a feature set called IMFCC is used to capture speaker specific information lying in higher along with capturing speaker specific information lying in lower frequency part of the spectrum by MFCC.

This paper is organized as follows. In section 2, pre-processing and feature extraction are described. In section 3, a brief overview of Gaussian Mixture Model (GMM) is given. The implementation and results are discussed in section 4 and concluding remarks are given in section 5.

2. Filter bank Design and Feature Extraction

2.1. Feature extraction

Features are the representatives of the speech signal about speaker specific information. The human perception of the frequency contents of sounds follows a subjectively defined nonlinear scale called the Mel scale [13]. The Mel scale is linear below 1KHz and logarithmic above 1KHZ [14]. The linear scale to Mel scale conversion is defined as,

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

where, f_{mel} is the subjective pitch in Mels corresponding to f, the actual frequency in Hz.
Two types of features are used in this project: Mel Frequency Cepstral Coefficients (MFCC), which describes the vocal tract characteristics and easy to extract and Inverted Mel Frequency Cepstral Coefficients (IMFCC), which captures the complementary information present in the high frequency part of the spectrum. The steps involved in determining these coefficients are illustrated in Fig. 1.

2.2. Filter bank Design

Let $\{y(n)\}$, n=1.....M, represent a frame of the preprocessed signal. First, $y(n)$ is converted to the frequency domain by an Ms point DFT which leads to an energy spectrum. This is followed by the construction of filter banks (Mel Scale and Inverted Mel Scale), each with unity height Gaussian filters The Gaussian MFCC filter bank is designed using the following expression:

$$\psi_i^{gMFCC} = e^{\frac{-(k-k_{bi})^2}{2\sigma_i^2}} \quad (2)$$

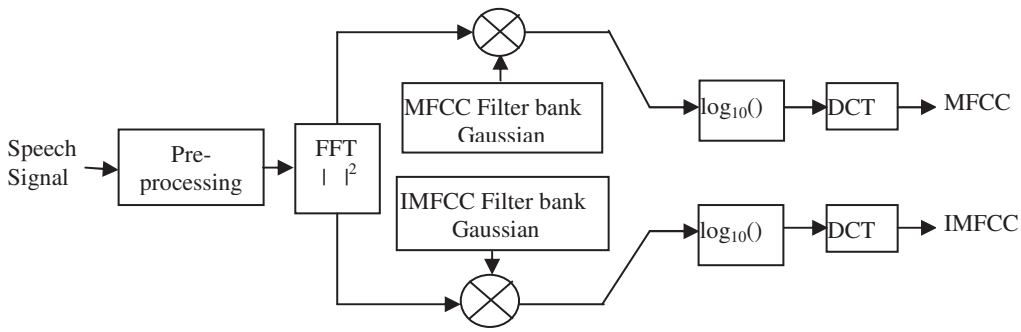

Fig .1 Steps involved in extraction of Gaussian MFCC and IMFCC features

In (2) σi is the standard deviation or square root variance and can be defined as,

$$\sigma_i = \frac{k_{b_{i-1}} - k_{b_i}}{\alpha} \quad (3)$$

where α is the parameter that controls the variance. The term k_{b_i} is a point between the i^{th} Triangular Filter's boundaries located at its base, as shown in Fig. 2(a), and considered as the mean of the i^{th} Gaussian Filter. Gaussian Mel scale filter bank (for MFCC) and the Inverted Gaussian Mel scale filter bank (for IMFCC) are shown in Fig.2 (b) and 2(c) respectively.

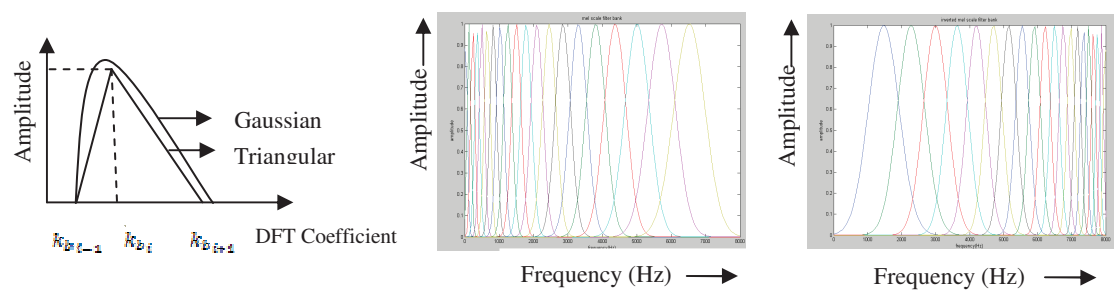

Fig .2 (a) Filter Bank Design, (b) Mel scale filter bank and (c) Inverted Gaussian Mel scale filter bank

In other words, we derive the Gaussian Filter bank from the boundary points of a Triangular filter bank. These boundary points can be using the expression,

$$k_{b_i} = \left(\frac{M_s}{F_s}\right) f_{mel}^{-1}\left[f_{mel}(f_{low})\right] + \frac{i\left(f_{mel}(f_{high}) - f_{mel}(f_{low})\right)}{Q+1} \quad (4)$$

where *Ms* is the number of points in the DFT, Fs is the sampling frequency, *flow* and *fhigh* are the low and high frequency boundaries of the filter bank, Q is the number of filters in the bank and f_{mel}^{-1} is the inverse of the transformation in (4) defined as,

$$f_{mel}^{-1}(f_{mel}) = 700\left[10^{\frac{f_{mel}}{2595}} - 1\right] \quad (5)$$

The Inverted Mel scale filter bank structure can be obtained by simply flipping the original filter bank around the midpointof the frequency range that is being considered. This flip-over is expressed mathematically as,

$$\psi_i'(k) = \psi_{Q+1-i}\left(\frac{M_s}{2} + 1 - k\right) \quad (6)$$

where, $\psi_i\ (k)$ is the response of the original MFCC filter bank.
These filter banks can be imposed on the spectrum obtained by taking FFT of the pre-processed signal as follows:

$$e(i) = \sum_{k=1}^{\frac{M_s}{2}} |Y(k)|^2 . \psi_i(k) \quad (7)$$

Finally, DCT is taken on the log filter bank energies $\left\{\{\log[e(i)]\}_{i=1}^{Q}\right\}$ and the final MFCC coefficients C_m can be written as,

$$C_m = \sqrt{\frac{2}{Q}} \sum_{i=0}^{Q-1} \log[e(i+1)] . \cos\left[m . \left(\frac{2l-1}{2}\right) . \frac{\pi}{Q}\right] \quad (8)$$

where $0 \le m \le R-1$, and R is the desired number of cepstral features. The same procedure can be used to extract IMFCC features as well.

3. Gaussian Mixture Model

One way to represent the variability in speech is probabilistically through multidimensional Gaussian pdf [15]. A different Gaussian pdf is assigned for each acoustic class. The Gaussian pdf of a feature vector for ith state is given by

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_i)'\Sigma^{-1}i(\vec{x}-\vec{\mu}_i)\right) \quad (9)$$

where, μ_i is the mean vector, $\sum_i$ is the covariance matrix, D is the dimension of the vector.
The probability of feature vector in any one of M acoustic class for a particular speaker model λ is represented by the union or mixture of different Gaussian pdf [15]. This is represented as

$$p\left(\frac{\vec{x}}{\lambda}\right) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad (10)$$

where, $\vec{x}$ is a D dimensional random vector, $b_i(\vec{x})$, i=1....M, are the component densities and pi, i=1....M, are the mixture weights.

The complete Gaussian mixture density is parameterized by [11]. posterior probability for acoustic class i is given by [12].

$$\lambda = \left(p_i, \vec{\mu}_i, \sum_i\right) \; i = 1,...,M \quad (11)$$

$$p\left(i/\vec{x}_t, \lambda\right) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \quad (12)$$

For speaker identification, each speaker is represented by a GMM and is referred by his/her model λ. The diagonal covariance matrices are sufficient for good approximations and to reduce the number of unknown variables to be estimated.

3.1. Maximum Likelihood Parameter Estimation

The aim of ML estimation [17] is to find the model parameters which maximize the likelihood of GMM. For a sequence of training vectors $x = (\vec{x}_1, \vec{x}_2, ..., \vec{x}_t)$, the GMM likelihood can be written as

$$p\left(\frac{x}{\lambda}\right) = \Pi_{t=1}^{T} p\left(\vec{x}_t/\lambda\right) \quad (13)$$

This expression is a nonlinear function of the parameters λ and so direct maximization is not possible. The ML parameterestimate is obtained iteratively using Expectation Maximization algorithm. The parameter is obtained at the convergence point of the algorithm.

3.2. Expectation Maximization (EM) Algorithm

The EM algorithm begins with an initial model λ to estimate a new model $\lambda 1$ .The new model then becomes the initial model and the process is repeated till convergence. A proper initialization must be

done for model parameter. For mean vector, randomly selected vectors from training data are chosen. We need to initialize the covariance matrix with large variance, to reduce the risk that the EM training gets struck in some local maximum. The mixture weights for all values of *I* is chosen as *1/M*. Number of mixture components should be neither too few nor too many.

3.3. Speaker Identification

Two separate models are created for each speaker during the training phase from the MFCC and IMFCC feature sets respectively, using GMM. During the test phase, MFCC and IMFCC features are extracted in a similar way from the test speech utterance as done in the training phase and are sent to their respective models. For each speaker, two scores are generated, one each from the MFCC and IMFCC models. A uniform weighted sum rule is adopted to fuse the scores from the two classifiers. A governing equation which describes fusing outputs of parallel classifiers methodology via weighted sum rule is expressed as follows:

$$S_{com}^{i} = w\sum_{t=1}^{T} \log p\left(x_{tMFCC} \middle| \lambda_{MFCC}\right) + (1-w)\sum_{t=1}^{T} \log p\left(x_{tIMFCC} \middle| \lambda_{IMFCC}\right) \tag{14}$$

The value of the weight w is chosen iteratively. The identity of the true speaker i_{true} is given by

$$t_{true} = \arg\max i\left(S_{com}^{i}\right) \tag{15}$$

4. Implementation and Results

The sequence of the steps involved in speaker identification system implemented in this work is clearly illustrated in Fig. 3. It can be seen from the figure that training and testing are done twice, one each for MFCC and IMFCC feature set. The implementation is started by preparing the database. The TIMIT database consists of 6300 signals, 10 signals spoken by each of 630 speakers. The speakers are spread among eight different directories based on their dialect region. For this system, speech signals from 120 speakers from the dialect region 1 & 4 of the TIMIT database are used. For the training phase, six out of the ten signals per speaker are used. For testing phase, the remaining four signals are used. The signals in the TIMIT database consist of an initial silence portions. The maximum and minimum amplitude out of the first 20 samples of this portion is determined. Samples whose amplitude is between one fourth of the maximum and one fourth of the minimum are considered as silence and are omitted. Next, the signal is pre-emphasized with a pre-emphasis factor equal to 0.96. The pre-emphasized signal is windowed using a 16 ms Hamming window with 8 ms frame shift.

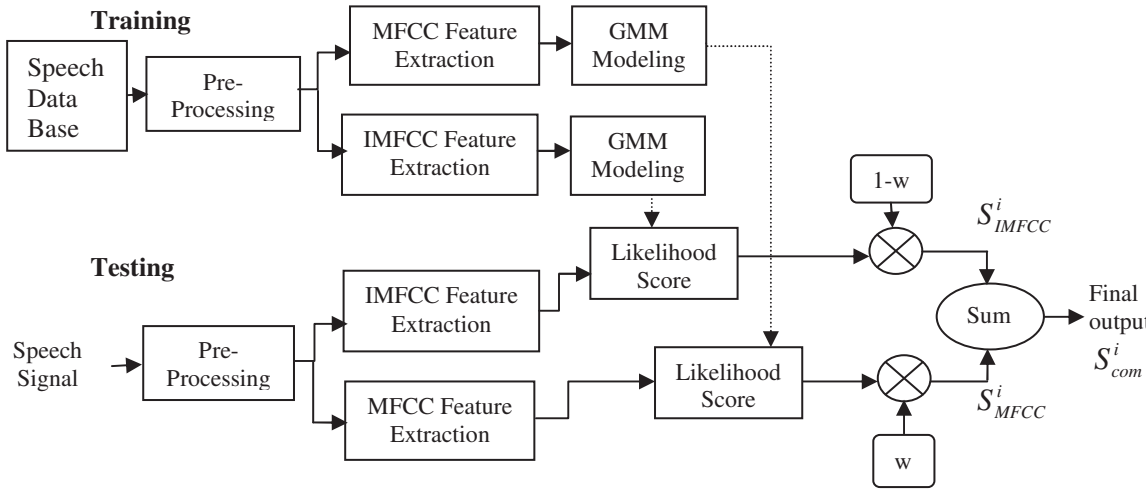

Fig .3 Steps involved in Speaker Identification System

The features are obtained by using two types of filter banks namely the Gaussian Mel scale filter bank (for MFCC) and the Inverted Gaussian Mel scale filter bank (for IMFCC). The feature vectors obtained are trained using Expectation Maximization Algorithm. A separate model for each speaker is created. In the testing phase, features are extracted from the incoming test signal and then the likelihood of these test features with each of the speaker models is determined. Thus we will have two likelihood values for each speaker, one each for MFCC and IMFCC. By using a weighted sum of these likelihood values as indicated earlier, a fused score is created for each speaker and the speaker with maximum score is identified as the correct speaker. In addition to this, the identification performance is tested separately for MFCC and IMFCC systems individually. The usage of weights in the fusion scheme can be changed to different values to test the system for optimum results. Here, we have used four different weights. The Table 1 shows the performance level of the proposed system for different number of mixtures and different weights.

From the table, it can be seen that the maximum performance is achieved at 16 mixtures. Also, it can be noted that the performance of the fused system exceeds the performance of the individual MFCC and IMFCC systems. The percentage of maximum performance is 93.88% and hence good identification with limited errors has been obtained.

TABLE 1
Performance of the Proposed System for different Number of Mixtures

No. Of Mixtures	GMM					
			FUSED			
	MEL	IMEL	w= 0.5	W=0.7	w = 0.77	w = 0.8
8	67.35	55.10	79.59	88.2	89.8	89.8
16	74.36	58.97	87.75	90.31	**93.88**	91.84
32	71.43	53.06	83.67	89.15	91.84	91.84

5. Conclusions

In this work, the use of Gaussian Mixture Model is evaluated for fusion based speaker identification. Higher performance level is achieved by fusing complementary information, present in the high frequency region of the spectrum, with the usual MFCC coefficients. Higher accuracy has been achieved while using 16 mixtures. The performance has been tested with different weights and the maximum identification achieved is 93.88% for a weight of w = 0.77. Further enhancements can be investigated by changing the number of filters in the filter banks, length of the frames and by using other different weighting combinations.

References

[1] Chang-Hoon Lee and Soo-Young Lee, "Noise-Robust Speech Recognition Using Top-Down Selective Attention with an HMM Classifier", *IEEE Signal Processing Letters*, July 2007, Vol. 14, No. 7.

[2] Douglas A. Reynolds and Richard C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Transactions on Speech and Audio Processing*, January 1995, Vol. 3, No. 1.

[3] Jialong He, Li Liu and Palm.G, "A discriminative training algorithm for VQ-based speaker identification," *IEEE Transactions on Speech and Audio Processing*, May 1999, Vol. 7, No. 3.

[4] R.V. Pawar, P.P. Kajave, and S.N. Mali, "Speaker Identification using Neural Networks**",** World *Academy of Science, Engineering and Technology*, December 2005.

[5] Ching-Tang Hsieh, Eugene Lai and You-Chuang Wang, "Robust Speaker Identification System Based on Wavelet Transform and Gaussian Mixture Model", *Journal of Information Science And Engineering* 19, 267-282 (2003).

[6] .Douglas O' Shaughnessy, "Speech communication Human and Machines", *IInd edition , Universities Press(India) Limited(*2001)

[7] B.Yeganarayana and P.Satyanarayana Murthy, "Source System Windowing For Speech Analysis And Synthesis", *IEEE Transactions on Speech And Audio Processing,* March 1996, Vol.4, No.2.

[8] Harrag A. Mohamadi., Serignat J.F., "LDA Combination of Pitch and MFCC Features in Speaker Recognition", *Proceedings of INDICON 2005*, 11-13 Dec., IIT Chennai, Indian, 2005, pp. 237-240,

[9] K. Sri Rama Murthy and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition", *IEEE Signal Processing Letters*, Jan. 2006, vol 13, No. 1, pp. 52-55.

[10] Yegnanarayana B., Prasanna S.R.M., Zachariah J.M. and Gupta C.S., "Combining evidence from source, suprasegmental and spectral features for a fixed-test speaker verification system", *IEE Trans. Speech and Audio Processing*, July 2005, Vol. 13, No. 4, pp. 575-582.

[11] Chakroborty S., Roy A., and Saha G., "Improved Closed set Text- Independent Speaker Identification by Combining MFCC with Evidence from Flipped Filter Banks" , *International Journal of Signal Processing*, 2007, Vol. 4, No. 2, Page(s):114-122.

[12] J.Kittler, M. Hatef, R. Duin, J. Mataz, "On combining classifiers", *IEEE Trans. Pattern Anal. Mach. Intell*. 20(1998) 226-239 .

[13] G.Senthil Raja, Dr.S.Dandapat, "Performance of Selective Speech Features for Speaker Identification", *Journal of the Institution of Engineers (India),* May 29, 2008, Vol. 89.

[14] Md.Rashidul Hasan,Mustafa Jamil Md.Golam Rabbani,Md.Saifur Rahman, "Speaker Identification using Mel Frequency Cepstral Coefficients", *3rd International conference on Electrical and computer engineering ICECE* 2004, December 2004.

[15] Ben Gold and Nelson Morgan, "Speech and Audio Signal Processing,Part-IV", *John Wiley & Sons*, 2002 , Chap. 14, pp. 189-203,

[16] Thomas F.Quatieri, "Discrete Time Speech Signal Processing Principles and Practice", *Pearson Education*, New Delhi, 2007.

[17] Douglas A. Reynolds, Richard C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Models" , IEEE Trans. on Speech and Audio Processing, January 1995, Vol. 3, No. 1.