Development and Implementation of Algorithm for Speaker recognition for Gujarati Language

Jigarkumar Patel[1], Arun Nandurbarkar[2]

[1] PG student, Electronics and Communication, L.D college of engg, Gujarat, India
[2] Associate Professor, Electronics and Communication, L.D college of engg, Gujarat, India

---

Abstract - *Modern speech understanding systems merge interdisciplinary technologies from Signal Processing, Pattern Recognition, Natural Language and Linguistics into a unified statistical framework. In this paper weighted MFCC(Mel frequency cepstral coefficients) and GMM(Gaussian Mixture Model) are implemented for Speaker Recognition in Gujarati Language. The experimental database consists of 30 speakers, 10 female and 20 male, collected in sound proof room. The result of this experiment certificates that this technique works better for speaker recognition for Gujarati language than only traditional MFCC with GMM.*

Key Words: speaker recognition; Mel frequency cepstral coefficients; feature extraction; weighted Mel frequency cepstral coefficients; Gaussian Mixture Model; maximum likelihood.

1. INTRODUCTION

Modern speech understanding systems merge interdisciplinary technologies from Signal Processing, Pattern Recognition, Natural Language and Linguistics into a unified statistical framework. These systems, which have applications in a wide range of signal processing problems, represent a revolution in Digital Signal Processing(DSP)[1][2]. Once a field dominated by vector-oriented processors and linear algebra bases mathematics, the current generation of DSP-based systems rely on sophisticated statistical models implemented using a complex software paradigm. Such systems now capable of understanding continuous speech input for vocabularies of several thousand words in operational environments.

Speech signal processing technology is an indispensable technology in the information society, and speaker recognition is an important research field of speech processing. Speaker recognition is also called the voiceprint recognition, which makes it possible to identify or verify the identity of the speaker using the speech feature. It combines the theories of various subjects, such as acoustics, phonetics, linguistics, physiology, digital signal processing, pattern recognition and artificial intelligence etc. Speaker recognition has a wide application prospect in the judicial identification, security Monitoring, e-commerce and other fields. The extraction of the Mel frequency cepstral coefficients is one of the popular approaches of feature extraction.

Speaker modeling is the main part of a speaker recognition system. The Gaussian mixture model (GMM) is the most common approach for speaker modeling in text-independent speaker recognition[4][5]. A general speaker recognition system, shown in Figure 1, consists mainly, of three stages, each stages are explained in next sections.

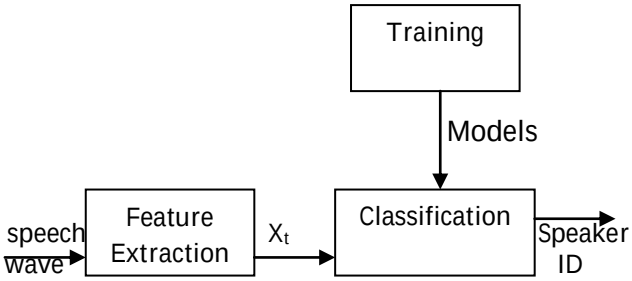

Figure 1. Speaker Recognition System[5].

2. The Feature Extraction

2.1 MFCC(Mel Frequency Cepstral Coefficients)

The purpose of feature extraction is to convert the speech waveform to a set of features for further analysis. Where appropriate information is estimated in a suitable form and size, from the speech signal to obtain a good representation of the speaker features, (Mel Frequency Cepstral Coefficients (MFCC features) are chosen in this paper because they are based on the perceptual characteristics of the human auditory system[4], figure 2 shows a block diagram of the steps in Mel feature extraction.

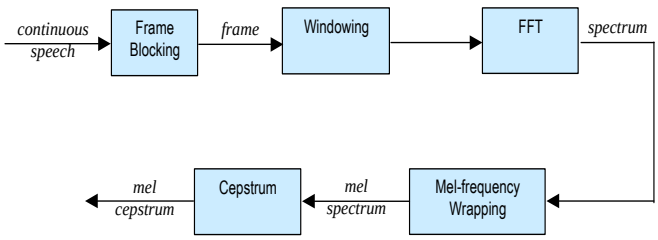

Figure 2. MFCC feature extraction block diagram

The input to the system is speech sampled at 11025Hz and converted to 16-bit digital format. The digital speech signal is then applied to a 256 samples (≈23ms) hamming window every 128 samples (≈12ms). Each individual frame is windowed so as to minimize the signal discontinuity at the beginning and end of each frame as shown in Figure 2. The Fast Fourier Transform (FFT) converts each frame of samples from the time domain into the frequency domain. The frequency scale is then converted from the hertz to the mel-scale using filter banks with frequency spaced linearly at low frequencies and logarithmically at high frequencies, and the logarithm is then taken. This stage is done in order to capture the phonetically important characteristics of speech in a manner that reflects the human perceptual system. The Discrete Cosine Transform (DCT) is then applied to the output to produce a cepstrum[4][5]. Therefore if we denote those mel power spectrum coefficients that are the result of the last step are $\tilde{S}_0, k = 0,2,..., K-1$, we can calculate the MFCC's, $\tilde{c}_n$, as

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \qquad n = 0,1,..., K-1$$

The first component m= 0 is excluded from the DCT since it represents the mean value of the input signal which carried little speaker specific information.

2.2 Weighted MFCC

The static characteristics of the voice are described by the Mel frequency cepstral coefficients. Changing voice signals are an important feature of speech signal, so we introduce the first order differential MFCC[3] parameters

$$d(n) = \frac{1}{\sqrt{\sum_{i=-k}^{k} i^2}} \sum_{i=-k}^{k} i \times c(n+i)$$

where *k* is a constant, usually take *k* =2 . Noise to some extent can be eliminated by the differential MFCC, so we can achieve the better performance.

3. Implementation based on Gaussian Mixture Model(GMM)

The GMM forms the basis for both the training and classification processes. The principle of GMM is to abstract a random process from the speech, then to establish a probability model for each speaker[4][5]. A Gaussian Mixture density is a weighted sum of M component densities as shown in figure 3.

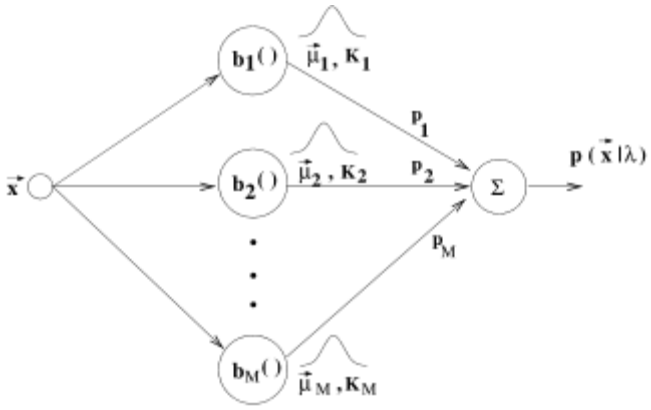

Figure 3. M probability densities forming a GMM

In the GMM model, the features distributions of the speech signal are modeled for each speaker as follows

$$P(x|\lambda) = \sum_{i=1}^{M} pi * bi(x)$$

where, $\sum_{i=1}^{M} pi = 1$

x is a random vector of D-dimension, $p(x/\lambda)$ is the speaker model; *pi* is the ith mixture weights; *bi*(x) is the ith pdf component that is formed by the ith mean μi and ith covariance matrix, where *i* = 1,2,3,...,*M*, and *M* is the number of GMM components, each density component is a D-variants Gaussian distribution given eq. below.

$$P(X|\mu,\Sigma) = \frac{1}{\sqrt[2]{2\pi}\,|\sum i|^{1/2}} e^{-(1/2)(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}$$

A statistical model for each speaker in the set is developed and denoted by λ. For instance, speaker *s* in the set of size *S* can be written as follows:

$$\lambda_s = \{p_i, \mu_i, \sum i\}$$, where i={1,...M} & s={1,...S}

3.1 ML Parameter Estimation

The ML(Maximum Likelihood) algorithm for GMM estimation[5] is shown in Figure 4.

To obtain an optimum model representing each speaker we need to obtain a good estimation of the GMM parameters. To this end, the Maximum-Likelihood Estimation (ML) approach, which is a very efficient method, can be used; where for a given of *T* vectors used for training, X=(*x*1,*x*2,...,*xT*), the likelihood of GMM can be written as.

$$p(X|\lambda_s) = \prod_{t=1}^{T} p(x_t|\lambda_s)$$

Since the GMM likelihood of the nonlinear function is impossible that maximizes directly, the ML estimations can be possible by using the EM algorithm iteratively.

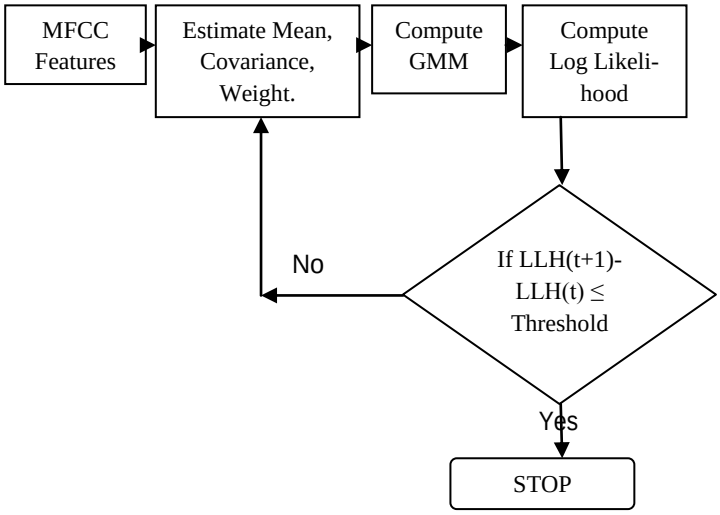

Figure 4. EM algorithm for GMM

The training phase consists of two steps, namely *initialization* and *expectation maximization (EM)*. The initialization step provides initial estimates of the means for each Gaussian component in the GMM model. The EM algorithm recomputed the means, covariances, and weights of each component in the GMM iteratively[5]. Each iteration of the algorithm provides increased accuracy in the estimates of all three parameters. The EM algorithm formulas are the following:

- weight

$$\mathbf{w} = \frac{1}{\mathbf{T}}\sum_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{p_i}\,(\mathbf{i}|\mathbf{x_t},\boldsymbol{\lambda})$$

- mean

$$\vec{\bar{\mu}} = \frac{\sum_{t=1}^{T} p(i|x_t\,,\,\lambda)x_t}{\sum_{t=1}^{T} p(i|x_t\,,\,\lambda)}$$

- covariance matrix

$$\bar{\sigma_i}^2 = \frac{\sum_{t=1}^{T} p(i|x_t\,,\,\lambda)x_t^2}{\sum_{t=1}^{T} p(i|x_t\,,\,\lambda)} - \bar{\mu_i^2}$$

- the likelihood of the posteriori of the i[th] class is given by posterior probability

$$p(i \mid \vec{x_t},\lambda) = \frac{p_i b_i(\vec{x_t})}{\sum_{k=1}^{M} p_k b_k(\vec{x_t})}$$

This process is repeated until convergence is achieved.

3.2 Classification based on GMM

In this stage, after the GMM models for each speaker are estimated, the target is to find the model with the maximum likelihood a posteriori for an observation sequence. The input to the classification system is denoted as

$$X = \{x1, x2, x3, \ldots, xT\}$$

The rule to determine if X has come from speaker s can be stated as

$$p(\lambda s \mid X) > p(\lambda r \mid X)\ r = 1, 2, \ldots, S\ (r \neq s).$$

Therefore, for each speaker s in the speaker set, the classification system needs to compute and find the value of s that maximizes $p(\lambda s \mid X)$ according to

$$\hat{S} = \arg \max_{1\le s\le S} P(\boldsymbol{\lambda}_s|X) = \arg \max_{1\le s\le S} \frac{p(X|\boldsymbol{\lambda}_s)\Pr(\boldsymbol{\lambda}_s)}{p(X)}$$

The classification is based on a comparison between the probabilities for each speaker. If it can be assumed that the prior probability of each speaker is equal, then the term of $p(\lambda s)$ can be ignored. The term $p(X)$ can also be ignored as this value is the same for each speaker, so $p(\lambda s \mid X) = p(X \mid \lambda s)$,
where,

$$\mathrm{p}(\mathrm{X}|\lambda_s) = \prod_{t=1}^{T} \mathrm{p}(\mathrm{x_t}|\lambda_s)$$

Practically, the individual probabilities, $p(x_t/\lambda_s)$, are typically in the range 10^{-3} to 10^{-8}, the result $p(x_t/\lambda_s)$,will underflow probability for all speakers will be calculated as zero. Thus, $p(X \mid \lambda s)$ is computed in the log domain in order to avoid this problem[2]. The likelihood of any speaker having spoken the test data is then referred to as the log likelihood.

The speaker of the test data is statistically chosen by

$$\hat{\mathrm{S}} = \arg \max_{1\le s\le S} \mathrm{p}(\mathrm{X}|\lambda_s) \xrightarrow{\text{take log}} \hat{\mathrm{S}} = \arg \max_{1\le s\le S} \sum_{t=1}^{T} \log p(\vec{x_t}|\lambda_s)$$

4. RESULTS AND ANALYSIS

4.1 speech signal

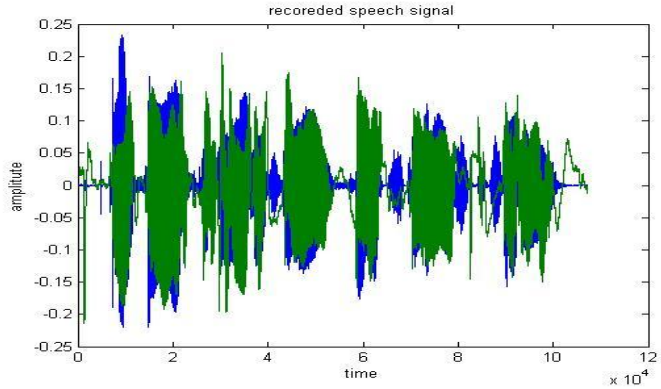

Figure 5. recorded speech signal of Gujarati Language

4.2 weighted MFCC of one frame

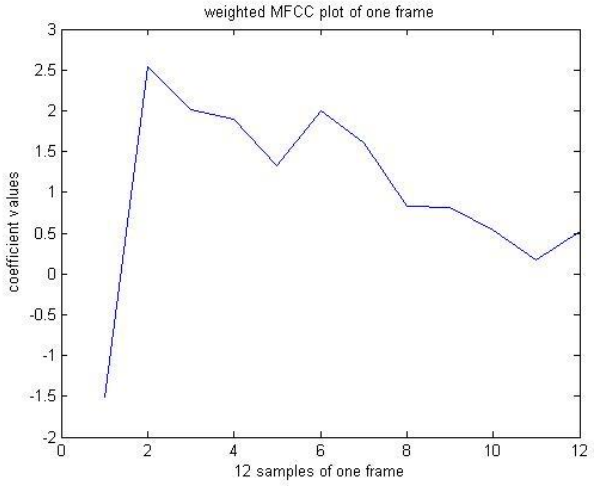

Figure 6. weighted MFCC of one frame

4.3 iterations of mean for one frame

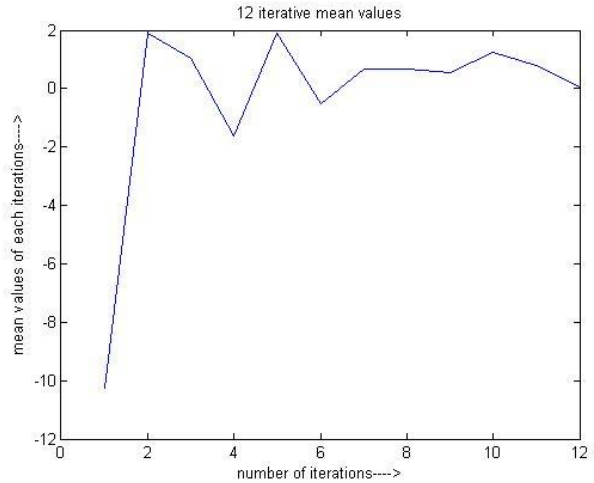

Figure 7. 12 iterative values of mean

4.4 iterations of covariance

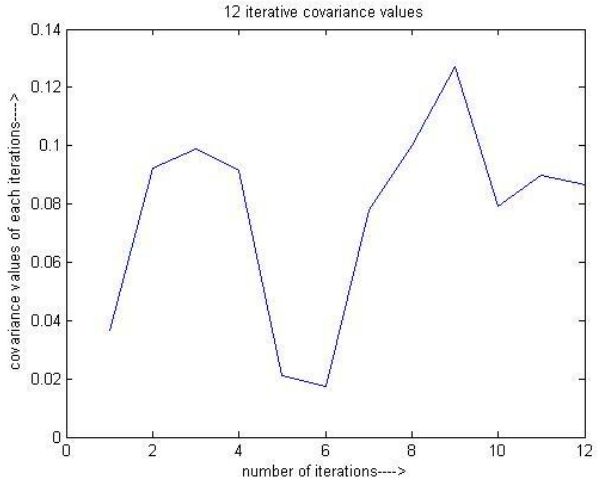

Figure 8. 12 iterative values of covariance

4.5 Gaussian Mixture Model of one speaker

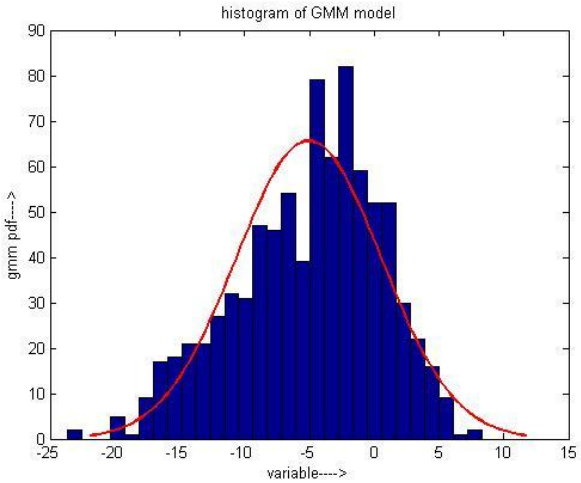

Figure 9. Gaussian Mixture Model pdf of one speaker

4.6 Analysis of results

Experimental results showing that with implementation of weighted MFCC for extraction of features from the speech signal of Gujarati Language, GMM gives the better results for speaker recognition. In this GMM model 12 iterations are used for achieving the approximate suitable Gaussian model with the help of 128 MFCC frames.

5. CONCLUSIONS

After analyzing the results, the experiment concludes that Weighted MFCC+GMM gives better efficiency of speaker recognition than the Traditional MFCC+GMM for Gujarati Language. The recognition rate is approximately 1% higher compared to the Tradition MFCC+GMM.

REFERENCES

[1] L.R. Rabiner and R.W. Schafer, " Digital Processing of Speech Signals", Printice Hall Signal Processing Series, Second Edition.

[2] Thomas F. Quatieri, " Discrete-Time Speech Signal Processing Principle and Practics", Pearson Education Signal Processing Series, First Indian Reprint.

[3] Zhang Wanli and Li Guoxin " The Research of Feature Extraction Based on MFCC for Speaker Recognition" published in 3rd International Conference on Computer Science and Network Technology in 2013.

[4] Douglas A. Reynolds and Richard *C.* Rose "Robust Text-Independent Speaker Identification Using Gaussian Mixture Model" published in IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 3, NO. 1, JANUARY 1995.

[5] Snani Cherifa and Ramdani Messaoud "New Technique to Use the GMM in Speaker Recognition System(SRS)" published in 978-1-4673-5285-7/13/$31.00 ©2013 IEEE.

[6] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors," *Speech Commun.*, vol. 52, no. 1, Jan. 2010, pp. 12-40.

[7] Zhang Wanli, Li Guoxin,"Speech Recognition Using Improved MFCC", 2012 International Conference on Electrical and Computer Engineering, Volume 11, pp.99-104, July, 2012

[8] Z.J.Wu and Z.G.Cao, "Improved MFCC-Based Feature for Robust Speaker Identification", TSINGHUA Science and Technology, vol.10,pp. 158-161, Apr. 2005.

[9] M. Hassan, M. Jamil, M. Rabbani, and M. Rahman, "Speaker identification using Mel frequency cepstral coefficients," in *Proceedings of the 3rd International Conference on Electrical &Computer Engineering*, pp. 565–568, 2004).

[10] Hassen Seddik, "text independent speaker recognition using the mel frequency cepstral coefficients and a neural network classifier",2004

BIOGRAPHIES

Jigarkumar Patel was born in Bayad, Gujarat, India in 1992. He received B.E(EC) from Vishwakarma Govt. Engg. college, Chandkheda in 2013 and pursuing M.E in Communication System Engg. from L.D College of Engg, Ahmedabad in 2015.

Arun Nandurbarkar was born in India,1970. He is working as a Associate Professor at L.D.College of Engineering, Ahmedabad.

