UACEE International Journal of Advances in Computer Science and its Applications – IJCSIA
Volume 3 : Issue 3 [ISSN 2250 – 3765]
Publication Date : 09 September 2013

Multi-Devices Hindi Speech Database for Speaker Identification using GMM.

[Sonu Kumar, Mahesh Chandra]

Abstract— **In this paper, we study the effect on speaker identification (SI) system when speech data is recorded on two different sensors, a HP Pavilion third generation laptop and a Samsung mobile (S3770K) both with built-in microphone in parallel in a closed room in noise free environment. The database contains 10 Hindi sentences (50-60 seconds speech) and one english sentence (7-8 seconds speech) of each 39 speakers (26 Male and 13 Female) in a reading style manner. Identification process adopts the methods of feature extraction based on Mel-frequency cepstrum coefficients (MFCC), linear predictive coding (LPC) coefficients. Gaussian mixture model (GMM) is used as a classifier. Our study shows that higher degradation in performance in case of mismatch of sensors during training and testing of data and MFCC performs better during matched conditions, LPC performs better than MFCC in mismatched conditions .**

Index Terms **— MFCC, LPC, GMM, SI, features, PIR**

I. INTRODUCTION

The main goal of this paper is to address the issues in speaker identification system during the mismatch between training and testing conditions. In addition, most of the speech databases available are in English language and very few studies have been reported for Indian languages. To overcome this constraint, we develop a multi-device, Hindi speech database.

Now these days' people are using laptop and mobile for secure transactions. For these powerful devices, there also exists a need for greater security. By considering this thing lots of researches have been done in this field to develop a reliable mechanism to control access of sensitive information stored on these devices. . The most used security mechanism is the text entered password. Although simple in implementation, this system is going to be difficult for a number of handicaps. Its effectiveness is highly depends on the use of hard to remember string / digit combinations which must be frequently changed. However, many users used simple password like his/her name, date of birth and mobile number if ever, changed providing little security. The size of keyboard and keypad of these devices is decreasing very rapidly; this improvement in technology is going to difficult to operate for handicaps and uneducated people. This problem has a simple solution by the use of integration of speaker identification technology for secure user logins. Speaker identification [6] is the process of determining who is speaking on the basis of individual information included in speech.

In this paper Gaussian mixture model [7] [8] is used as a classifier. Before constructing the GMM model for each speaker, speech signal is converted into a set of feature vectors which represent an individual speaker. In this paper LPC [2] [6] and MFCC [1] [3] have been used a feature extraction techniques.

The rest of this paper is organized as follows. Section 2 has brief explanation about the feature extraction techniques. Section 3 address the details of the Gaussian mixture model (GMM) .Experimental setup and results are given in section 4. Finally conclusions are drawn in section 5.

II. FEATURE EXTRACTION TECHNIQUES

Speech signal is a random and complex signal, so before processing digitized speech pre-processing is required. Pre-processing mainly contains three steps: pre-emphasis filtering, normalization and mean subtraction. Pre-processing technique is common in MFCC and LPC as shown in "Fig.1".

The digitized speech signal is passed through first-order high pass filter to spectrally flatten the signal. The response $H(z)$ of the filter is given by Equation (1) [2]

$$H(z) = 1 - az^{-1}, \quad 0.9 \le a \le 1.0 \qquad (1)$$

The words are spoken by different speaker having different amplitude for the same samples. In order to reduce amplitude variations from speech samples for all the words, we divide the total samples by the highest amplitude sample in the signal, this process is called normalization. Mean subtraction [3] removes the dc offset which are introduced due to the microphone used for segmenting and some other effect introduced at the time of recording. Framing is done to remove dynamic nature of speech signal hence we divide the speech signal into small size frames. The purpose of the windowing is to limit the time interval to be analyzed so that the properties of the waveform do not change appreciably. Windowing also serves to remove the signal discontinuities at the beginning and end of each frame. Hamming window is used for this purpose since it provides smoother spectrum. Hamming window is given by Equation (2).

$$W(n) = 0.54 - 0.46Cos\left[\frac{2\pi n}{N-1}\right] \qquad (2)$$

Where, $0 \le n \le N-1$
And N is the number of sample in a single frame

A. MFCC

MFCC is based on the variation of the human ear critical bandwidths with frequencies. The spectral coefficients of each frame are converted to Mel scale after applying a filter bank. The Mel-scale is a linear scale below 1000 Hz and logarithmic above 1000Hz.Equation (3) defined the Mel-scale

$$Mel(f) = 2595\log[1 + f/700] \quad (3)$$

Where $Mel(f)$ is the pitch in Mels corresponding to actual frequency f.

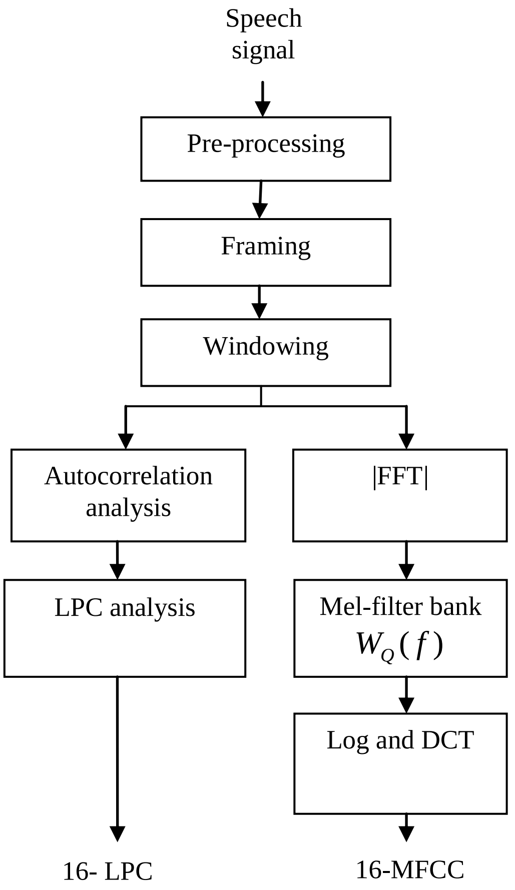

Figure.1. LPC and MFCC feature extraction

Fast-Fourier transform (FFT) is used to calculate the N-point discrete Fourier transform (DFT) of a Hamming-windowed input signal in an efficient manner, saving processing power and reducing computational time as shown in Equation (4).

$$|S(f)|^2 = \left|\sum_{n=1}^{N} s(n)e^{-j2\pi fn/N}\right|^2 \quad (4)$$

Where $1 \le f \le N$ and $s(n)$ is the pre-processed and windowed speech frame which is converted to frequency domain and leads to energy spectrum. Now pass this to a group of Mel- triangular filter bank having order 20 and unity height, uniformly scale in Mel scale. The output of Mel scale filters can be calculated by weighted summation of filter response

$W_k(f)$ and energy spectrum $|S(f)|^2$ as shown in Equation (5)

$$e(i) = \sum_{k=1}^{N/2} |Y(k)|^2 W_K(f) \quad (5)$$

Where i=1 to Q (order of Mel filter) and finally discrete cosine transform (DCT) is applied to the log of the Mel-spectral coefficients to obtain the Mel-Frequency Cepstral Coefficients.DCT has higher degree of spectral compaction and tends to have more of its energy concentrated in a small number of coefficient when compared to other transform.

$$C_m = \sqrt{\frac{2}{Q}}\sum_{l=1}^{Q-1}\log[e(i+1)].Cos\left[m\left(\frac{2l-1}{2}\right)\frac{\pi}{Q}\right] \quad (6)$$

Where m belongs to number of coefficients taken $1 \le m \le T$, T is desired coefficients and C_m are the final MFCC coefficients given by Equation (6).

Out of extracted coefficients, the first one is discarded as it represents the DC component. These cepstral features are most efficient features for speaker identification.

C. LPC

LPC is the most common techniques for low bit-rate speech coding and is a very important tool in speech analysis. Linear prediction is a method which predicts the nth sample of the signal by forming a linear combination of p previous based on LPC Model. The linear combination is usually optimized by minimizing the square of the prediction error. “Fig. 2” shows the LPC model[2] of speech synthesis, let *S* (n) be a speech sample as shown in Equation (7)

$$S(n) = \sum_{k=1}^{p} a_k S(n-k) + Gu(n) \quad (7)$$

We consider the linear combination of past P samples as $\bar{S}(n)$ as shown in Equation (8).

$$\bar{S}(n) \cong \sum_{k=1}^{p} a_k S(n-k) \quad (8)$$

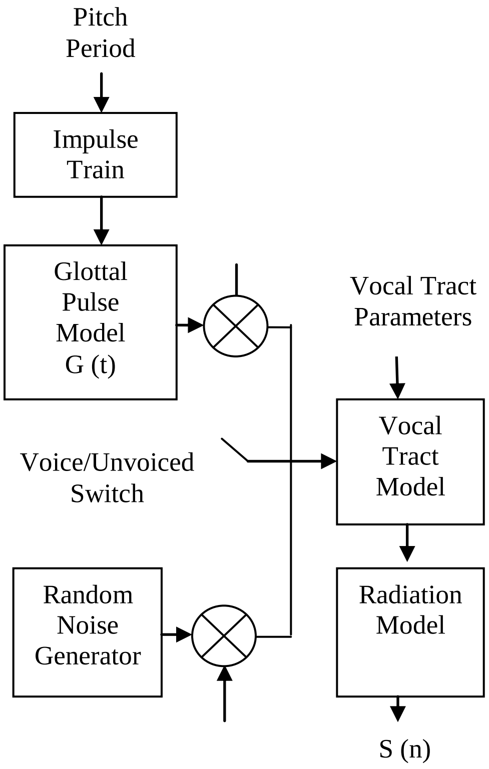

Figure 2.2: Speech Synthesis Model Based on LPC Model

The Prediction error e (n) is defined as,

$$e(n) = S(n) - \bar{S}(n)$$

$$e(n) = S(n) - \sum_{k=1}^{p} a_k S(n-k) \qquad (9)$$

By minimizing this sum of the squared differences between the actual speech samples and the linearly predicted ones, a unique set of predictor coefficients is determined as given in Equation (9).

LPC uses the Levinson-Durbin recursion to solve the normal equations that arise from the least-squares formulation. This computation of the linear prediction coefficients is often referred to as the autocorrelation method .

III. GMM

Gaussian Mixture Models are probability density models that comprise a number of component Gaussian functions. Gaussian mixture densities is a weighted sum of M component densities [1] given by equation,

$$p(X/\lambda_s) = \sum_{i=1}^{M} w_i^s * b_i^s(X) \qquad (10)$$

Where w_i^s are the mixture weights, i=1, 2….*M,* X is a D dimensional value data vector and $b_i^s(X)$ are the Gaussian densities for s speakers as given in Equation (11).

$$b_i^s(X) = \frac{1}{(2\pi)^{\frac{D}{2}} \left|\Sigma_i^s\right|^{0.5}} e^{\left(\frac{-1}{2}\left(X-\mu_i^s\right)^t \left(\Sigma_i^s\right)^{-1} \left(X-\mu_i^s\right)\right)} \qquad (11)$$

Where μ_i^s represent mean vector, Σ_i^s as covariance matrix.

During training, maximum likelihood parameters can be estimated using Expectation maximization algorithm (EM) [9]. This algorithm generates the highest value of log-likelihood after 5-20 iteration. After these iteration the model parameters converges to stable values. For a number of feature vectors $X = x_1, x_2 \ldots x_T$ the log-likelihood of a model λ_s for T frames utterance can be calculated as

$$L_s(X) = \log p\left(X/\lambda_s\right) = \sum_{t=1}^{T} \log p\left(X_t/\lambda_s\right)$$

The value of $L_s(X)$ is computed for all s speaker models enrolled in the system.

IV. Experimental setup and Results

Database contains ten Hindi sentences of 39 speakers (26 male and 13 female) recorded in parallel in a reading style using a two different devices a Samsung mobile and a laptop in a built-in microphone. Each sentence has 5-8 seconds speech that means 60-70 seconds speech in total for every speaker. A sampling rate of 16 KHz and 16-bit resolution with mono channel was taken. Speakers from age group of 18 to 30 years were chosen. All speakers uttered same sentences which were recorded in a noise-free environment

Before calculating the features, we divided the speech signal in frames of 25 ms with 15 ms overlapping. In our paper we have taken 20 triangular Mel filter bank for MFCC and LPC model order 15 for calculating 16 coefficient of each MFCC and LPC. Now the 39 GMM model were trained with the calculated frame based features. Feature vectors of first 150 frames of first sentence of each speaker were taken for testing; hence we got a total of 5850 features for testing.

"Fig.3" shows the setup of speaker identification process using GMM classifier [3].

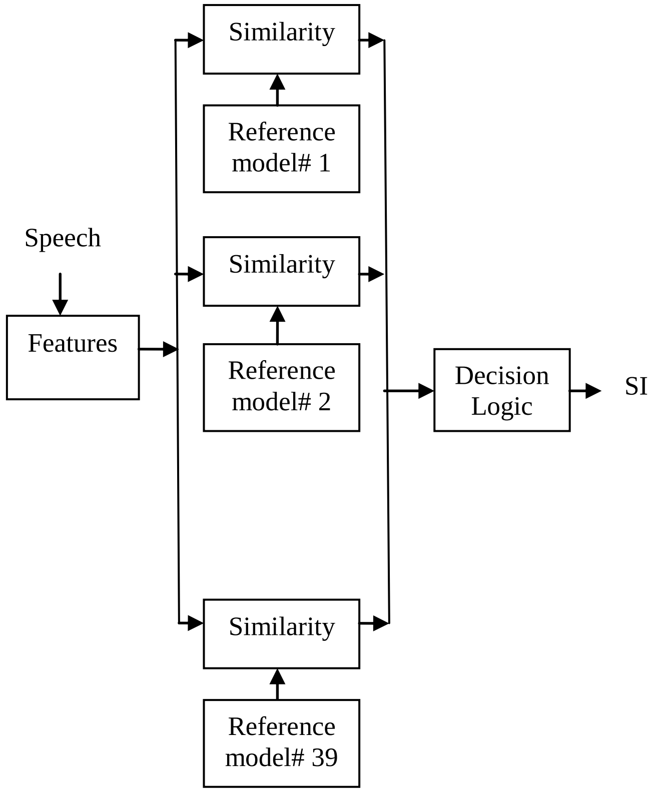

Figure 3. Speaker identification using GMM

Experiments were conducted to study the difference in identification rate by varying the Gaussian mixture densities, M.

$$\text{Percentage identification rate (PIR)} = \frac{\text{Number of utterances correctly identified}}{\text{Total number of utterances under test}}$$

According to the results present in Table I and Table II as the number of mixture component density is increasing PIR is increasing.

Table I. Results for Percentage Identification rate (PIR) using MFCC technique.

	Mixture component density(M)			***Average***
	L=4	***L=8***	***L=16***	
LL	82.45	86.85	88.37	85.90
MM	82.60	87.33	88.51	86.23
LM	33.91	34.57	31.81	33.43
ML	42.92	44.71	44.40	44.01

Where LL means training with laptop database and testing with laptop database. MM means training with Mobile database and testing with mobile database. LM means training with laptop database and testing with mobile database. ML means training with mobile database and testing with laptop database.

Table II. Results for percentage identification rate (PIR) using LPC technique.

	Mixture component density(M)			***Average***
	L=4	***L=8***	***L=16***	
LL	70.31	78.46	78.24	75.67
MM	79.45	76.77	80.00	78.74
LM	40.96	42.85	43.05	42.29
ML	50.50	57.64	57.54	55.23

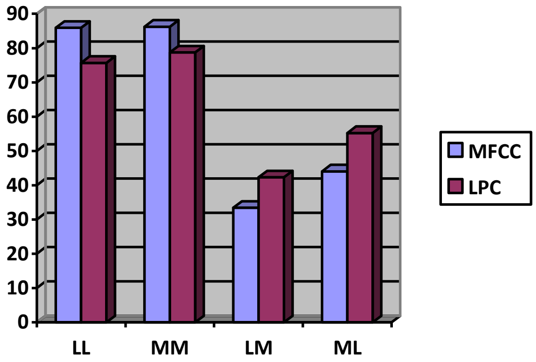

Figure 4. Average percentage identification rate of MFCC and LPC techniques.

V. Conclusion

On the basis of our study following conclusion comes out:

I. There is a huge degradation in identification rate When there is mismatch in training and testing.

II. MFCC performs better than LPC during matched condition i.e. 10.23% increase in identification rate in case of training and testing with laptop database and 7.5% in case of training and testing with mobile database.

III. LPC shows an improvement in identification rate during mismatch condition i.e. 8.86% increment in identification rate when training with laptop database and testing with mobile database and 11.22% increment when training with mobile database and testing with laptop database.

Acknowledgment

I would like to thank the students of MCA first year ,Birla Institute of Technology Mesra, Ranchi who have taken part as speaker in making database.

References

[1] S.Chakroborty, G.saha,“Improved text-independent speaker identification using fused MFCC &IMFCC feature sets based on Gaussian filter”, International journal of Information and Communication Engineering 5,vol 1,2009,pp.11-19.

[2] L. R. Rabiner and B. H. Juang, Fundamental of Speech Recognition, Englewood cliffs, NJ: Prentice Hall inc.,2000.

[3] Pawan kumar and Mahesh Chandra, “Speaker identification using Gaussian mixture models” ,MIT International journals of Electronics and Communication Engineering Vol. 1, No. 1, Jan. 2011, pp. 27-30.

[4] V.Manjusha, “Robust speaker verification for mobile transmission”, IEEE, ICSP 2010 proceedings ,pp. 518-521.

[5] Anurag jain, Nupur Prakash, S.S. agrawal , “Evaluation of MFCC for emotion Identififaction in Hindi speech”, IEEE, Vol. 2, 2011,pp. 189-192.

[6] J. Makhoul, “Linear prediction: A tutorial review,” Proc. of IEEE, vol. 63, no. 4, pp. 561-580, 1975.

[7] Reynolds, D. A. and Rose, R. C. "Robust text-independent speaker identification using Gaussian mixture speaker models”. IEEE Trans. Speech Audio Process. *3*, 1995, pp. 72–83.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, Series B, vol. 39, 1977, pp. 1-38,.

[9] T. K. Moon, "The Expectation-Maximization Algorithm,” IEEE *Signal Processing Magazine*, 1996, pp. 47-59.

Sonu Kumar

Department of Electronics and Communication

BIT, Mesra Ranchi, India

Dr. Mahesh Chandra

Department of Electronics and communication

BIT Mesra, Ranchi, India