International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 12, December-2014

Speaker Identification for Channel Mismatching Condition using Hindi Language

Sonu Kumar
Deptt of Electronics and communication
Dr. K.N. Modi Engineering College
Modinagar, India

Muketa Gupta
Deptt of Electronics and communication
Dr. K.N. Modi Engineering College
Modinagar, India

***Abstract*— This paper shows the result of Gaussian mixture model (GMM), when Hindi sentences are recorded in two Samsung mobiles (Samsung galaxy Trend, model No. S 7392) having same cost. A call is made by one mobile to another mobile for recording the voice of each speaker via wireless channel. Recording is done simultaneously in both the devices, in first one via built in headphone at transmitter end and in second one via wireless channel at receiver end. Sampling rate is set to be 8 KHz in both the cases. Several results are observed on matched and mismatched condition like training with transmitter end recorded speech and testing with receiver end recorded speech. MFCC is showing improvement in results as compare to LPC in all the cases. When trained with transmitter end Hindi database, there are 16.11 % and 12.54% increase in identification rate in matching condition as compare to mismatching condition for MFCC and LPC respectively. In case of training with receiver end database, there are 25.68 % and 30.68% increase in identification rate in matching condition as compare to mismatching condition for MFCC and LPC respectively**

Keywords—*GMM, MFCC, LPC, Speaker Identification, Features vectors*

I. INTRODUCTION

Speaker identification is the process to recognize a person on the basis of his/her voice. There are several techniques to perform speaker identification task. First of all we have to find out the features present in a speech then classification of these features are perform by using several methods like GMM, SVM, Neural network, Hidden Markov model (HMM), dynamic time warping (DTW) and Vector quantization (VQ). GMM are generative models i.e. the log likelihood is calculated for each model by comparing with previous model. SVM do not provide posterior probability outputs directly but the class labels. It divides the feature vector in different classes on the basis of maximizing the distance between the samples and classification function [1]. The DTW classifier provides the result on distance measurement between the feature vectors. Vector quantization a code book is made to compress the feature vectors to a small set of points.

The biggest challenge in speaker identification system is the mismatching of training and testing condition. There can be several types of mismatching conditions like mismatching of languages, environments, channels and recording devices etc. [1].These days the speaker identification researchers is using telephone speech for authentication of particular person in field of banking and other sectors. The banking services will provide secure services for remote speakers by authenticating the identity of speaker before allowing transaction and credit card payment. To make this system feasible, the banking authority should record the speech of the person. The person whose voice is stored in database can give instructions to the authority to do the transaction on a mobile phone from a far region by verifying his/her speech without going the bank. This can be possible only by a reliable and accurate system [2].

The main aim of this paper is speaker identification process that is divided into two parts mainly, training and testing phase. In training phase a model is trained from the speech samples of individual's speakers. In the testing phase a test sample of any speaker is used for comparing against the inbuilt models. For implementing both these phases, feature extraction is the very first step [3].

During feature extraction, different techniques are preferred these days like Mel-frequency Cepstral Coefficients (MFCC), Linear Prediction coefficients (LPC) and Delta MFCC (Δ- MFCC).

The structure of this paper is as follows: Section I includes the introduction of the topic, Section II provides an overview of different feature extraction technique like MFCC, LPC, Section III describes about GMM classifier, Experimental setup and results are presented in section IV, Section V concludes this paper.

II. FEATURE EXTRACTION

Several features can be extracted from each input voice sample for further processing. Most of speech recognition systems use a preprocessing method before implementing particular features extraction techniques. Using this method we can eliminates the effect produced by input device and microphones internal disturbances. The preprocessing and feature extraction techniques are described as follows.

A. Preprocessing

This process is implemented before using the actual feature extraction techniques. Each recorded voice is divided into frames of 25 ms with an overlapping of 10 ms with the adjacent frames. Now by applying a pre-emphasis filter, the lip radiation effects are eliminated. This filter [4, 5] is defined by equation (1)

$$S_n^{'} = S_n - 0.96 \times S_{n-1} \text{ , for } \quad n= 1, 2\ldots\ldots399 \qquad (1)$$

Where S_n is the n^{th} sample of the frame S and $S^{'}_0 = S_0$. Now the difference in loudness at the time of recording will be remove using normalization of every sample. After that Hamming window will be used on pre-emphasized frames as shown in equation (2)

$$S_i^h = S_i^{'} \times h_i, \quad \text{for} \quad i = 0,\ldots\ldots 399 \qquad (2)$$

With $h_i = 0.54 - 0.46 \times Cos(2\pi i/399).$

B. MFCC

After pre-emphasized, the Fourier transform will be applied to obtain the frequency spectrum of the speech signals. To get smooth spectrum, the spectrum will be multiplied with the Mel filter bank. The MFCC algorithm makes use of the vocal tract characteristics. The conversion of linear frequency to Mel frequency is shown in equation (3)

$$f_{mel} = 2595 \ln\left(1 + f/700\right) \qquad (3)$$

Where, f_{mel} is the pitch in mels corresponding to actual frequency, f in Hz. Let {y (n)}, n= 1……..M, represent a frame of the preprocessed signal. First, y (n) is converted to the frequency domain by an M point DFT which leads to an energy spectrum and followed by construction of filter banks each of which consists of a group of M triangular band pass filters [6,7].

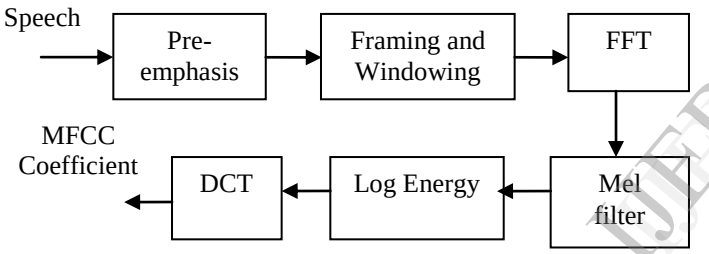


Fig.1: The MFCC Extraction Process

The output of each filter is taken to calculate the log energy E (m), where m=1, 2, 3….N. Finally calculate the discrete cosine transform to find out the sixteen MFCC coefficients using the equation (4).

$$C_i(n) = \sum_{m=1}^{M} E(m) Cos\left[\frac{\pi n(m-0.5)}{M}\right], \quad 0 \le m < M, \quad (4)$$

C. LPC

The idea behind this technique is based on the fact that a speech sample, S(n) can be approximated as linear combination of past speech samples, S(n-k). By reducing the sum of the squared difference between the actual speech samples and the predicted ones, we can find out a unique set of predictor coefficients [8, 9]. The prediction error, e (n) can be calculated by equation (6).

$$e(n) = S(n) - \sum_{k=1}^{p} a_k S(n-k) \qquad (5)$$

The value of LPC coefficients, a_k can be calculated by minimizing short term prediction error, E (n) as shown by equation (6).

$$E(n) = \sum_m [S(n) - \sum_{k=1}^{p} a_k S(n-k)]^2 \qquad (6)$$

III. GMM

A Gaussian Mixture Model is a probability density function represented as a weighted sum of Gaussian component densities. The Gaussian probability density function of a feature vector for i^{th} state is given by equation (7)

$$B_i(z) = \frac{1}{2\pi^{\frac{D}{2}} |\Sigma_i|^{0.5}} \exp\left(\frac{-1}{2}(z-\mu_i)^i \sum_i^{-1} (z-\mu_i)\right) \qquad (7)$$

Where μ_i is the mean vector, Σ_i is the covariance matrix, z is the D dimensional vector. The log-likelihood function for z feature vectors is defined as follows:

$$P\left(\frac{z}{\lambda}\right) = \sum_{i=1}^{N} w_i B_i(z) \qquad (8)$$

Where w_i is the mixture weights for i=1…M and B_i is the component densities. Collectively, the parameters of the density model are represented by $\lambda = (w_i, \mu_i, \Sigma_i)$ .Now the posterior probability for all the N classes is calculated by equation (9)

$$P\left(\frac{i}{z_T}, \lambda\right) = \frac{w_i B_i(z)}{\sum_{j=1}^{M} w_k B_k(z_T)} \qquad (9)$$

After that maximum likelihood will be estimated using iterative expectation-maximization (EM) algorithm [10, 11]. In this algorithm the likelihood will be calculated for first model and this model will become an initial model for next model and the whole process will repeat upto a certain threshold value for q to q+1 iterations, as shown below in equation(10)

$$P\left(\frac{z}{\lambda^{q+1}}\right) \ge P\left(\frac{z}{\lambda^{q}}\right) \qquad (10)$$

Maximum 8-12 iterations are sufficient for the convergence. Now the average of all likelihood models is calculated using equation (11)

$$\log P\left(\frac{z}{\lambda}\right) = \frac{1}{T} \sum_t \log P\left(\frac{z_t}{\lambda}\right) \qquad (11)$$

After calculating average likelihood we can find out the identification rate of the enrolled speakers individually as shown in Fig.2.

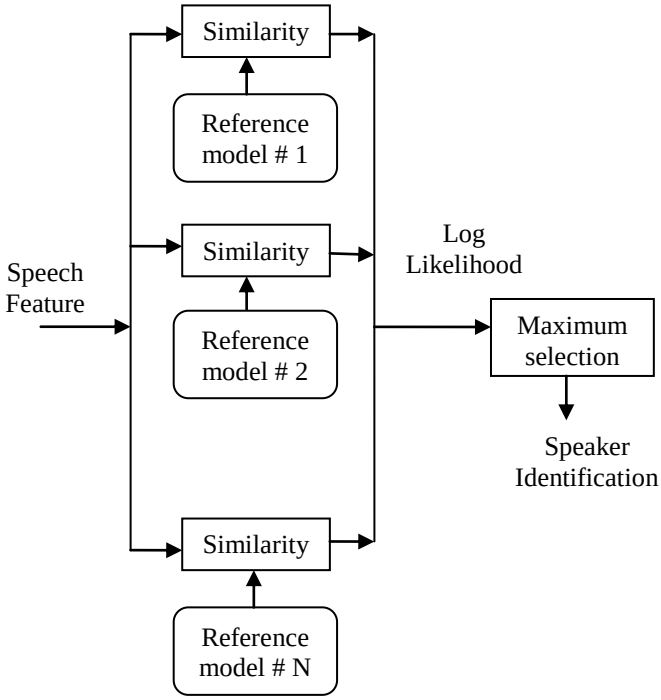


Fig.2: Speaker Identification using GMM

IV. Experimental Setup and Results

A. Establishment of speech database

The database consists of 23 speakers (16 male and 7 female) each is speaking 10 Hindi sentences of 6-8 seconds length. For testing phase we have taken first 100 frames of fifth sentence of each speaker. The database was recorded at 8 KHz sampling frequency. Two Samsung mobile (Samsung galaxy Trend, model No. S 7392) of same cost and model are used. One mobile is used as a transmitter to setup a call and another as a receiver to receive the call. After handling the call by the receiver, each speaker is repeating these sentences and automatic call recorder is used to record the call in both device simultaneously. Speakers from age group of 18 to 35 years are chosen. After making database the whole speech is cut in different individual sentences using Goldwave software. Here we have calculated 16 MFCC and LPC coefficients respectively.

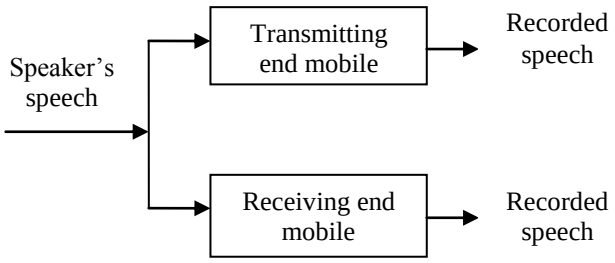


Fig.3: Recording setup diagram

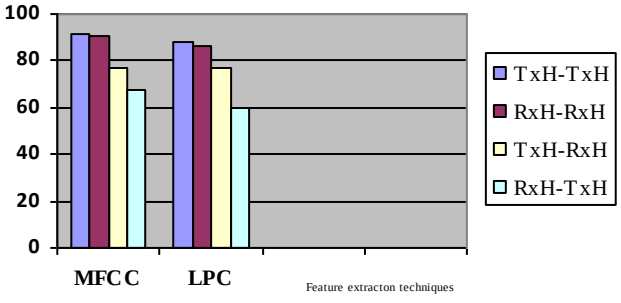
B. Analysis of results

This experiment is performed in the matched and mismatched conditions for Hindi language when training and testing with different databases i.e. Training with transmitter / receiver end Hindi (Tx-H / Rx-H) database and testing with transmitter / receiver end Hindi (Tx-H / Rx-H) database respectively as shown in Table. I. For testing only first 100 frames of the first sentence is taken.

TABLE.I SPEAKER IDENTIFICATION RATE IN MATCHED AND MISMATCHED CONDITIONS FOR HINDI LANGUAGE

Database taken for Training and Testing		Feature Extraction Techniques	
Training Database	*Testing database*	*MFCC*	*LPC*
Tx-H	Tx-H	91.43	87.69
Rx-H	Rx-H	90.56	86.17
Tx-H	Rx-H	76.70	76.69
Rx-H	Tx-H	67.30	59.73

Tx-H-Transmitter end recorded Hindi database, R_xH-Receiver end recorded Hindi database



V. CONCLUSION

On the basis of our study following conclusion comes out:

I. MFCC is showing improvement in results as compare to LPC in all the cases.

II. There are 16.11% and 12.54% increase in identification rate in matching condition in MFCC and LPC respectively as compare to mismatching condition when trained with transmitter end recorded Hindi database(Tx-H).

III. There are 25.68% and 30.68% increase in identification rate in matching condition in MFCC and LPC respectively as compare to mismatching condition when trained with receiver end recorded Hindi database(Rx-H).

ACKNOWLEDGMENT

I would like to thank the student and faculty members of Dr. K.N. Modi Engineering College, for contributing his/her voice in making database.

REFERENCES

[1] Hyson Seo, Hong-Goo Kang, Chi-Sang, "Robust session variability compensation for SVM speaker verification" IEEE transaction on audio , speech and language processing, Vol.19 No.6, August 2011, pp. 1631-1641.

[2] D.A. Reynolds, W.M. Campbell, Springer handbook of speech processing, chapter 13, pp.763.

[3] Hesham Tolba, "A high performance text independent speaker identification of Arabic speakers using CHMM-based approach" Alaxendria Engineering journal (2011), vol.50, pp-43-47.

[4] Md. Khademul Islam, Keikich Hirose, "On effectiveness of MFCCs and their statistical distribution properties in speaker identification" in Proceedings IEEE international conference on virtual environments, human computer interfaces and measurement systems, July 2013, Boston, USA.

[5] M Murugappan, N. Q. Baharuddin, "DWT and MFCC Based Human speech emotional classification using LDA", International Conference on Biomedical Engineering (ICoBE), February 2012, Penang.

[6] Chuan Xie, Xiaoli Cao, Lingling, "Algorith of abnormal audio recognition based on improved MFCC", in proc.of Engineering of IWIEE 2012, vol.29, pp. 731-737.

[7] Claude Turner, A.joseph, M. Aksu, "The Wavelet and Fourier transforms in feature extraction for text dependent, Filter based speaker recognition", Procedia Computer Science 2011, vol. 6,pp. 124-129

[8] J. Makhoul, "Linear prediction: A tutorial review," Proceedings. of IEEE., vol. 63, no. 4, pp. 561-580, 1975.

[9] M. Marvi, A. Harimi, "Estimation of LPC coefficients using Evolutionary Algorithms" Journal of AI and data mining, Vol. 1, No.2, 2013, pp.111-118.

[10] R. Shantha, S. Selva,"Fused Mel feature sets based text-independent speaker identification using GMM", in International conference on Communication Technology and System design, procedia engineering, vol.30, 2012, pp. 319-326.

[11] D. Reynolds, "Gaussian mixture models", MIT Lincoln Laboratory, USA.