Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 125 (2018) 880–887

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach

Ankur Maurya[a,*], Divya Kumar[a], R.K.Agarwal[b]

[a]*Motilal Nehru National Institute of Technology Allahabad, Allahabad-211004*
[b]*National Institute of Technology Kurukshetra, Kurukshetra-136119*

Abstract

Speaker recognition for different languages is still a big challenge for researchers. The accuracy of identification rate (IR) is great issue, if the utterance of speech sample is less. This paper aims to implement speaker recognition for Hindi speech samples using Mel frequency cepestral coffiecient–vector quantization (MFCC-VQ) and Mel frequency cepestral cofficient-Gaussian mixture model (MFCC-GMM) for text dependent and text independent phrases. The accuracy of text independent recognition by MFCC-VQ and MFCC-GMM for Hindi speech sample is 77.64% and 86.27% respectively. However, the accuracy has increased significantly for text dependent recognition. The accuracy of Hindi speech samples are 85.49 % and 94.12 % using MFCC-VQ and MFCC-GMM approach. We have tested 15 speakers consisting 10 male and 5 female speakers. The total number of trails for each speaker is 17.

Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications.

Keyword: Identification rate (IR), MFCC-GMM, MFCC-VQ.

1. Introduction

Speaker recognition is an investigated area for the researchers of speech processing. Speaker recognition is task of identifying person from properties of speech samples. The knowledge of acoustic phonetics and pattern matching

* Corresponding author
E-mail address: ankur.maurya.pbh@gmail.com

1877-0509
Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications
10.1016/j.procs.2017.12.112

techniques are needed in speaker recognition. Similar to face recognition, language recognition and speech recognition; speaker recognition corresponds to multidisciplinary problem. Most of the speaker recognition tasks depend on short engrossed speech or instructions from people. The speaker recognition task is of two kinds: text independent and text dependent. In text independent recognition, the input phrase and target phrases are same. While text dependent recognition involves same input and target phrases. There are many factors which affects speaker recognition system. Factors like session variability, noise, emotions, short phrases etc. Less amount of testing utterance makes speaker recognition difficult to work with good accuracy. Recently, the advancement has been made in this area is remarkable, the determination of gender by speech phrase can be done. Nevertheless, speaker recognition systems are far from impeccable. Speaker recognition has been a vigorous topic of research for several decades already. Successful application using this technology can already be found from the market. However, reliable recognition system requires fast and expensive hardware to operate in real time. The Speaker recognition task is usually attained in two stage processing: training and testing. The training process computes speaker-specific feature parameters from the stream of speech. The features are used to generate statistical models of different speakers. In the testing phase [1,6], speech samples from unknown speakers are compared using models and classifier techniques.

The work of speaker recognition started in 1960 [2]. Filter banks and spectrograms were used in the recognition system for pattern matching [2]. Bricker et al. [3] used auto correlation method for text independent phrases. The use of GMM models for text-independent speaker recognition was investigated by Reynolds et al. [4]. MFCC and segmental approach was studied in [5]. Speaker recognition was performed on many databases NIST [7, 8], Mercury and Orion datasets [9], NTT [10], SRI [11]. The work on noisy environment has not been explored significantly. Some of the speaker recognition work has been done on noisy environment and telephone speech [1, 2].

The performance and robustness of experiments are evaluated in terms of, identification rate (IR), accuracy, and error rate. The previous work on speaker recognition is done mostly on normal environment. This work aims to compute results where utterance (speech data) is shorter and dialects are in Hindi. The distance between different samples is calculated using maximum log likelihood ratio. The Section II in the paper briefly describes the fundamental component of work and also explains fundamental components and algorithmic steps of proposed framework followed by experimental set up. Results are given in section III following concluding remarks, given in section IV.

2. Fundamental Components

This part of paper contains the description of algorithm used in proposed approach. It also reveals their usefulness and robustness towards human's speech identification system. The algorithms used in our work are MFCC [12, 14], VQ [13] and GMM [15, 16].

A. *Mel Frequency Cepestral Cofficient*

The speech signal is first digitized and then converted into frames. The length of frames can be in range between 20 to 40 ms. After framing the speech signal is smoothed by Hamming window. MFCC involves the following five steps:

1) The Fast Fourier transformation of speech signal x(n) is done using a 30 ms Hamming window H(n). Speech signal and hamming window is used to calculate power spectral estimation. The power spectral estimation is given by equation (2.1).

$$x(k) = \sum_{n=0}^{n-1} \mathrm{x(n)H(n)e^{-j2\pi nk/N}} \qquad 0 \le \mathrm{n} \le \mathrm{N} \tag{2.1}$$

where N is 30 ms hamming window length having cosine impulse given by (2.2)

$$H(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & for \;\; n = 0, \ldots N-1 \\ 0 & else \end{cases} \tag{2.2}$$

To show the characteristics of vocal tract, square magnitude of DFT is calculated by equation (2.3)

$$X_{k=} |\dot{x}(k)|^2 \tag{2.3}$$

2) Apply triangular filters to power spectrum, energy is calculated in each filter bank channel is given by equation (2.4).

$$E_m = \sum_{k=0}^{k-1} \emptyset_m (k) X_k; \quad \text{m =1, 2, ….M} \tag{2.4}$$

where M is number of triangular filters (∅), which can have range 20 to 35 with condition is given by equation (2.5).

$$E_m = \sum_{k=0}^{k-1} \emptyset_m (k) X_k; \quad \text{m =1, 2, ….M} \tag{2.5}$$

3) Cepestral coefficient is calculated by computing discrete cosine transformation of log filter-bank energies

$$c_j = \sum_{m=1}^{M} \log_{10}(E_m).\cos j((m+0.5)\frac{\pi}{m}) \qquad j = 1, 2, \dots \dots L \tag{2.6}$$

The weighted value can be calculated by equation (2.7).

$$c_i = (1 + \frac{l}{2}\sin(\frac{\pi i}{l}))c_i \tag{2.7}$$

where c_i is cepestral coefficient of i^{th} order and liftering coefficient is set to 21 in our approach.

4) 13-Dimensional feature vector is produced after normalized frame energy.

5) First and second order derivative of 13 coefficient are calculated. These are append to original MFCC, 39 dimensional feature vectors will be produced for each frame.

B. *Vector Quantization*

Vector Quantization (VQ) is a classical technique for quantifying signal processing capabilities that can model probability density by distributing prototype vectors. It was used to compress data. It works by dividing a large set of points (vectors) into groups that have the same number of points closest to them. Each group is represented by the central point, as in the k-means and other pooling algorithms.

The d-dimensional vectors are mapped in the vector space R_k into a finite set of vectors $Q = \{q_i : i = 1, 2,..., n\}$ by vector quantizer. VQ works well for lossy data compression. The main elements in the mapping are code word- and code-book. The each vector q_i associated with quantizer is called a code-word or a code-vector and the set of all the code words are called a codebook. The region that belongs to nearest neighbour q_i is Voronoi region.

C. *Gaussian Mixture Model*

A Gaussian mixing model is a probabilistic model that assumes that all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. One can think of blending models such as the generalization of k-means clusters to incorporate information about the covariance structure of the data, as well as the centers of latent Gaussians. The important aspect of the any accent modeling is to collect and find the weight of the mean vector of each accent and mixture from the training speech utterance. The parameters of Gaussian Mixture Model are estimated with the help of maximum likelihood Gaussian Mixture Model. The probability density function of arbitrary shape is approximated. The GMM probability density function can be presented by the covariance matrix ($\sum$), Mean vector (μ_i) and parameter set of the mixture weight (K_i). The multidimensional Gaussian mixture model is given by equation (2.8)

$$K_i = \frac{1}{(2\pi)^{\frac{d}{2}} |\sum_i|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu_i)^T \sum_i^{-1}(x-\mu_i)\} \tag{2.8}$$

The proposed algorithm consists of feature extraction followed by classification methods. The analog signal is first converted into digital form. The features of Hindi speech signals is then calculated using Mel frequency cepestral coefficient. The extracted feature is then classified using Gaussian mixture model. The final result is calculated using maximum log likelihood function. The closest match will be the recognition result. There are various methods to calculate closeness of target and input samples like Euclidean distance, Manhattan distance and Mahalanobis distance. The algorithm for speaker recognition using MFCC-GMM approach is given below.

Algorithm (MFCC-GMM)

Initialization: s is signal to analyze.

fs is sampling rate (12500 Hz)

m is distance between two frames (100)

n is number of samples each frames (256)

p is number of filters in filterbank

***Step 1**: Record the audio using microphone and store it into train folder for training purpose.*

***Step 2**: Record the audio for testing using microphone and save it into test folder.*

***Step 3**: Put the signal into frames (s, fs, m, s)*

***Step 4**: Call weiner filter to remove noise (n, fs, m,, s)*

***Step 5**: Feature Extaraction is done using MFCC*

MFCC (n, fs, m s)

***Step 6** :Compute matrix for mel- space filterbank.*

Melfb (p, n, m, fs)

Compute MFCC of audio data for test and train samples

MFCC (train{i})

MFCC (test(j))

***Step 7**: Gaussian mixture model is used for classifying samples from MFCC.*

***7.1** GMM (MFCC_train {i})*

***7.2** GMM (MFCC_test {j})*

***Step 8**: Compute maximum likelihood ratio (7.1 and 7.2)*

(G, GMM_test(i, j))

***Step 9**: Pattern matching*

Min(likelihood ratio (MFCC_test(i, j))) && min (distance)< threshold

Result = recognized, else

Result= not recognized.

In proposed method, the first step is to analyze the signals. The speech signals received with the help of microphone is in analog form. It is not possible to extract the features from the signal in analog form, it has infinite

number of points. So, it is converted into digitize form. The digital signals are used for feature extraction using MFCC. Gaussian mixture model is used as classifier for the features selected in feature extraction step i.e. *GMM (MFCC_train {i})* and *GMM (MFCC_test {j}*. This training and testing data is calculated separately. Then maximum likelihood ratio is compared for training and testing data. The result will recognize as true, if the calculated likelihood ratio for training and testing data are close and less than a fixed threshold value. Threshold value is calculated experimentally. The flowchart of algorithm is given below.

Flowchart of MFCC-GMM:

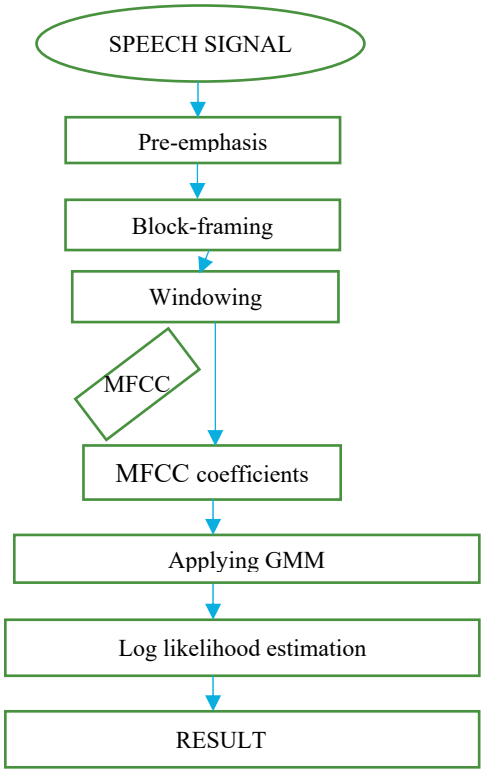

Figure 1: Flowchart of MFCC-GMM method

3. Results

The work is performed on dataset of 15 speakers which consists of 10 male (M) and 5 female (F). The work is performed on 10 numbers of males and 15 females. The tables give the recognition result using MFCC-VQ method and MFCC-GMM method for Hindi speech. The total number of trails for a speaker is 17. Table 1 and table 2 shows result for text independent and text dependent speaker recognition respectively. The graphs for text independent speaker recognition is shown in fig 2. Fig 3 shows the graph for text dependent recognition. The overall accuracy for both method is shown in fig 4. The accuracy of recognition is given by equation (3.1)

$$\textbf{Accuracy} = \frac{\textbf{no.of correct recognition}}{\textbf{total no.of trails}} \times \mathbf{100} \tag{3.1}$$

The accuracy is calculated by dividing number of correct recognition to total no. of trails by speaker. Correct recognition means the number of times the speaker is identified by system. The total number of trails for each

speaker can be taken anything but for more accuracy we have taken it 17. Let us consider, a speaker M1, the number of trails are 17. If the total number of correct recognition is 13, then the accuracy will be 76.47.

Table 1: Speaker recognition for text independent speech

Speakers	Techniques	M 1	M 2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	M 10	F 1	F 2	F 3	F 4	F 5	Average
Text independent	MFCC-VQ	70.59	76.47	64.71	82.35	88.24	70.59	82.35	88.24	76.47	82.35	88.24	70.59	76.47	76.47	70.59	77.64
	MFCC-GMM	76.47	82.35	82.35	82.35	94.12	94.12	82.35	94.12	88.24	94.12	88.24	82.35	88.24	82.35	82.35	86.27

Table 2: Speaker recognition for text dependent speech

Speakers	Techniques	M 1	M 2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	M 10	F 1	F 2	F 3	F 4	F 5	Average
Text dependent	MFCC-VQ	82.35	76.47	82.35	76.47	88.24	82.65	82.35	88.24	76.47	94.12	82.35	88.24	82.35	82.35	94.12	85.49
	MFCC-GMM	100	100	94.12	88.24	100	94.12	100	94.12	94.12	88.24	88.24	100	94.12	88.35	88.35	94.12

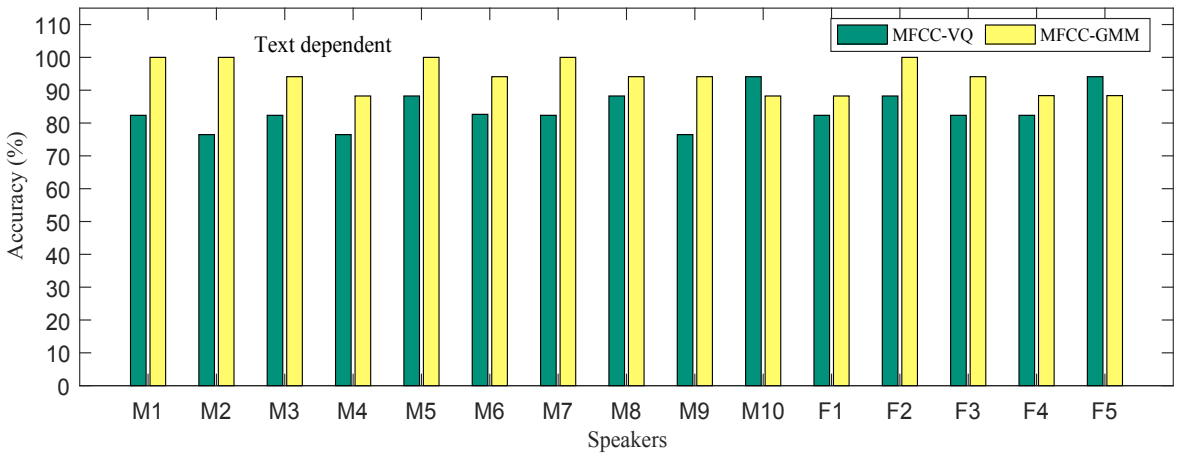

Figure 2: Graph showing accuracy of each speaker for text-dependent recognition

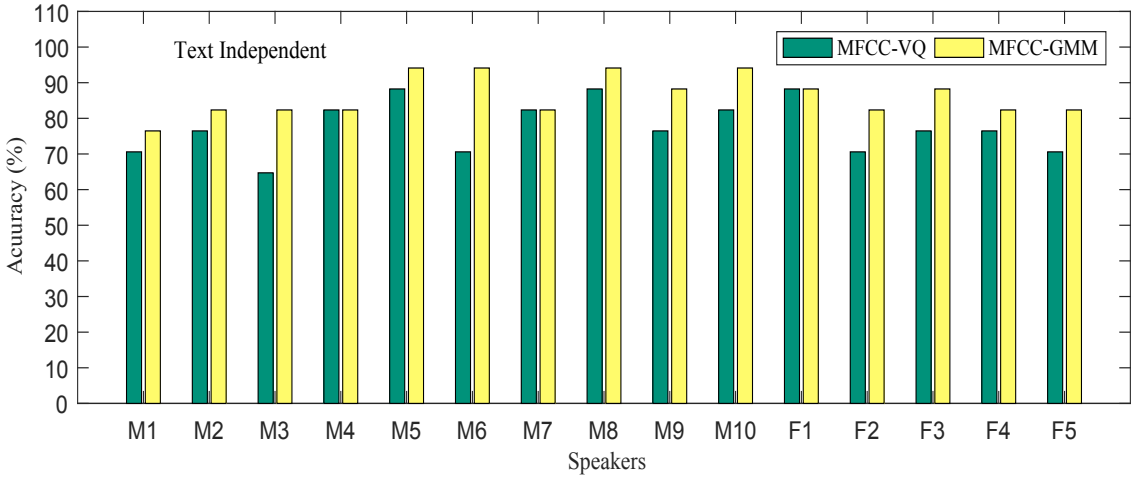

Figure 3: Graph showing accuracy of each speaker for text independent speaker recognition

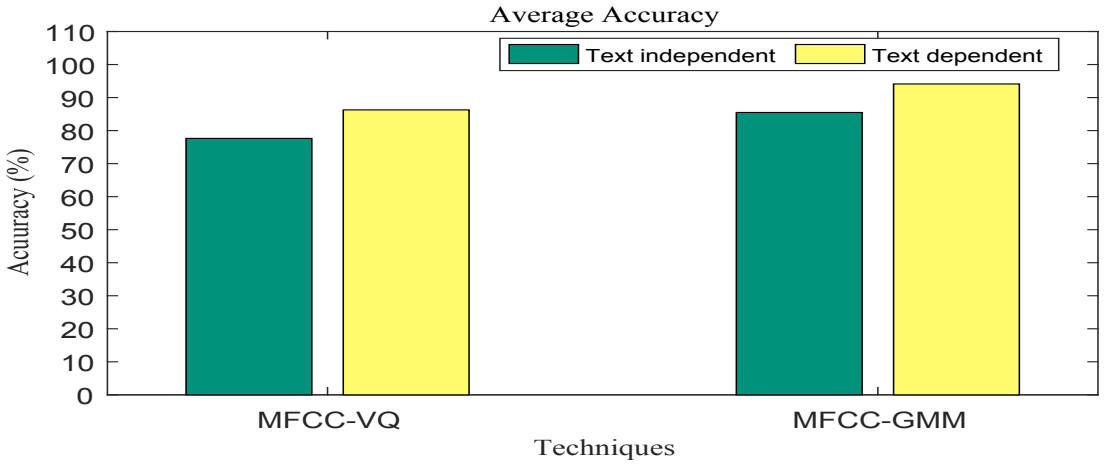

Figure 4: Graph showing overall accuracy of algorithm using MFCC-VQ and MFCC-GMM

4. Conclusion

The accuracy of speaker recognition for Hindi speech is still a big issue for researchers. The result we are getting using MFCC-GMM method for Hindi speech is remarkably good. There is a possibility that the speech can be recorded and can be used in place of the original speaker. Psychophysical studies show that there is a possibility that human speech may vary over a period of 2-3 years. Formulation of the ASR system is a challenging due to various effecting factors of human's speech like emotions, diseases, noises, session variably, etc. So, the training sessions are to be repeated to update the speaker specific codebooks in the database.

References

1. Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech, and Signal Processing, 29(2), 254-272..
2. Naik, J. M., Netsch, L. P., & Doddington, G. R. (1989, May). Speaker verification over long distance telephone lines. In Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on (pp. 524-527). IEEE
3. Pruzansky, S. (1963). Pattern-Matching Procedure for Automatic Talker Recognition. The Journal of the Acoustical Society of America, 35(3), 354-358.
4. Bricker P. D. (1971). Statistical techniques for talker identification. In Bell System Technical Journal, 50,. 1427-1454.
5. Bimbot F., Bonastre, J. F., Fredouille C., Gravier G., Magrin-Chagnolleau, I., Meignier, S., Reynolds D. A. (2004). A tutorial on text-independent speaker verification. EURASIP journal on applied signal processing, 2004, 430-451.
6. Kinnunen, T., Karpov, E., & Franti, P. (2006). Real-time speaker identification and verification. IEEE Transactions on Audio, Speech, and Language Processing, 14(1), 277-288.
7. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., & Dumouchel, P. (2008). A study of interspeaker variability in speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 16(5), 980-988.
8. Campbell, W. M., Campbell, J. P., Reynolds, D. A., Jones, D. A., & Leek, T. R. (2004). Phonetic speaker recognition with support vector machines. In Advances in neural information processing systems (pp. 1377-1384).
9. Elman J. L. (1990). Finding structure in time. In Cognative Science, 14, 176-211.
10. Specht, D. F. (1988, July). Probabilistic neural networks for classification, mapping, or associative memory. In IEEE international conference on neural networks (Vol. 1, No. 24, pp. 525-532).
11. Doddington, G. R. (2001, September). Speaker recognition based on idiolectal differences between speakers. In Interspeech (pp. 2521-2524).
12. Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing, 28(4), 357-366.
13. Gersho, A., & Gray, R. M. (1992). Vector Quantization and Signal Compression, Kluwer Academic. Norwell, MA.
14. Dao, V. L., Nguyen, V. D., Nguyen, H. D., & Hoang, V. P. (2016, December). Hardware Implementation of MFCC Feature Extraction for Speech Recognition on FPGA. In International Conference on Advances in Information and Communication Technology (pp. 248-254). Springer International Publishing.
15. Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. Digital signal processing, 10(1-3), 19-41
16. Bao, L., & Shen, X. (2016, April). Improved Gaussian mixture model and application in speaker recognition. In Control, Automation and Robotics (ICCAR), 2016 2nd International Conference on (pp. 387-390). IEEE.