Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 143 (2018) 267–276

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Conference on Advances in Computing and Communication (ICACC-2018)

Study of MFCC and IHC Feature Extraction Methods With Probabilistic Acoustic Models for Speaker Biometric Applications

Sithara A[a], Abraham Thomas[a], Dominic Mathew[a]

[a]*Rajagiri School of Engineering and Technology, Rajagiri Valley, Kochi 682039,India*

Abstract

Voice is an important human trait in natural human-to-human interaction / communication for identifying a person. So voice can be regarded as a biometric measure for recognizing or identifying the person similar to other biometric measures such as face, iris and fingerprints. Speaker recognition is a class of voice recognition where speaker is identified from the speech rather than the message. Automatic speaker recognition (SR) is an approach to identify people based on features extracted from speech utterances. The major task in any speaker recognition is to extract useful features and allow meaningful patterns of speaker models. This paper compares the performance of two feature extraction techniques Mel Frequency Cepstral Coefficient (MFCC) and Inner Hair Cell Coefficient (IHC) with two different modelling methods Gaussian Mixture Model - Universal background model (GMM - UBM) and i- vector approach. In this experiment speech samples of 600 speakers from TIMIT database with 10 utterances of each speaker are taken for identifying the speaker. A text independent speaker recognition system was implemented and this study resulted in an inference of MFCC feature outperforms IHC feature for both GMM and i vector. The performance of voiced speech IHC feature which simulates the physiological behaviour of human ear is better in terms of accuracy than full speech (voiced and unvoiced) in GMM.

Keywords: MFCC; IHC; GMM; i-vector; PLDA;

1. Introduction

Speech is a vocalized form of communication used by human beings which carries information such as message, language being spoken, the emotional state, gender etc. When we hear two different individuals through same channel, there is always a difference in the sound they produce due to their vocal tract shape or voice production organs. So voice can be regarded as a biometric measure for recognizing or identifying a person. SR can be used in various applications such as security and surveillance, forensic analysis, authentication, e-commerce etc. The speaker recognition

E-mail address: abrahamt@rajagiritech.edu.in, dominicmathew@rajagiritech.edu.in

10.1016/j.procs.2018.10.395

is one of the research area in audio processing and pattern recognition. Speaker recognition includes identification, verification (authentication), classification, and by extension, segmentation, tracking etc [1]. In this study the system aims at identifying the speaker from a set of known speakers i.e. the utterance of the test speaker is compared against all the available speaker models and the speaker ID of the model with the closest match is returned and there is no identity claim. When the audio of test speaker is selected from the known set it is called a closed set scenario. The test speaker could also be from outside the predefined group and it becomes an open-set scenario. The text dependent and text independent are the two modalities of automatic speaker recognition. In text dependent speaker recognition text of speech used for training and testing should be same where in text independent there is no constrain on the text. In this work performance evaluation of a text independent closed set speaker identification system is implemented. The primary step of any speaker identification is feature extraction followed by a robust speaker modelling technique for generalized representation of extracted features.
The block diagram of speaker identification system is shown in Fig. 1. In this experiment, the system is implemented

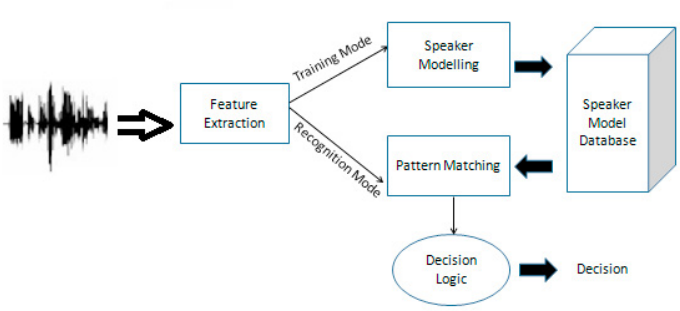

Fig. 1. Schematic diagram of Speaker Identification System

using probabilistic models like Gaussian Mixture-Universal background (GMM-UBM) and i-vector framework using both MFCC and IHC features. The distribution of feature vectors extracted from multiple speech utterances of different speakers are represented as a mixture or a weighted sum of multiple Gaussian distributions. Every speaker in the data base will be associated with a model that consist of a mixture of Gaussian whose parameters are estimated using Expectation Maximization algorithm. In i-vector approach speaker and channel variabilities are jointly represented in a single low dimensional space where i-vectors are estimated for every individual speaker. Automatic speaker identification system is implemented in two phases. First is the enrolment phase (training phase) where speaker models are built based on training features and second is identification phase (testing phase) where feature vectors of test speech utterance are compared against each of the trained speaker model and then identified the best match speaker model. The remainder of this paper is organized as follows. Section II presents the Feature extraction using MFCC and IHC section III describes speaker modelling methods, experiments and simulation results are given in Section IV and conclusion in section V

2. Feature Extraction

Feature extraction is the process of computing a sequence of features for each short-time frame of the input signal, with an assumption that such a small segment of speech is sufficiently stationary to allow meaningful modelling. In speech processing features are categorized into (1) short-term spectral (2) voice source (3) spectro-temporal (4) prosodic (5) high-level features based on their physical interpretations. Some of the short term spectral features are Mel-Frequency Cepstral Coefficient (MFCC) [2], Power Normalized Cepstral Coefficients (PNCC) [3], Relative Spectral Perceptual Linear Prediction (RASTA PLP) [4], Inner hair cell coefficient (IHC) [5].

2.1. Pre-processing

The speech signals sampled at a frequency of 16KHz is used as input signal for feature extraction. For MFCC and IHC features the pre-processing steps are the same. [6]

2.1.1. Pre-Emphasis

It is used to boost up the amplitude of high frequency signal and compensates for the human speech production process which tends to attenuate high frequencies. The first order high pass filter is applied to input signal and the response $H(z)$ is given by

$$H(z) = 1 - \alpha * (z^{-1}) \tag{1}$$

where is usually α between 0.9 and 1.In our method α is 0.97.

2.1.2. Framing

The speech signal is assumed to be stationary over short time frames 20-30ms. Hence input speech signal is divided into short frames of size 25ms with 15ms overlapping frames for our analysis.

2.1.3. Windowing

A window is needed to smooth the effect of using a finite-sized segment for the subsequent feature extraction by tapering each frame at the beginning and end edges. A Hamming window $\mathbf{w}[\mathbf{n}]$ function is given by equation

$$\mathbf{w}[n] = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}) \tag{2}$$

where $0 \leq n \leq N-1$, N is the window length. Signal is extracted by multiplying the value of the signal at time n, $\mathbf{s}[n]$ with window at time n, $\mathbf{w}[n]$ is given by equation

$$\mathbf{y}[n] = \mathbf{s}[n]\mathbf{w}[n] \tag{3}$$

2.2. MFCC Feature extraction

MFCC is an audio feature extraction technique which extracts speaker specific parameters from the speech [2]. The block diagram of MFCC feature extraction is shown in Fig. 2. Mel-Frequency Cepstral Coefficients (MFCC) is the most popular and dominant method to extract spectral features for speech by the use of perceptually based Mel spaced filter bank processing of the Fourier Transformed signal.

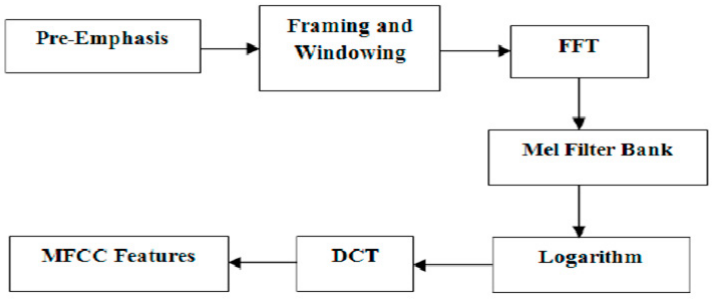

Fig. 2. Block diagram of MFCC Feature extraction

2.2.1. Fast Fourier Transform

The Fast Fourier Transform (FFT) is the commonly used algorithm for DFT and is applied on each frame to calculate the frequency spectrum, which is also called Short-Time Fourier-Transform (STFT), then we compute the power spectrum (periodogram).

2.2.2. Mel Filter Bank

The next step is applying triangular filters, on a Mel-scale to the power spectrum to extract frequency bands. Mel scale is a non linear frequency scale and we can compute mel frequency from frequency in hertz by the equation

$$Mel(f) = 1127 \ln(1 + \frac{f}{700}) \tag{4}$$

Mel-scale aims to mimic the non-linear human ear perception of sound, by being more discriminative at lower frequencies and less at higher frequencies. Each filter in the filter bank is triangular filter having a magnitude of 1 at the centre frequency and decrease linearly towards 0 till it reaches the centre frequencies of the two adjacent filters. Frequency domain filtering is obtained by element wise multiplication of power spectrum of signal with frequency response of the filter. In our work 26 overlapping filter banks are used.

2.2.3. Logarithm

Humans do not hear sound in a linear scale so next step is to compute the logarithm of magnitude spectrum.

2.2.4. Discrete Cosine Transform

Due to overlapped filter banks the filter energies are correlated. DCT is computed to decorrelate the filter bank energies. The cepstral values from second to fourteenth is taken and 13 Coefficient will represent information regarding vocal tract features.

2.2.5. Deltas and Double deltas

Speech signal is not constant from frame to frame. Features related to change in time are taken into account. The 13 delta or velocity and 13 double delta or acceleration features are obtained. The delta coefficients are obtained by equation

$$\delta(t) = c(t + 1) - c(t) \tag{5}$$

Double delta is computed from the delta at time t+1 to time t in a similar manner as we compute delta. In this experiment 39 MFCC coefficients including delta and double delta is computed.

2.3. Inner hair cell coefficients(IHC)

The meddis inner hair cell model describes the way transduction takes place in the inner hair-cell / auditory nerve on a level quite close to physiology [5]. The block diagram of IHC feature extraction technique is shown in Fig. 3

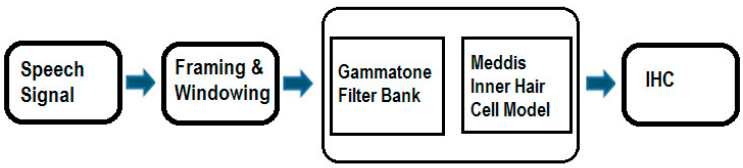

Fig. 3. Block diagram of IHC Feature extraction

2.3.1. Gammatone filter

The preprocessing stage is similar to that of MFCC and gammatone linear filter bank is used and the centre frequencies are linearly spaced in Equivalent Rectangular Bandwidth (ERB) between 200 Hz and 8000 Hz. A Gammatone filter is described by an impulse response that is the product of a gamma distribution and sinusoidal tone [7]. These are non uniform band pass filter banks that are designed to imitate the frequency resolution of human hearing system.

Gammatone filter bank simulate the motion of basilar membrane and are infinite impulse response filters is given by equation

$$h(t) = at^{n-1}e^{-2*b*t}\cos(2\pi f_c t + \psi) \tag{6}$$

n is the filter order, f_c is the centre frequency in Hz, $h(t)$ is the impulse response of filter, a is the filter gain, ψ is the phase in radians, b is the bandwidth. The ERB is a psychoacoustic measure of the width of the auditory filter at each point along the cochlea. Filters are equally placed from 100 Hz to 8000 Hz called equivalent rectangular bandwidth scale. A critical band or ERB filter models the signal that is present within a single auditory nerve cell or channel. Equivalent rectangular bandwidth scale is given by

$$ERB(f_c) = 24.7(\frac{4.3f_c}{1000} + 1) \tag{7}$$

2.3.2. Meddis Inner Hair Cell

The Meddis Inner hair cell model describes the transduction of sounds to neural signal taking place in the auditory nerve [5]. When sound reaches the ear, due to physical movement of inner hair cells a depolarization is caused which in turn results in a receptor potential. These potential will result in the release of neurotransmitters in the synaptic cleft which is a small gap between the hair cells and the auditory nerve. If the concentration of neurotransmitters is above a level an action potential is discharged (spike). A description of IHC modelling is as follows The neurotransmitters pass from inner hair cell to the cleft through permeable membrane. Permeability is the measure of movement of neurotransmitters from inner hair cell to the synaptic cleft and is not constant, it varies with the instantaneous amplitude of stimulus.

$$k(t) = \frac{g \times (s(t) + A)}{s(t) + A + B} \tag{8}$$

where $k(t)$ is the permeability, $s(t)$ is the instantaneous amplitude of stimulus, g is the maximum permeability, A is the lowest amplitude for which membrane is permeable, B is the specific rate at which permeability is approached Neurotransmitter concentration in the hair cell is given by equation

$$\frac{dq(t)}{dt} = y \times [1 - q(t)] \times r \times c(t) - k(t) \times q(t) \tag{9}$$

where $q(t)$ is the neurotransmitter level inside the hair cell, $k(t)$ is the permeability, y is the replenishment rate factor, r is the return factor from the cleft, $c(t)$ is the transmitter content of cleft, $y \times (1 - q(t)$ is the amount of neurotransmitter produced in the factory, $r \times c(t)$ is the re-uptake and $k(t) \times q(t)$ is the amount of neurotransmitters that passes to the synaptic cleft. Neurotransmitter concentration in the synaptic cleft is given by equation

$$\frac{dc(t)}{dt} = k(t) \times q(t) - l \times c(t) - r \times c(t) \tag{10}$$

l is the loss factor and r is the return factor, $l \times c(t)$ is the quantity of neurotransmitters lost from the cleft and $r \times c(t)$ is the amount of neurotransmitters return to the hair cell. Depending on the changes in the amount of neuro-transmitters inside the hair cell as well as in the synaptic cleft caused by an acoustic stimulus, a train of spike is produced. Spike probability is proportional to the amount of neurotransmitters in the synaptic cleft.

$$p(e) = h \times c(t) \tag{11}$$

$p(e)$ is the spike probability, $c(t)$ is the amount of transmitters in the synaptic cleft h is the proportionality factor. In our work 64 gammatone filters are used so 64 coefficients are extracted during each time frame.

2.4. Pitch and Formants

The voice source features reflects the speaker specific information. In this study pitch and formants are extracted to compare the performance of voice source features with short spectral and temporal features. Pitch is defined by fundamental frequency of vibration of vocal folds and formant is defined by resonant frequencies of vocal tract tube. The speaker recognition accuracy can be improved by appending pitch and formants with spectral and temporal features [8].

3. Probabilistic models

3.1. Gaussian Mixture Models

Gaussian Mixture Model is used for modelling the probability density function of a multi-dimensional feature vector [9]. An M component Gaussian densities as given by the equation:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} \pi_i b_i(\mathbf{x}) \tag{12}$$

$\mathbf{x}$ is a D-dimensional random vector, $b_i(\mathbf{x})$ with $i = 1, 2, ..., M$ are component densities, π_i with $i = 1, 2, ..., M$ are mixture weights with constraints $\pi_i \geq 0$, $\sum_{i=1}^{M} \pi_i = 1$ Each component density is a D-variate Gaussian function of the form shown in equation

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_i|^{1/2}} exp(\frac{-1}{2}(\mathbf{x} - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x} - \mu_i)) \tag{13}$$

with parameter vector μ_i and covariance matrix $\mathbf{\Sigma_i}$.We assume covariance matrix is a diagonal matrix. The complete Gaussian mixture density is parametrized by the mean vectors, covariance matrices and the mixture weights from all component densities. These parameters are collectively represented by the notation

$$\lambda = \{\pi_i, \mu_i, \mathbf{\Sigma}_i\}, i = 1, 2..M \tag{14}$$

Expectation Maximization (EM) algorithm is used for estimating GMM Parameters. Each speaker is represented by his/her model. For a sequence of T training vectors $X = \{ x_1, x_2, ...x_T\}$, GMM likelihood can be written as

$$P(X|\lambda) = \Pi_{t=1}^{T} p(x_t|\lambda) \tag{15}$$

The aim of Maximum Likelihood Estimation is to find the model parameters which maximize the likelihood of GMM given the training data. Universal Background Model (UBM) is trained using EM Algorithm which make use of all the speakers training speech utterances speaker specific GMM is created from the UBM using Maximum A Posteriori (MAP) Parameter estimation. During speaker identification GMM scoring is done where we calculate the log likelihood ratio of the test signal given the speaker model to the universal background model and one with maximum score is identified as the speaker.

3.2. i-Vector Approach

i Vector is a compact representation of speaker specific information based on joint factor analysis proposed by Dehak in May 2011. This method is based on the extraction of parameters (i-vectors) from a low dimensional space (Total variability Space) [10]. In this architecture no separate model for speaker and channel space is made. A single space named total factor space is constructed to model speaker and channel variability jointly [11]. The basic idea of i-vector approach is that each speaker and channel dependent GMM mean super vector M can be modelled as

$$\mathbf{M} = \mathbf{m} + \mathbf{Tw} \tag{16}$$

where $\mathbf{m}$ is a speaker and channel independent super vector whose values taken from UBM of size $CF \times 1$, $\mathbf{T}$ is a low rank matrix, which represent a basis of the reduced total variability space and $\mathbf{w}$ is i-vectors of size $R \times 1$. It is a vector with a prior of standard Gaussian distribution. Speaker frames of an utterance are represented by posterior estimation of $\mathbf{w}$ the new feature vector $\mathbf{w}$ is named total factor, often referred to as identity vector or i-vector. To compute the i-vectors Baum-Welch Statistics have to be estimated. Baum-Welch statistics are extracted using the UBM model. The posterior estimation of i-vector for a given utterance can be represented by the equation

$$\mathbf{w} = (\mathbf{I} + \mathbf{T}^t\mathbf{\Sigma}^{-1}\mathbf{N}(u)\mathbf{T})^{-1}\mathbf{T}^t\mathbf{\Sigma}^{-1}\tilde{\mathbf{F}}(u) \tag{17}$$

where $\mathbf{N}(u)$ is the diagonal matrix of dimension $CF \times CF$ whose diagonal blocks are $N_c(u)\mathbf{I}$, where $(c = 1, ...C)$, $\mathbf{I}$ is the identity matrix of dimension $F \times F$, $N_c(u)$ is a zero order Baum-Welch statistics, $\tilde{\mathbf{F}}(u)$ is a super vector of dimension

$CF \times 1$ obtained by concatenating all the centralized first order Baum-Welch statistics $\tilde{\mathbf{F}}_c(u)$. Σ is a diagonal covariance matrix of dimension $CF \times CF$ that is estimated during the training of $\mathbf{T}$.
If the dimension of the feature vectors is F and the number of mixture components in the GMM is C, then the vectors $\mathbf{M}$ and $\mathbf{m}$ has dimension of $CF \times 1$, $\mathbf{T}$ is of dimension $CF \times R$ and $\mathbf{w}$ has the size of $R \times 1$.

3.2.1. Channel Compensation Techniques

The i-vector frame work is used in this paper to account for channel compensation [12]. During i-vectors extraction no distinction is made between channel and speaker variability. To overcome this problem, channel variability has to be considered while constructing classifiers for speaker recognition using i-vector. The first approach is within-class covariance normalization (WCCN) where the inverse of the within-class covariance is used to normalize the linear kernel. The second approach is Linear Discriminative Analysis (LDA) attempts to define new optimal mapping that minimize the intra-class variance caused by channel effects, and to maximize the variance between speakers.

3.2.2. Classifiers

In this paper i-vector after channel compensation is followed by a simple variant of PLDA (Probabilistic Linear Discriminant Analysis) ie Gaussian PLDA [13]. For identifying speaker we compute the likelihood ratio of two i-vectors(test and enrolment) given the two hypothesis. The H_1 indicates that both i-vectors come from same speaker and H_0 indicates that both i-vectors come from different speaker.PLDA scoring $s_{lin}(\mathbf{w_1}, \mathbf{w_t})$ by

$$s_{lin}(\mathbf{w_1}, \mathbf{w_t}) = \frac{p(\mathbf{w_1}, \mathbf{w_t}|H_1)}{p(\mathbf{w_1}|H_0)P(\mathbf{w_t}|H_0)} \tag{18}$$

where $\mathbf{w_1}$ indicates enrolment i-vector, $\mathbf{w_t}$ indicates test i-vector. One with maximum scoring is identified as speaker. Given the Gaussian assumptions, the log likelihood ratio can be computed in closed form.

4. Experiment and Result

This study has used MATLAB software for implementation. Experiments are carried out on 600 different speakers of the TIMIT database, which consists of 10 speech utterances for each speaker. The data base consist of speech signals sampled at a frequency of 16KHz and of duration within 2-3 seconds. Out of the above mentioned 10 speech utterances 7 speech utterances are selected randomly for training and the remaining 3 are used for testing. In this work speech is normalized in the range of -1 to 1 and we assume the parameters zero crossing rate and energy for separating voiced and unvoiced part and threshold as 100 and 0.5 [14]. Thirty nine MFCC and sixty four IHC dimensional feature vectors are extracted for each time frame of full speech and voiced speech frames respectively. Training was done on short utterance of 2-3 sec full speech, voiced speech features and also by appending voice source features like pitch and formants with voiced speech features. In our work seven additional feature vector (one pitch and six formants per frame) are appended with the voiced speech features to improve the identification rate of voiced speech and the dimension of this new feature vector is 46 and 71 respectively. Testing is performed on short utterance and also on long utterance of 6-9 sec constructed by concatenating three speech test utterance. Accuracy is measured in terms of percentage of correctly identified speakers.

4.1. Simulation Results

Accuracy rate of speaker recognition system implemented using GMM and i-vector with MFCC and IHC features are shown in tables. From Table 1 and Table 3 it was understood full speech performance is better than voiced speech in both GMM and i-vector. Table 2 and Table 4 shows the performance for long utterance test signal with GMM and i-vector. Accuracy rate increases significantly when compared with short utterance test speech signal. Table 5 and Table 6 shows the IHC feature performance in GMM and found that voiced speech performance is better than full speech for both short and long utterances. In i-vector approach full speech is showing more accurate than voiced with IHC test feature which is shown in Table 7 and Table 8. In all cases it was found that by appending speaker specific informations like pitch and formants with voiced speech the accuracy increases than voiced speech and the new dimension of feature vector is 47 for MFCC and 71 for IHC. In our work GMM mixing coefficient selected is M= 32,and i-vector dimension is 100 with LDA dimension is min(100,number of classes-1).

Table 1. Accuracy Rate for Speaker Identification using GMM with MFCC Feature for Short utterances

No.of Speakers	Voiced Speech (%)	Voiced Speech + Voice source Features (%)	Full Speech (%)
100	90	92.33	97.66
200	87.16	91	97.5
300	87	88.33	95.66
400	81.12	86.58	95
500	81	85.26	94.80
600	78.88	83.27	94.22

Table 2. Accuracy Rate for Speaker Identification using GMM with MFCC Feature for long utterances

No.of Speakers	Voiced Speech (%)	Voiced Speech + Voice source Features (%)	Full Speech (%)
100	99	98	99
200	98	99.5	99.5
300	97.66	99	99.33
400	96.5	98.75	99
500	96.6	98.6	99.4
600	96.66	98.11	99.33

Table 3. Accuracy Rate for Speaker Identification using i-vector with MFCC Feature for Short utterances

No.of Speakers	Voiced Speech (%)	Voiced Speech + Voice source Features (%)	Full Speech (%)
100	72.66	80	93
200	64.88	73.88	90.16
300	59.44	70.11	86.22
400	58.08	67.16	85.08
500	57.16	66.66	82.13
600	57.01	62.83	80.8

Table 4. Accuracy Rate for Speaker Identification using i-vector with MFCC Feature for long utterances

No.of Speakers	Voiced Speech (%)	Voiced Speech + Voice source Features (%)	Full Speech (%)
100	97	96	100
200	96.5	97	100
300	94	97.33	99.3
400	93.5	95.75	99.25
500	95.2	95.4	99.2
600	93.33	94.33	99.16

Table 5. Accuracy Rate for Speaker Identification using GMM with IHC Feature for Short utterances

No.of Speakers	Voiced Speech (%)	Voiced Speech + Voice source Features (%)	Full Speech (%)
100	90	90.66	87.3
200	89.16	89.83	83
300	84.77	85.66	82.8
400	83.5	84.16	72.91
500	82	82.98	70.46
600	79.5	81.4	66.00

Table 6. Accuracy Rate for Speaker Identification using GMM with IHC Feature for long utterances

No.of Speakers	Voiced Speech (%)	Voiced Speech + Voice source Features (%)	Full Speech (%)
100	100	100	99
200	97	97	93.5
300	96.67	96.67	91
400	94.5	95	87.25
500	94	95	87.4
600	93	95.33	86.64

Table 7. Accuracy Rate for Speaker Identification using i-vector with IHC Feature for Short utterances

No.of Speakers	Voiced Speech (%)	Voiced Speech + Voice source Features (%)	Full Speech (%)
100	62	68	67.33
200	46	60	57.16
300	41.2	57.22	53.88
400	35.33	52.66	50.2
500	32.4	49.4	47.16
600	30.22	46.5	44.22

Table 8. Accuracy Rate for Speaker Identification using i-vector with IHC Feature for long utterances

No.of Speakers	Voiced Speech (%)	Voiced Speech + Voice source Features (%)	Full Speech (%)
100	87	94	97.00
200	76.5	87.5	90.5
300	72.67	84.33	89.33
400	67.5	80.75	85.75
500	68.2	80.2	83.6
600	65.17	79.17	79.5

5. Conclusion

Speaker Recognition Systems perform the task of identifying the persons from their voices. Speaker recognition systems has gained popularity in a wide range of biometric applications such as access control, forensic investigation, user authentication in telephone banking. This study focus on the performance of speaker recognition system using MFCC and IHC features with two different probabilistic models GMM and i-vector. It has been found that MFCC outperforms better than IHC for speaker Identification using GMM and i-vector approach. By appending voice source features pitch and formants with voiced Speech gives better results than voiced speech in both GMM and i-vector. i-vector performs well when longer utterances are used for testing. The performance of voiced speech IHC feature which simulates the physiological behaviour of human ear is better in terms of accuracy than full speech (voiced and unvoiced) in GMM. The accuracy reduces as the number of speakers increases due to increased number of targets.

References

[1] Homayoon Beigi. (2011) "Fundamentals of Speaker recognition."*Springer science and Business media.*

[2] Sirko Molau, Michael Pitz, Ralf Schluter and Hermann Ney. (2001) "Computing Mel frequency Cepstral Coefficients on the power spectrum." *IEEE Transactions on Audio, Speech and Language Processing*

[3] Chanwoo Kim and Richard M. Stern. (2016) "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition." *IEEE Transactions on Audio, Speech and Language Processing*, 1315–1329.

[4] Hynek Hermansky and Nelson Morgan. (1994) "RASTA Processing of Speech" *IEEE Transactions on Speech and Audio Processing*, 578–589.

[5] Thomas Ulrich Christiansen. (1986) "The Meddis Inner Hair-Cell Model" *Journal of the Acoustical Society of America, vol 79*

[6] Siddhant C. Joshi and Dr. A.N.Cheeran. (2014) "MATLAB Based Feature Extraction Using Mel Frequency Cepstrum Coefficients for Automatic Speech Recognition." *International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 6.*

[7] Malcolm Slaney.(1993) "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank." *Apple Computer Technical Report 35 Perception Group-Advanced Technology Group*

[8] K. Sreenivasa Rao and Shashidhar G. Koolagudi. (2013) "Robust Emotion Recognition using Pitch Synchronous and Sub-syllabic Spectral Features. " *Springer Briefs in Electrical and Computer Engineering, New York* pp 17–46

[9] D. A. Reynolds and R. C. Rose. (1995) "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models." *IEEE Trans. on Speech and Audio Processing, vol.3, No.1,* pp.72– 83.

[10] Mohammed Senoussaoui, Patrick J. Kenny, Najim Dehak, Pierre Dumouchel.(2011) "An i-vector Extractor for speaker Recognition with Microphone and Telephone Speech. " *CSAIL-MIT*

[11] Wei Li , Tianfan Fu and Jie Zhu.(2015) "An improved i-vector extraction algorithm for speaker verification Transient Flow." *EURASIP journal on Audio, Speech and Music Processing,* pp 1-9

[12] Najim Dehak, Patrick J. Kenny, Dehak, Pierre Dumouchel and Pierre Ouellet. (2011) "Front-End Factor Analysis for Speaker Verification." *Transactions on Audio, Speech and language Processing,vol.24,No.3*

[13] Padmanabhan Rajana b, Anton Afanasyeva, Ville Hautamakia and Tomi Kinnunena.(2014) "From single to multiple enrollment i-vectors: practical PLDA scoring variants for speaker verification. " *Digital Signal Processing*

[14] Dominic Mathew, V.D.Devessia and Tessamma Thomas. (2006) "A K-means Clustering Algorithm for Frequency Estimation and Classification of Speech Signals." *IEEEconference/ICSIP*.