Indian Journal of Science and Technology, Vol 9(28), DOI: 10.17485/ijst/2016/v9i28/74624, July 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Speaker Adaptation on Hidden Markov Model using MFCC and Rasta-PLP and Comparative Study

Shweta Bansal[1*], Atul Kumar[2] and S. S. Agrawal[1]

[1]KIIT College of Engineering, Gurgaon - 122102, Haryana, India; bansalshwe@gmail.com, ss_agrawal@hotmail.com
[2]Ansal University, Gurgaon - 122003, Haryana, India; atulkumar@ansaluniversity.edu.in

Abstract

This work compares the performance of the Mel-Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) features for developing a text-dependent speaker identification system. Continuously spoken Hindi speech sentences have been used to train the HMM models using HTK toolkit for each speaker separately. The experiments have been performed using a set of 200 continuously spoken sentences with vocabulary of 20000 isolated words using a database of 100 speakers. The results show an accuracy of 92.26% recognition when PLP features have been used and accuracy of 91.18% for MFCC features. A confusion matrix has been created for all the 20 test speakers based on the recognition scores obtained for each of these speakers and their confusion with other speakers. Performance has been compared in the closed set and open set conditions of testing and as it is expected, the performance in the closed set condition is far better than the open set. We propose that if PLP features are used in place of MFCC, they may provide improvement in speaker identification accuracy by reducing the cases of false acceptance.

Keywords: Hindi Speech, HMM, MFCC, RASTA-PLP, Speaker Identification

1. Introduction

A speech signal not only contains information about text message but about the identity of the speaker. In speaker recognition, therefore, an attempt is made to extract the features and parameters from the signal which are helpful to identify or verify a speaker[1]. In the past, several experiments have been performed using MFCC features[2] but less number of experiments using PLP features[3]. These results have shown varying performances. Richard M. Stern et al. have found WER of 35.1% for MFCC and 38.0% for PLP features, for NRL Spine database[4], but very few experiments have been done to compare the performance of speaker recognition using both MFCC and PLP features on the same database-particularly using Hindi database.

In the present paper, an attempt has been made to conduct speaker identification experiments from MFCC and PLP features. HTK toolkit[5] has been used to model the HMM classifier. The results obtained on using both the feature extraction techniques have been compared.

2. Overview of the System

Figure1 gives the overview of our approach for this study. The input speech is passed through front end processing unit. It includes the alignment of audio files and their transcription. Necessary processing is done to extract MFCC and PLP features from the speech signal. These feature vectors help to create the acoustic models containing all the feature vectors and duration of the words which exist in a particular sentence.

The lexicon module along with the language model plays the supportive role to find the system accuracy. The lexicon module contains all the lexicons (phones) contained in our database. Language model has been used to capture the possible variations in pronunciation uttered by a particular speaker, used to tied-states in HMM classifiers. We have tried to develop a phone-based trigram model to capture the sequence of phones in a word and then words in a particular sentence. With the help of all these inputs, the system follows the matching approach

*Author for correspondence

using Viterbi algorithm to find the best accuracy score of the words in a particular sentence[6]. The 1-best hypothesis of a sentence has been taken.

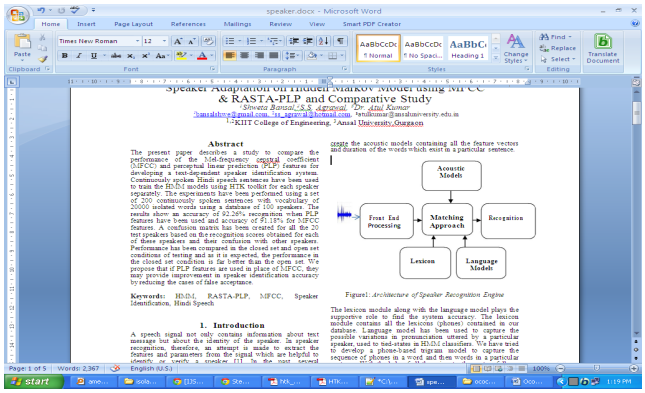

Figure 1. Architecture of speaker recognition engine.

3. Methodology

3.1 Database Collection

The present system uses a data-set of 200 phonetically rich Hindi sentences recorded by 100 native Hindi speakers (50 male and 50 female) within the age group of 18 to 30 years. Recording was carried out in a sound treated room environment having S/N > = 40 db. These were sampled at the rate of 16 bit-16 kHz. Eighty speakers were used for training the system and the other twenty speakers for testing the system in open-set condition. Similarly, the system has been trained for one hundred speakers and tested by using twenty speakers out of these 100 speakers in closed-set condition.

3.2 Pronunciation Dictionary

The following steps have been followed for the creation of pronunciation dictionary[7]:

Formation of Sentences

↓

Extraction of 20000 words from 200 sentences

↓

Total word tokens are 8500 out of 20000 words

↓

Create a dictionary of possible pronunciations of one word uttered by different speakers

3.3 Transcription

For parallel alignment of audio and text data of various speakers, phonetic transcription has been done manually. For alignment purpose, all wave files were converted into text for the 100 speakers.

3.4 Creation of Phone Sets

To capture the sound units in a word, phone-sets of Hindi have been created. 45 phones exist in our database, shown in Figure 2.

अ[ə], आ[ɑ], इ[ɪ], ई[i], उ[ʊ] ऊ [o], ए[e], ऐ[ɛ], अ[ə], क[k], ख[kʰ], ग[g], घ[gʰ], च[tʃ], छ[tʃʰ], ज[dʒ], झ[dʒʰ], प[p], फ[pʰ], ब[b], भ[bʰ], म[m], त[t], थ[tʰ], द[d], ध[dʰ], न[n], ट[ʈ], ठ[ʈʰ], ड[ɖ], ढ[ɖʰ], य[j], र[ɾ], ल[l], व[w], श[ʃ], ष[ʂ], स[s], ज़[z], ओ[ə], फ़[f], औ[ɔ:], ह[ɦ], ड़[ɽ], ण[ɳ]

Table 3: Perception of Hindi consonants spoken by non native speakers and perceived by Hindi natives

3. ACOUSTIC ANALYSIS

As the process of speech perception hinges on the acoustic spectrum, the formant frequencies provide enough information to distinguish the speech sounds. We have

Figure 2. Hindi phone set.

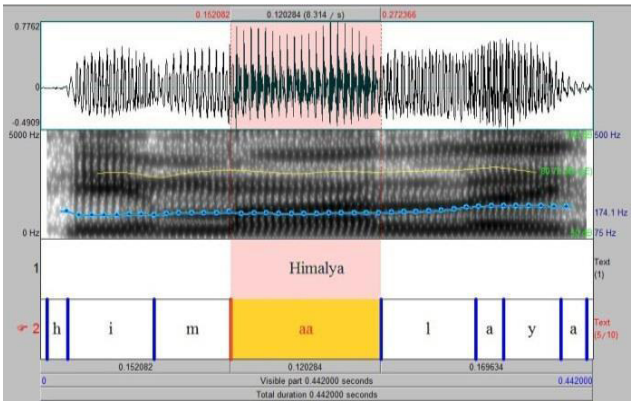

Figure 2. Hindi phone set.

3.5 Annotation at Phoneme Level

Speech database of 100 speakers were annotated at phoneme level using PRAAT software tool to capture the

duration of individual phonemes. Figure 3 represents the samples of annotated files of two Hindi words (Himalaya and Vadiyaan) at phoneme level.

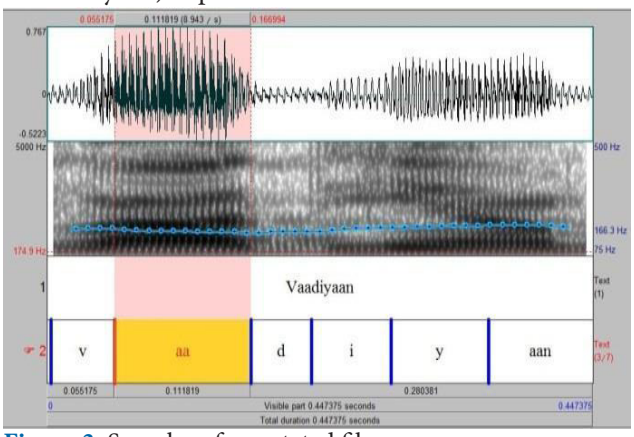

Figure 3. Samples of annotated file.

3.6 Parameterization

For extracting the relevant information from speech spectra MFCC and RASTA-PLP have been used for speaker diarization and compare their results using HTK[8].

3.6.1 Mel Frequency Cepstrum Coefficient (MFCC)

Mel Frequency Cepstrum Coefficients are widely used for parameterization[9]. For MFCC extraction, the source kind is wave and the target kind is MFCC_0[10]. The Hamming Window size used in the experiment has been taken to be 25 ms with a frame shift of 10 ms. Pre-emphasis co-efficients have been set as 0.97. Twenty-two band pass filters have been taken to capture all the co-efficients related to our database. For MFCC extraction, Logarithmic compression has been taken into consideration.

3.6.2 Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP)

RASTA-PLP is a technique of warping spectra to minimize the differences between speakers while preserving the important speech information[11]. It is an approach completely based on perceptual linear prediction.

The whole RASTA-PLP extraction is illustrated in the following figure:

For PLP extraction, the source format is wave and the target kind is PLP_0[11]. Similar to MFCC, in this case also the Hamming window size of 25 ms and 22 bandpass filters have been taken to capture all the co-efficients from the database. For PLP extraction, cubic root compression have been taken into account.

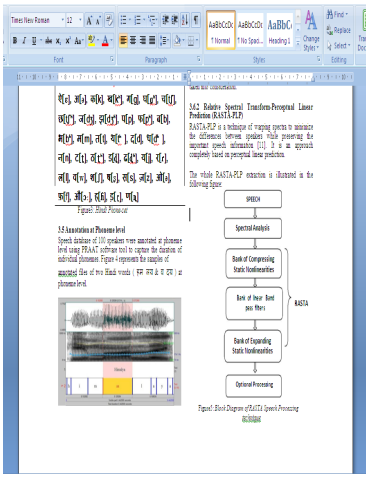

Figure 4. Block diagram of RASTA speech processing technique.

3.7 Speaker Modeling using HMM

We trained the recognition engine using MFCC and RASTA-PLP features for a database of 200 sentences for all the 100 speakers separately. For speaker identification, two models have been created - one is acoustic model which estimates the means and variances of a HMM in which each state output distribution is a single component Gaussian and the other is text-dependent language model[12] which is phone based trigram model to capture the sequence of phones in a word and then sequence of words in a particular sentence. To determine more accurate parameters, Baum-Welch re-estimation has been used[13].

We perform single-pass retraining in which the state/component occupation probabilities are calculated using an existing model and training set, but the new model parameters are calculated using a new training set[14].

3.8 Testing of the Recognition Engine

To recognize unknown test words, the likelihood of each model generating the word is calculated using Viterbi Algorithm[15] and the most likely model identifies the word. The testing is done separately with both MFCC and RASTA-PLP features.

4. Experimental Results

Figure 5 shows the performance of conducting the recognition tests. For example, it can be seen that if PLP features are used in place of MFCC, they may provide improvement in speaker identification accuracy and the cases of false acceptance may be reduced.

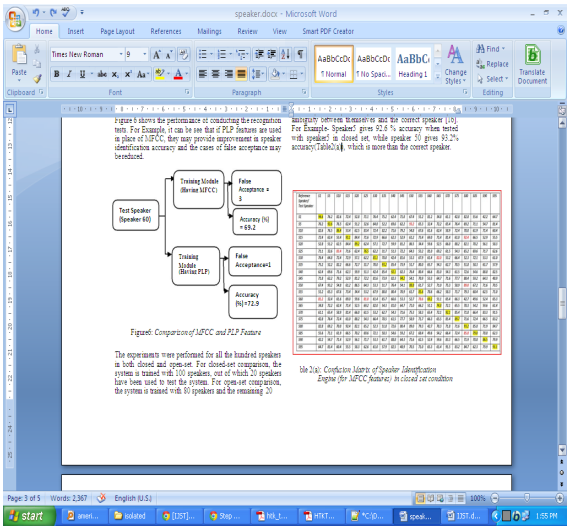

Figure 5. Comparison of MFCC and PLP feature.

The experiments were performed for all the hundred speakers in both closed and open-set. For closed-set comparison, the system is trained with 100 speakers, out of which 20 speakers have been used to test the system. For open-set comparison, the system is trained with 80 speakers and the remaining 20 speakers have been taken as a test data. Minimum threshold figure for accuracy has been considered as 60. Although the sentence contains varying number of words, the total number of words are 20000. The comparison with correct recognition and the false acceptance are shown in Table 1.

Table 1. Performance of system in closed and open set

Test	Training	Correct		Correct Rejection	
Speakers	Utterances	Recognition (%)		(%)	
		MFCC	RASTA-	MFCC	RASTA-
			PLP		PLP
20	20000	91.2	92.3	8.8	7.8
(Closed					
Set)					
20	16000	64.2	74.6	35.8	25.4
(Open					
Set)					

The above results show that performance of system is significantly better when RASTA-PLP is used, particularly in case of open-set comparison.

Table 2(a). Confusion matrix of speaker identification engine (for MFCC features) in closed set condition

Reference Speaker/ Test Speaker	S1	S5	S10	S15	S20	S25	S30	S35	S40	S45	S50	S55	S60	S65	S70	S75	S80	S85	S90	S95
S1	94.8	81.4	83.2	76.5	71.2	79.1	82.2	81.5	75.2	79.6	77.7	60.0	66.3	61.1	72.4	65.5	70.2	63.3	67.4	54.3
S5	81.4	83.6	72.0	75.4	63.3	52.8	75.1	76.2	80.0	71.3	91.2	66.1	52.9	74.4	78.3	81.2	76.3	72.4	69.1	83.7
S10	83.2	72.0	87.3	61.1	62.6	74.0	76.2	68.6	65.1	62.0	64.0	72.5	72.3	75.0	83.4	63.1	92.2	71.0	62.4	67.3
S15	76.5	75.4	61.1	92.2	86.2	74.3	75.1	70.3	69.1	70.2	68.6	75.2	71.3	75.5	86.4	66.6	91.2	68.0	66.6	59.0
S20	71.2	63.3	62.6	86.2	90.3	63.3	57.4	73.7	60.4	83.1	87.0	36.2	59.9	53.1	66.2	89.0	83.1	71.2	60.1	62.4
S25	79.1	52.8	74.0	74.3	63.3	80.0	62.3	41.9	53.1	73.2	66.2	55.4	62.0	70.3	62.3	59.2	69.2	70.6	72.3	64.3
S30	82.2	75.1	76.2	75.1	57.4	62.3	83.4	72.2	45.1	82.6	55.5	69.1	62.1	83.1	54.0	68.1	55.5	73.0	53.9	71.0
S35	81.5	76.2	68.6	70.3	73.7	41.9	72.2	94.0	86.2	74.2	51.7	82.2	68.3	56.3	63.1	72.1	60.2	62.2	62.1	68.4
S40	75.2	80.0	65.1	69.1	60.4	53.1	45.1	86.2	99.1	83.4	78.0	81.4	69.1	82.2	55.2	62.3	74.1	59.9	70.3	85.0
S45	79.6	71.3	62.0	70.2	83.1	73.2	82.6	74.2	83.4	92.0	56.3	72.0	58.2	65.4	72.5	79.0	82.2	64.0	64.3	57.1
S50	77.7	91.2	64.0	68.6	87.0	66.2	55.5	51.7	78.0	56.3	90.2	62.3	54.3	72.4	76.2	62.2	90.1	69.6	72.3	75.1
S55	60.0	66.1	72.5	75.2	36.2	55.4	69.1	82.2	81.4	72.0	62.3	83.6	79.2	68.6	59.9	73.3	81.2	72.2	65.4	72.3
S60	66.3	52.9	72.3	71.3	59.9	72.0	62.1	68.3	69.1	58.2	54.3	79.2	72.9	52.4	66.6	69.1	50.6	59.2	52.9	66.6
S65	61.1	74.4	75.0	75.5	53.1	70.3	83.1	56.3	82.2	65.4	72.4	68.6	52.4	83.0	73.4	69.0	71.4	61.2	59.9	62.9
S70	72.4	70.3	83.4	86.4	66.2	62.3	54.0	63.1	55.2	72.5	76.2	59.9	66.6	73.4	90.2	85.0	72.9	71.3	85.0	89.6
S75	65.5	81.2	63.1	66.6	89.0	59.2	68.1	72.1	62.3	79.0	62.2	73.3	69.1	69.0	85.0	90.3	72.0	74.6	69.1	84.0
S80	70.2	76.3	92.2	91.2	83.1	69.2	55.5	60.2	74.1	82.2	90.1	81.2	50.6	71.4	72.9	72.8	92.4	86.2	73.1	85.2
S85	63.3	72.4	71.0	68.0	71.2	70.6	73.0	62.2	59.9	64.0	69.6	72.2	59.2	61.2	71.3	74.6	86.2	70.2	78.0	64.1
S90	67.4	69.1	62.4	66.6	60.1	72.3	53.9	62.1	70.3	64.3	72.3	65.4	52.9	59.9	85.0	69.1	73.1	78.9	87.4	80.4
S95	54.3	83.7	67.3	59.0	62.4	64.3	71.0	68.4	85.0	57.1	75.1	72.3	66.6	62.9	89.6	84.0	85.2	64.1	80.4	92.5

Table 2(b). Confusion matrix of speaker identification engine (for PLP features) in closed set condition

Reference Speaker/ Test Speaker	S1	S5	S10	S15	S20	S25	S30	S35	S40	S45	S50	S55	S60	S65	S70	S75	S80	S85	S90	S95
S1	94.6	76.2	82.6	72.4	52.0	71.1	76.4	75.2	62.4	71.8	67.4	51.2	81.2	34.8	61.1	42.0	82.8	55.6	42.2	64.7
S5	76.2	92.6	76.5	62.4	51.2	32.6	64.8	52.2	69.6	62.2	93.2	65.3	32.4	72.2	65.4	76.4	69.2	71.1	54.7	81.4
S10	82.6	76.5	86.4	53.4	62.5	83.4	72.4	82.2	71.6	79.2	54.0	67.6	61.6	62.4	58.9	72.4	70.0	61.9	71.4	60.4
S15	72.4	62.4	53.4	91.1	84.4	71.6	72.9	66.6	62.3	52.9	61.2	73.4	69.0	71.4	81.4	61.0	92.4	66.5	52.9	55.5
S20	52.8	51.2	62.5	84.4	89.2	62.4	57.1	72.7	59.9	81.2	86.5	34.4	59.6	52.5	66.0	88.2	82.1	70.2	56.1	58.3
S25	71.1	32.6	83.4	71.6	62.4	76.5	62.2	31.7	51.3	72.2	64.3	53.2	81.0	69.2	61.5	54.3	65.2	69.6	71.7	62.6
S30	76.4	64.8	72.4	72.9	57.1	62.2	81.1	70.0	42.4	81.6	53.3	67.9	61.4	82.0	53.2	66.4	52.3	72.1	53.3	61.0
S35	75.2	52.2	82.2	66.6	72.7	31.7	70.0	93.2	85.4	73.9	51.7	80.0	65.7	54.3	62.7	70.5	51.8	58.3	61.7	57.9
S40	62.4	69.6	71.6	62.3	59.9	51.3	42.4	85.4	92.1	82.3	76.4	80.4	66.6	81.0	54.3	61.5	72.6	54.6	80.8	82.5
S45	71.8	62.2	79.2	52.9	81.2	72.2	81.6	73.9	82.3	94.2	54.1	70.9	53.3	64.7	71.6	77.7	80.4	59.2	64.3	49.9
S50	67.4	93.2	54.8	61.2	86.5	64.3	53.3	51.7	76.4	54.1	88.0	61.7	52.7	71.0	75.3	58.9	89.0	67.2	71.6	70.5
S55	51.2	65.3	67.6	73.4	34.4	53.2	67.9	80.0	80.4	70.9	61.7	81.6	70.6	66.2	58.3	71.7	79.3	68.4	62.5	71.0
S60	81.2	32.4	61.6	69.0	59.6	81.0	61.4	65.7	66.6	53.3	52.7	70.6	69.2	51.1	65.4	66.3	42.7	49.6	52.4	65.3
S65	34.8	72.2	62.4	71.4	52.5	69.2	82.0	54.3	81.0	64.7	71.0	66.2	51.1	74.8	72.1	65.5	70.3	54.2	59.6	61.4
S70	61.1	65.4	58.9	81.4	66.0	61.5	53.2	62.7	54.3	71.6	75.3	58.3	65.4	72.1	92.1	85.4	71.8	66.4	83.3	91.5
S75	42.8	76.4	72.4	61.0	88.2	54.3	66.4	70.5	61.5	77.7	58.9	71.7	66.3	65.5	85.4	89.7	71.6	72.4	66.5	83.2
S80	82.8	69.2	70.0	92.4	82.1	65.2	52.3	51.8	72.6	80.4	88.0	79.3	42.7	70.3	71.8	71.6	93.2	85.2	71.9	84.7
S85	55.6	71.1	61.9	66.5	70.2	69.6	72.1	58.3	54.6	59.2	67.2	68.4	49.6	54.2	66.4	72.4	65.8	79.0	70.8	62.3
S90	42.2	54.7	71.4	52.9	56.1	73.7	53.3	61.7	80.8	64.3	71.6	62.5	52.4	59.6	83.3	66.5	73.9	70.8	86.5	79.9
S95	64.7	81.4	60.4	55.5	58.3	62.6	61.0	57.9	82.5	48.9	70.5	71.0	65.3	61.4	91.5	83.2	84.7	62.3	79.9	91.1

4.1 Confusion Matrix

In order to find out the individual performance of test speakers, confusion matrices has been made for the results obtained using MFCC and PLP features. These are shown in Table 2 (a and b) and Table 3 (a and b) for the closed set and open set experiments respectively. The confusion matrix show that in some cases the accuracy is higher for a test speaker even though the reference speaker is not same and providing ambiguity between themselves and the correct speaker[16]. For example – Speaker 5 gives 92.6% accuracy when tested with speaker 5 in closed set,

while speaker 50 gives 93.2% accuracy (Table 2(a)), which is more than the correct speaker.

Table 3(a). Confusion Matrix of Speaker Identification Engine (for MFCC features) in open set condition

Reference Speaker/ Test Speaker	S1	S5	S10	S15	S20	S25	S30	S35	S40	S45	S50	S55	S60	S65	S70	S75	S80	S85	S90	S95
S1	52.6	42.8	67.2	43.4	52.2	51.0	43.4	46.9	51.3	51.0	46.9	51.2	50.9	34.8	51.1	42.8	62.8	50.6	42.2	44.7
S2	66.2	72.8	66.5	62.4	51.2	32.6	64.8	52.2	69.6	62.2	82.4	65.3	32.4	72.2	65.4	70.4	69.2	71.1	54.7	81.4
S3	42.8	55.5	56.4	52.9	62.5	53.4	43.7	61.3	46.6	52.4	54.8	32.7	42.8	52.9	53.4	61.6	50.2	51.4	32.7	42.8
S4	32.6	54.3	53.4	71.9	43.2	51.3	43.2	46.4	62.3	52.9	61.2	73.4	69.0	71.4	61.5	61.0	72.4	66.5	52.9	55.5
S5	42.8	33.4	61.5	72.8	76.2	62.4	57.1	48.9	56.3	61.2	54.5	34.4	41.2	53.6	44.2	54.3	54.9	32.1	54.6	42.3
S6	71.1	32.6	83.4	71.6	62.4	76.5	62.2	31.7	51.3	72.2	64.3	53.2	81.0	69.2	61.5	54.3	65.2	69.6	71.7	62.6
S7	76.4	64.8	52.4	62.3	57.1	51.9	61.1	60.0	42.4	62.3	54.1	62.1	61.0	72.0	53.2	61.8	52.3	60.0	53.3	61.0
S8	56.5	43.2	44.4	32.3	61.7	31.7	70.0	63.2	62.4	73.9	51.7	59.0	54.5	54.3	62.7	63.4	51.8	58.3	61.7	56.3
S9	62.4	49.6	48.4	51.3	49.0	51.3	42.4	61.3	52.4	50.0	50.3	51.5	50.9	51.3	43.3	48.6	52.4	43.9	52.1	55.3
S10	71.8	62.2	79.2	52.9	71.3	72.2	71.3	73.9	77.6	74.2	54.1	70.9	53.3	64.7	71.6	70.4	72.3	59.2	64.3	48.9
S11	62.4	61.6	54.8	61.2	66.5	64.3	53.3	51.7	64.3	54.1	68.8	61.7	52.7	71.0	62.4	58.9	69.6	67.2	71.6	62.3
S12	61.2	65.3	66.3	61.8	34.4	52.7	65.3	67.1	71.9	70.9	61.7	70.2	66.6	66.2	58.3	68.9	66.9	62.1	62.5	66.3
S13	81.2	32.4	61.6	69.0	59.6	81.0	61.4	65.7	66.6	53.3	52.7	78.6	69.2	51.1	65.4	66.3	42.7	49.6	52.4	65.3
S14	44.8	72.2	62.4	64.3	43.2	45.6	32.7	57.8	62.3	64.7	71.0	45.5	51.0	64.8	62.3	54.3	54.4	55.3	56.3	54.4
S15	61.1	60.4	58.9	54.8	34.7	45.9	53.2	32.8	58.3	54.5	61.3	32.7	54.3	45.9	62.1	85.4	31.8	57.7	45.6	43.9
S16	42.8	56.4	62.3	61.0	58.2	54.3	54.9	43.1	57.7	54.8	58.8	52.9	61.3	51.9	52.3	59.7	41.8	45.8	46.2	48.3
S17	32.8	35.5	41.6	42.1	33.3	46.4	42.3	41.5	32.6	30.0	49.2	39.3	32.7	30.0	41.0	41.6	43.2	45.8	31.9	34.7
S18	55.6	71.1	61.9	66.5	70.2	69.6	72.1	58.3	54.6	59.2	67.2	68.4	49.6	54.2	66.4	72.4	85.8	79.0	78.8	62.3
S19	42.2	54.7	71.4	52.9	56.1	71.7	53.3	61.7	54.8	64.3	64.9	62.5	52.4	59.6	63.3	66.5	62.8	59.1	66.5	62.0
S20	64.7	70.4	60.4	55.5	58.3	62.6	61.0	57.9	72.6	48.9	70.5	71.0	65.3	61.4	62.3	63.2	54.7	55.0	52.3	71.1

Table 3(b). Confusion matrix of speaker identification engine (for P features) in open set condition

Reference Speaker/ Test Speaker	S1	S5	S10	S15	S20	S25	S30	S35	S40	S45	S50	S55	S60	S65	S70	S75	S80	S85	S90	S95
S1	62.8	52.3	61.2	52.4	52.2	61.0	59.4	56.5	51.3	51.0	59.9	61.3	50.4	60.2	62.8	61.3	62.8	50.6	42.2	44.7
S2	66.2	76.3	66.5	62.4	51.2	32.6	64.8	52.2	69.6	62.2	72.4	65.3	32.4	72.2	65.4	70.4	69.2	71.1	54.7	72.3
S3	52.8	56.4	53.9	57.6	62.5	54.8	57.2	51.3	53.4	59.2	54.8	32.7	42.8	52.9	53.4	61.6	50.2	51.4	32.7	42.8
S4	32.6	54.3	53.4	76.9	43.2	51.3	43.2	46.4	62.3	52.9	61.2	73.4	69.0	71.4	61.5	61.0	72.4	66.5	52.9	55.5
S5	42.8	33.4	61.5	72.8	76.2	62.4	57.1	48.9	56.3	61.2	54.5	34.4	41.2	53.6	44.2	54.3	54.9	32.1	54.6	42.3
S6	71.1	32.6	80.3	71.6	62.4	76.5	62.2	31.7	51.3	72.2	64.3	53.2	81.0	69.2	61.5	54.3	65.2	69.6	71.7	62.6
S7	76.4	64.8	52.4	62.3	57.1	51.9	77.1	60.0	42.4	62.3	54.1	62.1	61.0	72.0	53.2	61.8	52.3	60.0	53.3	61.0
S8	56.5	43.2	44.4	32.3	61.7	31.7	70.0	73.2	62.4	73.9	51.7	59.0	54.5	54.3	62.7	63.4	51.8	58.3	61.7	56.3
S9	62.4	49.6	48.4	51.3	49.0	51.3	42.4	61.3	72.4	50.0	50.3	51.5	50.9	51.3	43.3	48.6	52.4	43.9	52.1	55.3
S10	71.8	62.2	70.2	52.9	71.3	72.2	71.3	73.9	77.6	78.2	54.1	70.9	53.3	64.7	71.6	70.4	72.3	59.2	64.3	48.9
S11	62.4	61.6	54.8	61.2	66.5	64.3	53.3	51.7	64.3	54.1	78.8	61.7	52.7	71.0	62.4	58.9	69.6	67.2	71.6	62.3
S12	61.2	65.3	66.3	61.8	34.4	52.7	65.3	67.1	71.0	70.9	61.7	71.2	66.6	66.2	58.3	68.9	66.9	62.1	62.5	66.3
S13	71.2	32.4	61.6	69.0	59.6	71.0	61.4	65.7	66.6	53.3	52.7	71.6	72.2	51.1	65.4	66.3	42.7	49.6	52.4	65.3
S14	44.8	72.2	62.4	64.3	43.2	45.6	32.7	57.8	62.3	64.7	71.0	45.5	51.0	74.8	62.3	54.3	54.4	55.3	56.3	54.4
S15	61.1	60.4	58.9	54.8	34.7	45.9	53.2	32.8	58.3	54.5	61.3	32.7	54.3	45.9	72.1	65.4	31.8	57.7	45.6	43.9
S16	42.8	56.4	62.3	61.0	58.2	54.3	54.9	43.1	57.7	54.8	58.8	52.9	61.3	51.9	52.3	69.7	41.8	45.8	46.2	48.3
S17	32.8	35.5	41.6	42.1	33.3	46.4	42.3	41.5	32.6	30.0	49.2	39.3	32.7	30.0	41.0	41.6	63.2	45.8	31.9	34.7
S18	55.6	71.1	61.9	66.5	70.2	69.6	72.1	58.3	54.6	59.2	67.2	68.4	49.6	54.2	66.4	72.4	82.4	80.0	78.8	62.3
S19	42.2	54.7	71.4	52.9	56.1	71.7	53.3	61.7	54.8	64.3	64.9	62.5	52.4	59.6	63.3	66.5	62.8	59.1	76.5	62.0
S20	64.7	70.4	60.4	55.5	58.3	62.6	61.0	57.9	72.6	48.9	70.5	71.0	65.3	61.4	62.3	63.2	54.7	55.0	52.3	73.2

On the basis of confusion matrix, the speakers can be considered as nearest and farthest speakers (having highest and lowest accuracy).

Table 4. Performance of nearest and farthest speakers

Nearest Speakers	Accuracy (%)	Farthest Speakers	Accuracy (%)
5 and 50	92.2	5 and 60	42.2
55 and 60	78.9	20 and 55	34.8
80 and 85	86	-	-

Recognition performance has been also computed by comparing speech samples of 43 male and 57 female speakers as shown in Figure 6 (a, b).

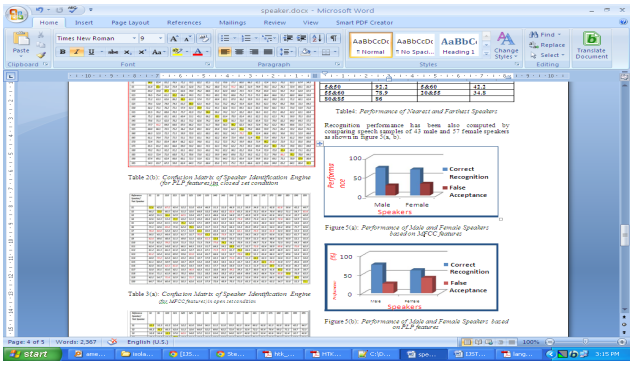

Figure 6(a). Performance of male and female speakers based on MFCC features.

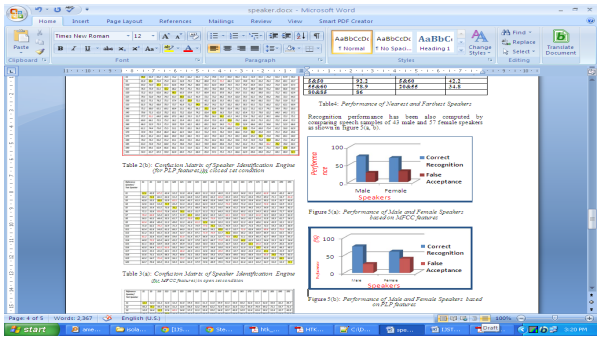

Figure 6(b). Performance of male and female speakers based on PLP features.

5. Conclusions

- Performance of PLP is better than the MFCC, particularly in the case of open set.
- The performance increases as the number of utterances in training samples increases.
- Performance of male speakers is found better than female speakers.
- On comparison with the earlier experimental results conducted by different researchers, it is

concluded that here the sentence level recognition give about

- 91.2% accuracy for MFCC and 92.3% for PLP which is better than the earlier experiments[16].

6. References

1. Rabiner LR, Juang BH. Fundamentals of speech recognition. PTR Prentice-Hall; 1993.
2. Agrawal SS, Bansal S, Pandey D, Tayal H. A Hidden Markov Model (HMM) based speaker identification system using mobile phone database of NATO (National Atlantic Treaty Organization) Words; ICA-Montreal, Canada. 2013.
3. 3. Yuan J, Liberman M. Speaker identification on the Scotus Corpus. Available from: http://www.ling.upenn.edu/~jiahong/publications/c09.pdf
4. 4. Stern RM, Ponce P, Singh R. Feature extraction for robust automatic speech recognition using synchrony of zero crossings. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.407.1507&r ep=rep1&type=pdf
5. HTK Hidden Markov Toolkit Version 1.4 Manual. Cambridge University Engineering Department, Speech Group; Cambridge. 1992.
6. Viterbi algorithm. Available from: http://en.wikipedia.org/wiki/Viterbi_algorithm
7. Reddy V, Das PK. A HMM based text-prompted speaker verification system using HTK. Oriental COCOSDA; 2004 Nov 17-19. p. 313–6.
8. Shrawankar U, Thakare V. Techniques for feature extraction in speech recognition system: A comparative study. Available from: http://arxiv.org/ftp/arxiv/papers/1305/1305.1145.pdf
9. Skowronski MD, Harris JG. Human factor cepstral coefficients. Cancun, Mexico; 2002 Dec.
10. Aida–Zade KR, Ardil C, Rustamov SS. Investigation of combined use of MFCC and LPC features in speech recognition systems. Proceedings of World Academy of Science, Engineering and Technology; 2006 May; 13. ISSN 1307-6884.
11. Hermansky H. Perceptual Linear Predictive (PLP) analysis of speech. J Acous L Soc Am. 1990 Apr; 87(4).
12. Hermansky H, Morgan N, Bayya A, Kohn P. RASTA-PLP speech analysis. Technical Report (TR-91-069). International Computer Science Institute; Berkeley, CA. 1991.
13. Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Published in Processing, IEEE Transactions on Acoustics, Speech and Signal. 1980 Aug; 28(4):357–66.
14. Gauvain JL, Lamel LF, Adda G, Adda- Decker M. (1994): Speaker-independent continues speech dictation. Speech Communication. 1994; 15:21–37.
15. Gupta R. Speech recognition for Hindi. M. Tech. Project Report. Department of Computer Science and Engineering, Mumbai, India: Indian Institute of Technology Bombay; 2006.
16. Kral P, Jezek K, Jedlicka P. Evaluation of automatic speaker recognition approaches. WESPAC X; Beijing, China. 2009.
17. Petr Motlicek. Feature extraction in speech coding and recognition. Report of PhD Research Internship in ASP Group, OGI-OHSU; 2001/2002
18. Useful links. 1. http://www.nist.gov/speech, 2. http:// www.voxforge.ac.in, 3. http:// www.phon.ucl.ac.uk