Automatic Speaker Recognition System in Urdu using MFCC & HMM

Shaik Riyaz, Bathula Lakshmi Bhavani, S.Venkatrama Phani Kumar

ABSTRACT--- Speech is one of the most common ways of communication between users and it is also serves to recognize the individual. In this paper, an automatic speaker recognition system with Mel-Frequency Cepstral Coefficients (MFCC) and Hidden Markov Model (HMM) is proposed to recognize the identity of the users using Urdu utterances. MFCC is a very popular feature extraction approach to extract features with human auditory behavior. In the view of feature size and to increase the efficiency, acoustic precise feature extraction is carried with Vector quantization (VQ). HMM will make the recognition process simple and much more realistic. Performance of the proposed model is evaluated on a dataset with 250 isolated Urdu words uttered by twenty speakers, out of which eight speakers are male and twelve speakers are female. The proposed model outperforms with 96.4% of accuracy when compared with other models.

Keywords: Hidden Markov Model (HMM), Mel-Frequency Cepstral Coefficients (MFCC), Vector Quantization.

I. INTRODUCTION

The advanced mechanism give ease of using the knowledge on a thing by computer a unit of that is an artificial intelligent application of that system is Biometric Identification [15]. Here the sound is nothing but the audio produced from human. The voice recognition is divided into two parts they are speaker’s recognition and speech recognition. The main objective of speaker recognition is extracting characteristic of speaker. The speaker identification is dependent on the human earing scheme involves audio processing. The architecture is developed for voice recognizing in which the data in voice is compared with one another.

This architecture is used for recognizing the arriving speech through matching it with the saved voice in database. Speaker recognition task is divided into three parts namely: “speech recognition”, “speaker recognition” and “language recognition”. In this, the process starts with fetching each individual person’s voice signal that has-been stored using microphone. The input signal at the audio card via system converts analog signals to digital signals to do further processing in an easiest manner. The MFCC is appropriate to earing features of human. MFCC is utilized to improve extraction parameters, a procedure that transform speech signals into various features.

MFCCS depends on familiar deviation of human auditory critical bandwidth with frequency, filters are sequentially arranged at low frequencies (lesser than 1000 hz) and logarithmically at high frequencies (larger than 1000 hz) to capture phonetically significant features of speech.

In the present scenario, Speech is the most important mode of communication amongst humans as well as machines. The advancement in Automatic Speech Recognition(ASR) techniques have been enhanced machines to effectively communicate with machines. In the world, most widely spoken language is English. The ASR research has been worked much efficiently only for English language other than remaining languages. The speech recognition task plays much less progress in Urdu language. The Urdu is also one of the largest spoken languages in world and is also considered as national language of Pakistan. The Phonetics and phonology of Urdu language differs from English language. The dataset for Urdu language is development on ASR and it is also considered as fundamental requirement. The dataset is nothing but the combination of medium scale vocabulary of Urdu words.

The feature set is Mel Frequency Cepstral Coefficient have been extracted and the features are to be proven and it is to be most efficient for ASR. The ASR framework is based on CMU sphinx and HMM to perform the recognition. The results should be compared with previous work on the dataset. The recognition for the frame work is higher than the accuracy observed on same dataset in literature.

The main objective of this work is to develop speech resources which can be used in dialogue prediction and enabling the users to access online health related information[16]. By using these resources, the voice recognition for Urdu language is also developed. The voice recognition system for Urdu is developed spontaneously using HMM based vocabulary automatic speech recognition. The training data set needed for the architecture is divided into two categories. They are as follows: a phonetically valuable utterance dependent upon corpus read out by native speakers to give continuous audio data and impulsive communicative data from stored speech.

II. LITERATURE SURVEY

For a long time the interaction of computer human communication is being done. Still due to a significant number of limitations those systems leads to multiple malfunctions. In the following sections we discuss some of the challenges in speaker recognition system.

In this paper, the author present an approach to implement an automatic speech recognition system of Urdu isolated

Revised Manuscript Received on February 11 , 2019.

Shaik Riyaz Vignan's Foundation for Science, Technology & Research , AP, India.(E-mail: riyaz.shaik62@gmail.com)

Bathula Lakshmi Bhavani Vignan's Foundation for Science, Technology & Research, AP, India.

S.Venkatrama Phani Kumar Vignan's Foundation for Science, Technology & Research, AP, India.

words. The dataset used in this approach is 250 isolated words uttered by 20 speakers among them eight speakers are male and twelve speakers are female .In this Mfccs are widely used for speech recognition task. The classification algorithms used in this are Language model dictionary, acoustic model.

In this paper the advancement of an isolated word recognizer for Indian language kannada. The voice data was taken from 14 speakers using audacity [1]. The features are derived using both MFCC and LPCC methods using HCopy. The recognition was done using HVite tool. It performs passing of token algorithm on the given data and gives result.

In this paper the author discussed about the speaker independent isolated word recognition algorithm for the southern Indian Language named Kannada [1]. The data sets used in this paper are: three female speakers with recording from regional kannada broadcast news of duration of about 10 min. The feature extraction is estimation and re-estimation of HRest function of HTK tool. The algorithm for test samples is Viterbi coding algorithm. It has been utilized for finding the most matching word from dictionary.

In this paper, the author discussed about MFCC-GMM based accent recognition for telugu speech signals [2]. The datasets used in this paper are: Thirteen native local speakers are identified from each region and their speeches are stored using a rich in quality smart phone in locked area. Among these thirteen testing samples, thirteen average factors are extracted from three accents separately. In this the classification method used is GMM testing and it is done for each speech signal.

In this paper, the speech data from sample of population have collected. In this the feature extraction used is maximum livelihood linear transform on features that has been extracted using LDA [3]. The method used is Speaker adaptive training. The classification methods used in this paper are: GMM-HMM and DNN-HMM hybrid system.

In this paper, the recordings was done for two speakers(male). The identification dictionary is a collection of Hindi digits [4]. The feature extraction method used in this paper is: Vector quantization and Linear prediction coding and they are used to gain the feature vector from windowed input speech. The classification methods used are: Swaranjali had used a feed forward model of HMM for recognition.

In this paper, the Feed Forward Back Propagation Neural Networks for automatic speech identification for Arabic letters with four vowels has been investigated [5]. The datasets used in this paper were: four speakers were participated by uttering all Arabic letters with their four vowels three times each one firstly and secondly five times. The Speech signal of each letter is recorded and the features are extracted from recorded corpus by using different methods such as: Linear Predictive Codes, Perceptual Linear Prediction. The performance has been improved when we use PLP method followed by Principal Component Analysis technique.

In this paper, the author discuss about building a speech for corpus for a rare but significent Indian dialect chattisgarhi [6]. The 100 unique chattisgarhi words are chosen from English to Chattisgarhi vocabulary. In this paper the feature extraction method used were: Mel frequency Cepstral coefficient as it is most suitable feature extraction technique for speech processing. The HMM,SVM,ANN classifiers are used for the purpose of recognizing speech .

In this paper, the author implement a technique for identifying spoken words in Bangla. The voice detection algorithm is used to suppress the silence parts and take only speech signal part [7]. The feature extraction methods used in this paper were: MFCC, GMM and LPC model. The dynamic time warping, Posterior probability function and Euclidean distance for measuring distance between feature matrix and reference.

In this paper, the author discuss about continuous speech recognition model and it is presented using kaldi tool kit [8]. In this the data is split into acoustic data and language data. In this paper, both MFCC and PLP features were extracted.

In this proposed work we discuss a type of system where the variation of speaker is calculated using the MFCC and VQ features applied on the HMM.

III. DATASET, FEATURE EXTRACTION AND CLASSIFIER SELECTION

In this section, proposed method is described. This includes the speech dataset that we have used, feature extraction method and Classifier used.

A. Urdu Speech Dataset

The Urdu speech Dataset used in our experiment is developed by [9]. The dataset contains 250 isolated words uttered by twenty speakers, out of which eight speakers are male and twelve speakers are female. The average length of an audio file is 0.5 seconds and the average file size is 16kb. The recording was done in a noise free studio using Sony Linear PCM Recorder at a sample rate of 441 OOHz and then the wav files are converted to mono (with single channel) at a sampling rate of 16000Hz.

B. Feature Extraction

a) Mel Frequency Cepstral Coefficients (MFCCs): Mel Frequency Cepstral Coefficients (MFCCs) are the most common feature extraction method for speech identification task. This feature set is based upon the human perception of hearing. As speech is produced by human vocal tract, therefore vocal tract acts is a filter for speech production. The envelope of speech produced by human vocal tract is a presentation of the short-term periodogram of speech. MFCCs tend to determine the envelope of the speech. Therefore, MFCCs are popularly used as features for speech recognition. The key steps for MFCC extraction are outlined in Fig. 1. In our work, we calculate the MFCCs using Sphinx train, the Sphinx supplementary package.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

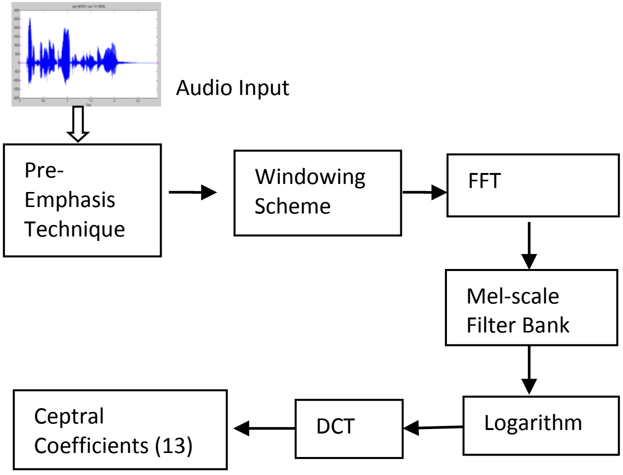

Figure 1: MFCC Extraction Procedure

b) Vector Quantization:

Vector quantization converts the vectors from a big set into small region collection and these collections are called the codebook [15]. For each specified speaker, codebook is produced and while performing analysis the Euclidian distance among the acoustic vector of examine input signal and the specified codebook is measured. The speaker with the lowest Euclidian distance is taken.

Vector quantization is totally dependent on the manner of block coding and it is called as lossy compression of data. Lowest Euclidian distance gives the length from the closest codebook, computed during analysis phase of the speaker identification system. The speaker corresponds to the lowest Euclidian distance has to be picked and examined.

Working Steps of VQ: Consider, The input speech signal is D={d1,d2,d3…di} where, i is the number of raw speech data. The speech sample are converted number of speech feature vector X={x1, x2… xL} by the feature extraction (i.e, MFCC).

And the feature vector X is partitioned into a set of mutually exclusive convex regions S={s1, s2…sN}, where N ≤ L For each region a centroid is assigned C={c1, c2…cN} by the Euclidean distance. The collection of these centroids C={c1, c2…cN} is known as a codebook.

C. Classification Methods

a) *Dynamic Time Wrapping (DTW):* DTW algorithm compares the features of an unknown word with features of one reference template. DTW is used because of simplicity of its hardware implementation, direct approach and speed of training. DTW principle is totally dependent on Dynamic Programming. It is utilized for calculating the resemblance among two time series which might differ in either time or speed [11-14]. This approach is utilized to measure an best alignment among the two times series, if one time series might be "warped" nonlinearly by stretching or shortening it along its time axes. This warping among two time series can now be utilized to measure corresponding regions among the two time series or to calculate the resemblance among the two time series[9]-[11]. Using DTW we arrange two sequences by an "n-by-m matrix", where the (ith, jth) item of matrix has the distances d(qi, cj) amongst the two points qi and cj is developed. Then, the exact length among values of two arrangements is determined using the Euclidean distance calculation as presented in "Eq. (1)".

"d(qi, cj) = d(qi, cj) 2" (1)

Every matrix item (i, j) refers to affiliation among the points qi and cj. Subsequently collected length is calculated by "Eq. (2)" .

"D (i, j) = min[D(i-1, j-1),D(i-1, j),D(i, j -1)]+ d(i, j)" (2)

b) Hidden Markov Model (HMM): HMM is now truly common in the audio speech processing community and is getting approval for conversation systems. HMMs provide a helpful method to model the dynamics of audio utterance. They give a firm arithmetical composition for the question of getting HMM features from audio observations. For accomplishing the success on HMM methods we usually apply the following steps [17]

1. Initialize a medium collection of L audio groups to perform modeling, such as phonemes or words; which are known as audio groups V = {vl ,v2, ...,vL}.
2. For each group, get a considerable size collection (the training set) of tagged speech which are familiar to be available in this group.
3. Considering every training set, workout the computation problem to get a optimal model li for every group vi (i = 1, 2, . . . , L).
4. While identification, we calculate P(O/l) (i = 1, 2, . . . , L) for the unidentified voice observation and identify the audio signal that has formed O as group v j.

c) Gaussian Mixture Model: GMM statistical speaker model is formed following the extraction of features [18]. The condition is when one normal distribution is unsuccessful in that moment a limited mixture models and their distinctive feature estimation methods may be computed by a broad kind of the probability density functions (pdf). To compute the features of GMM for a collection of MFCC features extracted from training audio, to attain a maximum likelihood (ML) estimate we utilize an iterative expectation-maximization (EM) algorithm. A fundamental distribution to be followed is by utilizing predefined shared variety utilized in building a mixture. Gaussian distribution is without a doubt one among the most profitable and most popular allocation in performing important role in statistics and in additional places of utilizations[10][11].

IV. EXPERIMENTAL RESULTS

The Proposed method was done in MATLAB. The dataset contains 250 isolated words uttered by twenty speakers, out of which eight speakers are male and twelve speakers are female. The detailed description of audio signals stored in database is as shown in below table 1.

Table 1: Urdu Database Description

Parameter	Characteristics
Language	Urdu
No. of Speakers	20

Type of Speech	Speech reading
Condition of Recording	A normal room condition
Length of Audio	60-90 seconds
Type of Audio	Mono
Sampling Format	16-bit
Sampling Frequency	44.1 KHz

The following Pre-processing steps are done on the above audio dataset:

1. Segmentation of the words from the utterance and noise removal is accomplished by utilizing standard Adobe Audition Software.
2. The next step is the pre-emphasis of the signal to improve the energy of the higher frequency contents.
3. Since the speech features are mostly based on the energy, hence, to negate the loudness effect, the speech samples are normalized.

HMM can accurately model the statistical deviation in spectral features. Hence, HMM-based techniques have attained considerably improved recognition accuracies than GMM-based techniques; in GMM the observed probabilities are formed as a weighted sum of Gaussian probability densities. The comparison of HMM and GMM techniques on the above dataset is as shown in the Table 2.

Table 2: Performance evaluation of MFCC+GMM & MFCC+VQ+HMM

Data Models	MFCC + GMM	MFCC+VQ + HMM
Recognition accuracy with Guntur Urdu accent	91.9	92.8
Recognition accuracy with Hyderabad Urdu accent	96.2	96.7
Recognition accuracy with Uttar Pradesh accent	98.4	99.5
Overall	95.5	96.4

During this work, various combinations of feature extraction methods and classification techniques have been verified. The first method used in this research is MFCC with Euclidean distance the accuracy is not good because we are considering only distance parameter. The second process is MFCC with DTW with the mel frequency cepstrum objective is to mimic the action of human hearing. Performance is improved. The third is the MFCC with GMM process which drastically improves performance. In the final proposed method we use VQ and HMM which gives the optimal performance. The speaker modeling was done using Vector Quantization (VQ). A VQ codebook was developed by gathering the training feature vectors of every speaker and then placed in the speaker database as shown in figure 2.

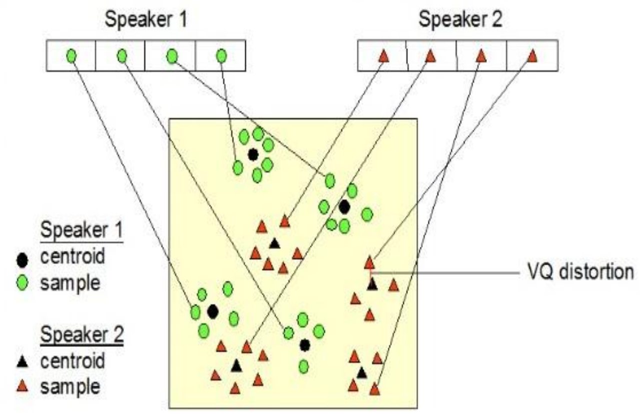

Figure 2: Vector Quantization codebook formation.

Overall performance of all the four methods are as shown in the below Table 3.

Table 3: Performance evaluation of various approaches

Methods	TAR	FRR
MFCC +ED	85.6%	14.4%
MFCC + DTW	89.3%	10.7%
MFCC + GMM	95.5%	4.4%
MFCC+VQ+HMM	96.4%	3.6%

V. CONCLUSION

In this work, authors have 39 ceptral features from a window with midterm features of 1 sec duration and short-term features from 4 m sec window duration. These 39 features include 13 coefficients, 13 delta features and 13 delta-delta features. Vector Quantization is applied to reduce the feature vector size and as well as it was applied to concise the feature vector. Further, Hidden Markov Model is applied to classify the feature vectors using probability estimation. Performance of the proposed model is estimated on Urdu utterances and achieved 96.4% recognition accuracy with minimum no. of features.

REFERENCES

1. A. Thalengala and K. Shama (2016), “Study of sub-word acoustical models for Kannada isolated word recognition system,” Int. J. Speech Technol., vol. 19, no. 4, pp. 817–826.
2. K. Mannepalli, P. N. Sastry, and M. Suman (2016), “MFCC-GMM based accent recognition system for Telugu speech signals,” Int. J. Speech Technol., vol. 19, no. 1, pp. 87–93.
3. P. Mandal, S. Jain, G. Ojha, and A. Shukla (2015), “using deep neural network,” vol. 1, no. Mmi, pp. 1241–1245.
4. T. Pruthi, S. Saksena, and P. K. Das (2000), “Swaranjali: Isolated word recognition for Hindi language using VQ and HMM,” Int. Conf. Multimed. Process. Syst., pp. 13–15.
5. M. Hassine (2015), “Hybrid Techniques for Arabic Letter Recognition,” Int. J. Intell. Inf. Syst., vol. 4, no. 1, p. 27.

6. A. Shaukat, H. Ali, and U. Akram (2016), “Mel Frequency Cepstral Coefficients (MFCCs) Model Dictionary :,” 2016 XXI Symp. Signal Process. Images Artif. Vis., pp. 135–139.
7. M. Hossain, M. N. Bhuiyan, and S. Engineer (2013), “Automatic Speech Recognition Technique for Bangla Words,” vol. 50, pp. 51–60.
8. P. Upadhyaya, O. Farooq, M. R. Abidi, and Y. V. Varshney (2018), “Continuous Hindi speech recognition model based on Kaldi ASR toolkit,” Proc. 2017 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2017, vol. 2018–Janua, no. 0, pp. 786–789.
9. R. Bharti and P. Bansal (2015), “Real Time Speaker Recognition System using MFCC and Vector Quantization Technique,” Int. J. Comput. Appl., vol. 117, no. 1, pp. 25–31.
10. D. A. Reynolds, A Gaussian mixture modeling approach to text independent speaker identification, Ph.D. thesis, Georgia Institute of Technology, Atlanta, Ga, USA, September 1992.
11. Bharti W. Gawali, Santosh Gaikwad, “Marathi Isolated Word Recognition System using MFCC and DTW Features”, ACEE, Vol. 01, No. 01, Mar 2011.
12. K. X. Huang, A. Acero, and H. Wuenon (2005), Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Pearson.
13. Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, “Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques”, Journal of Computing, Volume 2, Issue 3, March 2010, ISSN 2151-9617.
14. R. Bharti and P. Bansal (2015), “Real Time Speaker Recognition System using MFCC and Vector Quantization Technique,” Int. J. Comput. Appl., vol. 117, no. 1, pp. 25–31.
15. Charisma A, Hidayat MR, Zainal YB (2017). Speaker Recogn L tion Using Mel-Frequency Cepstrum Coefficients and Sum Square Error. 3rd Int Conf Wirel Telemat 2017;(27-28 July):160-163.
16. Raza A, Hussain S, Sarfraz H. An ASR System for Spontaneous Urdu Speech. Proc Orient …. 2010:1-6. http://www.cslhr.nu. edu.pk/gccs /spring2010/papers/Agha.pdf.
17. Srinivasan A. Speech Recognition Using Hidden Markov Model 2 Analysis using Wave Surfer 3 Vector quantization. 2011;5(79):3943-3948.
18. Aroon A, Dhonde SB. Speaker Recognition System using Gaussian Mixture Model. Int J Comput Appl. 2015;130(14):975-8887.