ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Accent Recognition for Malayalam Speech Signals

Aswathi Sanal, Mary Nirmala George

M.Tech Student, Department of ECE, Viswajyothi College of Engineering & Technology,Vazhakulam, Kerala, India

Asst. Professor, Department of ECE, Viswajyothi College of Engineering & Technology,Vazhakulam, Kerala, India

ABSTRACT: Speech is the most natural and efficient way of communication between humans. Speech recognition systems find their applications in our daily lives and have huge benefits for those who are suffering from some kind of disabilities.Malayalam is a South Indian language spoken predominantly in the state of Kerala. Thiruvanathapuram,Thrissur and Ernakulam are the three different accent corpora used for the accent recognition task. This paper presents an approach to extract features from speech signal of spoken words using the Mel-Scale Frequency Cepstral Coefficients. It is a nonparametric frequency domain approach which is based on human auditory perception system. Then this feature vector set obtained are classified in the classification phase using Gaussian mixture model (GMM) classifiers. During classification stage, the input feature vector data is trained using information relating to known patterns and then they are tested using the test data set. All this implementation is build in Matlab.

KEYWORDS: Accent recognition, Mel Frequency Cepstral Coefficients, Gaussian mixture model (GMM)

I.INTRODUCTION

Speech signals are naturally occurring signals and hence, are random signals. These information carrying signals are functions of an independent variable called time. Speech recognition is the process of automatically recognizing certain word which is spoken by a particular speaker based on some information included in voice sample. It conveys information about words, expression, style of speech, accent, emotion, speaker identity, gender, age, the state of health of the speaker etc [4]. There has been a lot of advancement in speech recognition technology, but still it has huge scope.

The speech signal hides the knowledge regarding accent with in the words based on the area of the speaker for which he belongs. Accent recognition helps speech recognition systems. The performance of automatic speech recognition systems can be increased, if the speaker's accent or dialects detected before the recognition of speech by adapting the suitable Automatic Speech Recognition (ASR) acoustic and/or language models. Speech recognition mainly focuses on training the system to recognize an individual's unique voice characteristics.

In this work, speech features have been explored to recognize three major accents of Malayalam. The accents under consideration are Thriruvanathapuram, Thrissur and Ernakulam. The accent of a given language is a pattern of pronunciation of a language used by the community of native speakers. Accent specific information in speech is contained at the entire segmental, sub segmental and supra segmental level. Only a very few works have been reported in Malayalam language. Malayalam is the most common language of Kerala with several variations. Malayalam has at least fourteen different accents with speaker population varying from one accent to another. In this work, explored spectral features of speech for recognizing the spoken dialects. Mel Frequency Cepstral Coefficients (MFCC) is used as the spectral features as it is less complex in implementation and more effective and robust under various conditions. MFCC is explained in the coming sections.

II.RELATED WORK

Several experiments have been reported that intonation patterns and pitch characteristics are utilized as features to minimize the error rate in the area of accent based speech recognition. Gaussian likelihoods dominate the total computational since the evolution in signal processing. Depending upon the amount of training data, it is

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

important to select the proper Gaussian mixtures. Because the size of databases available for Indian languages is small, higher range Gaussian mixtures cannot be applied to them.

In the work on the English sentence accent detection demonstrated that the prominence model better matches the auditory system and the efficiency increases to the inter-human agreement about 86 %[1]. A system for English speech recognition is designed by taking speeches from two child groups of native English speakers and Japanese-speakers with different levels of proficiency in English. This resulted in an error rate of approximately 13 % on native data and 20 % on Japanese non-native data.

In the study of impact of native language on accent in regional Indian Languages five regional Indian languages Bengali, Manipuri, Kashmiri, Urdu and Hindi were considered. Spectral features and Gaussian mixture models (GMM) have been used. Besides, the results of carrying out Language and Accent Identification tasks, it has been importantly found out that the extent of impact of the three native languages on Hindi was different. Kashmiri affected the accent much more than Bengali or Manipuri. Different approaches were presented which uses a number of streams of information in the acoustic signal. The system recognizes the local dialects or accent.

The accent of the speaker identified by inspecting the phonetic, frame-based acoustic and phonotic features and high level Prosodic features by the comparison of generative and discriminative techniques. It was reported that the features like mel-warped cepstrum coefficients and varied parameterizations of LPC are very useful in the area of speaker identification and speech recognition. An auditory-based feature extraction algorithm was proposed in which the features used were cochlear filter Cepstral Coefficients (CFCCs) which are defined based on the newly developed transform called auditory transform plus a set of modules that emulates the signal processing functions in the cochlea.Under non-noisy conditions, the MFCC and CFCC features performance is almost same[1]. When an input speech signal with a signal-to noise ratio of 6 dB is considered, the efficiency of the MFCC features fell down to 41.2 %, but the cochlear filter Cepstral Coefficients (CFCCs) still have shown an accuracy of 88.3 %.

A study was conducted on accents by using MFCC algorithm and RASTA-PLP algorithm to extract short-time spectrum features of each speech segment based on structured information. The experimental results indicate that MFCC and sub-segment splicing feature structured method of RASTA-PLP can be used in Chinese accent detection study with efficiencies of82.1 and 80.8 % relatively. For the speaker recognition an algorithm was proposed in which the mel-frequency Cepstral Coefficients (MFCCs) features were used for speaker accent recognition. They reported that k-nearest neighbors yield the highest average test accuracy compared to SVM. In the work done for automatic classification system based on accent dependent parallel phoneme recognition for foreign accents in Australian English speech[1]. In this work continuous speech was used and the classifier to discriminate between indigenous speakers of Australian English and two migrated speaker teams is the statistical model HMM. The best accent recognition efficiencies reported were 76.6 and 85.3 % for the three accents and accent pair classification tasks.

III.SYSTEM ARCHITECTURE

The proposed algorithm follows three steps. In the first step, preprocessing of the data is done where the voice signal is converted from analog to digital, i.e. sampling is done. The second step proceeds with calculating the parameters that will be used for recognition purpose. This phase involves feature extraction and is referred to as training module. In the last step, the process of feature matching which involves the testing procedure is carried out. This module is known as testing module. The following figure shows the steps carried out for the proposed system. In the following sections of this chapter, the detail about each of these steps is discussed including the implementation procedure.

A. Pre-processing

The foremost step is the generation of the speech data corpus which is followed by the preprocessing phase, however, the speech signal captured decides the recognition accuracy base on the microphone's quality. Analog signal has a number of variable frequency components. In this phase, the analog signals captured using a microphone is digitized according to the Nyquist theorem, which says that the signal must be sampled more than twice the highest frequency. For normal microphones, sampling rate of 16 KHz or more is preferred. The analog signal cannot be directly applied in the computer as it understands only digital data. It is necessary to sample the analog signal into the discrete-time signal,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

which the computer can use to process. The sampled data is then further fed into the training module for processing where feature extraction is done.

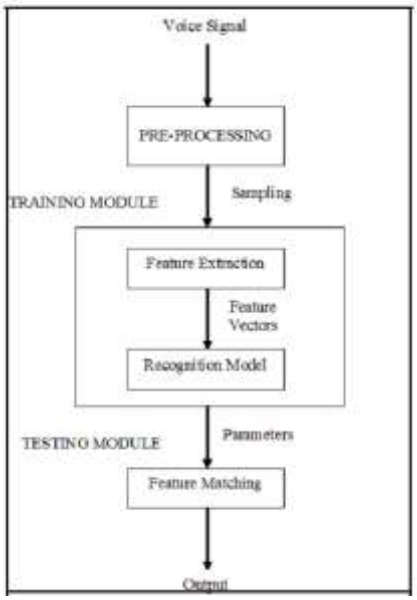

Fig1.Design of the proposed system

B. Feature Extraction

In this phase, the sampled data speech signal is processed further for feature extraction. This generates feature vectors that are used by the recognition model for training the system. A number of feature extraction techniques are available; however Mel Frequency Cepstrum Coefficients have been used here. In automatic speech recognition, using 20 MFCC coefficients is very common but 13 coefficients are considered to be enough for encoding speech [21]. The procedure for feature extraction is as follows:

1. Pre-emphasis: Noise has a greater effect on the higher modulating frequencies than the lower ones. Hence, higher frequencies are artificially boosted to increase the signal-to-noise ratio. Pre-emphasis process performs spectral flattening using a first order finite impulse response (FIR) filter [4]. This process will increase the energy of signal at higher frequency [8].

2. Framing: In this step, the speech samples are segmented into small frame size within the length of 20 to 40msec The number of samples used for each frame is 256. To calculate the number of frames, the total numbers of samples in the input voice file are divided by 128.

3. Windowing: Discontinuities at the beginning and end of the frame are likely to introduce undesirable effects in the frequency response. Hence, each row is multiplied by window function. A window alters the signal, tapering it to nearly zero at the beginning and the end. Hamming window is used as it introduces least amount of distortion [21]. This implementation uses Hamming window of length 256. The following function represents the Hamming window function:

$$h[n] = \begin{Bmatrix} 0.54 - .46\cos\left(\frac{2\pi n}{N}\right), 0 \le n \le N-1 \\ 0 \qquad\qquad\qquad , otherwise \end{Bmatrix} \quad (1)$$

4. DFT Computation: DFT is the next step, which converts each frame into frequency domain from time domain. It is done to speed up the processing [8]. It is represented by the following equation:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{j2\pi kn}{N}}, 0 \le k \le N-1 \quad (2)$$

Here, x(n) represents input frame of 256 samples and X(k) represents its equivalent DFT. We use 256-point FFT algorithm to convert each frame of 256 samples into its equivalent DFT [21].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

5. Mel Frequency Filter Bank: The Mel frequency filter bank is applied to the Fourier transformed frame obtained from each sample. Since Mel scale helps to space windows equally, the design and implementation becomes easier. Mel-frequency analysis of speech is based on human perception experiments. It has been proved that human ears are more sensitive and have higher resolution to low frequency compared to high frequency. Hence, the filter bank is designed to emphasize the low frequency over the high frequency. The following relation defines the relation between frequency of speech and Mel scale; it converts linear scale frequency into Mel scale frequency.

$$\text{Mel}(f) = 1127 * \log_{10}(1 + f/700) \quad (3)$$

In this work, the number of filters used is filter bank are 26. A total of 42 MFCC parameters include twelve original, twelve delta (First order derivative), twelve delta-delta (Second order derivative), three log energy. In this work, 13 coefficients are extracted and used for analysis.

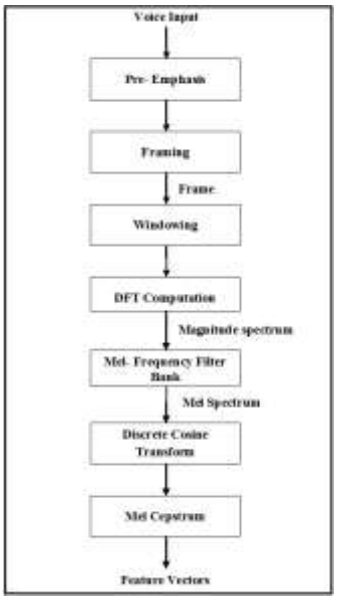

Fig 2.MFCC feature extraction

6. Discrete Cosine Transform: This is the final step in feature extraction that leads to the generation of feature vectors. The conversion of log Mel spectrum into time domain using Discrete Cosine Transform (DCT) is carried out in this step. The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called feature vectors. Therefore, each input utterance is transformed into a sequence of feature vector.

C.Gaussian Mixture Models (GMM)

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system.

GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum *A Posteriori* (MAP) estimation from a well-trained prior model. The Gaussian mixture model for speech representation assumes that a M component mixture model with windowing function weights $P(\omega m)$ and the mixture components in the input voice sample contains Gaussian components.

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. GMMs are often used in biometric systems, most notably in speaker recognition systems, due to their capability of representing a large class of sample distributions. Given training vectors consisting of MFCC feature vectors and a GMM configuration wish to estimate the parameters of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

the GMM, λ, which in some sense best matches the distribution of the training feature vectors. There are several techniques available for estimating the parameters of a GMM[4]. By far the most popular and well-established method is maximum likelihood (ML) estimation. The aim of ML estimation is to find the model parameters which maximize the likelihood of the GMM given the training data.

For a sequence of T training vectors X = {x_1, . . . ,x_T}, the GMM likelihood, assuming independence between the input data vectors, can be written as,

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda) \quad (4)$$

Unfortunately, this expression is a non-linear function of the parameters λ and direct maximization is not possible and hence the iterative Expectation-Minimization algorithm(EM) was used for training as well as matching purposes. In the "Expectation" step, calculate the probability that each data point belongs to each cluster.In the "Maximization" step, re-calculate the cluster means and covariance based on the probabilities calculated in the expectation step.

IV. IMPLEMENTATION

The proposed work is implemented in MATLAB® 2010b and the system is developed on Windows 7 32-bit operating system. The system is designed to recognize accent of three different district of Kerala as mentioned above .

A. Speech Corpus

The speech database has been recorded in a noise free room environment using the recording tool *Audacity 2.1.0* which is it latest version and the files are stored in *.wav* format. The speech data is recorded using a unidirectional microphone by keeping an approximate distance of 5-10cm between the mouth and the microphone and no noise reduction mechanism has been applied to the speech files. Sampling rate used for recording is 16 KHz on mono- channel. A total of 65 speakers from each district are recorded the data.The age of speakers is between 20 to 50 years.The following table gives the details about that.

B. Pre-processing

The speech files are sampled at a sampling frequency of 16 KHz where each file has a length of 1 to 3 sec. However, this much length is too large to be analyzed by the system. Hence, framing is done. Usually it takes in a frame of the speech signal every 1-3sec and performs certain spectral analysis [8]. The steps followed in this module are discussed below:

1. The speech file is read from the directory using the command *'wavread'*.
2. Next step is to determine the length of the speech signal using the *'plot'* command which plots the graph of the speech signal.
3. Now the start and end of the wave is determined and silence is removed from the sound, keeping only the uttered sentence to increase the accuracy.
4. Using the command *'wavwrite'*, a new speech file is created.
5. The above steps are repeated for all the speech files before proceeding further.

C. Training module

A speech signal cannot be directly fed into the system for analysis. It has to be represented in a more efficient and compact form. The original speech signal is converted into a series of feature vectors using feature extraction technique MFCC. MFCC performs a series of steps to generate the feature vectors that are used for training the system. The following procedure is implemented:

1. Pre-emphasis follows the process of filtering. The FIR filter is used to flatten the speech signal. The number of filters used by this system is 26. The command *'filter'* is used here.
2. Framing is the next step in this procedure where the speech signal is divided into a number of frames.
3. Windowing is done to minimize the discontinuities in the frames. Each frame is multiplied by a windowing function.
4. After the windowing, Fast Fourier Transformation (FFT) is calculated for each frame to extract frequency components of a signal in the time-domain.
5. The above step is used to calculate Mel Filter Bank which generates the mel spectrum coefficients that are further converted into time domain as mel spectrum coefficients are real numbers using the DCT procedure.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

6. The final step is the generation of Mel Frequency Cepstrum Coefficients. In this work, 13 coefficients are obtained for each frame. The command used for feature extraction is *'melcepst'*. A matrix is obtained with the number of frames as rows and mel cepstrum coefficients as columns.

D. Recognition Model

The system is trained using Gaussian Mixture Model . GMM initialization is done using a prototype model. The GMM prototype gives the model topology specification, number of states used, transition parameters and the output distribution parameters. The number of states varies according to the input data.. Parameters are re-estimated repeatedly until the reestimation for the training data converges.

E. Testing Module

This is the last phase in speech recognition which performs feature matching, i.e., to find the most probable sequence that matches the outputs obtained from the training module. Once the system model is generated, it is used for the recognition of an unknown utterance, this is known as testing.

V.RESULTS AND DISCUSSION

A basic accent recognition system which recognizes accent was thus developed. The system was trained in an environment with minimum ambient noise. One of the shortcomings of the system is that the performance degrades in the presence of noise. The system gives the most accurate results when implemented in the environment where it was trained. The system performance was analyzed by increasing the number of training iterations for the EM algorithm, including setting a threshold on the likelihood difference between steps.That, however, proved to have little benefit in practice; neither the execution time nor the amount of misclassification rate showed any mentionable improvements over just fixing the number of iterations. The reason why the execution time did not show any significant improvements is because most of the execution time is spent during feature extraction, and not in training. A total of two utterances of speakers from the collected database were considered in the accent recognition system.

The MFCC features were extracted from these speech signals as feature vectors, which were further used for the training the system in the proposed work. Even for the testing speech samples these feature vectors were extracted and used to recognize the accent. The procedure and methodology of extracting MFCC is described in detail in the previous section. After extracting the MFCC features, GMM are used to model and identify the accent of the test speech signal.

VI. CONCLUSIONS

1. Accent based MFCC features of Malayalam speech for both training and testing database were extracted successfully.
2. The recognition accuracy using MFCC–GMM method is 89 %.

REFERENCES

1 KasiprasadMannepalli,PanyamNarahariSastry,MalojiSuman, "MFCC-GMM based accent recognition system for Telugu speech", International Journals of Speech Technol, 2016.

2 Zhao YunXue, Zhang Long,ZhengShiJie and Zhang Wei "Chinese Accent Detection Research Based on Features Structured",International Journal of Hybrid Information Technology Vol.8, No. 5 (2015), pp. 303-316

3 NiluSingh,R.AKhan,RajShree"MFCC and Prosodic Feature Extraction Techniques:A Comparative Study"International Journal of Computer Applications (0975 – 8887)Volume 54– No.1, September 2012

4 Tanja Schultz, Qin Jin, KornelLaskowski, Alicia Tribble, Alex Waibel"Speaker, Accent, And Language Identification Using Multilingual Phone Strings" International Journal of Computer Applications (0975 – 8887)Volume 54– No.1, September 2009

5 Munish Bhatia, Navpreet Singh, Amitpal Singh, "Speaker Accent Recognition by MFCC Using KnearestNeighbour Algorithm: A Different Approach"International Journal of Advanced Research in Computer and Communication EngineeringVol. 4, Issue 1, January 2015

6 SantoshGaikwad,BhartiGawali,K.V.Kale,"Accent Recognition For Indian English Using Acoustic Feature Approach", International Journals of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Computer Application, 2013, vol. 63.

7 CiniKurian and KannanBalakrishnan, "Automated Transcription System For Malayalam Language", International Journals of Computer Application, 2011, vol. 19.

8 HemakumarGandPunithaP."Speaker Accent and Isolated Kannada Word Recognition"American Journal of Computer Science and Information Technology

9 Georgina Brown"Automatic Accent Recognition Systems and the Effects of Data on Performance"Department of Language and Linguistic Science University of York, UK.

10 Carlos Teixeira, Isabel Trancoso and Ant´onio Serralheiro "Accent Identification"Instituto de Engenharia de Sistemas e Computadores.

11 SantoshGaikwad,BhartiGawali,K.V.Kale,"Accent Recognition For Indian English Using Acoustic Feature Approach", International Journals of Computer Application, 2013, vol. 63.

12 Matthew Seal, Matthew Murray, ZiyadKhaleq"Accent Recognition with Neural Network" International Journal of Computer Applications Volume 45– No.1, September 2010.

13 R. K. Aggarwal and M. Dave"Using Gaussian Mixtures for Hindi Speech Recognition System" International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 4, December, 2011.

14 Tao Chen, Chao Huang, Eric Chang and JingchunWang"Automatic Accent Identification Using Gaussian Mixture Models" Microsoft Research China.

15 MohamadHasanBahari,RahimSaeidiy Hugo Van hamme,David Van Leeuweny "Accent Recognition Using I-vector, Gaussian Mean SupervectorAndGaussian Posterior Probability Supervector For Spontaneous Telephone Speech"Center for processing speech and images, KU Leuven, Belgium.

16 Douglas Reynolds"Gaussian Mixture Models" MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA.

17 M.A.Anusuya, S.K.Katti"Classification Techniques used in Speech Recognition Applications: A Review"Int. J. Comp. Tech. Appl., Vol 2 (4), 910-954.

18 Ramesh Sridharan"Gaussian mixture models and the EM algorithm".

19 FadiBiadsy "Automatic Dialect and Accent Recognition and its Application to Speech Recognition"Submitted in partial fulllment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences.[20 Matthew Nicholas Stuttle, Hughes Hall"A Gaussian Mixture ModelSpectral Representation for Speech Recognition"Cambridge University Engineering Department

21 HarpreetKaur,RekhaBhatia"speech recognition system for Punjabi language" International Journal of Advanced Research in Computer ApplicationsandSoftware Engineering Volume 5 August 2015.

BIOGRAPHY

Aswathi Sanal was born in Kerala. She took her B.Tech in Electronics and Communication Engineering from SreeNarayanaGurukulam College of Engineering, Kadayiruppu(2015). She is doing her M.Tech in VLSI and Embedded System at Viswajyothi College of Engineering & Technology, Muvattupuzha. Her fields of interest include VLSI, Digital System Design and Embedded System.

Mary Nirmala George received her Bachelor of Technology (B Tech) from M G university in 2007. From year 2008-2010,she did her Master of Technology(M.Tech) in Embedded Systems from DOEACC centre Calicut. She is working as Assistant Professor in Viswajyothi College of Engineering and Technology for the last 5 years. Her field of interest include VLSI & Signal processing applications for Embedded system