COMPARATIVE STUDY OF MFCC AND LPC FOR MARATHI ISOLATED WORD RECOGNITION SYSTEM

Leena R Mehta [1], S.P.Mahajan [2], Amol S Dabhade [3]

Lecturer, Dept. of ECE, Cusrow Wadia Institute of Technology, Pune, Maharashtra, India [1]

Associate Professor, Dept. of ECE, College of Engineering , Pune, Maharashtra, India [2]

PG Student [SP], Dept. of ECE, College of Engieering, Pune, Maharashtra, India [3]

ABSTRACT: This Paper presents Marathi database and isolated word recognition system using Mel-frequency cepstrum coefficients (MFCC) and vector quantization (VQ) technique. It also compares the performances of MFCC and LPC features under VQ environment. Marathi speech database is recorded in noisy environment aiming language learning tool as an application. The database consists of simple Marathi words starting with both vowels and consonants. Each word has been repeated 10 times by one male and one female speaker. This paper presents comparative plots of MFCC and LPC features.

Keywords: Marathi database, Feature extraction, LPC, MFCC, VQ, Recognition

I.INTRODUCTION

The Speech is the most prominent and natural form of communication between humans. There are various spoken Languages thought the world. Marathi is an Indo-Aryan Language, spoken in western and central India. There are 90 million of fluent speakers all over world. However; there is lot of scope to develop systems using Indian languages which are of different variations. Some work is done in this direction in isolated Bengali words, Hindi and Telugu .The amount of work in Indian regional languages has not yet reached to a critical level to be used it as real communication tool, as already done in other languages in developed countries. Thus, this work was taken to focus on Marathi language [1]. It is important to see that whether Speech Recognition System for Marathi can be carried out similar pathways of research as carried out in English. Current computer interfaces like keyboards also assume a certain level of literacy from the user. It also expects the user to have certain level of proficiency in English. In our country where the literacy level is as low as 50% in some states, if information technology has to reach the grass root level; these constraints have to be eliminated. In this paper we are presenting work consists of the creation of Marathi speech database and its speech recognition system for isolated words.

The paper is divided into six sections. Section 1,gives Introduction. Section 2 deals with details of creating Marathi speech database. Section 3 focuses on Feature extraction using MFCC and LPC, Section 4 covers vector quantization and section 5 deals with results and conclusion followed by section 6 with the References.

II.MARATHI SPEECH DATABASE

The Collection of utterances in proper manner is called the database. We have selected basic ‘Anklipi’ developed by renowned publication. It is basic book for beginners. For accuracy in the speech recognition, we need a collection of utterances, which are required for training and testing[1]. The generation of a corpus of Marathi Vowels, words and sentences as well as the collection of speech data are described below. The vocabulary size of the database consists of

- Marathi Vowels: 120 samples
- Marathi consonants: 360 samples

1.Speech data collection:

Speaker Selection

Database was recorded with one male and one female speaker of 25-35yrs. age. Mother tongue of both the speakers was Marathi.

Data Collection

Each speaker was asked to speak the 48 words with 5 utterances of every word. Total 480 utterances of the words were recorded.

2. Recording procedure [5]:

The isolated words were recorded using built in microphone of laptop using the PRAAT speech Software. The data will be recorded in closed rooms where background noise was present. The recording of the Speech data in such noisy environment will be useful in future for developing a robust automatic speech recognition system.

III.FEATURE EXTRACTION

The general methodology of speech classification involves extracting discriminatory features from the speech data and feeding them to a pattern classifier. Different approaches and various kinds of speech features were proposed with varying success rates. The features can be extracted either directly from the time domain signal or from a transformation domain depending upon the choice of the signal analysis approach. Some of the speech features that have been successfully used for speech classification include Mel-frequency cepstral coefficients (MFCC), linear predictive coding (LPC). Few techniques generate a pattern from the features and use it for classification by the degree of correlation. Few other techniques use the numerical values of the features coupled to statistical classification method.

A.LINEAR PREDICTION COEFFICIENT

LPC (Linear Predictive coding) analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue. In LPC system, each sample of the signal is expressed as a linear combination of the previous samples. This equation is called a linear predictor and hence it is called as linear predictive coding [3].The coefficients of the difference equation (the prediction coefficients) characterize the formants.Speech signal recorded using PRAAT and sampled at 16 KHz, is processed for extracting the features in MATLAB. The basic steps of LPC processor include the following [4]:
1. *Preemphasis*: The digitized speech signal, $s(n)$, is put through a low order digital system, to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing. The output of the preemphasizer network is related to the input to the network, $s(n)$, by difference equation:

$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1)$$

2. *Frame Blocking*: The output of preemphasis step, $\tilde{s}(n)$, is blocked into frames of N samples, with adjacent frames being separated by M samples. If x (n) l is the l *th* frame of speech, and there are L frames within entire speech signal, then

$$x_l(n) = \tilde{s}(Ml + n)$$

3. *Windowing*: After frame blocking, the next step is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. If we define the window as $w(n)$, $0 \le n \le N-1$, then the result of windowing is the signal:

$$\tilde{x}_l(n) = x_l(n)w(n) \quad \text{where } 0 \le n \le N-1$$

Typical window is the Hamming window, which has the form

$$w(n) = 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right] \qquad 0 \le n \le N{-}1$$

Autocorrelation Analysis: The next step is to auto correlate each frame of windowed signal in order to give

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m) \qquad m = 0,1,\ldots,p$$

where the highest autocorrelation value, p, is the order of the LPC analysis
4. LPC Analysis: The next processing step is the LPC analysis, which converts each frame of $p+1$ autocorrelations into LPC parameter set by using Durbin's method. This can formally be given as the following algorithm:

$$E^{(0)} = r(0)$$

$$k_i = \frac{r(i) - \sum_{j=1}^{i-1} \alpha_j^{i-1} r(|\, i-j \,|)}{E^{i-1}} \qquad 1 \le i \le p$$

$$\alpha_i^{(i)} = k_i$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad 1 \le j \le i\text{-}1$$

$$E^{(i)} = (1-k_i^2)E^{i-1}$$

By solving above equations recursively for $i = 1,2,\ldots,p$, the LPC coefficient, a_m, is given as
a= $\alpha_m^{(p)}$

$$c_m = \sum_{k=m-p}^{m-1} \left(\frac{k}{m}\right) \cdot c_k \cdot a_{m-k} \qquad m > p$$

B.MEL FREQUENCY CEPSTRUM COEFFICIENTS

Mel Frequency Cepstral Coefficients (MFCC) is one of the most commonly used feature extraction method in speech recognition. The technique is called FFT based which means that feature vectors are extracted from the frequency spectra of the windowed speech frames.The Mel frequency filter bank is a series of triangular bandpass filters. The filter bank is based on a non-linear frequency scale called the mel-scale. A 1000 Hz tone is defined as having a pitch of 1000 mel. Below 1000 Hz, the Mel scale is approximately linear to the linear frequency scale. Above the 1000 Hz reference point, the relationship between Mel scale and the linear frequency scale is non-linear and approximately logarithmic [4]. The following equation describes the mathematical relationship between the Mel scale and the linear frequency scale

$$f_{Mel} = 1127.01 \ln\left(\frac{f}{700} + 1\right)$$

The Mel frequency filter bank consist of triangular bandpass filters in such a way that lower boundary of one filter is situated at the center frequency of the previous filter and the upper boundary situated in the center frequency of the next filter. A fixed frequency resolution in the Mel scale is computed, corresponding to a logarithmic scaling of the repetition frequency, using Δf_{Mel} = ($f_{H\ mel}$ − $f_{L\ mel}$)/ (M + 1) where $f_{H\ mel}$ is the highest frequency of the filter bank on the Mel scale, computed from using equation given above, fL $_{mel}$ is the lowest frequency in Mel scale, having a corresponding and M is the number of filter bank. The values considered for the parameters in the present study are: f_{max}=16KHz and f_{min} =0 Hz. The center frequencies on the Mel scale are given by:

$$f_{cm\,(Mel)} = f_{L(Mel)} + \frac{m(f_{H(Mel)} + f_{L(Mel)})}{M+1}, 1 \le m \le \mathrm{M}$$

The center frequencies in Hertz, is given by

$$f_{cm} = 700\left(e^{\frac{f_{cm\,(Mel)}}{1127.01}} - 1\right)$$

Above Equation is inserted into equation of f_{mel} to give the Mel filter bank. Finally, the MFCCs are obtained by computing the discrete cosine transform of using

$$c(l) = \sum_{m=1}^{M} X'(m)\cos(l\frac{\pi}{M}(m - \frac{1}{2}))$$

For l = 1, 2, 3, ….., M where *c(l)* is the l^{th} MFCC.
The time derivative is approximated by a linear regression coefficient over a finite window, which is defined as

$$\Delta\, c_t(l) = \left[\sum_{K=2}^{2} k\ c_{t-k}(m)\right].G,\ 1 \le l \le M$$

where is the l^{th} cepstral coefficient at time t and G is a constant used to make the variances of the derivative terms equal to those with the original cepstral coefficients.

III.FEATURE MATCHING METHOD (VECTOR QUANTIZATION)

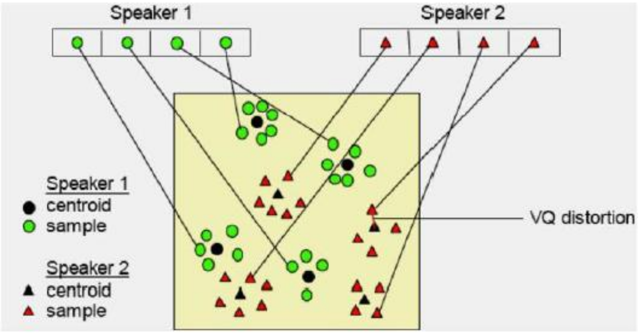

Fig.1: Vector quantization codebook formation

The fig.1 above explains vector quantization method for speaker identification based on Euclidean distance. The problem of speech recognition belongs to a much broader topic in scientific and engineering so called pattern recognition [3]. The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called patterns and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The classes here refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can be also referred to as feature matching. Furthermore, if there exists, some set of patterns that the individual classes of which are already known, then one has a problem in supervised pattern recognition. This is exactly our case since during the training session, we label each input speech with the ID of the word. The remaining patterns are then used to test the classification algorithm; these patterns are collectively referred to as the test set [4]. If the correct classes of the individual patterns in the test set are also known, then one can evaluate the performance of the algorithm. The state-of-the-art in feature matching techniques used in speech recognition includes Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). In this paper, the VQ approach is used, due to ease of implementation and high accuracy. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a codeword. The collection of all code words is called a codebook. Figure 1 shows a conceptual diagram to illustrate this recognition process. In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from the speaker 2.

A.LBG ALGORITHM

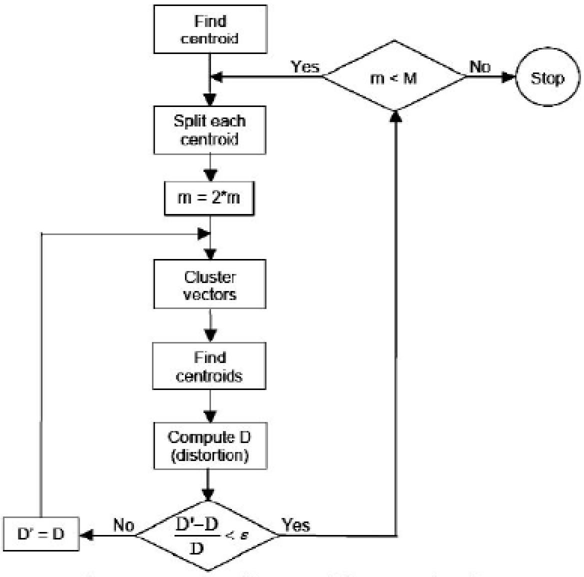

Fig.2: Flow diagram of the LBG algorithm

Fig. 2 shows flow diagram of LBG algorithm. It is explained stepwise below. After the enrolment session, the acoustic vectors extracted from input speech of a speaker provide a set of training vectors. As described above, the next important step is to build a word-specific VQ codebook for this speaker using those training vectors. There is a well-known algorithm, namely LBG algorithm [Linde, Buzo and Gray, 1980], for clustering a set of L training vectors into a set of M codebook vectors. The algorithm is formally implemented by the following recursive procedure [6]:

1. Design a 1-vector codebook; this is the centroid of the entire set of training vectors (hence, no iteration is required here).
2. Double the size of the codebook by splitting each current codebook yn according to the rule where n varies from 1 to the current size of the codebook, and ε is a splitting parameter (we choose ε=0.01).
3. Nearest-Neighbour Search: for each training vector, find the codeword in the current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest codeword).
4. Centroids Update: update the codeword in each cell using the centroids of the training vectors assigned to that cell.
5. Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold
6. Iteration 2: repeat steps 2, 3 and 4 until a codebook size of M is designed. Intuitively, the LBG algorithm designs an M-vector codebook in stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the code words to initialize the search for a 2-vector codebook, and continues the splitting process until the desired M-vector codebook is obtained. Figure 2 shows the detailed steps of the LBG algorithm. "Cluster vectors" is the nearest-neighbour search procedure which assigns each training vector to a cluster associated with the closest codeword. "Find centroids" is the centroid update procedure. "Compute D (distortion)" sums the distances of all training vectors in the nearest-neighbour search so as to determine whether the procedure has converged.

IV. RESULTS AND DISCUSSIONS

One male and one female speaker recorded the words in Marathi 'Anklipi'. Some of the MFCC and LPC Features extracted of the Marathi words are shown in the figures below. The training set for the vector quantizer was obtained by recording utterances of a set of Marathi words.The vector quantizer for each of the words was trained with 5 utterances of the word for the 2 speakers. The results of comparison of both the features for few words are as shown in Table 1 and Table 2 below.

WORD	SPEAKER 1	SPEAKER 2
AAI	75%	73%
ANANAS	78%	74%
BAL	80%	78%
KSHATRIYA	81%	80%
AVERAGE	78.5%	76.25%

Table 1: recognition accuracy for LPC feature

WORD	SPEAKER 1	SPEAKER 2
AAI	98%	99%
ANANAS	100%	100%
BAL	100%	100%
KSHATRIYA	100%	100%
AVERAGE	99.5%	99.75%

Table 2: recognition accuracy for MFCC feature

Table 1 and 2 shows that recognition accuracy is more with MFCC. So MFCC can be thought of as better feature for Marathi language tutor application in speech recognition.

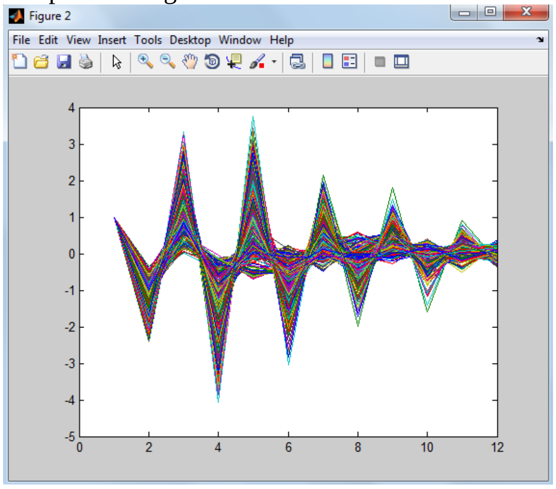

Fig.3 Plot of LPC features of the word 'aai'

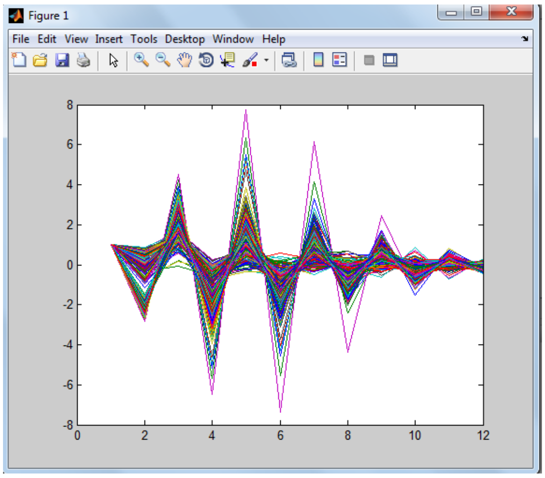

Fig.4 Plot of LPC features of the word ‘ananas’

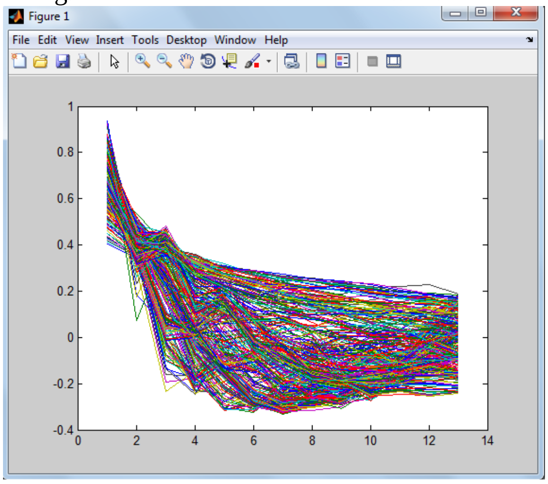

Fig. 5 Plot of MFCC features of the word ‘aai’

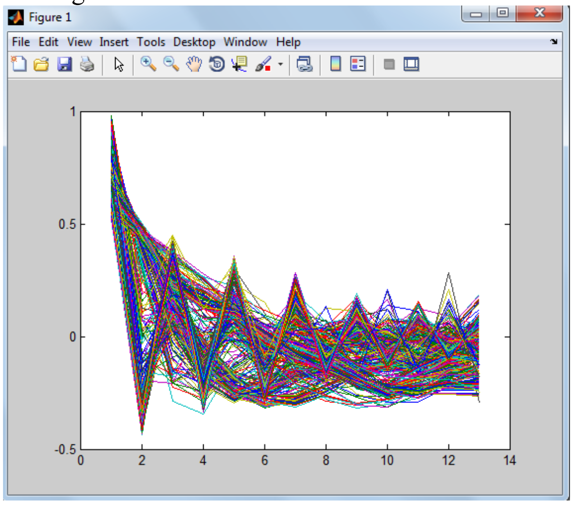

Fig.6 Plot of MFCC features of the word ‘ananas’

From fig.3, fig. 4, fig.5, fig.6, it is proved that MFCC is better choice for this application of speech recognition.

V. CONCLUSION

This paper has discussed an effective method for feature extraction of isolated Marathi words. It presents a Marathi database and isolated word recognition system based on Mel-frequency cepstral coefficient (MFCC) and vector quantization as recognition method. It also compared the recognition systems of LPC and MFCC features.

In recent years there has been a steady movement towards the development of speech technologies to replace or enhance text input called as Mobile Search Applications. Recently both Yahoo! and Microsoft have launched voicebased mobile search applications. Future work can include improving the recognition accuracy of the individual words by combining the multiple classifiers.

ACKNOWLEDGMENT

It is my pleasure to get this opportunity to thank my respected Guide **Dr. S. P. Mahajan** who imparted valuable knowledge specifically related to Speech Processing.

REFERENCES

[1] Bharti W. Gawali, Santosh Gaikwad, Pravin Yannawar, Suresh C.Mehrotra "*Marathi Isolated Word Recognition System using MFCC and DTW Features*" Proc. of Int. Conf. on Advances in Computer Science, Vol. 1, pp. 21-24, 2010.

[2] Rabiner L. and Juang B.H., "*Fundamentals of Speech Recognition*". *New York:Prentice Hall Publishers,1993.*

[3] Tarun Pruthi, Sameer Saksena, Pradip K Das, 1993, "*Swaranjali: Isolated Word Recognition for Hindi Language using VQ and HMM*" Journal of Computing and Business Research

[4] Kayte Charansing Nathoosing " *Isolated Word Recognition forMarathi Language using VQ and HMM*." science Research Reporter 2, Vol. 2, pp. 161-165, April 2012.

[5] *'http: //www.fon.hum.uva.nl/praat'* cited on 5/12/2013

[6]. Y. Linde, A. Buzo & R. Gray, "*An algorithm for vector quantizer design*", IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.