ISSN 2320 - 2602
Volume 4 No.7, July 2015
International Journal of Advances in Computer Science and Technology
Available Online at http://www.warse.org/ijacst/static/pdf/file/ijacst01472015.pdf

Speaker Identification in Odiya using Mel Frequency Cepstral Coefficients and Vector Quantisation

Pamela Chaudhury[1] **H.K Tripathy**[2]

Asst. Prof, Department of CSE, Silicon Institute of Technology, Bhubaneswar, India [1]
Assoc Prof, Department of CSE, KIIT, Bhubaneswar, India [2]

Abstract:. Automatic Speaker Identification technology has recently been implemented in several of commercial areas successfully. Speaker identification comes under Speaker recognition and is gaining significance for voice based biometrics. It is used in appliances that understand voice commands, provides security to confidential information,etc.
In this paper we have built reference model for each speaker using the acoustic features .Testing has been done by comparing the features of the test sample with the reference model.We have used MFCC technique for feature extraction .For creating the reference model Vector Quantization(VQ) with LGB (Linde, Buzo, and Gray)was used. For testing VQ distortion was used.The system achieved 85% accuracy.

Keywords: speaker recognition, Odiya digits,codebook, MFCC, Vector quantisation

1. INTRODUCTION

Speaker recognition is the process of automatically identifying the speaker on the basis of individual information included in speech signals.It gained its significance due to a growing popularity in voice biometrics. [1]Speaker recognition is a method of automatically identifying a speaker from a recorded or a live speech signal by analyzing the speech signal parameters It can be divided into Speaker Identification and Speaker Verification. Speaker verification is a method of accepting or rejecting the identity claim of an individual speaker. Speaker identification determines which registered speaker provides a given utterance from amongst a set of known speakers. Speaker verification accepts or rejects the identity claim of a speaker[2]Figureure 1. shows the block diagram of speaker identification. It has two phases: - Enrollment session or Training phase and Operation session or testing phase.

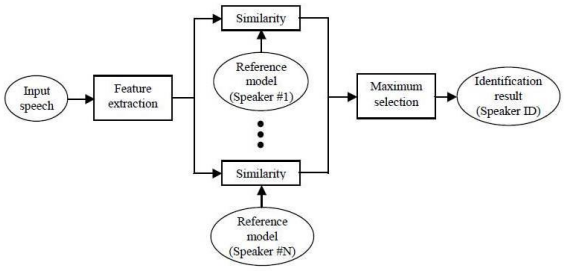

Figure. 1. Speaker identification

In the enrollment phase, each registered speaker provides samples of their speech so that the system can develop a reference model for a speaker. During the testing phase, a speech sample is matched with stored reference model and speaker recognition is done. Speaker recognition process is based on features of speech of each speaker and uses different algorithms to authenticate a speaker. Formant frequencies, amplitude, pitch, phonetic emphasis etc. are the various speech parameters that are often used in speaker authentication systems.[3]

The sample speech is converted to system readable format. This enables the system to processes the data. The data processing consists of feature extraction and feature matching which form the core of a speaker recognition system. In this paper we have taken words in Odiya and then considered the problem of feature extraction and feature matching.
Various features of the spectrum of the speech that are used include the real cepstral coefficients (RCC), Adaptive Component Weighting (ACW), LPCC, LPC ,PLP,and MFCC.In this paper we have used MFCC to extract features and it is achieved by transforming speech signal into frequency domain. Hence it is less prone to noise. Also this method is used because MFCC mimics the human ear behavior. Features that are extracted have to be matched with the help of feature matching techniques. There are several feature vector models that are used.The Dynamic Time Warping (DTW),neural networks (NN), Hidden Markov Model (HMM), and Vector quantization (VQ)etc. are some of the methods.These methods use complex mathematical functions effectively.[4]

2. SPEAKER IDENTIFICATION SYSTEM

First speech signals are recoded using a microphone in a noise free environment for different speakers. The sampling rate used is 8000 Hz. It is stored in the system in the .wav format. It is converted into a format which the system can process for authentication purpose.This system consists of two main modules feature extraction and feature matching.

For extracting feature a detailed step by step procedure is followed. The features of a speaker will eventually form the codebook .In feature matching a test speech sample is matched with the codebook .

In speaker identification systems, the utterance is compared against multiple speech samples in order to determine the best match . We have used Speaker Identification system and compared each Oriya speech sample with all the speech samples of all speakers already recorded and stored in the system.[5]

3. FEATURE EXTRACTION OF SPEECH SAMPLE

The purpose of feature extraction is to convert the continuous speech waveform into a set of features for analysis. Feature

extraction for our experiment consists of the following steps as shown in Figureure 2.

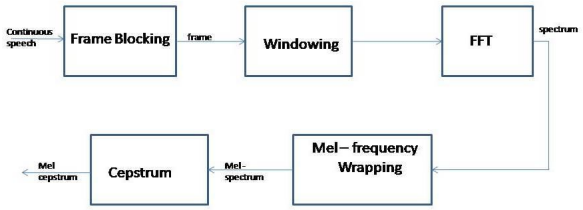

Figure. 2. Feature extraction

3.1 FRAME BLOCKING

Continuous speech is divided into frames of N samples.Frames adjacent to eachother are separated by M samples, where $M < N$. Overlapping frames are used to decrease loss of information. It helps in maintaining correlation between the frames adjacent to each other. The 1st frame consists of N samples and the 2nd frame begins M samples after the first and overlaps it by $N - M$ samples. The values of N=256 and M=100 is used in our experiment as it provides resolution in time and frequency.

3.2 WINDOWING

Windowing , each individual frame is done to minimize the signal discontinuities at the beginning and the end of each frame. This minimizes the spectral distortion and provides better results. We use Hamming Window to taper the signal. The Hamming Window is given as:

$$w(n) = 0.54 - 0.46cos\frac{2\pi n}{N-1} \qquad (1)$$

3.3 FAST FOURIER TRANSFORM

In this method, each frame of N samples is converted from time domain to frequency domain. This important mathematical tool eliminates the redundant calculations and helps to analyze the spectral properties of a signal. Figure 3 displays the power spectrum and logarithmic power spectrum of the signal.

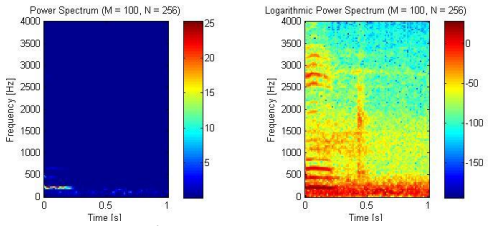

Figure. 3. Power Spectrum

3.4 MEL FREQUENCY WRAPPING

Our perception ability of frequency contents of sounds do not be follow a linear scale. Therefore each tone with frequency, f, measured in Hertz, a subjective pitch is measured on a scale known as the mel scale. The mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz[6].

$$Mel(f) = 2595 * \log(1 + \frac{f}{700}) \qquad (2)$$

To simulate the subjective spectrum a filter bank, is spaced uniformly on the mel-scale. Such a filter bank has a triangular band pass frequency response. Figure.4 shows a Mel spaced filter bank.

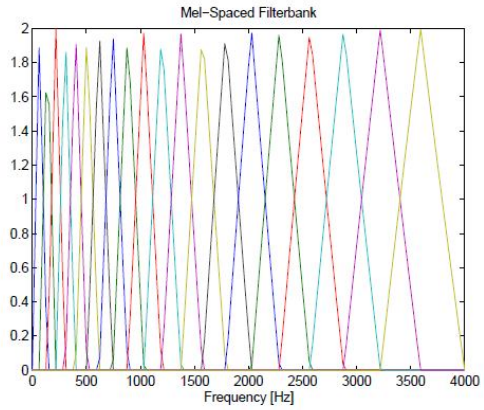

Figure. 4. MEL FILTER BANK

3.5 CEPSTRUM

The final step of feature extraction is to calculate cepstrum which is done by converting the log mel spectrum back in time domain.Inverse Fourier Transform is computed of the logarithm of the power spectrum of a signal. The formula is given by:

$$C(n) = ifft(\log|fft(s(n)|) \qquad (3)$$

The above formula converts the log Mel spectrum into time domain called Mel frequency cepstrum coefficients using Discrete Cosine Transform. The cepstral representation of the speech spectrum gives a very good representation of the local spectral properties of the signal for the each speech utterance . Each sample speech is transformed in to a sequence of acoustic vector.Figure 5. indicates the power spectrum modified through Mel cepstrum filter.

4. CODEBOOK FORMATION

The features vectors obtained using MFCC are of very large vector space. In order to make computations simpler we have used Vector Quantisation and LBG (Linde Bruzo and Gray)algorithm to produce codebook.

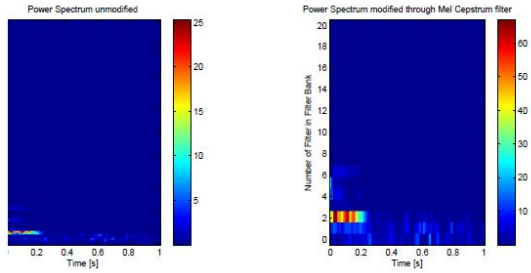

Figure. 5. Power spectrum modified

Each region is a cluster whose center is represented by its centroid and is also referred as codeword.All the centroids or code words together form a codebook.The codebook is a collection of feature vectors smaller in dimension than the feature vector achieved after MFCC.The resultant codebook is simpler and accurate. This is achieved by vector quantization.[7] Vector quantization is the process of mapping vectors from a large vector space to a smaller and finite number of regions or clusters. Each cluster and can be represented by its centroid.[8].

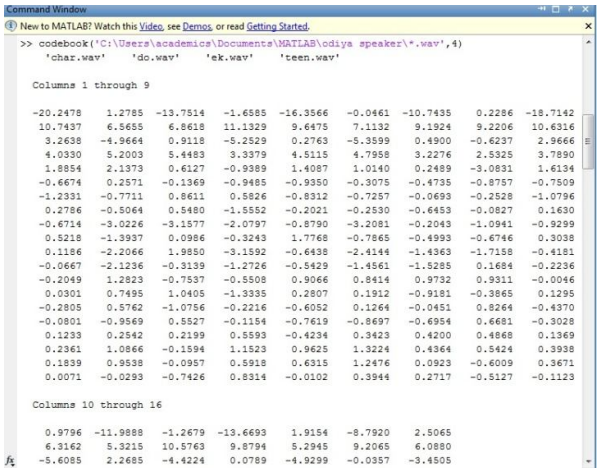

Figure. 6. (Snapshot of part of a codebook generated for a speaker)

5. VECTOR QUANTISATION

Vector Quantization is the of process of mapping vectors from a very large vector space to a small number of regions in that space. VQ is used because it makes
computations simpler by reducing the dimensionality of the feature vector. A vector quantizer maps p-dimensional vectors in the vector space *Sp* into a finite set of vectors.
Y = *yi* : *i* = 1, 2, ...,*N* Each vector *ti* is a codeword .The set of all these codewords is codebook. Associated with each codeword, *ti*, is a nearest neighbor region called Voronoi region, and it is defined by:

$$v_i = x \in S^P : \|x - y_i\| \leq \|x - y_j\| \qquad (4)$$

for all j ≠ i.
The codeword is closest in Euclidean distance from the input vector. The Euclidean distance is given by:

$$d(x, y_i) = \sqrt[2]{\sum_{j=1}^{k}(x_j - y_i)^2} \qquad (5)$$

Where *xj* is the *jth* component of the input vector, and *yij* is the *jth* component of the codeword *yi*.

6. OPTIMISATION USING LBG ALGORITHIM

LBG algorithm [Linde, Buzo and Gray, 1980], is used for clustering a set of training vectors into a set of M codebook vectors. Given a set of N training feature vectors, t1, t2, tN of each speaker, a partitioning of the feature vector space, P1, P2,..., PM is determined.For each speaker the whole feature space P, is represented as P = P1 U P2 U...U PM.Each partition, Pi, forms a nonoverlapping
region.Every vector inside Pi is represented by the corresponding centroid vector, ci, of Pi. Every iteration moves the centroid vectors such that the accumulated distortion between the feature vectors is smaller.[9] The algorithm is formally implemented by the following procedure:-

1. Design a single vector codebook.It is the centroid of all training vectors.
2. Codebook size is increased to twice by splitting each current codebook according
   to the rule:

$$y_n^+ = (1 + \varepsilon) \quad (6)$$
$$y_n^- = (1-\varepsilon) \quad (7)$$

Where n ranges from 1 to the present size of the codebook, and ϵ is a splitting parameter (ε = 0.01)

3. For each training vector, a centroid is found in the current codebook that is closest to the training vector, and the vector is assigned to the corresponding cluster.
4. Centroid Update: Recompute the centroid after adding the feature vector.
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.

7. FEATURE MATCHING

The speaker recognition is a kind of pattern recognition problem. The pattern recognition problem classifies data into a number of categories or classes. The data consists of acoustic vectors that are extracted from an input speech. The classes in the experiment refer to different speakers.Speaker recognition is a problem in the domain of supervised pattern recognition.During the training session, the labelling of each input speech is done with an ID of the speaker . VQ codebook is generated for each speaker by clustering feature vectors of the speakers. The Figureure 7 shows codebook for speaker1 and speaker 2.

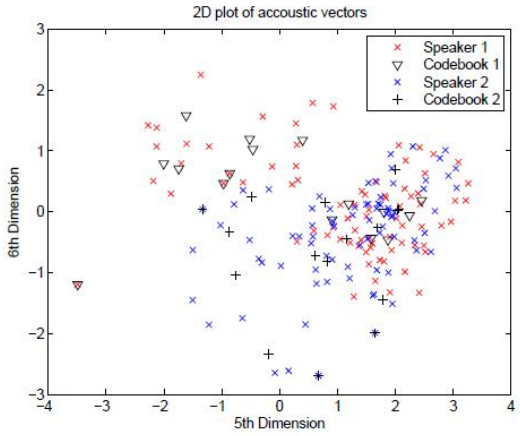

Figure. 7. Codebook of speaker1 and speaker 2

Feature matching is done using VQ- Distortion.Any speaker who has to be identified gives his/her speech sample. From the sample speech features are extracted.It is vector quantised to get vector Q. The distance d of vector Q to the closest codeword in the codebook called VQ-distortion is calculated. VQ distortion is the Euclidian distance between the two vectors and is given by the formula:

$$d(x, y_i) = \sqrt[2]{\sum_{j=1}^{k}(x_j - y_{ij})^2} \qquad (8)$$

In the identification phase,the unknown speaker is identified to be speaker S whose distortion d is minimum from the codebook of speaker S.

8. ANALYSIS AND CONCLUSIONS

In our experiment we have trained the system with five different speakers for some specific words in Odiya language. Identification of each registered speaker who provided their speech samples in the training phase was done. In the training phase we built a speaker specific reference model .Then it was

compared with test samples . The system developed is moderately tolerant to background noise . We have used Odiya words for training the system ek (one), dui (two), teen (three), char (four).The system uses the MFCC feature extraction technique , for speaker modelling uses VQ and LGB algorithim and VQ distortion method for pattern matching .About85% success rate is achieved during the experiment.

The experiment is carried out in almost noise free environment. The system correctly identified the speaker trained for a particular word by comparing the input speech for that word against the stored reference model for that word.

9. FUTURE WORK

Identification of speaker developed has been found to be very successful and achieved a high accuracy and also a high learning rate. The standard deviation of about 15% from actual results has been calculated in our experiments. In near future, we not only try to compare other feature extraction methods but also we propose to build an application where we use speaker identification as a biometrics for people speaking Odiya language.

References

[1]. Reynolds, D:An overview of automatic speaker recognition, Proceedings of the International Conference on Acoustics, Speech and Signal Processing 2002.

[2] Rabiner,L.,Schafer,R.:Introduction to Digital Speech Processing

[3] Furui,S.: Recent advances in speaker recognition. AVBPA97, pp. 237–251, 1997

[4] Sedd ik, H.; Rahmouni, A.; Sayadi, M.:Text independent speaker recognition using the Mel frequency cepstral coefficients and a neural network classifier, 1st International Symposium on Control, Communications and Signal Processing, Proceedings of IEEE 2004 pp 631-634.

[5] Jurafsky,D., Martin,J.:Speech and Language Processing

[6] Campbell,J.P. :Speaker recognition: A tutorial, Proceedings of the IEEE, vol. 85, pp. 1437–1462, September 1997.

[7] Zhonghua,F.,Rongchun,Z.:An overview of modeling technology of speaker recognition,IEEE Proceedings of the International Conference on Neural Networks and Signal Processing Volume 2, Page(s):887 891, Dec. 2003.

[8] Nakai, M., Shimodaira, H., Kimura, M.,:A fast VQ codebook design algorithm for a large number of data, IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 1, Page(s):109 112, March 1992.

[9] Linde,Y.,Buzo,A.,and Gray.,R.: An algorithm for vector quantizer design, IEEE Trans. Commun., vol. COM-28, no. 1, pp. 84-95, 1980.