Speaker Recognition Using MFCC Front End Analysis and VQ Modeling Technique for Hindi Words using MATLAB

Nitisha
M.Tech. (pursuing)
Hindu College of engineering
Sonipat, Haryana
India

Ashu Bansal
Assistant professor
Hindu College of engineering
Sonipat, Haryana
India

ABSTRACT
Automatic Speaker Recognition technology has recently been implemented in large number of commercial areas successfully. Speaker recognition is being used in voice based biometrics; voice controlled appliances, security control for confidential information, remote access to computers and many more interesting areas. This paper introduces text dependent systems that have been trained for a particular user. All speaker recognition systems contain two main modules: feature extraction and feature matching. Here, we have used MFCC technique for feature extraction and Vector Quantization model for feature vectors modeling.
There are mainly two important tasks to be performed in speaker recognition process: one is training phase and other is testing phase. During the training phase, the input speech features are extracted and the corresponding feature vectors are modeled using modeling techniques. These feature vectors are stored as reference templates. They are then compared with the entered speech signals during the testing phase and thus how helps in identification of voice. [13]

General Terms
This Document contain following keywords:
Automatic speaker recognition
MFCC: Mel-Frequency Cepstrum Coefficients
VQ: Vector Quantization
Feature Extraction

Feature Recognition Keywords
Automatic speaker recognition, MFCC: Mel-Frequency Cepstrum Coefficients, VQ: Vector Quantization, Feature Extraction, Feature matching.

1. INTRODUCTION
Automatic *Speaker recognition* is the process of automatically recognizing "who is speaking" on the basis of individual information included in speech waves. [1]
Two major applications of speaker recognition technologies and methodologies are speaker identification and speaker verification. [13]

Speaker identification is used to identify the speaker. It involves selection between a set of known voices. Speaker identification is broadly classified into closed set identity and open set identity. Closed set identification process assumes that all speakers are known to the systems whereas in open system identification possibility may be there that speaker is not among the speakers known to the system. In this process the user does not claim an identity.

Speaker verification on the other hand authenticates a particular speaker and hence user claims an identity. System performs the task of accepting or rejecting identity claim. Speaker may belong to the closed set or open set.

Identification vs. Verification

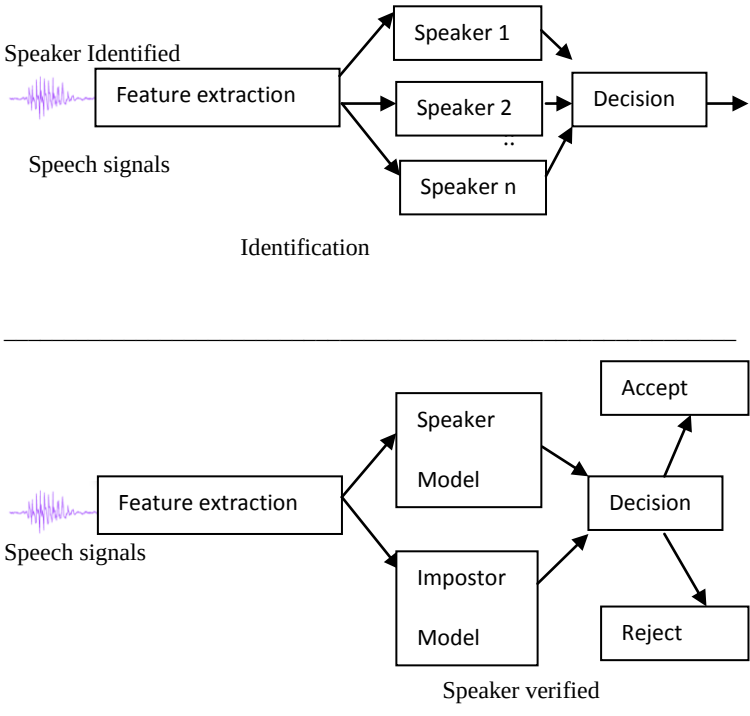

Fig. [i] Identification vs. Verification

Automatic speaker recognition system may be classified into text dependent and text independent recognition systems.
In text dependent recognition systems, systems know the text spoken by person. Fixed or prompted phrases are used by the speakers. Same text must be used for enrollment and verification. They generally are more accurate and improve system performance as system has knowledge of spoken text.
In text independent recognition, system does not know the text to be spoken by person. Text entered during enrollment and test is different. It is more flexible system but also more difficult problem. [13]
Recognizing natural speech is a challenging task as we are unable to see the speech signals. Amplitude, pitch, phonetic emphasis etc. are the various speech parameters that are used in speech recognition systems. These parameters form the basis of these recognition systems as they vary from speaker to speaker.
Most audio features are mainly based on the statistical characteristics are related to the temporal and spectral features of each frame. Statistics like mean, standard deviation or covariance etc. are used to describe characteristics of audio. [15]
The speaker recognition system basically involves speech analysis or speech feature extraction, modeling of these extracted feature vectors using different modeling technique during the training session Testing or pattern matching of the stored template is then performed with the speaker samples during test phase.
Thus, there are two main problems that have been considered for this research paper. The first is features extraction issue, whereas the second is speech features matching in speaker dependent systems.
Historically, various features of the spectrum of the speech include the real cepstral coefficients (RCC), LPC, LPCC and MFCC. Other speech features are PLP, Adaptive Component Weighting (ACW) and other wavelet based features did not have that much popularity. [18]
Feature extraction is also referred to as speech analysis or front end analysis. MFCC is the most frequently used technique especially in speech recognition and speaker verification application. MFCC is more preferred in feature extraction as it generates the training vectors by transforming speech signal into frequency domain. Therefore it is less prone to noise. Also this method is used because human hearing is based on frequency analysis. [5] MFCC mimics the human's ear behavior. [17]
Features that are extracted are needed to be matched with the help of feature matching techniques. Most commonly feature matching techniques uses feature vector models includes the Hidden Markov Model (HMM), Dynamic Time Warping (DTW)neural networks (NN) and Vector quantization (VQ).These methods involve complex mathematical functions, but essentially, they take the information known to the system to figure out the information hidden from it. Thus these are the most commonly used pattern recognition techniques in speech and speaker recognition field.

2. SPEAKER RECOGNITION SYSTEM

Speaker recognition system is different from that of speech recognition system. These two terms are often confused. Speech recognition system deals with the process of converting speech into text whereas Speaker recognition system is the process of identifying the speaker based on their acoustic characteristics.

Speaker recognition system may either have to simply detect the speaker of entered speech signals and also it may authenticate a particular speaker to a system as in speaker verification systems.
Acoustic signals or speech signals are entered into the system using microphone, sound card etc. These are then stored as analog signals (in the form of .wav files) which are further readable by various electronic systems. The analog signals are then digitized to the digital signals. Due to the high dynamic range of speech signals they can suffer from additive noise so a simulink block sets are used to extract and analyze acquired data.
In isolated word recognition systems, two major activities are performed: one is training the system in which the system is put to learn the reference patterns. Reference patterns are used to represent different speech sounds (by word) that constitute the vocabulary of the application. Each reference is learnt from a set of spoken words and acts as a template using statistical properties of the pattern of speech. Another major activity is the testing the system in which speech signal or new recorded speech signal is identified using the stored reference patterns or the template obtained from the training. [2]
In this paper we are considering two main problems: first is features extraction issue, in which we will use MFCC technique for extraction of various speech features and the second one is feature vector modeling. Speech feature vectors extracted using MFCC technique needs to be modeled into compressed form. Modeling of speech feature vectors is actually the process of compressing of the speech feature vectors. The compressed feature vectors are easier to store and compare and hence are very useful in pattern recognition. Vector Quantization data compression scheme is used in feature vector modeling in this paper. [14]
In this paper speaker identification system using MFCC and VQ with LBG algorithm technique has been implemented.

3. SPEECH FEATURE EXTRACTION

Acoustic characteristics of speech signals are obtained with the help of feature extraction. This stage is often referred as speech processing front end. The main goal of Feature Extraction is to simplify recognition by summarizing the vast amount of speech data without losing the acoustic properties that defines the speech [12].

Speech recognition procedure used in our experiment involves following main steps:

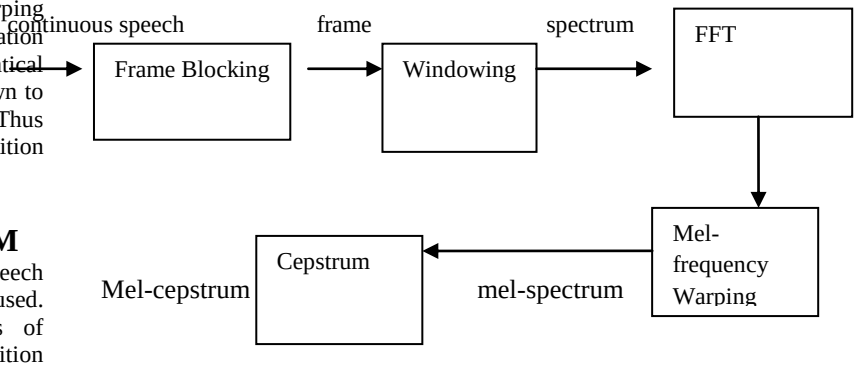

***Fig*. (Ii) FEATURE EXTRACTION STEPS**

Speech recognition system involves preprocessing activity as its first step. The normalization of various extracted speech signals features is done in preprocessing step. It is then followed by selection of an accurate data length of sampled signal so that the mathematical calculation becomes easy. Specification of the limits is given for the signal we wanted to analyze. Only specific sample from the set of given samples are analyzed. It is very necessary to trim the speech signal in its tightest limits in order to avoid errors in the modeling of subsequent utterance of the same words. Therefore a threshold is applied at both the ends of the word being analyzed. This signal thus obtained is almost free from noise and other background disruptions. Once the signal becomes noise free the next step is to extract the speech signals.

Before speech pattern comparison, the spoken word features will be extracted. For this, construct a function using data acquisition toolbox in MATLAB. Noisy or no spoken signal needs to be removed before the real analysis. Not whole of the speech signals are analyzed but particular number of samples from entire set of samples of voice signals is taken. The features may be either of the type time domain or the frequency domain. Some of the features are better studied in time domain whereas other characteristics of digital signals can be better studied in frequency domain.

In this paper we will study the frequency domain features of speech feature extraction. [10]

3.1 Frame Blocking

Framing involves division of the digitized speech signal to be broken into smaller frames with some time lengths In this step the continuous speech signal is blocked into i frames of N samples, with adjacent frames being separated by N ($N<M$). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by N - M samples and so on. This process continues until all the speech is accounted for within one or more frames. [3]

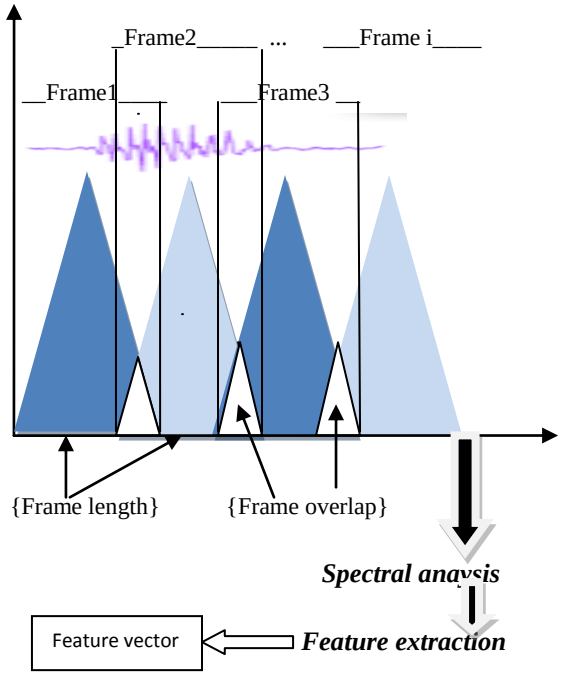

***Fig.(iii)*frame blocking in feature extraction**

3.2 Windowing

A Window is a way of modifying the input signal so that there are no sudden discontinuities in it. Even if we are not using a window function, when we sample a signal, we impose a rectangular function on it. Windowing is done at each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is for the minimization of the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as:

$w(n), 0 \le n \le N-1,$ where N is the number of samples in each frame,

Then, the result of windowing is the signal can be given as:

$$y_l(n) = x_l(n)w(n), \quad 0 \le n \le N-1$$

3.3 Fast Fourier Transform

FFT is a very important mathematical tool. The FFT algorithm eliminates the redundant calculations and enables to analyze the spectral properties of a signal. It offers rapid frequency domain analysis and processing of digital signals, and investigation of digital systems.[3]

Discrete Fourier transform (DFT) is itself a Sequence rather than a function of continuous variable and it corresponds to equally spaced frequency samples of discrete time Fourier transform of a signal. FFT also allows time domain signal processing operations to be performed equivalently in frequency domain. In both domains FFT has considerable reduction in computation time.

In1965, Cooley &Turkey described a method of computing in which they showed DFTs requiring a number of operations proportional to $N\log_2 N$, which is referred to as FFT. The FFT is claimed to be a $N\log_2(N)$ algorithm whereas DFT requires no. of operations equals to N^2. Therefore the speed up is of order of $(\log_{2N})/N$ where time series sample length is. Thus

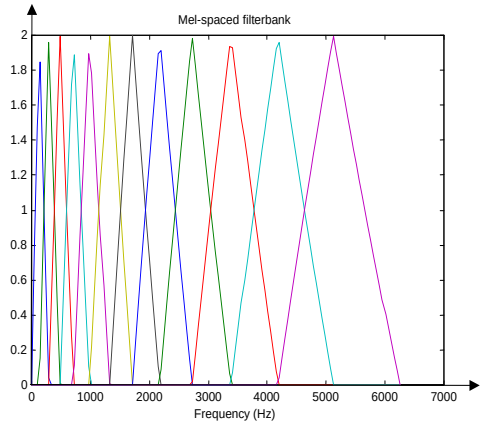

Fig. (iv) mel scale

FFT algorithms are mainly helpful in computing the DFT &

also helpful in correlation analysis. [8]

Fourier series gives the representation of the periodic sequence. Discrete Fourier transforms corresponding to a finite length sequence can be said to be used for Transforming

discrete time sequence *x(n)* of finite length into discrete frequency sequence *X[k]* of finite length. This means that by using DFT, the discrete time sequence *x(n)* is transformed into corresponding discrete frequency sequence *X[k]*.

The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of *N*, samples $\{x_n\}$, as follow:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \qquad k = 0,1,2,\ldots,N-1$$

When X_k's are generally complex numbers and we only consider their absolute frequencies here. The resulting sequence $\{X_k\}$ is interpreted as follow: positive frequencies $0 \le f < F_s/2$ correspond to values $0 \le n \le N/2-1$, while negative frequencies $-F_s/2 < f < 0$ correspond to $N/2+1 \le n \le N-1$. Here, F_s denote the sampling frequency.

3.4 Mel Frequency Warping

According to psychological studies, human perception of the frequency contents of sounds for speech signals are found not be following a linear scale. Thus for each tone with an actual frequency, *f*, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The *mel-frequency* scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the mel-scale. That filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The number of mel spectrum coefficients, *K*. Note that this filter bank is applied in the frequency domain, thus it simply amounts to applying the triangle-shape windows as in the Figure (iv) to the spectrum.

A useful way of thinking about this mel-wrapping filter bank is to view each filter as a histogram bin (where bins have overlap) in the frequency domain.

3.5 Cepstrum

In this final step, we convert the log mel spectrum back to time. The result is called the mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). Therefore if we denote those mel power spectrum coefficients that are the result of the last step are $\tilde{S}_0, k = 0,2,\ldots,K-1$, we can calculate the MFCC's, $\tilde{c}_n$, as

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K} \right], \qquad n = 0,1,\ldots,K-1$$

4. MODELING FEATURE VECTORS USING VECTOR QUANTIZATION TECHNIQUE

Here, VQ approach will be 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 codewords is called a *codebook*.

Figure (v) 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. In the training phase, using the clustering algorithm described a *speaker-specific* VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result codewords (centroids) are shown in Figure (v) by black circles and black triangles for speaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the *total VQ distortion* is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified as the speaker of the input utterance. [5]

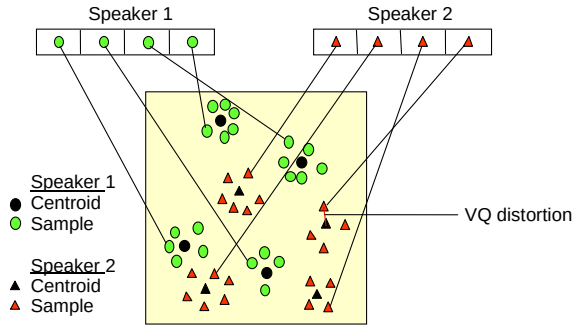

fig.(v)Conceptual diagram representing vector quantization codebook formation

LBG Algorithm [Linde, Buzo and Gray, 1980], may be employed for clustering a set of L training vectors into M codebook vectors. [5]

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 speaker-specific VQ codebook for this speaker using those training vector. The codebook is hence obtained for a speaker.

5. FEATURES MATCHING

The objects are generically called patterns and in this case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The goal of pattern recognition is to classify objects into one

of a number of categories or classes 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. This can also be said to be a testing phase.
Furthermore, if there are some set of patterns that the individual classes of which are already known, then one has a problem in supervised pattern recognition. These patterns comprise the training set and are used to derive a classification algorithm. The remaining patterns are then used to test the classification algorithm; these patterns are collectively referred to as the test set. If the correct classes of the individual patterns in the test set are also known, then one can evaluate the performance of the algorithm

6. ANALYSIS AND CONCLUSION

In our experiment we trained the system for a specific user for some specific words. We then identified that particular speaker by testing each registered speaker who provided their speech samples in the training phase.
In the training phase we actually build a reference model for a particular speaker and compared that stored reference model against the input speech and this is how recognition of a speaker is done.

The system developed is speaker dependent and moderately tolerant to background noise hence it is a very efficient system. The main strengths of the developed system are that it provides quite accurate results. A high degree of accuracy has been achieved by using this system setup in the previous works.

We used Hindi words for training the system ek (one), do (two), teen (three), char (four)…and the system have been found to possesses a great degree of learning and recognizing accuracy.
The system uses the MFCC feature extraction technique and Vector Quantization method for pattern matching combination which proved to be a great success. About 90% success rate is achieved during the experiment.
The experiment is carried out in 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.

7. FUTURE RECOMMENDATIONS

Isolated word recognition system developed here has been found to be very successful and achieved a very high learning accuracy and also a high learning rate. The standard deviation of about 10% from actual results has been calculated in our experiments.
In near future, we will try to apply combination of various other pattern matching techniques (VQ, DTW, HMM etc) along with other extraction techniques including MFCC, delta MFCC, deltadeltaMFCC, etc and conclude better results.

8. REFERENCES

[1] "DSP Mini-Project: An Automatic Speaker RecognitionSystem"http://www.ifp.uiuc.edu/~minhdo/teaching/speaker_recognition

[2] Nitisha Garg, Ashu Bansal, Anu Taneja "Understanding Speech Recognition System and Recognizing Hindi Language Numerals Using MATLAB" IJECWKI

[3] S K Hasnain "Recognizing Spoken Urdu Numbers Using Fourier Descriptor and Neural Networks with Matlab"IEEE

[4] Muhammad Salman Haleem "Voice Controlled Automation System"IEEE

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

[6] Mohammad A. M. Abu Shariah1, Raja N. Ainon1, Roziati Zainuddin1, Othman O. Khalifa2 "Human Computer Interaction Using Isolated-Words Speech Recognition Technology"IEEE

[7] www.Howstuffworks.com

[8] Digital signal processing by salivahanan, salivahanan, vallavaraj, gnanapriya

[9] J. Holmes and W. Holmes. "Speech Recognition and Synthesis" Taylor and Francis, 2001.

[10] Lawrence Rabiner, Biing-Hwang Juang – "Fundamentals of Speech Recognition"

[11] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy and Kong-Pang Pun – "An Efficient MFCC Extraction Method in Speech Recognition "The Chinese University of Hong Kong, Hong, IEEE - ISCAS, 2006

[12] Woszczyna, M.: "JANUS 93: Towards Spontaneous Speech Translation", IEEE

[13] www.Wikipedia.com/Speaker_recognition.htm

[14] www.Wikipedia.com/Vector Quantization.htm

[15] Rui Cai†, Lie Lu‡, Hong-Jiang Zhang‡, and Lian- Hong Cai† "IMPROVE AUDIO REPRESENTATION BY USING FEATURE STRUCTURE PATTERNS" IEEE

[16] Talal Bin Amin, Iftekhar Mahmood "Speech recognition using Dynamic Time warping" IEEE

[17] Ali Zulfiqar, Aslam Muhammad, Martinez Enriquez A.M "a speaker identification system using MFCC Features with VQ Technique". IEEE

[18] Md Afzal Hossan Thesis on "Automatic Speaker Recognition Dynamic Feature Identification and Classification using Distributed Discrete Cosine Transform Based Mel Frequency Cepstral Coefficients and Fuzzy Vector Quantization", RMIT