ABSTRACT- This paper presents a study of robust speaker recognition for regional Indian accents. Features of the speech signal are extracted in the form of MFCC coefficients and Dynamic Time Warping (DTW) has been used as features matching techniques. The database consists of the key words. Each key word has been stored in database by the ten speakers, eight male speakers and two female speakers. The recognition results are tested for clean and noisy test data. Average accuracy for clean data is good while that for noisy data is poor. We face problem in noise environment to detect correct utterance. We are going to review different papers and find out different techniques to design our ASR control system for regional Indian accents using MFCC and DTW in noise environment.

Keywords- Speech recognition; Dynamic Time Wrapping (DTW); MFCC; Mel frequency; Automatic Speech Recognition (ASR); Hindi Key words Recognition; Mel Frequency Cepstral coefficient (MFCC); Dynamic Time Wrapping (DTW).

I. INTRODUCTION

The term ‘mel’ denotes some kind of measurements of perceived frequency or pitch of a tone. It is not linearly in accordance with the physical frequency of the corresponding tone, as the human auditory system obviously does not perceive pitch in linear manner. Besides, it has been found that the perception of some frequency by the auditory system, say f, is greatly influenced by the energy in the critical band of the frequencies around f. The bandwidth of the critical band varies with the perceived frequency. The advantages are capable of being immune to noise and easy to warp frequency to a non-uniform scale, such as mel scale [2]-[3]. The mapping between the real frequency scale (Hz) and the perceived frequency scales (mels) is approximately linear below 1KHz and logarithmic at higher frequency, and such an approximation is usually adopted in speech recognition. The suggested formula that models their relationship is described as [1]

\[ f_{mel} = \frac{2595 \cdot f_{Hz}}{1 + \frac{f_{Hz}}{700}} \]

When the user says something, this is known as an utterance. An utterance is any stream of speech between two periods of silence. Utterances are fed to the speech engine to be processed then human voice is converted into digital signal representing each level of signal at every discrete time step. The digitized speech samples are then processed using MFCC to produce voice features. After that, the coefficient of voice features can go through DTW to select the pattern that matches the database and input frame in order to minimize the resulting error between them. The popularly used cepstrum based methods to compare the pattern to find their similarity are the MFCC and DTW. The MFCC and DTW features techniques can be implemented using MATLAB 7.5. This work was taken to focus on Hindi key word recognition. [4] In this paper we are presenting work consists of the creation of key word database and its speech recognition system for home automation.

The main goal in this paper is to enhance speaker recognition system performance at lower FARs with the help of an accent classification system, even when evaluated on a realistic noisy dataset.

A) Speech production process:

For efficient extraction of features, it is necessary to understand the speech production mechanism in human beings. Then the conversion of the message into a language code takes place in which the talker converts the message into sets of phoneme sequences corresponding to the sounds that make up the words. Along with that the talker also determines the duration, loudness of sounds and pitch associated with it. Once the language code is chosen, the talker must execute a series of neuromuscular commands to cause the vocal cords to vibrate and to shape the vocal tract such that the proper sequence of speech sounds is produced. Depending on which speech sound you articulate, the speech signal can be excited in three possible ways:

- **Voiced excitation:**
  The glottis is closed. The air pressure forces the glottis to open and close periodically thus generating a periodic pulse train (triangle shaped). This “fundamental frequency” usually lies in the range from 80Hz to 350Hz [5].

- **Unvoiced excitation:**
  The glottis is open and the air passes a narrow passage in the throat or mouth. This results in a turbulence which generates a noise signal. The spectral shape of the noise is determined by the location of the narrowness.

- **Transient excitation:**
  A closure in the throat or mouth will raise the air pressure. By suddenly opening the closure the air pressure drops down immediately.

Deepika Vashisht1, Swati Sharma2, Lucky Dogra3
1Research Scholar, Universal Institute of Engineering and Technology, Lalru, Mohali.-Punjab (India)
2Associate Professor, Universal Institute of Engineering and Technology, Lalru, Mohali.-Punjab (India)
3Assistant Professor, Universal Institute of Engineering and Technology, Lalru, Mohali.-Punjab (India)
1deepikavashisht18@gmail.com, 2drswatisharma.ugi@gmail.com, 3lucky.dogra@yahoo.com
With some speech sounds these three kinds of excitation occur in combination. The spectral shape of the speech signal is determined by the shape of the vocal tract. By changing the shape of the pipe (and in addition opening and closing the air flow through your nose) you change the spectral shape of the speech signal, thus articulating different speech sounds.

![Diagram of Human Speech Production Process]

**II. ISSUES FOR SPEAKER RECOGNITION**

**A) Body Movement:**
While a person is in conversation his/her body movement plays a crucial role in order to express his/her feeling to the other person like hand movement, eye contact, body posture etc. But while it comes to ASR system, this system completely misses these body movements because it just works on speech signals not on the body movements. This is one the issue with [5] ASR system but researchers are working on it and one of the technique which is developed till now is from the field of Multimodality in which studies are carried for including body movement to human-computer interaction.

**B) Noise:**
Noise is one of the biggest issue from which ASR system has to undergo. Noise is available everywhere in our surrounding and to overcome with it while you are in conversation is impossible because while you are in conversation it is possible that a radio plays in background or some other person speaks in background and due to this ASR system will not be able to work properly and results will not be up to the mark. Any kind of unwanted sound is called Noise and this unwanted sound can come from anywhere. Noise is a thing which can’t be removed completely but its effect could be reduced. Different techniques are available which works on noise reduction and different filters were also available for the same purpose. These [6] filters reduces the effects of noise from speech signals which results in good efficiency of the system.

**C) Spoken Language ≠ Written Language:**
Spoken language is very much different from written language as while you are speaking there is no consideration of grammatical mistakes while in written language, grammatical mistakes are considered. Written communication is one way communication while spoken communication is based on feedback. So, in ASR system we have to consider these differences between both of these two ways of communication. Other than this, disfluences also creates problem for ASR system. In normal An Introduction To Speech Recognition And Its Approaches Proceedings of [7] speech hesitations, repetitions of words, change of subject, slip of tongue etc. are not considered but ASR system has that kind of behavior which considered these kind of disfluences. So, these are also taken into consideration.

**D) Continuous Speech:**
Different Speaker have different way of speaking. Some speaks with slow speed with some gap in words but some have fast way of speaking with no gap between the words. Speech as no natural pauses within the words boundaries. In speech mainly pauses occurs in syntactic level like after a phrase. This causes a problem to speech recognition of how a waveform will translated into a sequence of words? The simplest way to overcome with this problem is to give clear pause between words so that system could able to understand all these pauses clearly and based on that translation could be done but this technique only works in case of shorts commands only.

**E) Regional and Social Dialects:**
Regional and Social dialects includes features of vocabulary pronunciation and grammar which differ based on the geometrical region of speaker whereas social dialects are distinguished by features of pronunciation, vocabulary and grammar based on social group of the speaker. ASR system have an issue with this change in vocabulary and pronunciation as same word have more than one way of pronunciation which causes problem for the system.

**F) Speaking style:**
Everyone has their own way of speaking, their own personality. Each individual have different style of speaking while you are in different situations like we do not speak the way we speak in bank with our friends and with our family. Emotions effects our way of speaking. Because of these emotions our style of speaking gets changed like if we are sad we drop our pitch and speak more slowly and if we are angry then we speak in high pitch and our speed also varies. So, these emotions have an effect on our way of speaking. ASR system has to consider this effect of emotions on speaker’s voice then only good results could be obtained.

**G) Time:**
Time means, how much time an individual takes for completion of a particular sentence. An individual has different time for completion of a particular sentence in different emotional states. If we talk about a secretary, a good secretary has a speed of typing 300 keys utters in a minute while average speed is 150 words per minute. So, typing rate is not high enough to transfer a stream of spoken words into a readable form in real time and this is one of the issue with ASR system which has to be overcome for obtaining good efficiency of the system.
H) Message Transfer:
The main goal behind speech to text transfer is to give people access to spoken words and auditory events with realization of original sound event. However for deaf people or people with limited access to spoken languages, transfer of spoken words into written text may not be very helpful. Written form become useless in case of unknown words or if sentences are too complex. A speech to text provider has to know the audience as well as they should know which phrases or words should be exchanged so that they could be easily understood and have less complexity. Speech to text provider should know techniques of how to make languages more accessible which transferring information is preserved.

I) Real Time Presentation of the Written Text:
Reading means that words are already written and different people will read that in their respective reading speed. But this is not possible in case of real time speech to text conversion. In this case, text is written and control of reading speed shifts between speaker and to the Speech to text provider. If the text is not written means that it is not available in advance and new words are produced continuously, at that time Reader must follow production process very carefully for the purpose of using real time abilities of Speech to text transfer. Because of this connection between writing and reading, presentation of written text must be optimally adapted to the reading needs of the audience.

III. MEL-SCALE FREQUENCY CEPSTRAL COEFFICIENT (MFCC)
Mel Frequency Spectral Coefficient (MFCC) was introduced by Davis and Mermelstein in the 1980 AD. It is very common and one of the best method for feature extraction method especially for automatic speech and speaker recognition system. An application of hand gesture recognition, MFCC was used as feature extractor by converting input image into 1D signal with SVM classifier [12]. The MFCC coefficients can be used as audio classification features to improve the classification accuracy, is used for the music features, and then BPNN algorithm recognizes the music classes [13].
MFCC is used in speaker verification with speaker information like, contents and channels [11]. MFCCs are a feature widely used in automatic speech and speaker recognition. There is computation for extracting the cepstral features parameters from the Mel scaling frequency domain. Feature Extraction refers to the process of conversion of sound signal to a form suitable for the next stages to use. Feature extraction may include extracting parameters such as amplitude of the signal, energy of frequencies, etc. Recognition involves mapping the given input (in form of various features) to one of the known sounds. This may involve use of various knowledge models for precise identification. Knowledge models refer to models which help the recognition system. Mel-frequency cepstrum coefficients (MFCC) are well known features used to describe speech signal. They are based on the known evidence that the information carried by low frequency components of the speech signal is phonetically more important for human perception than carried by high-frequency components. Technique of computing MFCC is based on the short term analysis, and thus from each frame a MFCC vector is computed. MFCC extraction is similar to the cepstrum calculation except that one special step is inserted, namely the frequency axis is warped according to the Mel scale. Summing up, the process of extracting MFCC from continuous speech is illustrated in Figure 2 given bellows.

A) Setup for recording (Hindi key words):
For key words recording samples are taken at home with noise and without noise. To achieve a high accuracy for key word recognition in home automation system we use 10x12 rooms for recording some key words without noisy sound (fan off, tap off, no cooking in kitchen) and some key words with internal noise environment (fan on, tap on, cooking in kitchen). The sampling frequency for all recording was 16000 HZ and duration of recording was two seconds. We use i-ball microphone for voice recording. The distance of mouth to microphone is about 5-10 cm. The speech data is recorded with the help of wavrecord command, apply pre-emphasis and store in database using wavwrite command.

B) Create vectors for key words (Database):
After recording we computed these key words and carried out end point detection using threshold levels. If any frame is with volume less than the threshold it is labeled as silence. After end point detection we reduce the noise by calculating mean silence energy (noise energy), maximum energy for entire utterance, zero crossing rate thresholds, lower and upper thresholds. After getting good quality of signal we calculate Mel frequency cepstral

![Figure 2. Mel Frequency Cepstral Coefficient Flow Chart](apm.com)
coefficients. These coefficients are then stored as a vector in the database.

C) Pre-emphasis:
Pre-emphasis refers to increase, within a band of frequencies, the magnitude of some (usually higher) frequencies with respect to the magnitude of the others (usually lower) frequencies in order to improve the overall SNR. Hence, this step processes the passing of utterance through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency. We use these commands to pre-emphasis utterance.

\[
\begin{align*}
\text{a} &= 0.95; \\
\text{Speech In} &= \text{filter}([1, -a], 1, y);
\end{align*}
\]

The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. Moreover, it can also amplify the importance of high-frequency formants. Figure 3 shows pre-emphasis for Hindi key word “AAGE”.

D) Framing:
The process of segmenting the speech samples into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M < N). Typical values used are M = 100 samples, sampling frequency (Fs)=16000 HZ, Time duration (Tf)=0.025 sec. Hence number of samples per frame is 400 (Fs*Tf). Figure 4 shows the framing of Hindi key word “AAGE”.

E) Hamming Window:
The Hamming window applied to the signal is to prevent the non-continuity in the two ends of the audio frame, and to avoid the influence of front and back audio frames when analyzed. The non-continuity of the audio frames can be eliminated by multiplying the Hamming window with the audio frames, because this process can make each audio frame more centered on the frequency spectrum. The following Equations (1) and (2) are the function of Hamming window [5] and the result of the multiplication from Hamming window and audio frame, respectively.

\[
\begin{align*}
W(n) &= 0.54 - 0.46 \cos \left[ \frac{2 \pi n}{L-1} \right]; \quad 0 \leq n \leq L-1, \\
W(n) &= 0; \quad \text{otherwise} \quad (1) \\
F(n) &= W(n) \times S(n) \quad (2)
\end{align*}
\]

In which \(S(n)\) is a frame of speech signal, \(W(n)\) is the Hamming window, and \(F(n)\) is the result of audio frame multiplied by Hamming window as shown in Figure 5.

\[
\begin{align*}
\text{Figure 5. Hamming window x-axis (400 samples) and y-axis (amplitude).}
\end{align*}
\]

F) Fast fourier transform:
In this step we convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse \(U[n]\) and the vocal tract impulse response \(H[n]\) in the time domain. This statement supports as shown below in the “Eq. (3)” [2].

\[
Y(w) = \text{FFT} [h(t) \ast X(t)] \\
Y(w) = H(w) \ast X(w) \quad (3)
\]

If \(X(w)\), \(H(w)\) and \(Y(w)\) are the Fourier Transform of \(X(t)\), \(H(t)\) and \(Y(t)\) respectively. Figure 6 shows FFT for Hindi key word “AAGE” and its power spectrum.

\[
\begin{align*}
\text{Figure 6. Compute FFT and Power Spectrum.}
\end{align*}
\]

G) Mel filter bank processing:
The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale as shown in figure 7 is then performed. Figure 7 shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter’s magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters [7]-[8].
Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency \( f \) in HZ:

\[
F_{\text{Mel}}(f) = \{2595 \times \log_{10} [1 + f/700]\} 
\]

(4)

**H) Discrete cosine transform:**

In this step convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of these coefficient are called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

Delta energy and delta spectrum: The frame-to-frame information in voice signal changes. Therefore, there is a need to add features related to the change in cepstral features over time. 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame for a signal \( x \) in a window from time sample \( t_1 \) to time sample \( t_2 \), is represented as shown below in “Eq. (5)” [2].

\[
\text{Energy} = \sum X^2[t] 
\]

(5)

Where \( X[t] = \text{signal} \)

Each of the 13 delta features represents the change between frames corresponding to cepstral or energy feature, while each of the 39 double delta features represents the change between frames in the corresponding delta features.

**I) Feature matching (DTW) scores:**

DTW algorithm is based on Dynamic Programming. This algorithm is used for measuring similarity between two time series which may vary in time or speed. This technique also used to find the optimal alignment between two times series if one time series may be “warped” non-linearly by stretching or shrinking it along its time axis. This warping between two time series can then be used to find corresponding regions between the two time series or to determine the similarity between the two time series.

To align two sequences using [9]-[11] DTW, an \( n \)-by-\( m \) matrix where the \((i, j)\) element of the matrix contains the distance \( d(q_i, c_j) \) between the two points \( q_i \) and \( c_j \) is constructed. Then, the absolute distance between the values of two sequences is calculated using the Euclidean distance computation as shown in “Eq. (6)”.

\[
d(q_i, c_j) = d(q_i, c_j)^2 
\]

(6)

Each matrix element \((i, j)\) corresponds to the alignment between the points \( q_i \) and \( c_j \). Then, accumulated distance is measured by “Eq. (7)” [2].

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

(7)

When we combine these MFCCs they produce a unique voice print for that word. With the help of dynamic time wrapping algorithm we can find out minimum global distance between reference and stored vectors. This specrogram represent for ‘AAGE’ keyword in Figure 9. The dark color represents the utterance for ‘AAGE’ key word. These features are extracted and store in a vector for comparison with reference key word word. We take 10 neighbored distances with reference key word. If distance between stored and reference key word greater then 5, then give the output otherwise try again.

**IV. CONCLUSION**

So far we have seen that we have got a good level of accuracy also we will like to add more formant features and try to get a better level of accuracy as it is known that the more formant features are added the more accent features are added. In this paper, speaker recognition using Mel-frequency cepstral coefficients (MFCC) have been used as speaker features. A novel method to recognize/identify speakers was developed by including a new set of features. The algorithm was evaluated using dataset and the results showed on the average 5% improvement over the performance of our previous work [4]. Our future work will be to evaluate the proposed feature set for an even noisier database and observe the impact on its performance. We need to study the effect of adding more formants to see if that will provide better accent discrimination. Also we need to investigate ways to improve the computational time complexity of the proposed algorithm.

**REFERENCES**


AUTHORS

Deepika Vashisht, Research Scholar, in Electronics and Communication Engineering Universal Institute of Engineering and Technology, Lalru, Mohali, Punjab (India).

Dr Swati Sharma is an academic professional having an experience of over twelve years. She is an alumnus of Panjab University, Chandigarh having expertise in the areas of Wireless communication and Digital Signal Processing. Dr Swati Sharma has guided over seven post graduate research scholars and has about fifteen publications in renowned national and international journals. She has attended and presented papers in various national and international conferences held in India and abroad. Dr Swati has also organised various sponsored Faculty/Staff Development Programs. Presently she is working as Dean Academics in Universal Institute of Engineering and Technology, Lalru Distt. Mohali. She has headed and organised various national events during her tenure in the institution.

Email: drswatisharma.ugi@gmail.com

Lucky Dogra has obtained his Bachelor of Technology degree in Electronics & Communication Engineering from CT Institute of Engineering Management and Technology, Jalandhar and Master of Technology degree in VLSI Design from C-DAC, Mohali under Punjab Technical University in 2011 and 2014 respectively. He is working as an Assistant Professor in Electronics and Communication department at Universal Group of Institutions, Lalru, Punjab. He has more than 2 years of experience. His major research interest is digital integrated circuits design.

Email: lucky.dogra@yahoo.com