IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. VII (May-Jun. 2014), PP 79-85
www.iosrjournals.org

Design and Development of Automatic Speech Recognition of Isolated Marathi Words for Agricultural Purpose

Smita B. Magre[1], Ratnadeep R. Deshmukh[2]
[1, 2](Department of Computer Science and IT, Dr. B. A. M. University, Aurangabad – 431004, India)

Abstract : *Speech is a natural mode of communication for people. Yet people are so comfortable with speech, they would also interact with the computers via speech, and various interfacing devices such as keyboards and pointing devices. Outstanding work in speech recognition and computing has produced the commercial speech recognition systems for voice driven computing and word-processing systems. The analysis within the space of speech recognition lots of work is done for English language and European language and the opposite hand very little work done for Indian language and really little in Marathi. Only Because of This We Develop Automatic Speech Recognition of Isolates Marathi Words For Agriculture Purpose. For developing ASR system we use hybrid feature extraction technique i.e. MFCC with RASTA and for recognition Dynamic Time Wrapping is used.*

Keywords: *DTW, MFCC, RASTA, Speech Recognition.*

I. INTRODUCTION

Speech is that the most prominent and natural variety of communication between humans. There are various spoken languages throughout the world. Communication among the human being is dominated by spoken language, therefore it's natural for people to expect speech interfaces with computer [1]. Speech Recognition is the ability of a computer to recognize general, naturally flowing utterances from a wide variety of users. It recognizes the caller's answers to move along the flow of the call [2]. ASR Technology that allows a computer to identify the words that person speaks into a microphone or telephone and convert it to written text. The ultimate goal of ASR research is to allow a computer to recognize in real time, with 100% accuracy, all words that are intelligibly spoken by any person, independent of vocabulary size, noise and speaker characteristics.

Speech recognition is an automatic conversion of spoken works into system based mostly readable text format. This technology permits a pc to acknowledge words that are spoken into mike or a telephone. This technique is additionally known as as Automatic Speech Recognition [3].

II. METHODOLOGY

Step 1: We developed a text corpus by making a collection of various words used in agriculture.
Step 2: Speakers are selected from Aurangabad district, both male and female.
Step 3: We have recorded 100 different words having 3 utterances of each from 100 different speakers.
Step 4: After the collection of speech samples Features are extracted using Mel Frequency Cepstral Coefficient (MFCC) in combination of RASTA.
Step 5: word recognition using DTW (Dynamic Time Warping).

A. Selection of the Speakers:

- One hundred speakers were selected from all over Aurangabad district.
- Selected in the speakers whose native language is Marathi and also those speakers whose native language is not Marathi.
- The speakers were selected to cover the maximum variation of the language of the district.
- The speakers are classifying on the basis of gender i.e. male and female.

B. Recording Procedure:

- We used PRAAT software for recording the speech. PRAAT is a very versatile tool to do speech analysis. It offers a good range of ordinary and non-standard procedures, together with spectrographic analysis, articulative synthesis, and neural networks [4].We used Sennheiser PC360 and Sennheiser PC350 headset for recording the speech samples. The PC360 and PC350 headsets are having noise cancellation facility and the signal to noise ratio (SNR) is less.

C. Steps for Recording the Speech Samples:

Step 1: *Selected speakers were asked regarding any problem with reading or speaking the Marathi words.*
Step 2: *Speakers were given the basic information about the headset used and when to speak the word.*
Step 3: *The sampling frequency was set to 24 KHz with 16 bit in Mono sound type.*
Step 4: *The speaker was asked to read each word and the recorded sample was saved as .wav file.*
Step 5: *Step 4 was repeated for all 300 utterances that were recorded from the speaker. All the steps were repeated for all the 100 speakers*

D. Data Collection Statistics:

- We have target 100 speakers from Aurangabad district. The Speech samples are recorded from 100 speakers out of which 50 were Male and 50 were Female in Noisy environment.
- As the speakers were asked to speak the 100 words with 3 utterances of each word we collected total 300 speech samples from each speaker. We collected total 30,000 utterances in all of the 100 words from 100 speakers.

III. IMPLEMENTATION

A. Feature Extraction

The main goal of the feature extraction step is to compute a saving sequence of feature vectors providing a compact representation of the given input signal. The feature extraction is usually performed in three stages.
The first stage: The speech analysis or the acoustic front end, It performs spectro temporal analysis of the signal and generates raw features describing the envelope of the power spectrum of short speech intervals.
The second stage: Compiles an extended feature vector composed of static and dynamic features.
Finally, the last stage: Transforms these extended feature vectors into more compact and robust vectors that are then supplied to the recognizer [5].

A block diagram of the structure of an MFCC processor is given in Figure 3.1. Waveforms themselves, MFFCs are shown to be less susceptible to mentioned variations [6]. The overall process of the MFCC is shown in Figure 3.1.

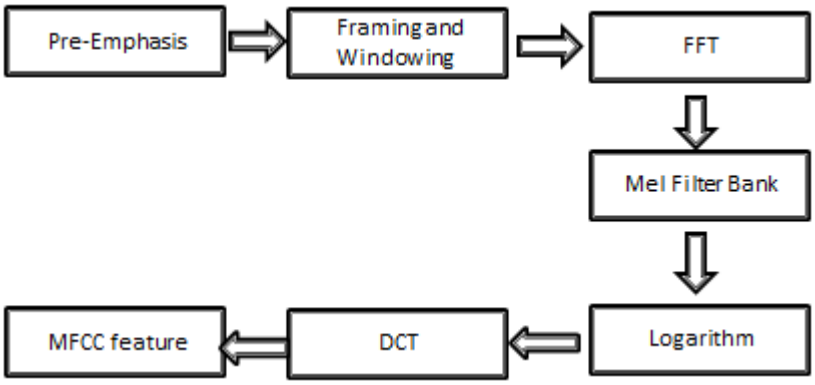

Figure 3.1 Block Diagram of MFCC Processing.

Step 1: Pre-emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

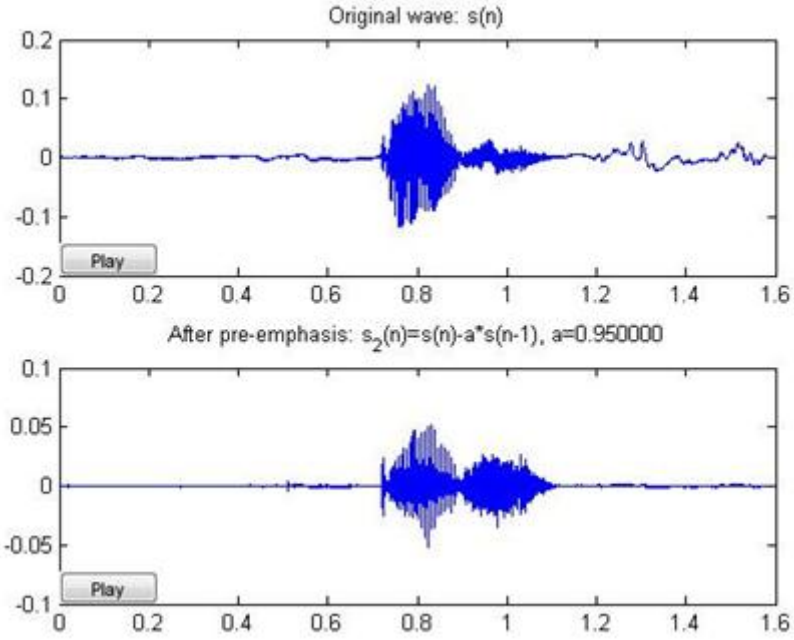

Figure 3.2 Pre-emphasis Signal of Word Tulas.

Step 2: Framing

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) 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.

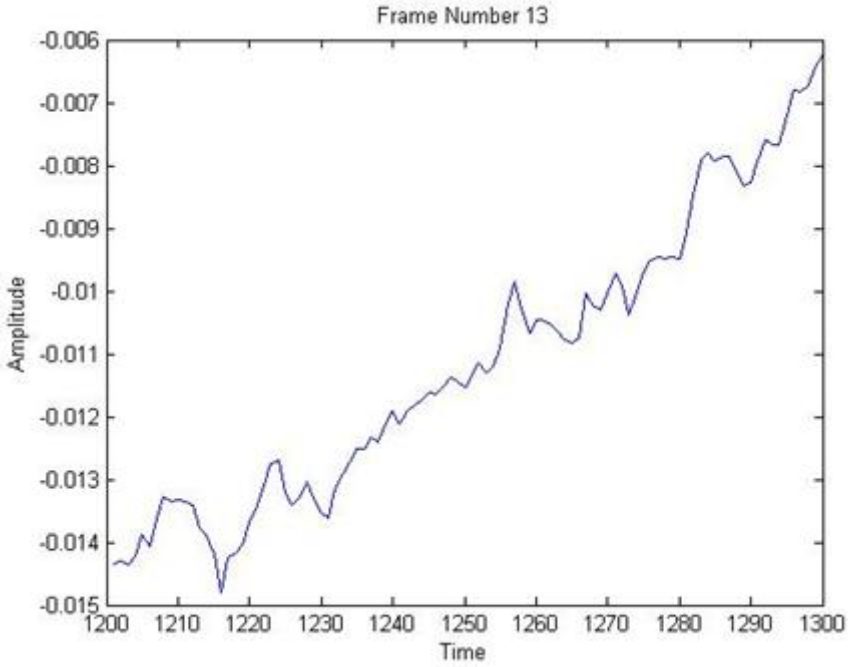

Figure 3.3 Framing

Step 3: Hamming Windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines.

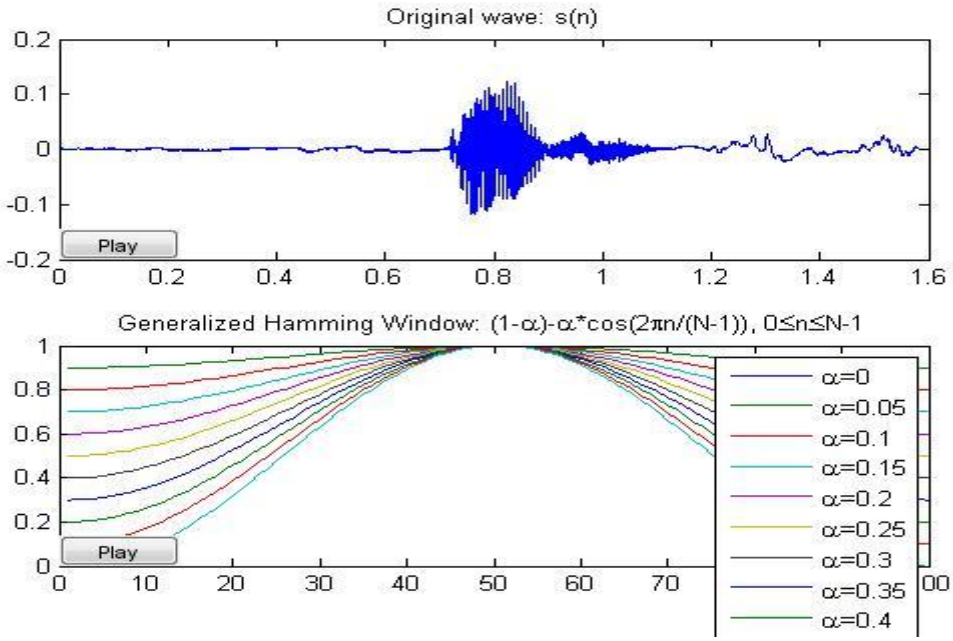

Figure 3.4 Hamming Windowing for Word Tulas.

Step 4: Fast Fourier Transform

To 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.

Step 5: 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 3.5 is then performed.

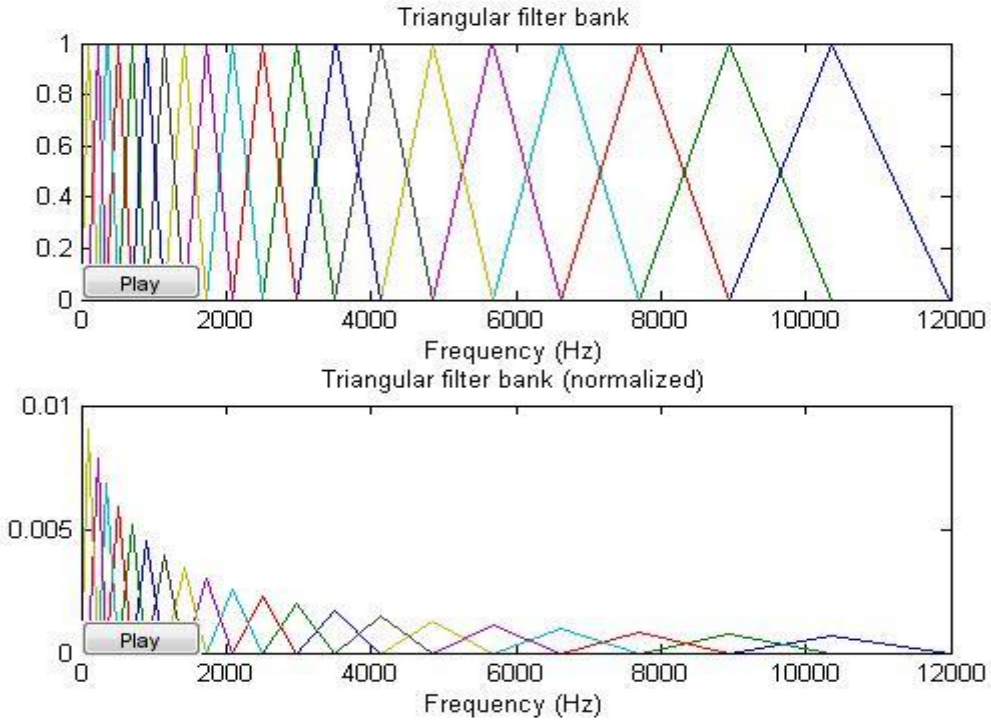

Figure 3.5 Mel Scale Filter Bank for Word Tulas.

This figure 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. Then, each filter output is the sum of its filtered spectral components.

Step 6: Discrete Cosine Transform

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT).

Discrete Cosine Transform (DCT) is being used to achieve the Mel-cepstrum coefficients. In a frame, there are 24 Mel-Cepstral coefficients, out of 24 only 13 coefficients have been selected for the recognition system.

B. RASTA (Relative SpecTral) Processing

We have to use the combination of MFCC and RASTA because we record the database in noisy environment and MFCC values are not very robust in the presence of additive noise, RASTA is find out Feature in Noisy data. The RASTA algorithm is a common piece of a speech-recognition system's front-end processing. It originally was designed to model adaptation processes in the auditory system, and to correct for environmental effects. Broadly speaking, it filters out the very low-frequency temporal components (below 1Hz) which are often due to a changing auditory environment or microphone. High frequency temporal components, above 13 Hz, are also removed since they represent changes that are faster than the speech articulators can move. The first input to this routine is an array of spectral data, as produced by the MFCC routines. Each row contains one "channel" of data; each column is one time slice. The fs parameter specifies the sampling rate, 100Hz in many speech recognition systems. The original RASTA filter is defined only for a frame rate of 100Hz. This code is equal to the original at 100Hz, but scales to other frame rates. Here the RASTA filter is approximated by a simple fourth order Butterworth bandpass filter [7]. The figure 3.6 shows the basic blocks of combination of MFCC and RASTA.

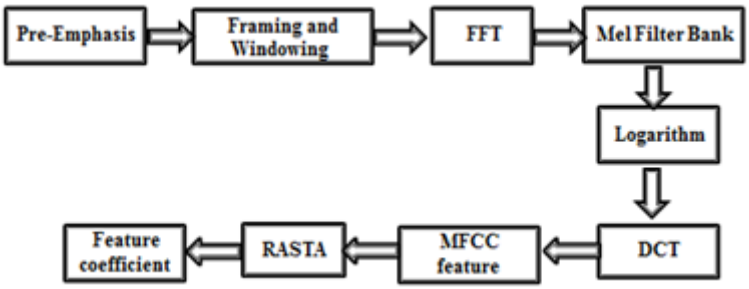

Figure 3.6 Block Diagram of MFCC+RASTA Processing.

MFCC analysis have the property that they turn a row of spectrogram into a small number of independent ,random coefficient this makes it to difficult view the result directly. RASTA then process these coefficient in time. Appling RASTA to the output of the MFCC routine and then reconstructing the original filter bank representation by inverting the discrete-cosine transform [8].

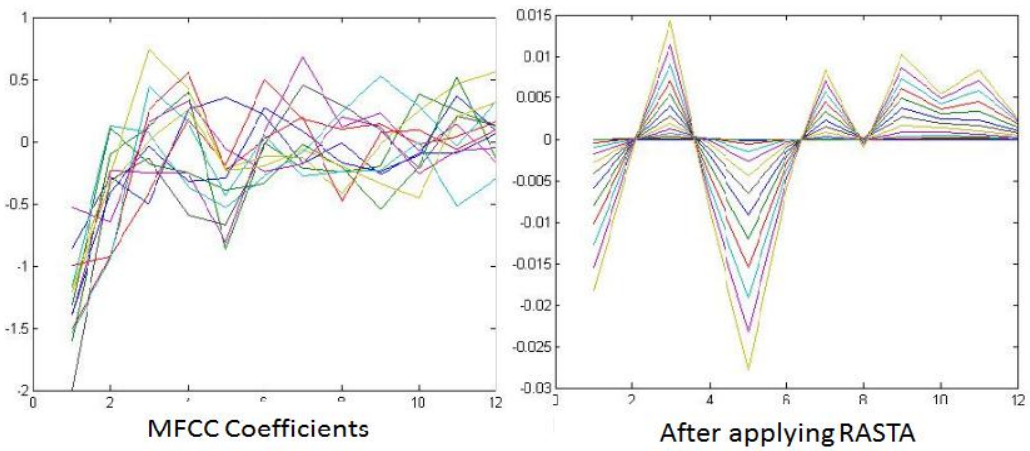

Figure 3.7 Graphical Representations of MFCC and RASTA.

IV. ISOLATED WORD RECOGNITION USING DTW

A. Dynamic Time Warping

DTW is a methodology that enables a computer to search out associate best match between two given sequences (e.g. time series) with bound restrictions. The sequences are} "warped" non-linearly within the time dimension to see a measure of their similarity freelance of bound non-linear variations within the time dimension.DTW has been applied to temporal sequences of video, audio and graphics information so, any data which may be become a linear sequence are often analyzed with DTW [9].

In Speech Recognition system DTW is usually used to confirm whether or not the two spoken waveforms represent an equivalent phrase. in a speech wave, the length of every spoken sound and therefore the interval between sounds are permitted to vary, however the speech must be similar.

Advantages of DTW:

- Reduced storing space for the reference template.
- Finding the optimal path.
- Increased recognition rate.
- If the error is too great then it stops the process.

B. Word Recognition

Word Recognition is a process where word uttered by the user has to be recognized by the speech recognition system. For recognition purpose we use Dynamic Time Warping Algorithm as Cost calculation unit. All the Trained data are put into Reference Frames one after the other. Now these Reference Features and Test Features are acts as inputs to the Dynamic Time Warping Algorithm program. Following figure 4.1 shows the flow of word recognition process.

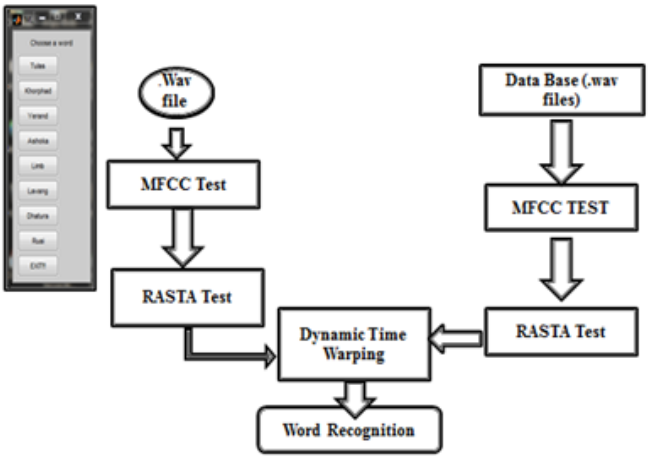

Figure 4.1 Isolated Word Recognition System.

This is a basic diagram of our ASR system which is basically divided into two parts database side and test side we have collected or recorded various words which are used in agricultures side. We have recorded 100 words, each word having 3 utterances. Among these first 2 utterances we have stored in database and 3rd utterance we are going to use as a test file. Then extract the feature using the combination of MFCC and RASTA. And compare the 3rd utterance with the two utterances which are stored in database. Same procedure we are going to use for all the files which are stored in our database one by one. After extraction of these files DTW is to compare these files and measure its similarity by calculating minimum distance between them.

Table 4.1 Distance Matrix of Ayurvedic Plants Using Dynamic Time Wrapping.

	Tulas	Khorphad	Yerand	Ashoka	Limb	Lavang	Dhatura	Ruai
Tulas	0.4019	0.931	0.522	0.759	0.634	0.738	0.751	0.673
Khorphad	0.728	0.574	0.589	0.680	0.720	0.692	0.785	0.705
Yerand	0.694	0.714	0.433	0.586	0.702	0.694	0.742	0.606
Ashoka	0.737	0.773	0.671	0.392	0.843	0.824	0.795	0.748
Limb	0.645	0.777	0.592	0.583	0.570	0.818	0.773	0.591
Lavang	0.689	1.007	0.598	0.753	0.819	0.526	0.749	0.741
Dhatura	0.662	0.768	0.506	0.526	0.742	0.718	0.504	0.567
Ruai	0.668	0.780	0.604	0.593	0.654	0.788	0.757	0.456

V. CONCLUSION

The main aim of this research work was to develop a speech database and automatic speech recognition system of isolated words for agriculture purpose in Marathi language. To developing ASR we use combination of MFCC and RASTA because it gives better and accurate result and for recognition purpose we use Dynamic Time Wrapping approach. It increases the recognition rate.

Acknowledgements

The author would like to thank the university authorities for providing the infrastructure to carry out the research. This work is supported by university commission.

REFERENCES

[1] Pukhraj P. Shrishrimal, Ratnadeep R. Deshmukh and Vishal B. Waghmare, "Indian Language Speech Database: A Review," International Journal of Computer Applications, Volume 47, No.5, 2012.
[2] C. Vimala and V. Radha, "A Review on Speech Recognition Challenges and Approaches," World of Computer Science and Information Technology Journal (WCSIT), Volume 2, No. 1, 2012.
[3] S.Ananthi and P. Dhanalakshmi, "Speech Recognition System and Isolated Word Recognition based on Hidden Markov Model (HMM) for Hearing Impaired", International Journal of Computer Applications, Volume 73, No. 20, 2012.
[4] P. V. Lieshout, "PRAAT Short Tutorial," Volume 4.2.1, 2003.
[5] W. Ding and G. Marchionini, "A Study on Video Browsing Strategies," Technical Report, University of Maryland at College Park 1997.
[6] M. Lindasalwa, B. Mumtaj and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," Journal of Computing, Volume 2, Issue 3, 2010.
[7] H. Hermansky and N. Morgan, "RASTA Processing of Speech," IEEE Transaction And Processing, Volume 2, No.4, 1994.
[8] H. Zhao, L. Hu, X. Peng, G. Wang, F. Yu and C. Xu, " An Improving MFCC Features Extraction Based on Fast ICA Algorithm Plus RASTA Filtering," Journal of Computers, Volume 6, No. 7, 2011.
[9] Dynamic Time Warping-Wikipedia, the Free Encyclopedia.