Design and Development of Continuous Marathi Speech Recognition System for Agriculture Purpose

Pratik Kurzekar[1], Shriniwas Darshane[2], Nikhil Salvithal[3], [4]Nitin Maske

[1]*Department of Computer Science and Engineering, SVERI's College of Engineering, Pandharpur*

[2]*Department of Computer Science and Engineering, SVERI's College of Engineering, Pandharpur*

[3]*Department of Computer Science and Engineering, SVERI's College of Engineering, Pandharpur*

[4]*Department of Computer Science and Engineering, SVERI's College of Engineering, Pandharpur*

[1]*pkkurzekar@coe.sveri.ac.in, svdarshane@coe.sveri.ac.in, nnsalvithal@coe.sveri.ac.in, nmmaske@coe.sveri.ac.in*

Abstract: *The research in the area of speech recognition for the English language and European languages has reached up to a critical level to be used for a real communication tool. On the other hand, the research for Indian languages has not yet reached to develop an application. The research for Indian languages has been carried out now in various institutes and research labs but, the research is more concentrated towards the development of applications in Tamil, Telugu, and Hindi, but the problem that is faced while doing the research is about the database for developing the application. The present study attempts to fulfill the need for the database.*

Keywords: *Continuous Speech Recognition (CSR), Automatic Speech Recognition (ASR), Mel-frequency Cepstral Co-efficient (MFCC).*

1. Introduction

Speech recognition is a widely researched topic around the world. Many scientists and researchers are busy doing works on speech recognition. Most of the languages within the world have speech recognizers of its own. But our mother tongue Marathi has not enriched with speech recognizers. Small research works have been carried on Marathi's speech recognizer, but it does not have a great outcome. Implementing continuous speech recognizers for Marathi is our primary goal throughout the thesis work. But implementing a recognizer is a huge task within a short period. [1].

Already various efforts have been made towards Automatic Speech Recognition (ASR) in Indian languages like Tamil, Telugu, Marathi, and Hindi. The recognition performance is acceptable in noise-free condition, but it degrades dramatically in the presence of noise. Investigation of sturdy options for recognition of abuzz speech remains a lively space of analysis. If the recognition system is implemented in public areas where several external noises such as aircraft noise, the speech of other persons, etc. are present, then the performance of the system degrades because the system is trained with clean speech and testing is performed in a noisy environment. So there is a need to analyze the performance of the system in noisy as well as noise-free conditions. The analysis approach used in this paper first adds controlled amounts of various types of noises to clean training data with a specified signal to noise ratio. The recognition performance is analyzed with noise suppression techniques applied to reduce the noise from the corrupted noisy signal [2].

The speech could be a natural mode of communication for folks. We learn all the relevant skills throughout the time of life, while not instruction and we continue to rely on speech communication throughout our lives. It comes thus naturally to the USA that we do not understand; however, advanced development speech is. The speech is the most prominent & primary mode of communication among human

beings. The communication among human-computer interaction is called a human-computer interface. Speech has the potential of being a crucial mode of interaction with a laptop.

The speech could be a primary mode of communication among persons and conjointly the foremost natural and economical style of exchanging info among humans in speech. So, it is only logical that the next technological development to be natural language speech recognition for Human-Computer Interaction. Speech Recognition may be outlined because of the method of changing a speech signal to a sequence of words by suggests that formula enforced as a bug. The speech process is one of the exciting areas of the signal process. The goal of the speech recognition area is too developed technique and system to develop for speech input to machine [3].

Speech recognition is a pattern recognition task in which acoustical signals are examined and structured into a hierarchy of subword units (e.g., phonemes), words, phrases, and sentences. Each level could offer extra-temporal constraints, known word pronunciations, or legal word sequences, which can compensate for errors or uncertainties at lower levels. This hierarchy of constraints will best be exploited by combining choices probabilistically in the slightest degree lower levels and creating distinct choices solely at the very best level. The structure of a speech recognition system is illustrated in Figure. The elements are as follows [4].

1.1 Types of Speech Recognition:

Speech recognition systems can be classified or separated into several different classes. The classification describing which types of utterances recognition ability they have. These classes are classified as follows:

1.1.1 Classification based on utterances:

Isolated Words: Isolated word recognizers usually require each utterance to have entirely on both sides of the sample window. It accepts single words or single utterance at a time. These systems have "Listen/Not-Listen states," where they require the speaker to wait between utterances.

Connected Words: Connected word systems are similar to isolated words, but it allows separate utterances to be "run-together" with a minimal pause between them.

Continuous Speech: Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize unique methods to determine the utterance boundaries.

Spontaneous Speech: At a basic level, it can be considered as a speech that is natural-sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs," and even slight stutters [5].

1.1.2 Classification based on Vocabulary size

Small Vocabulary: The speech recognition systems which can recognize limited and given a set of vocabulary (i.e., a few hundred words or sentences) are known as limited vocabulary speech recognition system.

Medium Vocabulary: The speech recognition system which can recognize a considerable number of vocabularies (i.e., a few from few hundred up to few thousands of

words or sentences) such systems are known as medium vocabulary speech recognition system.
Large Vocabulary: The speech recognition system which can recognize a large number of vocabularies (i.e., more than a few thousands of words or sentences) such systems are known as large vocabulary speech recognition system [6].

1.1.3 Classification based on Speaker mode

Speaker Dependent: Speaker dependent speech recognition systems learn the unique characteristics of a single person's voice, in a way similar to voice recognition. The system is trained using the training dataset, and it may use templates.

Speaker Independent: In speaker-independent speech recognition systems, there is no training of the system to recognize a particular speaker. So the stored word patterns must be representative of the collection of speakers expected to use the system. The word templates are derived by first obtaining a large number of sample patterns from a cross-section of talkers of different sex, age-group, and dialect, and then clustering these to form a typical pattern for each word.

Speaker Adaptive: In speaker adaptive speech recognition systems, the uses the speaker-dependent data and adapt to the best-suited speaker to recognize the speech and decrease the error rate by adaption [7].

2. Design and Development of Database

MFCC is a standard method for feature extraction in speech recognition tasks. MFCC transforms the speech signal, which is the convolution between glottal pulse and the vocal tract impulse response into a sum of two components known as the Cestrum. This computation is carried out by taking the inverse DFT of the logarithm of the magnitude spectrum of the speech frame.
The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information. The goal of automatic speaker reorganization is to analyze, extract, characterize, and recognize information about the speaker's identity. The performance of the speaker reorganization system depends on the technique employed in the various stages of the speaker reorganization system. The state of art of speaker reorganization system mainly used segmental analysis.

2.1 Methodology:

For the proposed study, we developed our Text corpus of consecutive sentences regarding different kinds of crops yielded in the Marathwada region in the Marathi language, which are specific to agriculture. The speech data was then collected from different speakers using the developed text corpus. The methodology followed by us for the proposed work is shown in figure 1.

Step 1:

We developed a text corpus using various blog articles over the internet. The selected words were checked for the typographic error. The words were distributed in various categories.

Step 2:

The 2 villages were selected from each tehsil of a district based on the crops taken in the district. There are 36 tehsils in four districts, such as Aurangabad, Jalna, Beed, and Osmanabad. The city was considered as two villages known as the old city and the new city.

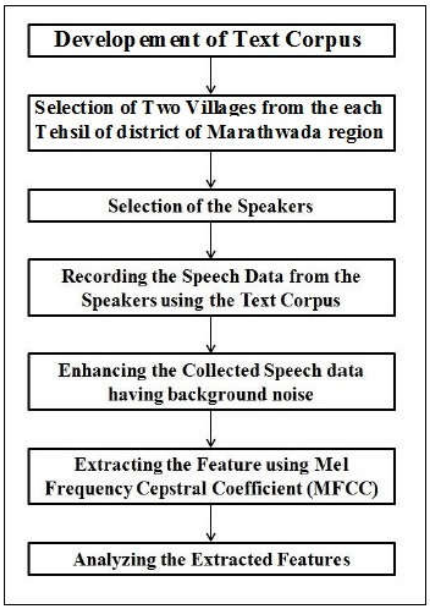


Figure 1. Methodology adopted for the proposed work

Step3:

20 Speakers were selected from each village, out of which 10 were male, and 10 were female.

Step 4:

The Speech samples were recorded from a total of 743 speakers from 72 villages in a normal environment.

Step 5:

After the collection of speech samples, the speech samples having noise were enhanced using spectral subtraction.

Step 6:

Features were extracted using Mel Frequency Cepstral Coefficient (MFCC).

Step 7:

The extracted features were analyzed.

2.2 Text Corpus

A text corpus is very crucial for language modeling, Continuous Speech Recognition (CSR), language synthesis, and speaker recognition. The text corpus should be developed in such a manner that using minimum sentences that will cover the maximum phonetic variations of a language for which the speech application will be developed. The phonetic aspects like phones, diphones, triphones, fricatives, etc. of the language should be covered so that when it is used, it will help to cover maximum sounds in limited words.
The development of text corpus is very tedious work as we need to check the typographic error and the grammar of the text that we have collected in the corpus. The text corpus should be grammatically correct and phonetically rich. The text that would be collected needs to be checked from a language expert so that if any error is there, it would be

corrected. There are few text corpora in the Marathi language developed by LDC-IL (Linguistic Data Consortium for Indian Languages). The developed text corpus contains a general-purpose sentence, text. The text corpus developed by LDC-IL was not of our use, so we decided to develop our word set/text corpus for the work and develop the speech database accordingly.

2.3 Selection of the Text for developing the text corpus

For developing a Speech database, the essential requirement is of grammatically correct, and phonetically rich text corpus would be recorded from various speakers. The text corpus for the proposed work was generated using various agriculture-related websites. The words were selected from the Blog articles published on the Internet. We select minimum sentences that will cover maximum variations [8], [9]. We select those crops that have major commodities in that area.

2.4 Text Corpus developed

The developed Text corpus consists of 15-20 sentences related to crop, which are specifically related to agriculture. We select those crops from that specific area having a major commodity.
The developed text corpus should be phonetically reached and grammatically correct. Below we show the example of developed text corpus, which is phonetically reached and grammatically correct.

Table 1. List of Taluka's in Aurangabad District, Name of villages and Major Commodities

Sr. No.	Taluka	Name of Villages		Major Commodities
1	Aurangabad	1.Malewada	2.Jaipur	Cotton, Maize, Gram, Wheat
2	Paithan	1.Bidkin	2.Paithan	Cotton, Maize, Gram, Wheat
3	Gangapur	1.Kaigaon	2.Gangapur	Jowar, Cotton, Wheat
4	Vaijapur	1. Lasur sta.	2.Vaijapur	Cotton, Maize, Gram, Wheat
5	Kannad	1.Kannad	2.Devegaon	Jowar, Wheat, Cotton
6	Khultabad	1.Mhaismal	2.Borgaon	Cotton, Maize
7	Phulambri	1.Phulambri	2.Pathri	Cotton, Maize, Gram, Wheat
8	Sillod	1.Sillod	2.Ajantha	Jowar, Wheat, Cotton
9	Soegaon	1.Soegaon	2.Nandgaon	Cotton, Maize

2.5 Standard for Speech Data Capturing

The standards set by LDC-IL for collecting the speech sample for the development of an isolated word recognition system in a studio environment/desktop microphone is as follows:

Sampling frequency:

The sampling frequency needs to be 16,000 Hz (As per standard, the sampling frequency should be multiple of 8 kHz), and it should 16 bit.

File format:

The file format used for saving the recorded speech file is .wav format. The sound should be recorded in mono, not in stereo.

2.6 Speech Data Collection Procedure:

Selection of the Speakers:
The speakers were selected to cover the maximum variation of the language of the district. One hundred speakers were selected from all over the district. The selected speakers were in the age group 18 above. The speakers were selected in the age groups whose native language is Marathi and also those speakers whose native language is not Marathi. Literacy was another criterion that was considered during the selection of a speaker.

Recording Procedure:

We used PRAAT software to record the speech. We used a Sennheiser PC360 and Sennheiser PC350 headset for recording the speech samples. The PC360 and PC350 headsets are having a noise cancellation facility, and the signal to noise ratio (SNR) is low. The steps followed for recording the speech samples was as follows:

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 recording was done at a sampling frequency of 16 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 a .wav file.

Step 5:

Step 4 was repeated for all app. 60 utterances that were recorded from the speaker. The procedure was repeated for all the 750 speakers.

3. Feature Extraction

The focus of the proposed study is the development of a Standard speech database and using that developed database for the development of an Automatic Speech Recognition System. For developing an Automatic Speech Recognition system, we need to extract the feature from the acquired/recorded speech and then apply the recognition algorithm. However, initially, we need to enhance the acquired/recorded speech signal sometime before we can extract the feature as it may contain noise. Different techniques are used for speech signal enhancement and speech feature extraction.

3.1 Feature Extraction:

Feature extraction is a fundamental pre-processing step to pattern recognition and machine learning problem. In feature extraction, the input data is transformed into a set of features that provides the relevant information for performing the desired task without the need for the full-size information using the reduced set. The speech recognition technique

is having a background from the DSP, i.e., Digital signal processing. DSP has been the center of progress in speech processing during the complete development of speech processing and speech recognition systems [10].
The objective to be achieved with feature extraction is to untangle the speech signal into the different acoustically identifiable components and to obtain the set of the feature with a low rate of change to keep the computation feasible. The feature extraction for speech recognition can be divided into the spectral analysis, parametric transformation, and statistical modeling.

3.2 Mel Frequency Cepstral Coefficient (MFCC):

The Mel Frequency Cepstral Coefficient is well known and most widely used feature extraction method in the speech area. The MFCC is similar to the human auditory perception system. For each tone with actual frequency f measured in Hz, a subjective pitch is calculated known as the 'Mel Scale.'
The FB in the implementations of defines the number of filters present in the Filter bank for the MFCC by the corresponding author. These implementations consider different sampling rates. To compute the features using MFCC, the steps that are followed are Pre-emphasizing, Framing and Windowing, Fast Fourier Transform, Mel-Frequency Filter Bank, Logarithm, and Discrete Cosine Transform. The block diagram of the MFCC feature extraction method is shown in figure 2.

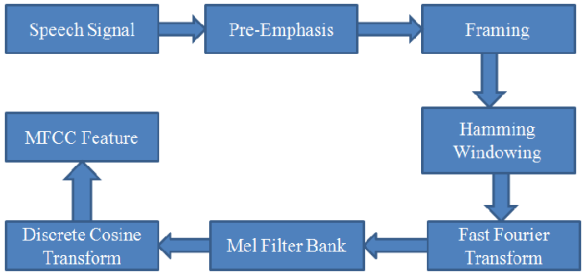


Figure 2. Block Diagram MFCC Feature Extraction Process

Pre-emphasizing:

The speech signal is first pre-emphasized with the pre-emphasis filter $1\text{-}az^{-1}$ to flatten the signal spectrally.

Framing and Windowing:

A speech signal is considered to remain stationary for approximately 20 ms. Dividing a discrete signal s[n] into frames in the time domain truncating the signal with a window function w[n]. This is done by multiplying the signal, consisting of N samples. The signal is generally segmented in the frame of 20 to 30 ms; then, the frame is shifted by 10 ms so that the overlapping between two adjacent frames is 50% to avoid the risk of losing the information from the speech signal. After dividing the signal into frames that contain nearly stationary signal blocks, the windowing function is applied. For the proposed work, the frame length was set to 25 ms, and the frame was shifted by 10 ms.

Fast Fourier Transform:

Fast Fourier Transform converts each frame N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples {xn}, as follows:

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

In general, X_k is complex numbers, and we only consider their absolute values (frequency magnitudes). The resulting sequence $\{X_k\}$ is interpreted as follow: positive frequencies $0 \leq f < Fs/2$, correspond to values $0 \leq n < N/2-1$, while negative frequencies $-Fs/2 < f < 0$ corresponds to $N/2+1 \leq n \leq N-1$. Here, Fs denote the sampling frequency. The result after this step is often referred to as a spectrum or periodogram. To obtain a good frequency resolution, a 512 point Fast Fourier Transform (FFT) is used.

Mel Frequency Filter Bank:

A filter bank is created by calculating several peaks, uniformly spaced in the Mel-scale, and then transforming back to normal frequency scale where they are used as peaks for the filter banks.

Feature Extraction Using MFCC:

For the proposed work, we used the Mel Frequency Cepstral Coefficient (i.e., MFCC) as the feature extraction technique. We have discussed the MFCC technique in the earlier section. The different values initialized during the computation of MFCC are as follows:

Table 2. Different parameters and Values for computation of MFCC

Parameter	Values
Sampling frequency	16000 Hz
Window type	Hamming Window
Window length:	25 millisecond
Step time (N)	10 millisecond
Number of coefficient	13 (1 Energy and 12 standard coefficient)
Min Frequency	0 (lowest band edge Mel filters (Hz))
Max Frequency	4000 (The highest band edge of Mel filters (Hz) set)
FFT:	512 point FFT

The features extracted for the developed Continuous speech database is shown below. The followed features are for two speaker's one male and one female. The features shown below are for the 1 utterances of the first sentence, both the 1st male and 1st female speaker. The table consists of the 13 features and the number of the frame to be shown is 12. However, the number of frames that are calculated varies according to the length of the speech signal. The Mean and standard deviation for the complete calculated MFCC was calculated. The Images show the plot of the 13 MFCC feature for 13 frames and the spectrogram having formats for the said utterance.

The detailed architecture of the graphical view of the continuous speech recognition system is explained two-state view. The two states, because of this system, show the performance of correct recognition as well as confusion matching. The architecture of the continuous speech recognition system, as shown in the following figure.

The figure is divided into 6 state views. The first view explained the login of the system means the system is started**.** The detail graphical views of every state point are explained below. The first part represents the login of a continuous speech recognition system. This part is independent. The second and six parts represent the selection of test samples and the performance of the system. These parts are dependent on each other because the performance of the system depends on the selection of the test sample.

The third and fourth part represents the extraction of MFCC features and training of features. Training and feature extraction processes are totally dependent on each other because we cannot train the system until we got the features. Recognition is the fifth and two-state procedures because it recognizes the system for correct as well as confusion matching.

Steps followed to recognize the sentences in continuous speech are shown below. The following GUI shows test sample, extracted MFCC features, training process, recognition of sentence, and total performance of the system. The steps are as follows:

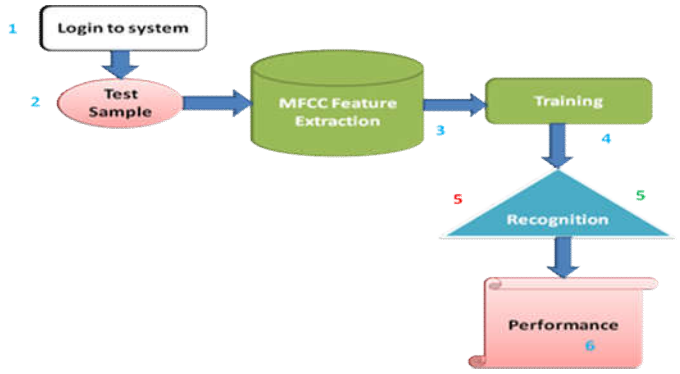
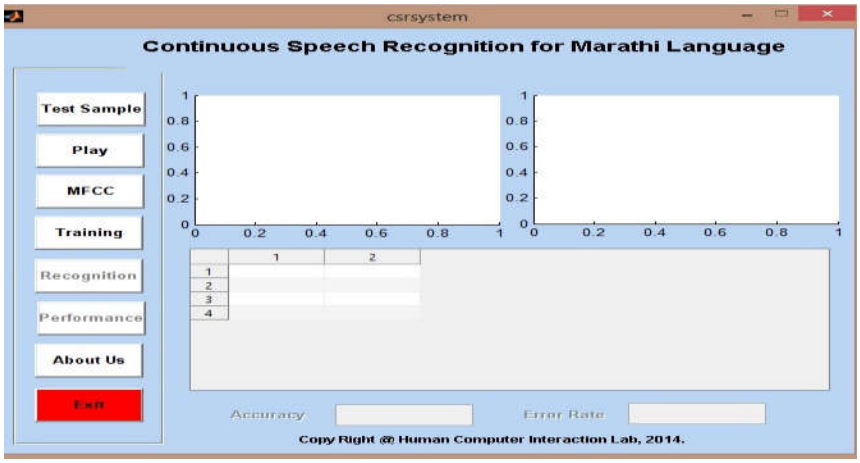
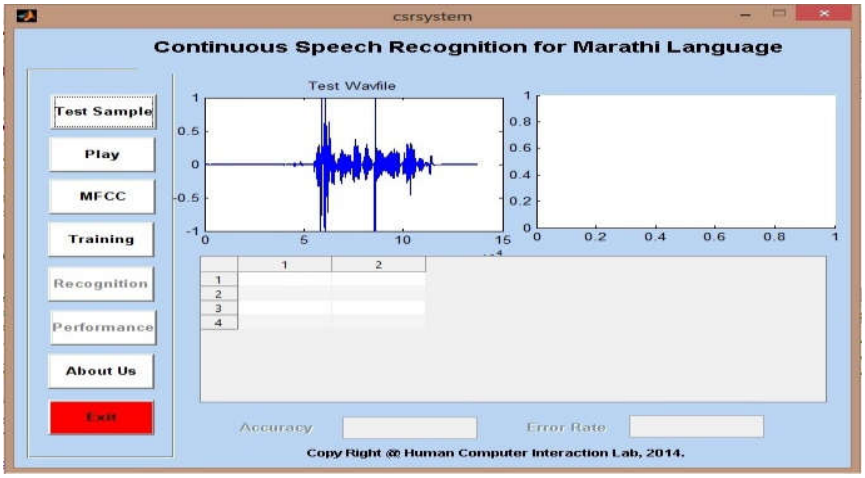


Figure 4. Architecture of Continuous Speech Recognition System

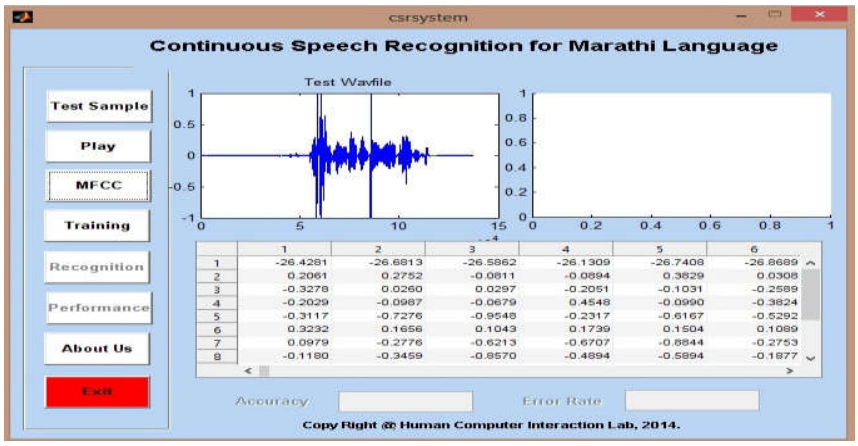
Step 1: Opening of the GUI:



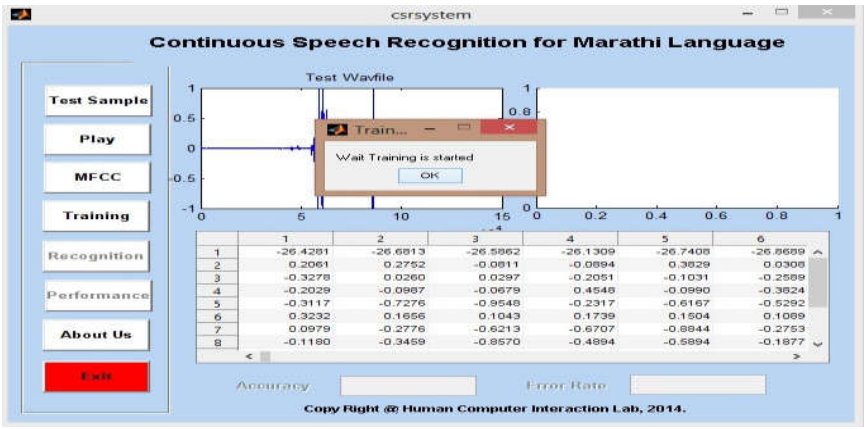
Step 2: Select Test Sample



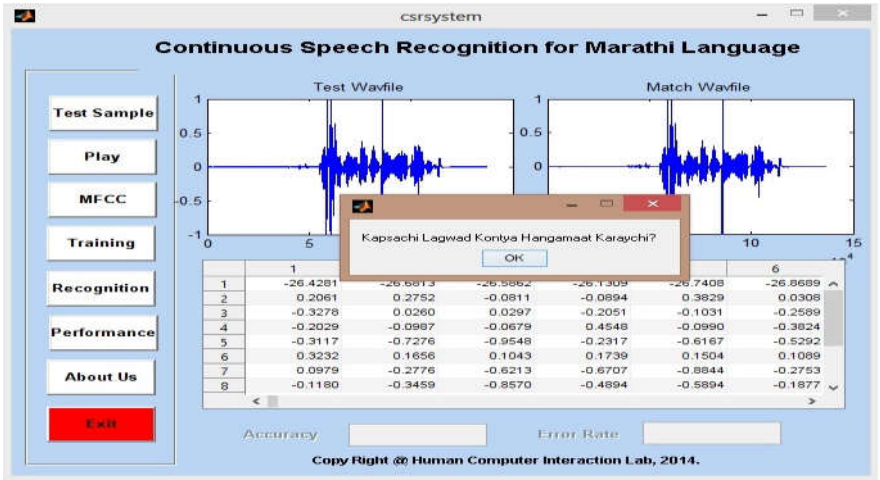
Step 3: Extraction of MFCC Features



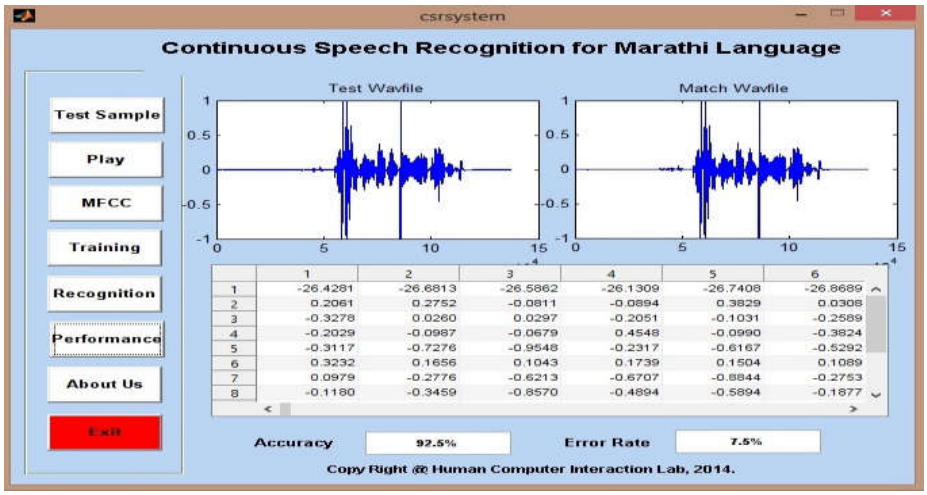
Step 4: Starting of Training Process:



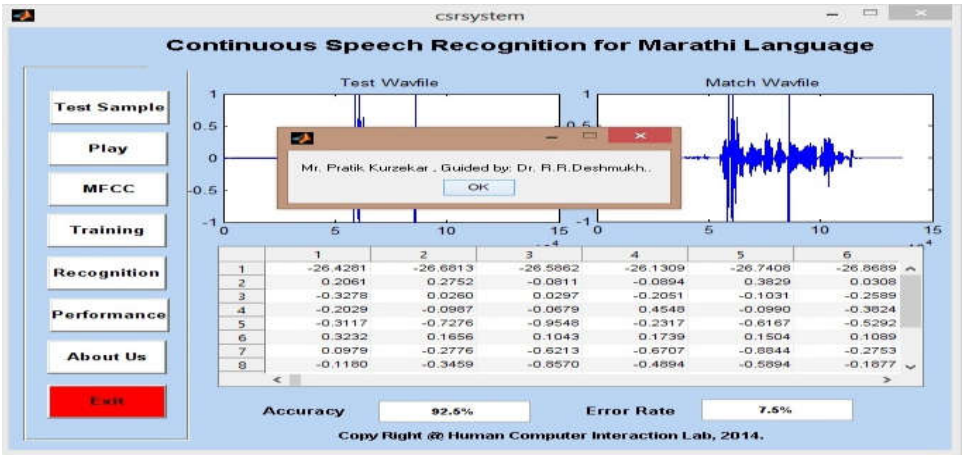
Step 5: Recognition of Test Sample:



Step 6: Performance of System:



Step 7: About Us:



3.3 Performance analysis

Euclidean distance formula is used for analysis of extracted features. The Euclidean distance formula is given by,

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Where,
d(p,q)= Distance between points p and q.
P and q are the two points in the Euclidean space.

3.4 Word Error Rate (WER):

Word Error Rate is given by following formulae
WER= $\frac{S+D+I}{N} * 100\%$
Where,
S=Number of Substitution
D=Number of deletion
I=Number of Insertion
N= Number of words in correct Sentence

3.5 System accuracy:

System accuracy is given by,
Accuracy= $(N - D - S - I) / N \times 100$
Where,
N=Number of words or sentences correctly recognized.
D = Number of unrecognized/missed words (Deletion errors)
S = Number of times a word was misrecognized as another word (Substitution errors).
I = Number of extra words inserted between correctly recognized words (Insertion errors).

Table 3. Table Showing Crops, Accuracy and Error Rate

Crop	Accuracy	Error Rate
Gram	94%	6%
Cotton	90%	10%
Wheat	93%	7%
Maize	93%	7%

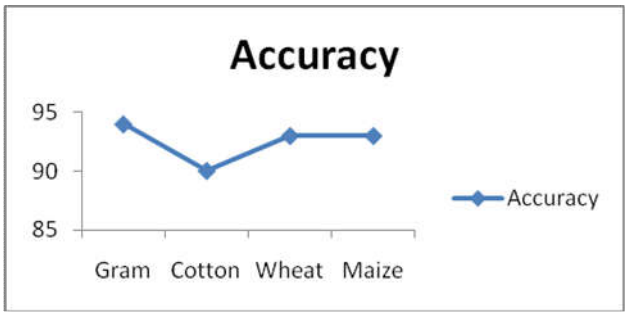


Figure 5. Crop wise Accuracy

Figure6. Crop wise Error rate

4.Conclusion

After doing the literature survey we developed the speech database of Continuous Speech for agriculture purposes in Marathi language as no such database is available till date.
After the completion of the literature survey for the feature extraction technique we selected Mel Frequency Cepstral Coefficient (MFCC) as it is very much closer to the human auditory perception system.
We have used the mean and standard deviation techniques for the accuracy at the speaker level. We observed that there is lots of difference between the uttered words.

REFERENCES

[1] *http://sistemic.udea.edu.co/wp-content/uploads/2013/introSR.pdf*

[2] *http://en.wikipedia.org/wiki/Speech_recognition*

[3] *Pukhraj Shrishrimal, Dr. R. R. Deshmukh, Vishal Waghmare "Indian Language Speech Database: A Review" International Journal of Computer Applications (0975 – 888) Volume 47– No.5, June 2012 pp.17-21*

[4] *Tejas Godambe, Samudravijaya K., "Speech Data Acquisition for Voice based Agricultural Information Retrieval", presented at the 39th All India DLA Conference, Punjabi University, Patiala, 14-16th June 2011.*

[5] *Santosh K. Gaikwad, Bharti Gawli, Pravin Yannawar, "A Review of Speech Recognition Technique", International Journal of Computer Applications (0975– 8887) Volume 10, No.3, November 2010.*

[6] *M. A. Anusuya, S. K. Katti, "Speech Recognition by Machine: A Review", International Journal of Computer Science and Information Security (IJCSIS), Vol. 6, No. 3, pp. 181-205, 2009.*

[7] *X. D. Huang, "A Study on Speaker - Adaptive Speech Recognition", Proc. DARPA Workshop on Speech and Natural Language, pp. 278-283, February 1991.*

[8] *www.agrowon.com cited on 10/03/2012*

[9] *http://lab.cgpl.iisc.ernet.in/Atlas/.../Cropnames_Indianlanguages.pdf cited on 11/03/2012*

[10] *Lawrence R. Rabiner and Ronald W. Schafer, "Digital Processing of Speech Signals, Signal Processing", Prentice-Hall, Englewood Cliffs, New Jersey, 1978.*