Speech to Text Conversion for Multilingual Languages

Yogita H. Ghadage, Sushama D. Shelke

***Abstract*—The current work presents a multilingual speech-to-text conversion system. Conversion is based on information in speech signal. Speech is the natural and most important form of communication for human being. Speech-To-Text (STT) system takes a human speech utterance as an input and requires a string of words as output. The objective of this system is to extract, characterize and recognize the information about speech. The proposed system is implemented using Mel-Frequency Cepstral Coefficient (MFCC) feature extraction technique and Minimum Distance Classifier, Support Vector Machine (SVM) methods for speech classification. Speech utterances are pre-recorded and stored in a database. Database mainly divided into two parts testing and training. Samples from training database are passed through training phase and features are extracted. Combining features for each sample forms feature vector which is stored as reference. Sample to be tested from testing part is given to system and its features are extracted. Similarity between these features and reference feature vector is computed and words having maximum similarity are given as output. The system is developed in MATLAB (R2010a) environment.**

***Index Terms*—Mel Frequency Cepstral Coefficients (MFCC), Minimum Distance Classifier, Speech Recognition, Speech-To-Text (STT), Support Vector Machine (SVM).**

I. Introduction

Speech Recognition is the procedure of extracting essential information from input speech signal to make accurate decision about the corresponding text. Speech signal conveys very rich information, such as speaker information, linguistic information which has inspired many researchers to develop the system that automatically process the speech e.g. speech enhancement, speech synthesis, speech compression, speaker recognition, speech recognition and verification. Speech recognition can be further classified as speaker dependent and speaker independent[1]. Computer follows human voice commands with the help of speech recognition mechanism and understand human languages i.e. it acts as good interface for human computer interaction. Generally today's speech recognition technologies are designed for English language. So that illiterate rural communities or educationally under-privileged people are being kept away of computer technology. If the processing of computer technology in native language is made possible i.e. if computer technologies can understand the native language then it will be easy to use computer technologies for illiterate people, people from rural communities or educationally under-privileged. Marathi is a native language of Maharashtra. In a day to day life while speaking we use English words, i.e. most of the time we mix English with native language. So author has designed Multilingual Speech-To-Text conversion system. In which Marathi, English, Marathi-English mix speech has given focus.

The objective of the proposed system is to design and implement Speech-To-Text conversion system for Marathi, English, Marathi-English mix languages. The system has developed for small database which contains 10 Marathi sentences, 3 English, 2 mix sentences. This work is based on MFCC, SVM & Minimum Distance Classifier[2].

The outline of the paper is as follows. Section II gives a brief overview of the system. Section III describes about the speech database. Section IV explains the MFCC feature extraction. Section V says about the pattern classification. Section VI describes about the experimental setup. Section VII discuss about the result. Section VIII,] concludes the paper. Section IX says about the future work.

II. System Overview

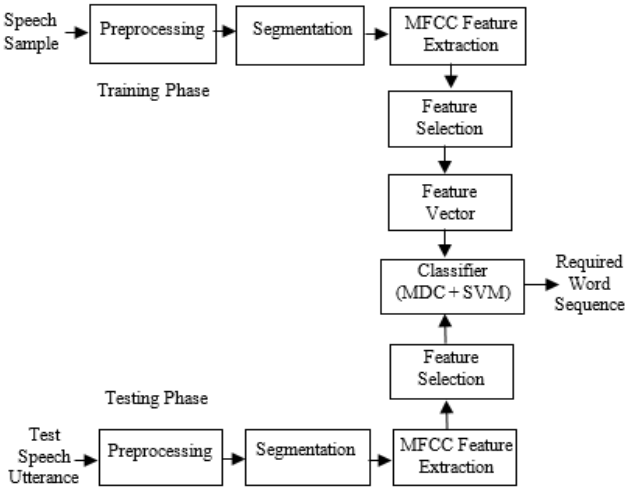

Fig. 1. Block diagram of system

Yogita H. Ghadage is with the Electronics and Telecommunication Engineering Department, NBN Sinhgad School of Engineering, Pune. (e-mail: ghadagehyogita@gmail.com).
Sushama D. Shelke is with the Electronics and Telecommunication Engineering Department, NBN Sinhgad School of Engineering, Pune. (e-mail: sushama.shelke@sinhgad.edu).

978-1-5090-0395-2/16/$31.00 ©2016 IEEE

Fig. 1 shows a block diagram of Speech-To-Text conversion system. The system operation is divided into two phases i.e. training and testing. First in training phase speech utterances of each sentences is recorded. Speech signal is preprocessed and segmented into words. For each word acoustic features are extracted using MFCC method. Such features for each word forming feature vector is stored for reference. In testing phase the speech utterance to be tested is preprocessed, segmented into words and features are extracted for each words. These features are compared with the reference feature vector stored during training phase. This is done by using combination of SVM and Minimum Distance Classifier. The word having minimum difference is given as recognized word.

III. SPEECH DATABASE

Database is the crucial point in the automatic Speech-To-Text conversion system. For any automatic speech recognition system first step is to configure the database. The proposed system is implemented on self-generated database[3]. The whole database is divided into two parts. One is training database and another is testing database.

Marathi language sentences: 10
English language sentences: 3
Marathi-English mix sentences: 2
Total sentences: 15
Speakers: 4 (2-Male, 2-Female)

A. Training Database

The training database contains recorded speech utterances by 4 different users for 10 Marathi sentences, 3 English sentences and 2 mix sentences. Here each sentence is uttered 10 times by each user, i.e. 40 utterances of each sentence are used to train the system i.e. total 600 samples are used to train the system. Sentences used in the formation of database are mentioned below in table1, 2 and 3.

B. Testing Database

The testing database also contains recorded speech utterances by 4 different users for 10 Marathi, 3 English and 2 English-Marathi mix sentences. Here each sentence is uttered 20 times by each user i.e. total 1200 samples are used to test the system. Each sample of training and testing databases is recorded with sampling frequency of 8 kHz.

TABLE I
MARATHI DATABASE

Sr. No.	Test sample	Marathi Sentence	Corresponding English transcription
1	Test 1	तू कुठे आहेस	Tu kuthe aahes
2	Test 2	मला माफ कर	Mala maaf kar
3	Test 3	आज डॉक्टर येणार आहेत का	Aaj doctor yenar aahet ka
4	Test 4	मी व्यस्त आहे	Mee vyasta aahe
5	Test 5	मला उशीर होइल	Mala ushir hoil
6	Test 6	नंतर फोन करते	Nantat fon karate
7	Test 7	थोडया वेळाने फोन करते	Thodya velane fon karate
8	Test 8	मी तुझी वाट बघत आहे	Mee tuzi vat baghat aahe
9	Test 9	तू कधी येणार आहेस	Tu kadhi yenar aahes
10	Test 10	शभु प्रभात	Shubha prabhat

TABLE II
ENGLISH DATABASE

Sr. No.	Test sample	English Sentence
1	Test 11	India is my country
2	Test 12	Record the speech
3	Test 13	Can you wait

TABLE III
MARATHI-ENGLISH MIX DATABASE

Sr. No.	Test sample	English-Marathi mix Sentence
1	Test 14	Hi kasha aahes
2	Test 15	Sorry usher zala

IV. MFCC FEATURE EXTRACTION

In any automatic speech recognition system first and the most important step is to extract features. i.e. To identify useful components of speech signal that are used to identify the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc[4].

Two main purposes of feature extraction are: first is to compress the input speech signal into features, and second is to use these features which are insensitive to speech variations, changes of environmental conditions and independent of speaker.

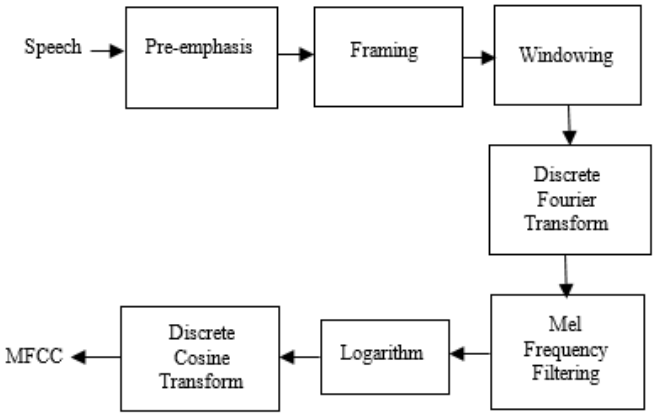

Fig. 2. Block diagram of MFCC

Steps of MFCC feature extraction are as follows:

A. *Pre-emphasis*

Pre-emphasis is applied to spectrally flatten the input speech signal. First order high pass FIR filter is used to pre-emphasize the higher frequency components.

B. *Framing*

An audio signal is constantly changing, so to simplify things we assume that on short time scales the audio signal doesn't change much. So we frame the signal into 20-40 ms frames. Hamming window is applied on each frame and it rid of some information at start and at end of frame so to reincorporate this information back into our extracted features overlapping is applied on frames[5].

C. *Windowing*

To avoid or reduce the unwanted discontinuities in speech segment and distortion in spectrum introduced by framing process windowing is performed. Mostly Hamming window is used in speech recognition.

D. *Discrete Fourier Transform (DFT)*

Spectral estimation can be done by DFT. FFT is very efficient algorithm to implement DFT. The magnitude frequency response of each frame is obtained after FFT execution. i.e. Spectral coefficients of the speech frames are complex numbers containing both magnitude and phase information. The phase information is usually discarded for speech recognition and only the magnitude of the spectral coefficients is retained.

E. *Mel Frequency Filtering*

Normally each tone with an actual frequency 'f' is measured in Hz. For speech signal, the ability of human ear to understand frequency contents of sounds does not follow a linear scale. So that for every tone a subjective pitch is measured on a scale called the 'Mel' scale. Below 1000 Hz, Mel frequency scale is a linear frequency spacing and above 1000 Hz it is a logarithmic spacing.

The following formula gives the transformation of a given linear frequency 'f' Hz into corresponding 'Mel' frequency.

$$Mel(f) = 2595 * \log 10\left(1 + \frac{f}{700}\right) \quad (1)$$

As non-linear characteristics of the human auditory system in frequency are approximated by Mel filtering in the same way natural logarithm is used to approximate the loudness non-linearity i.e. the relationship between the human's perception of loudness and the sound intensity is approximated by natural logarithm. Multiplication in frequency domain, become simple addition after the logarithm.

The log Mel filter bank coefficients are computed from the filter output as:

$$s(m) = 20 \log 10 \sum_{k=0}^{N-1} \left| X(k) H(k), 0 < \mu < M \right| \quad (2)$$

Where, M is the no. of Mel filters (20 to 40),
X (k) is N-point FFT of specific window frame of the input speech signal, H (k) is the Mel filter transfer function.

F. *Discrete Cosine Transform (DCT)*

For transforming the Mel coefficients back to time domain discrete cosine transform is performed. The result of this step is called the Mel Frequency Cepstral Coefficients (MFCC).

The inverse Fourier transform of the log magnitude of Fourier transform of the signal is called as Cepstrum. As coefficients of the log Mel filter bank are real and symmetric, we can replace the inverse Fourier transform operation by DCT to generate the Cepstral coefficients.

The smooth spectral shape or vocal tract shape is represented by lower order Cepstral coefficients. While the excitation information is represented by higher coefficients.

The Cepstral coefficients are the DCT of the M filter outputs obtained as:

$$\sum_{m=0}^{M-1} s(m) = \cos\left[\frac{\pi m(m - 1/2)}{M}\right] \quad (3)$$

Typically first 13 Cepstral coefficients are used. Generally MFCC coefficients are less correlated than the log Mel filter bank coefficients this is the biggest advantage.

V. PATTERN CLASSIFICATION

A. *Minimum Distance Classifier (MDC)*

In speech recognition or STT conversion there are mainly two phases first is training phase and second one is testing phase. For classification, during training phase zero crossing points (ZCP) corresponding to the different words are pre computed and stored as reference ZCPs[6].

Minimum distance classifier computes Euclidean distance between the zero crossing points of the uttered word and zero crossing points of words from database. The word having least Euclidean distance is declared as uttered word.

Euclidean distance is given as:

$$d^2(\bar{x}, \bar{p}) = \sum_{1}^{k} (x_i - p_i)^2 \quad (4)$$

Where, $\bar{x}$ and $\bar{p}$ ZCP database.

i.e. $\bar{x}$ is a ZCP vector of uttered word.

$\bar{p}$ is a ZCP vector of different words.

i varies from 1 to k (i.e. no. of ZCPs of a particular word).

The sum of squares of the difference between the individual zero crossing points is computed to calculate the Euclidean distance i.e. distance between the uttered word and all words in the database is found out. The word in the database with least distance is declared as the uttered word.

B. Support Vector Machine (SVM)

SVM is one of the effective method of pattern classification. SVM use linear and nonlinear separating hyper-planes for data classification. First input is mapped into a high dimensional space and then with the help of hyper-plane it distinguishes the classes.

The inner product, kernel which is caused by the high dimensional mapping is a crucial aspect of opting SVMs successfully i.e. a high dimensional feature space is implicitly introduced by a computationally efficient kernel mapping and in a high dimensional feature space SVM finds a separating surface with a large margin between training samples of two classes . And large margin implies a better generalization ability. SVM uses discriminative approach. The classification of any fixed length data vectors is possible by SVM. It cannot be readily applied to task involving variable length data classification.

The support vector classifier uses the function:

$$f(x) = (\lfloor \alpha * K_s(x) \rfloor) + b \quad (5)$$

Where, $K_s(x) = [k(x, s_1), \ldots . k(x, s_d)]^T$ is a vector of evaluation of kernel functions centered at the support vectors.

$f(x) = (\lfloor \alpha * K_s(x) \rfloor) + b$ Which are usually subset of the training data.

The classification rule is defined as:

$$q(x) = \{1 \, for f(x) \geq 0$$
$$\{2 \, for f(x) < 0 \quad (6)$$

And multiclass classification function and rule is defined as:

$$f_y(x) = (\alpha_y * k_S(x)) + b_y, y \in Y \quad (7)$$

$$Q(x) = \arg\max f_y(x), y \in Y \quad (8)$$

C. SVM-MDC Combination

The proposed system uses the combination of SVM and MDC for classification. It translates large class problems into small class problems i.e. multiclass problems are converted into binary problems. So solving these problems becomes easy. MDC is mainly used for coarse tuning and SVM performs fine tuning.

VI. EXPERIMENTAL SETUP

The system is trained with the training database and the recorded speech utterances stored in test database are used to test the system. All utterances are recorded at 8 kHz of sampling frequency. Duration for sentences is from 3sec-5sec.

The input speech signal is given to the MFCC which converts it into feature vectors. Minimum distance classifier and support vector machine techniques are used for classification purpose[7-9].

The trained speech samples are saved as reference models into database. After that each segmented speech sample of test speech signal is passed over reference models and minimum distance is computed. Each word recognition is done by using minimum distance and SVM model. The whole system is implemented and tested in MATLAB software.

VII. RESULTS

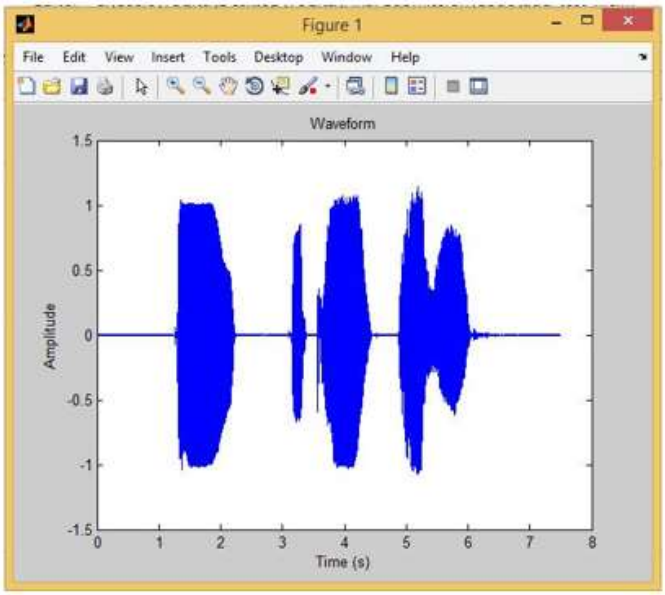

Fig. 3. Speech Waveform of 'Tu kuthe aahes'

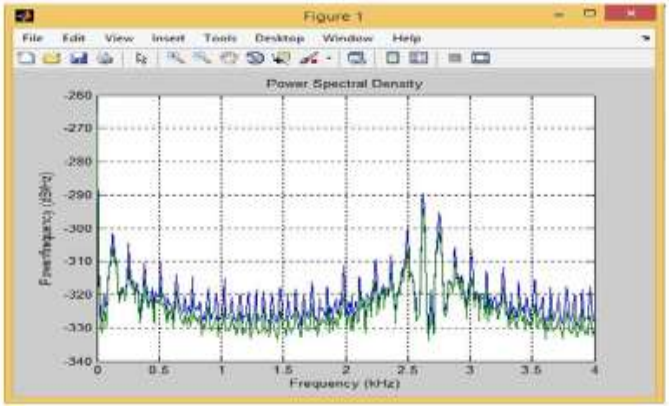

Fig. 4. Power Spectral Density of signal 'Tu kuthe aahes'

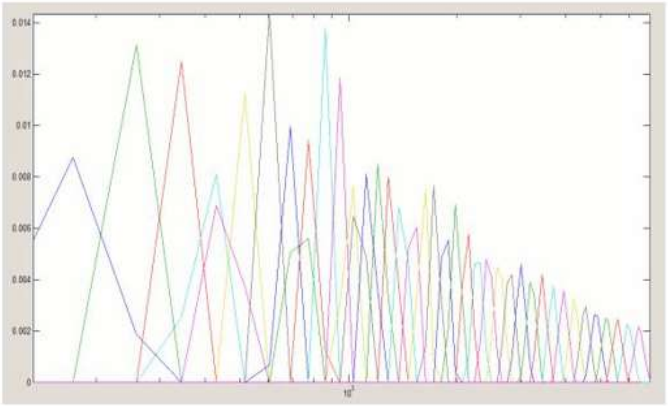

Fig. 5. MFCC Filter weights

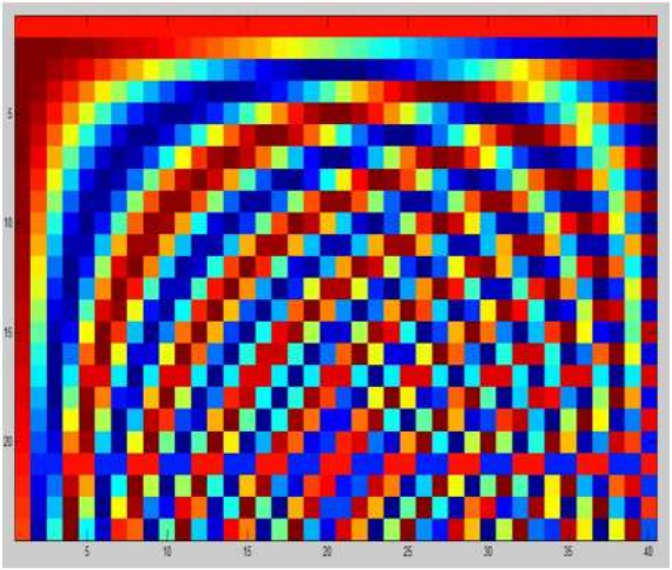

Fig. 6. MFCC Discrete Cosine Transform Matrix

TABLE IV
MARATHI-ENGLISH MIX DATABASE

Sr. No.	Language	% Accuracy
1	Marathi	93.625%
2	English	91.6667%
3	Marathi- English mix	90.625%

VIII. CONCLUSION

The % accuracy of the proposed system for Marathi language of 93.625% is achieved using MFCC for feature extraction, Minimum Distance classifier and SVM combination for classification. The proposed system achieved the higher accuracy compared to the using MFCC-feature extraction technique & CDHMM-classifier, which gives accuracy of 88.80% for Marathi language. The proposed system achieved 91.6667% accuracy for English and 90.625% accuracy for English-Marathi mix languages.

IX. FUTURE WORK

This connected word speech recognition system is developed for speaker independent Marathi, English, English-Marathi mix languages. This work may extend for other regional languages and may extend towards the real time connected word speech recognition for multilingual speech.

REFERENCES

[1] Priyanka P. Patil, Sanjay A. Pardeshi, "Marathi Connected Word Speech Recognition System," *IEEE First International Conference on Networks & Soft Computing*, pp 314-318, Aug. 2014.

[2] M.A.Anusuya, S.K.Katti, "Speech Recognition by Machine: A Review," *International Journal of Computer Science and Information Security*, vol.6, no.3, 2009.

[3] Mathias De Wachter, Mike Matton, Kris Demuynck, Patrick Wambacq, "Template Based Continuous Speech Recognition," *IEEE Transs. On Audio, Speech & Language Processing*, vol.15, issue 4,pp 1377-1390, May 2007.

[4] Vikram.C.M., K.Umarani, "Phoneme Independent Pathalogical Voice Detection Using Wavelet Based MFCCs, GMM-SVM Hybrid Classifier," *IEEE International Conference on Advances in Computing, Communications and Informatics*, pp 929-934, Aug. 2013.

[5] V.Naresh, B.Venkataramani, Abhishek Karan and J.Manikandan, "PSOC based isolated speech recognition system," *IEEE International Conference on Communication and Signal Processing*, pp 693- 697, April 3-5, 2013, India.

[6] Taabish Gulzar, Anand Singh, Dinesh Kumar Rajoriya and Najma Farooq, "A Systematic Analysis of Automatic Speech Recognition: An Overview," *International Journal of Current Engineering and Technology*, vol.4, no.3, June 2014.

[7] Santosh V. Chapaneri, "Spoken Digits Recognition using Weighted MFCC and Improved Features for Dynamic Time Warping," *International Journal of Computer Applications*, vol.40, no.3, Feb. 2012.

[8] Rashmi C. R., "Review of Algorithms and Applications in Speech Recognition System," *International Journal of Computer Science and Information Technologies*, vol. 5(4), pp 5258-5262, 2014.

[9] Shivanker Dev Dhingra, Geeta Nijhawan, Poonam Pandit, "Isolated Speech Recognition Using MFCC and DTW," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol.2, Issue 8, Aug. 2013.