ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 3, Issue 12, December 2016

Isolated Marathi Word Recognition Systems

[1] Avinash Bhute [2] B.B. Mesharm

[1] [2] Verrmata Jijabai Technological Institute, Matunga, Mumbai, India

Abstract: -- **Despite the fact that Marathi language is most common languages worldwide, there has been only a little research on Marathi speech recognition as compared to other languages such as English and Chinese. The speech recognition algorithms are very useful for designing efficient and accurate speech recognition systems. In this paper, we have proposed the automatic speech recognition systems by using MFCC features compare by DTW template matching algorithm. Our proposed systems divided into three phases, in first phase, the input signal is preprocessed and reduce the noise. In second phase the features are extracted from the speech signal using MFCC algorithm. For improving the recognition accuracy, we have added the delta coefficient. And last, the features of each word in test database are compared to the training database using dynamic time warping (DTW) algorithm. The proposed systems achieve the recognition rate of about 96.5%, which is outperforming than other approaches.**

Keywords:— **Marathi Speech Recognition Systems, MFCC, DTW, Isolated Words**

I. INTRODUCTION

The ability to recognize human speech has always been an area of interest of people, because of the large range of applications in almost every segment of society. The development of science and technology made visible improvements in the capabilities and the quality of recognition of the human speech using a kind of terminal devices. There are approximately 7106 languages currently spoken around the world, the majority of which have only a small number of speakers. About 4 billion of the earth's 6.5 billion people, or over 60% of the earth's population, speak one of the 30 languages as their native tongue [1]. Despite the fact that Marathi language is fifteenth (15th) most common languages worldwide, Forth (4th) in India, about 70 million in 2001 and 90 million people speak Marathi as a first language [2]. There has been only a little research on Marathi speech recognition as compared to other languages such as English and Chinese.

However, the automatic speech recognition (ASR) has received a great deal of attention by many researchers for a decade, which essentially permits a computer to acknowledge spoken words recorded by its microphone. Speech recognition is employed in different application domains ,includes interfacing with deaf people, home automation, healthcare, robotics, agriculture, and much more. Actually, various approaches were adopted for speech recognition, which are mainly found in three categories, Template-based such as dynamic time warping (DTW), statistics-based such as hidden Markov models (HMMs) and neural network-based such as artificial neural networks (ANNs).

This paper has proposed an efficient DTW-based speech recognition system for isolated words of Marathi language. The outline of our work is as follows: The input speech data have been undergoing a preprocessing phase, where not only the noise reduction but also normalization has been performed. The speech/non speech regions of input signals are detected and segmented into manageable segment for facilitating the other tasks. Then the Mel-frequency cepstral coefficient (MFCC) Feature are extracted. The delta coefficient is added to MFCC to improve the recognition results. Finally, Dynamic time warping algorithm a pattern matching algorithm is used for detecting the similar pattern.

The rest of the paper is organized is as follows: The Section II has been discussed our proposed approach towards Marathi speech recognition systems, section III discussed the experimental result and analysis followed by Conclusion in Section IV.

II PROPOSED MSRS SYSTEMS

The proposed Marathi speech recognition system (MSRS) consists of stages like database repository, preprocessing, segmentation, feature extraction and finally pattern matching. The Collection of utterances, sentences in the proper manner is stored in

DB repository which leads accuracy of the speech recognition. Preprocessing enhances the signal characteristics which leads to improve the recognition result. The segmentation used to detect the acoustic changes in speech. The Feature Extraction is retaining useful information of the speech signal while discarding redundant and unwanted information i.e. noise. The pattern matching is used for level of similarity between two time series speech signal. Based on the similarity between the test word, and the trained word of different utterances results is obtained.

2.1 Database Repository

We maintained the repository of utterances which are required for training and testing of the dataset. The generation of a corpus of Marathi isolated words sentences and a collection of speech data is as below: The speakers are selected from 22-25 in age group, who can speak Marathi language fluently. We have built the repository of 100 utterances which is Marathi isolated words and digits, recorded from 35 different speakers (20 male and 15 female) each word trice. These utterances are recorded in a normal environment, the sampling frequency of 16KHz without noise and echo. The database generation process is defined in following algorithm.

```
Algorithm:
Creating the Database
Input: Speech Signals
Output: DB-Database Repository
Pseudo code:
BEGIN
Generate the DB;
Do
Read SpeechFile;
Segment the Speech Signal;
Extract the Features;
Save into DB;
While(SpeechFile<=no.of speechFiles in dir)
END;
```

Due to the speech has an overall spectral tilt of 5 to 12 dB per octave, a pre-emphasis filter of the form 1-0.99 z-1 is normally used. This first order filter will compensate for the fact that the lower the spectral peaks of the sound spectrum contain more energy than the higher. If it weren't for this filter the lower the spectral peaks would be preferentially modeled with respect to the higher the spectral peaks.

2.2 Preprocessing

Preprocessing is the first step of speech signal processing, which involves the conversion of analog speech signal into a digital form [9].The speech i.e the continuous time signal is sampled at discrete time points to form a sample data signal representing the continuous time signal. Then samples are quantized to produce a digital signal. The problem of distinguishing speech signals from other non-speech signals is becoming increasingly important as ASR systems are being applied to an increasing number of real-world multimedia applications.

For achieving the more accurate results, pre-processing is required to enhance some signal characteristics through removing the noise which probably affects the quality of the recorded speech. To achieve this, firstly pre-emphasis has been done and followed by normalization of signal.

In pre-emphasis the high frequency contents of the input signal are emphasized in order to flatten the signal's spectrum. Because speech has an overall spectral tilt of 5 to 12 dB per octave, a pre-emphasis filter of the form 1-0.99 z-1 is normally used. In this paper, the pre-emphasizer is defined as

$$H(z) = 1 - a * z^{-1} \qquad (1)$$

$$H(z) = 1 - 0.95z^{-1} \qquad (2)$$

where, z refers to the Fourier transform of the speech signal. The high amplitude pulses in the signal are reduce due to the filter which improves the accuracy of stored word features. As shown in fig1.

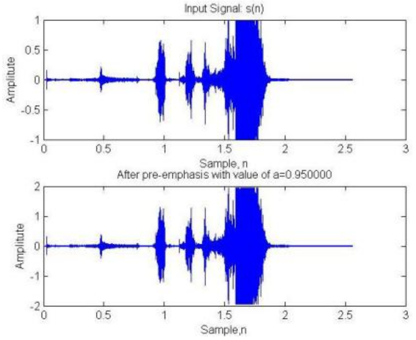


Fig1. The word "एकोणवीस" after pre-emphasis

The sensitivity of speech is differ due to loudness in speech and due to use of different microphone which requires to normalize the speech signal [4-6].

$$s_1(n) = \frac{x_{pre-emp}(n) - Mean(x_{pre-emp}(n))}{Max\left(\left|x(n) - Mean\left(x_{pre-emp}(n)\right)\right|\right)} \qquad (3)$$

Normalized the signal is shown in Fig. 2.

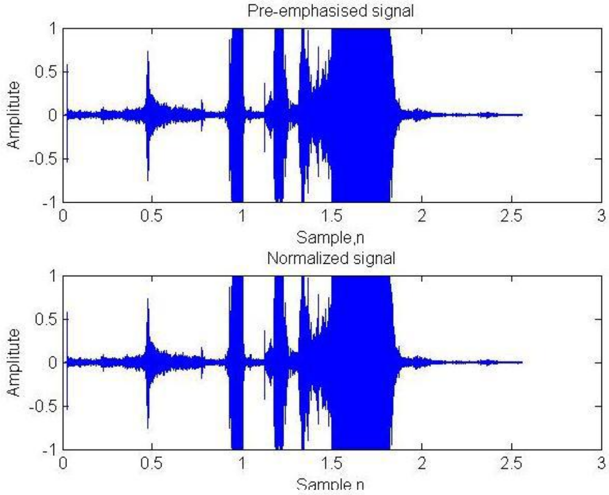


Fig 2. The word "एकोणवीस" after Normalization

2.2 End Point Detection

The major problem of speech recognition system is to detect the word start and boundary points. To specify the speech and non speech regions the short-term Energy and zero crossing rate is commonly used. The signals are segmented into non-overlap frames has a width of 20ms. The algorithm utilizes energy to acquire the reference points. The required characteristics of an ideal word boundary detector are: reliability, robustness, accuracy, adaptation, simplicity, real-time processing and no a priori knowledge of the noise. All these issues are solved by this algorithm [10].

Short-Term Energy and Zero crossing rate

The amplitude of nonspeech segments is noticeably lower than that of the speech segments. The short-time energy of speech signals reflects the amplitude variation. In a typical speech signal we can see that its certain properties considerably change with time The short-term energy is significantly increased in speech regions. However, it can be calculated according to [6] is as follows:

$$E_{S_1}(m) = \frac{1}{L}\sum_{n=m-L+1}^{m} {S_1}^2(n) \qquad (4)$$

where m - frame number, L- frame length, and n - frame index. The zero crossing rates tend to have larger values in non-speech regions. This gives a good indication of speech existence. It can be calculated according to [5]:

$$Z_{S_1}(m) = \frac{1}{L}\sum_{n=m-L+1}^{m} \frac{|\mathrm{sgn}(S_1(n)) - \mathrm{sgn}(S_1(n-1))|}{2} \qquad (5)$$

$$\mathrm{sgn}(S_1(n)) = \begin{cases} +1: S_1(n) \geq 0 \\ -1: S_1(n) < 0 \end{cases} \qquad (6)$$

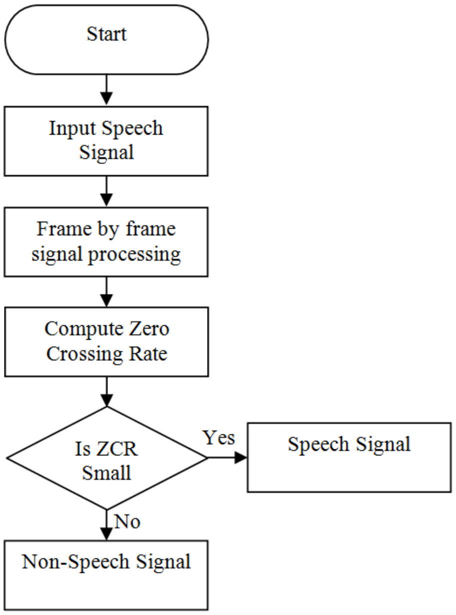
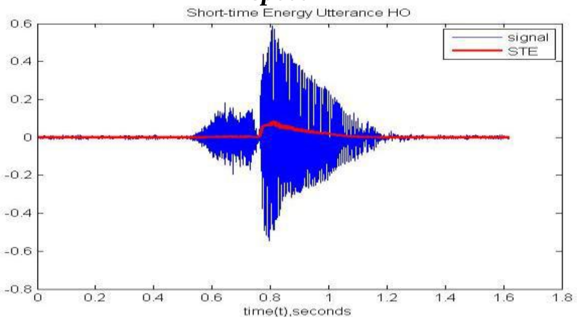
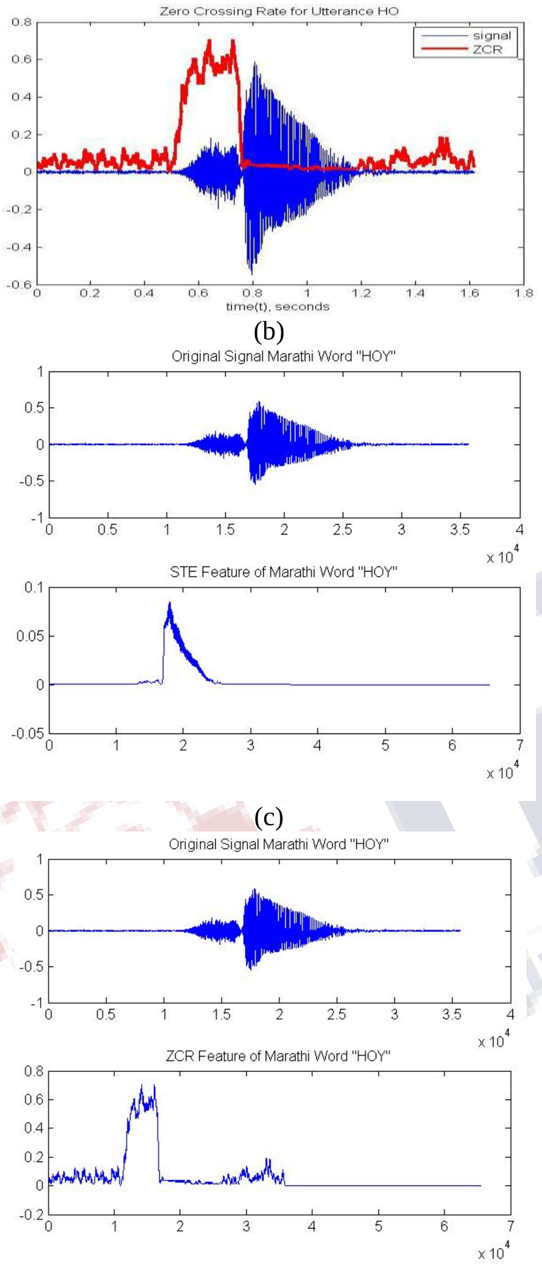


Fig 3. Classification of input signal speech and non-speech



(a)



(b)

(c)

(d)

Fig 4. Detection of Speech Non-speech signal using Short time Energy and Zero Crossing Rate (a) and (c) Short time Energy for utterance HOY w.r.t. input signal and (b) and (d) Zero Crossing Rate for utterance HOY w.r.t. input signal.

2.4 Feature Extraction

Feature extraction involves transforming the signal into a form appropriate for the models used for further processing. The goal is to find a set of properties of an utterance that have acoustic correlates in the speech signal, that is, parameters that can somehow be computed or estimated through processing of the signal waveform,. Such parameters are termed features [10]. For Isolate word recognition, we have extracted acoustic features Mel-scale frequency cepstral coefficient (MFCC) along with ZCR and STE. Step by step procedure for feature extraction is as follows:

1. Frame Blocking

In our experiments, the input speech signal (x1(n)) is segmented into J frames of S samples for each one with an overlapping ratio of 31.25%, so that adjacent frames are separated by S1 samples (where S<S1). The chosen values for S and S1 are 320 samples and 100 samples, respectively which were so appropriate. Hence, the output signal contains J vectors of length S, which corresponds to x1(s; j), where s=0,1,2,…,S- 1 and j=0,1,2,…,J-1. If the sample rate is 16 kHz and the frame size is 320 sample points, then the frame duration is 320/16000 = 0.02 sec = 20 ms. Additional, if the overlap is 160 points, then the frame rate is 16000/(320-160) = 100 frames per second [11].

2. Hamming Windowing

Each frame has to be multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame Applying hamming window, to the output signal i.e. framed signal, helps in reducing discontinuity at both ends of each frame. If the signal in a frame is denoted by s(n), n = 0,…N-1, then the signal after Hamming windowing is s(n)*w(n), where w(n) is the Hamming window defined by:

$$Ham(n, a) = (1 - a) - a\cos(2pn/(N-1)),\quad 0 \leqq n \leqq N-1$$

$$\mathrm{Ham}(n) = 0.68 - 0.32\cos\frac{2\pi n}{N-1},\quad 0 \leq n \leq N-1 \quad (6)$$

where i is sample index and N - the length of a frame (in samples). By applying Ham (p) to x1(p; j) for all frames, then x2 (p; j), which refers to the windowed signal. The generalized curve of hamming window is as shown in figure.

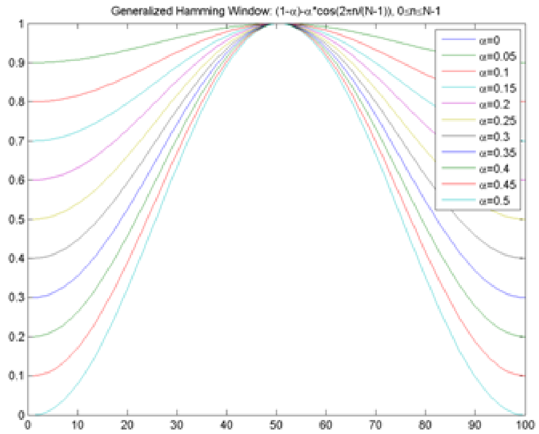


Fig 5. Generalized Hamming Window

3. Fast Fourier Transform

The characteristic of speech signal, i.e. Spectral analysis shows that different timbres in speech signal correspond to the different energy distribution over frequencies. We have performed an FFT to obtain the magnitude, frequency response of each frame. When we perform FFT on a frame, to convert windowed signal time domain to frequency domain. The frame length here is a power of 2 (N=2i), hence the output signal is 2 X (n; j).

4. Mel Filter Bank

According to the fact that human perception of speech frequencies is nonlinear, a Mel-scale is used so that for each tone with a frequency F measured in Hz, a subjective pitch is measured on a Mel-scale according to following formula [3-6]:

$$F_{mel} = 1125\, log_{10}\left(1 + \frac{F_{Hz}}{700}\right) \qquad (7)$$

Once finding the magnitude of X2 (n; j) and using the Mel scale filter bank (which consists of 20 triangular-band-pass filters which have an equal spacing before 1 kHz and logarithmic scale after 1 kHz), the Mel spectrum coefficients are found as the summation of the filtered results as the following:

$$Mel_v = \sum_{n=0}^{N-1} |X_2(n,j)|\ TF_v^{mel}(n) \qquad (8)$$

Where TF_v ^{mel} is the nth triangular filter.

5. Discrete cosine transform

In this step, we apply DCT on the 20 log energy Ek obtained from the previous step to have L mel-scale cepstral coefficients. The formula for DCT is as follows:

$$C_m = \sum_{k=1}^{N} S * cos\left(\frac{m(k-0.5)*p}{N}\right) * E_k \qquad (9)$$

m =1,2,3……L

where N is the number of triangular bandpass filters, L is the number of mel-scale cepstral coefficients. we set N=20 and L=12. Since we have performed FFT, DCT transforms the frequency domain into a time-like domain called quefrency domain. The obtained features are similar to cepstrum, thus it is referred to as the mel-scale cepstral coefficients, or MFCC.

6. Short-term energy

The energy within a frame is an important feature that can be easily obtained.The cepstral coefficients do not capture the energy. To increase the coefficients derived from Mel-cepstrum, the log of the signal energy feature is used. Alternatively, for every frame, the following energy term is added.

$$E_j = log \sum_{p=0}^{P-1} x_2^2(p,j) \qquad (10)$$

7. Delta Cepstrum

The speech signal is not constant, i.e. the slope of formants may change from stop burst to release [12]. Hence, it is worth adding these slopes in the features These are called delta features. It is also advantageous to have the time derivatives of (energy+MFCC) as new features, which shows the velocity and acceleration of (energy+MFCC). The equations to compute these features are

$$\Delta IC_l(m) = \frac{\sum_{i=1}^{C} i(IC_l(m+i) - IC_l(m-i)}{2\sum_{i=1}^{C} i^2} \qquad (11)$$

where $\bar{IC}$ (m) l is the mth MFCC coefficient.

2.5 Pattern Matching

The simplest way to recognize an isolated word sample is to compare it against a number of stored word templates and determine the best match [7,8]. DTW is an instance of the general class of algorithms and known as dynamic programming. Its time and space complexity is merely linear in duration of speech sample and the vocabulary size. The algorithm makes a single pass through a matrix of frame scores while computing locally optimized segment of the global alignment path. Let the features extracted from the word be A1;A2;A3; : : :AM and B1;B2;B3; : : :BN. Then DTW matching score between the two sequences is calculated using the equation:

$$D(i,j) = min\begin{cases} D(i-1,j-1) \\ D(i,j-1) \\ D(i-1,j) \end{cases} + d(i,j) \qquad (12)$$

Where D(i; j) is the score in aligning the ith element of A with jth element of B. The dynamic time warping algorithm provides a procedure to align in the test and reference patterns to give the average distance associated with the optimal warping path. The results of DTW are given in fig.

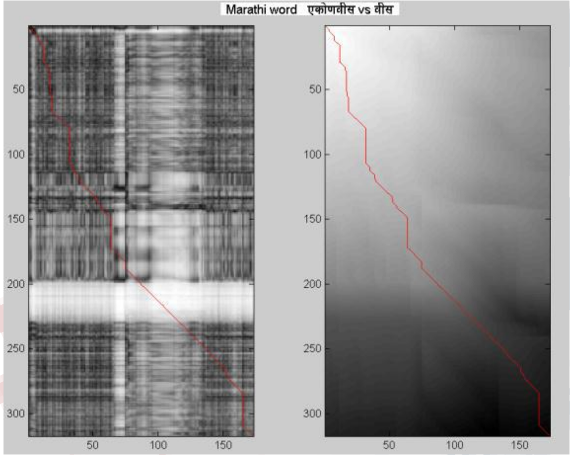


Fig 6. DTW Matching Score for Marathi Word एकोणवीस and "वीस"

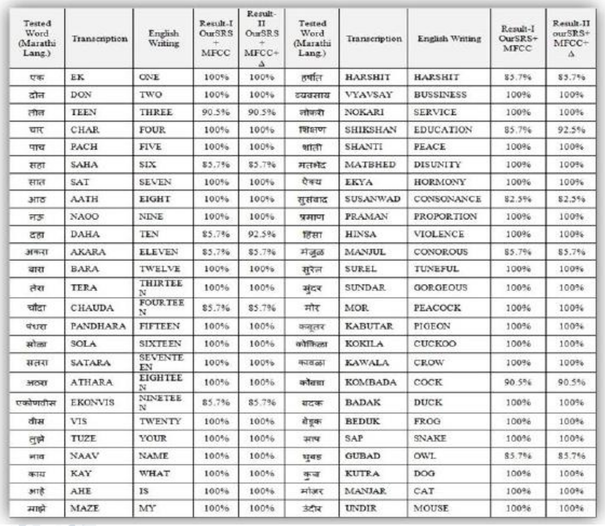
III PROPOSED IMPLEMENTATION DETAILS AND RESULTS

To evaluate the performance of the proposed system, recorded samples were split into training and testing sets whereas out of fifty, 34 words were used for training and the 16 words used for testing. The minimum number of tests made to recognize Marathi word is eight. The recognition rate of each word is calculated by using following formulae:

$$RecogRate = \frac{Number\ of\ Correctly\ Recognized\ Word}{Number\ of\ Tested\ Words} \qquad (13)$$

Table I describes the recognition rate for a speech sample of the tested words in the database. For every test word, the recognition rates are calculated using two different combinations of features i.e Our speech recognition systems with MFCC and Our speech recognition systems with MFCC and delta coefficient. The recognition rate of Our SRS with MFCC is clearly observed. Furthermore, adding delta coefficients to the feature set improves the recognition rate significantly. C was given a value of 3 in order to have a good accuracy with a relatively faster response.

Table I: Recognition rate of different feature sets



Tested Word (Marathi Lang.)	Transcription	English Writing	Result-I OurSRS + MFCC	Result-II OurSRS + MFCC+ Δ	Tested Word (Marathi Lang.)	Transcription	English Writing	Result-I OurSRS+ MFCC	Result-II ourSRS+ MFCC+ Δ
एक	EK	ONE	100%	100%	हर्षित	HARSHIT	HARSHIT	85.7%	85.7%
दोन	DON	TWO	100%	100%	व्यवसाय	VYAVSAY	BUSSINESS	100%	100%
तीन	TEEN	THREE	90.5%	90.5%	नोकरी	NOKARI	SERVICE	100%	100%
चार	CHAR	FOUR	100%	100%	शिक्षण	SHIKSHAN	EDUCATION	85.7%	92.5%
पाच	PACH	FIVE	100%	100%	शांती	SHANTI	PEACE	100%	100%
सहा	SAHA	SIX	85.7%	85.7%	मतभेद	MATBHED	DISUNITY	100%	100%
सात	SAT	SEVEN	100%	100%	ऐक्य	EKYA	HORMONY	100%	100%
आठ	AATH	EIGHT	100%	100%	सुसंवाद	SUSANWAD	CONSONANCE	82.5%	82.5%
नऊ	NAOO	NINE	100%	100%	प्रमाण	PRAMAN	PROPORTION	100%	100%
दहा	DAHA	TEN	85.7%	92.5%	हिंसा	HINSA	VIOLENCE	100%	100%
अकरा	AKARA	ELEVEN	85.7%	85.7%	मंजुळ	MANJUL	CONOROUS	85.7%	85.7%
बारा	BARA	TWELVE	100%	100%	सुरेल	SUREL	TUNEFUL	100%	100%
तेरा	TERA	THIRTEEN	100%	100%	सुंदर	SUNDAR	GORGEOUS	100%	100%
चौदा	CHAUDA	FOURTEEN	85.7%	85.7%	मोर	MOR	PEACOCK	100%	100%
पंधरा	PANDHARA	FIFTEEN	100%	100%	कबुतर	KABUTAR	PIGEON	100%	100%
सोळा	SOLA	SIXTEEN	100%	100%	कोकिळा	KOKILA	CUCKOO	100%	100%
सतरा	SATARA	SEVENTEEN	100%	100%	कावळा	KAWALA	CROW	100%	100%
अठरा	ATHARA	EIGHTEEN	100%	100%	कोंबडा	KOMBADA	COCK	90.5%	90.5%
एकोणवीस	EKONVIS	NINETEEN	85.7%	85.7%	बदक	BADAK	DUCK	100%	100%
वीस	VIS	TWENTY	100%	100%	बेडूक	BEDUK	FROG	100%	100%
तुझे	TUZE	YOUR	100%	100%	साप	SAP	SNAKE	100%	100%
नाव	NAAV	NAME	100%	100%	घुबड	GUBAD	OWL	85.7%	85.7%
काय	KAY	WHAT	100%	100%	कुत्रा	KUTRA	DOG	100%	100%
आहे	AHE	IS	100%	100%	मांजर	MANJAR	CAT	100%	100%
माझे	MAZE	MY	100%	100%	उंदीर	UNDIR	MOUSE	100%	100%

IV. CONCLUSION

Finding efficient automatic speech recognition techniques for Marathi words is of great interest since the research efforts remain limited. In this work, the robustness of MFCC combined with DTW algorithm are clearly noticed. The recognition rate is obviously improved. At the same time, delta coefficients help in improving the overall recognition accuracy. Many experiments were conducted to choose the best parameters that maximize the improvements of Marathi speech recognition. The speech recognition accuracy improvement than other approaches.

REFERENCES

1. Ethnologue: Languages Of The World, Seventeenth Edition (2014) Dallas, Texas: SIL International & Wikipedia.Org

2. "Check Out New Ethnologue",Ethnologue,30-04-2014, Retrieved On 13-Dec 2014.

3. Lawrence Rabiner, Biing-Hwang Juang, Fundamentals Of Speech Recognition, Upper Saddle River, New Jersey: Prentice Hall, USA, 1993

4. X. Huang, A. Acero, And H.-W. Hon, Spoken Language Processing, Upper Saddle River, New Jersey: Prentice Hall, USA, 2001.

5. Mikael Nilsson And Marcus Ejnarsson, "Speech Recognition Using Hidden Markov Model (Performance Evaluation In Noisy Environment)", Masters Thesis, Department Of Telecommunications And Signal Processing, Belkinge Institute Of Technology, Ronneby, Sweden, March 2002.

6. B. Gold And N. Morgan, Speech And Audio Signal Processing, New York, New York: John Wiley And Sons, USA, 2000.

7. Wei HAN, Cheong-Fat Chan, Chiu-Sing Choy And Kong-Pang PUN, "An Efficient MFCC Extraction Method In Speech Recognition", 0-7803-9390-2/06/2006 IEEE

8. K.K. Paliwal, Anant Agarwal And Sarvajit S. SINHA "Amodification Over Sakoe And Chiba's Dynamic Time Warping Algorithm For Isolated Word Recognition." Signal Processing 4.

9. Kale, Kaustubh R. "Isolated word, Speech recognition using Dynamic Time Warping towards smart appliances". Project towards EEL 6825: Pattern Recognition.

10. <www.cnel.ufl.edu/~kkale/6825Project.html> (May 15, 2002). 10. Tan, Li. "A study of tone classification for Thai vowels." Prince of Songkla University, (2004).

11. Roger Jong, "Audio Signal processing and Recognition", http://mirlab.org/jang/books/audiosignalprocessing/index.asp

12. B.S. Jinjin Ye, "Speech Recognition Using Time Domain Features From Phase Space Reconstructions", Masters Thesis, Department of Electrical and Computer Engineering, Marquette University, Milwaukee, Wisconsin, May 2004

13. Darabkh, K., A. Khalifeh, Baraa A. Bathech, and Saed W. Sabah. "Efficient DTW-Based Speech Recognition System for Isolated Words of Arabic Language." In Proceedings of International Conference on Electrical and Computer Systems Engineering (ICECSE 2013), pp. 699-702. 2013.

14. Avinash N Bhute, B.B. Meshram, "IVSS: Integration of Color Feature Extraction Techniques for Intelligent Video Search Systems" in proceeding of IEEE 4th International Conference on Electronics Computer Technology (ICECT 2012) pp. 113-118

15. A.N. Bhute, B.B. Meshram "Multimedia Indexing and Retrieval Techniques: A Review." International Journal of Computer Applications IJCA,vol. 58, no. 3 November 2012, ISSN 0975–8887, pp. 35-43