Speech Processing for Marathi Numeral Recognition using MFCC and DTW Features

Siddheshwar S. Gangonda*, Dr. Prachi Mukherji**

*(Smt. K. N. College of Engineering,Wadgaon(Bk), Pune, India).
sgangonda@gmail.com
**(Smt. K. N. College of Engineering,Wadgaon(Bk), Pune, India).
prachimukherji@rediffmail.com

ABSTRACT

Numeral recognition remains one of the most important problems in pattern recognition. It has numerous applications including those in reading postal zip code, passport number, employee code, form processing and bank cheque processing, postal mail sorting, job application form sorting, automatic scoring of tests containing multiple choice questions and video gaming etc. To the best of our knowledge, little work has been done in Indian language, especially in Marathi as compared with those for non Indian languages. This paper has discussed an effective method for recognition of isolated Marathi numerals. It presents a Marathi database and isolated numeral recognition system based on Mel-Frequency Cepstral Coefficient (MFCC) used for Feature Extraction and Distance Time Warping (DTW) used for Feature Matching or to compare the test patterns.

Keywords - **Feature Extraction, Feature Matching, Mel Frequency Cepstral Coefficient (MFCC), Dynamic Time Warping (DTW), Numeral, Recognition.**

1. INTRODUCTION[1]

India is multilingual country of more than 1 billion population with 18 constitutional languages and 10 different scripts. Devnagari, an alphabetic script, is used by a number of Indian languages. A numeral analysis is done after taking an input through microphone from a user. The digitized speech samples are then processed using MFCC to produce numeral features. After that, the coefficient of numeral features can go trough 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 can be implemented using MATLAB. Thus, this work focuses on Marathi language. The aim of this paper is to build a numeral recognition tool for Marathi language. This is a isolated word speech recognition tool. We have used MFCC and DTW algorithms. The section 1 gives the introduction, section 2 gives the literature survey, section 3 gives the methodology, section 4 gives the marathi isolated numeral recognition system, section 5 gives the results followed by conclusions in section 6.

2. LITERATURE SURVEY [2]

Designing a machine that mimics human behavior, particularly the capability of speaking naturally and responding properly to spoken language, has intrigued engineers and scientists for centuries. The research in automatic speech recognition by machine has attracted a great deal of attention over the past five decades.

In the late 1960's, Atal and Itakura independently formulated the fundamental concepts of Linear Predictive Coding (LPC), which greatly simplified the estimation of the vocal tract response from speech waveforms. By the mid 1970's, the basic ideas of applying fundamental pattern recognition technology to speech recognition, based on LPC methods, were proposed by Itakura, Rabiner and Levinson and others.The idea of the hidden Markov model appears to have first come out in the late 1960's. Another technology that was introduced in the late 1980's was the idea of artificial neural networks (ANN).

At present, due to its versatile applications, numeral recognition is the most promising field of research. Our daily life activities, like reading postal zip code, passport number, employee code, form processing and bank cheque processing etc. involves numeral recognition. However some works for south Asian languages including Hindi have also been done but no one provides efficient solution for Marathi language. The lack of effective Marathi numeral recognition system and its relevance has motivated us to develop such small size vocabulary system.

3. METHODOLOGY[3]

A numeral analysis is done after taking an input through microphone from a user. The design of the system, involves manipulation of the input audio signal. At different levels,

different operations are performed on the input signal such as Pre-emphasis, Framing, Windowing, Mel Cepstrum analysis and Recognition (Matching) of the spoken word. The voice algorithms consist of two distinguished phases. The first one is training sessions, whilst, the second one is referred to as operation session or testing phase as described in fig.1.

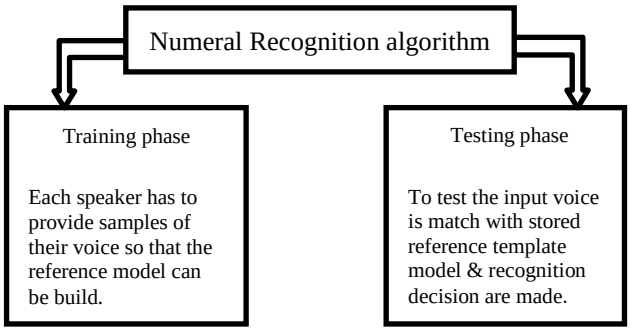


Fig.1: Numeral Recognition algorithm.

4. MARATHI ISOLATED NUMERAL RECOGNITION SYSTEM [1]

The Speech is the most prominent and natural form of communication between humans. There are various spoken Languages throughout the world. Marathi is an Indo-Aryan Language, spoken in western and central India. There are 90 million of fluent speakers all over world. It presents a work that consists of the creation of Marathi numeral database and its recognition system for isolated words.

4.1 Marathi Numeral Database

For accuracy in the numeral recognition, we need a collection of utterances, which are required for training and testing. The Collection of utterances in proper manner is called the database. The age group of speakers selected for the collection of database ranges from 22 to 35. Mother tongue of all the speakers was Marathi. The vocabulary size of the database consists of:

• Marathi Numerals: 0-9.

1) *Acquisition Setup*: To achieve a high audio quality the recording took place in the 10 x 10 rooms without noisy sound and effect of echo. The Sampling frequency for all recordings was 11025 Hz in the Room temperature and normal humidity. The speaker were Seating in front of the direction of the microphone with the Distance of about 12-15 cm. The speech data is collected by taking input through microphone from a user.

2) *Feature Extraction*: It is very important to extract the features in such a way that the recognition of numerals becomes easier on the basis of individual features of the numerals.

4.2 Numeral Recognition System

There are several kinds of parametric representation of the acoustic signals. Among of them the Mel-Frequency cepstral Coefficient (MFCC) is most widely used. We have developed the recognition system using MFCC and DTW.

4.2.1 Feature extraction (MFCC)

It is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. It consists of various steps given below in the fig.2.

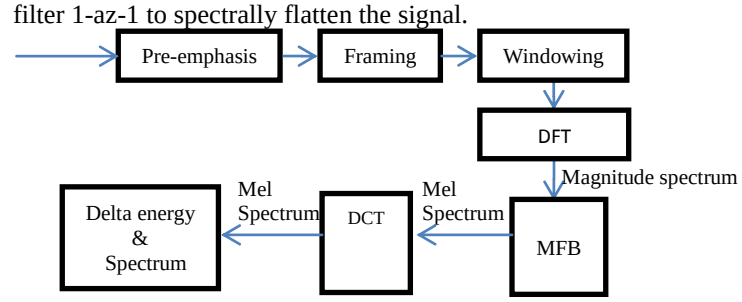
i. Speech Signal

The excitation signal is spectrally shaped by a vocal tract Equivalent filter. The outcome of this process is the sequence of exciting signal called speech.

Voice input

ii. Pre-emphasis

The speech is first pre-emphasis with the pre-emphasis filter 1-az-1 to spectrally flatten the signal.



Output

Fig.2: MFCC Block Diagram.

iii. Framing and Windowing

A speech signal is assumed to remain stationary in periods of 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 frame is shifted 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.

iv. Fourier Transform

To obtain a good frequency resolution, a 512 point Fast Fourier Transform (FFT) is used.

v. Mel-Frequency Filter Bank

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

vi. Discrete Cosine Transform

As the Mel-Cepstrum coefficients contain only real parts, the Discrete Cosine Transform (DCT) can be used to achieve the Mel-Cepstrum coefficients. There were 24 Coefficients out of that only 13 coefficients have been selected for the recognition system.

vii. Delta Energy and Delta Spectrum

The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time. The energy in a frame for a signal x in a window from time sample t1 to time sample t2, is represented as shown below in "(1)".

$$\text{Energy} = \Sigma x^2[t] \qquad (1)$$

Where X[t] = signal.

4.2.2 Feature matching (DTW) [4]

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. The Fig.3. shows the example of how one times series is 'warped' to another.

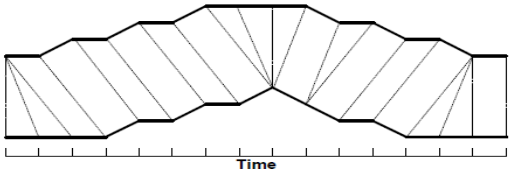


Fig.3: Warping between two time series.

To align two sequences using DTW, an *n*- by- *m* matrix where the (*i*th, jth) 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 "(2)".

$$d\,(qi,cj) = (qi - cj)^2 \qquad (2)$$

Each matrix element (*i, j*) corresponds to the alignment between the points *q*i and *c*j. Then, accumulated distance is measured by "(3)".

$$D\,(i, j) = \min[D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i, j) \qquad (3)$$

This is shown in Fig.4, where the horizontal axis represents the time of test input signal, and the vertical axis represents the time sequence of the reference template. The path shown results in the minimum distance between the input and template signal. The shaded area represents the search space for the input time to template time mapping function. Any monotonically non decreasing path within the space is an alternative to be considered. Using dynamic programming techniques, the search for the minimum distance path can be done in polynomial time P (t), using equation below:-

$$p\ t\ = o[N^2\ V] \qquad (4)$$

Where, N is the length of the sequence, and V is the number of templates to be considered.

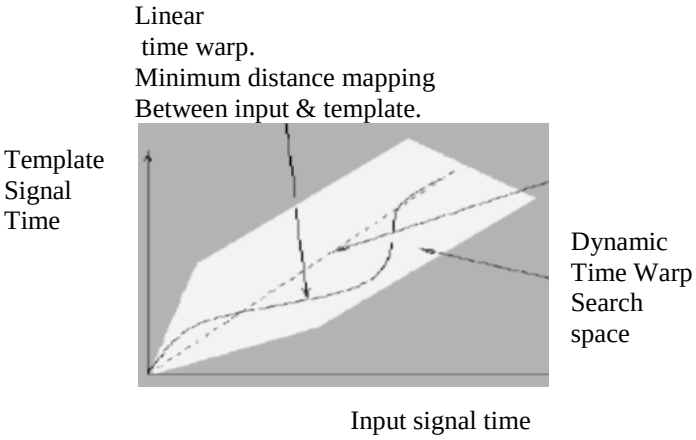


Fig.4: Dynamic time warping (DTW).

Theoretically, the major optimizations to the DTW algorithm arise from observations on the nature of good paths through the grid. These are outlined in Sakoe and Chiba and can be summarized as follows-

Monotonic condition:- The path will not turn back on itself, both i and j indexes either stay the same or increase, they never decrease.

Continuity condition:- The path advances one step at a time. Both i and j can only increase by 1 on each step along the path.

Boundary condition:- The path starts at the bottom left and ends at the top right.

Adjustment window condition:- A good path is unlikely to wander very far from the diagonal. The distance that the path is allowed to wander is the window length r.

Slope constraint condition:- The path should not be too steep or too shallow. This prevents very short sequences matching very long ones. The condition is expressed as a ratio n/m where m is the number of steps in the x direction and m is the number in the y direction. After m steps in x you must make a step in y and vice versa.

The purpose of DTW is to produce warping function that minimizes the total distance between the respective points of the signal.

5 RESULTS

Few people recorded the number zero to nine in Marathi. Some of the MFCC Features extracted of the Marathi Numerals are shown in the figures below.

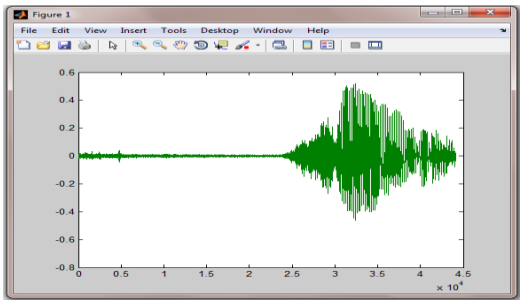


Fig.4: Mel Frequency Cepstrum Coefficients (MFCC) of shunya.

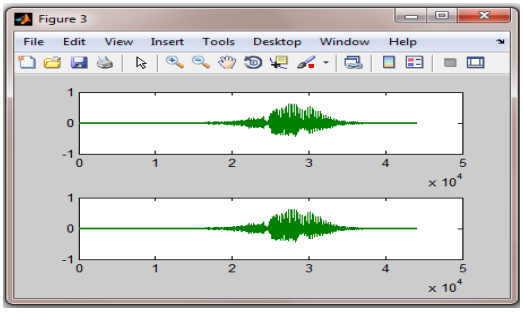


Fig.5: Plot for FFT & IFFT of y[shunya].

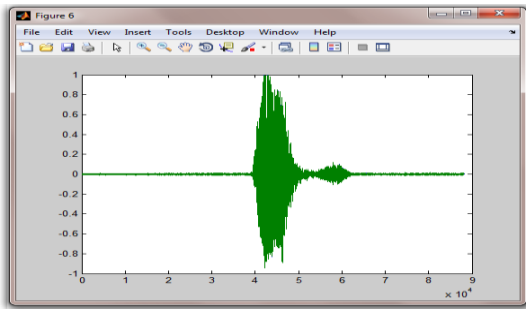
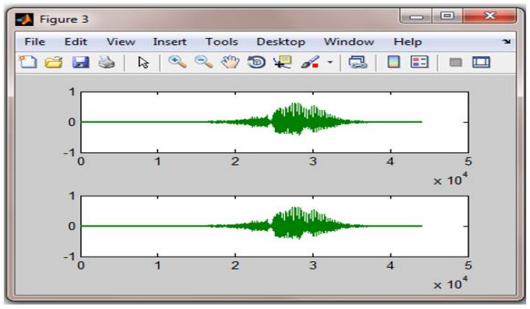


Fig.6: Mel Frequency Cepstrum Coefficients (MFCC) of pach.



. Fig.7: Plot for FFT & IFFT of y[pach].

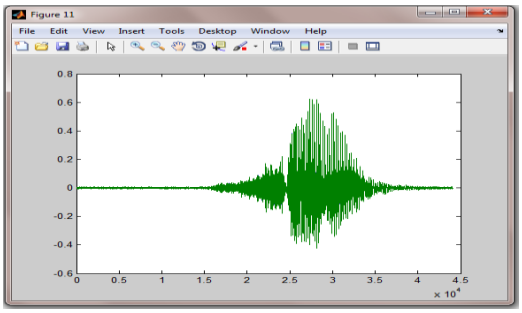
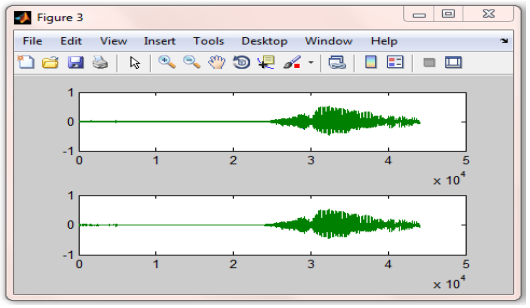


Fig.8: Mel Frequency Cepstrum Coefficients (MFCC) of sat.



. Fig.9: Plot for FFT & IFFT of y[sat].

6 CONCLUSIONS

This report has discussed an effective method for recognition of isolated Marathi numerals in Devnagari script. It presents a Marathi database and isolated numeral recognition system based on Mel-frequency cepstral

coefficient (MFCC) and Distance Time Warping (DTW) as features.

In recent years there has been a steady movement towards the development of speech technologies to replace or enhance text input called as Mobile Search Applications. Recently both Yahoo! and Microsoft have launched voice-based mobile search applications. Future work can include improving the recognition accuracy of the individual numerals by combining the multiple classifiers.

ACKNOWLEDGMENTS

It is my pleasure to get this opportunity to thank my beloved and respected Guide **Dr. Prachi Mukherji** who imparted valuable basic knowledge of Electronics specifically related to Speech Processing.

REFERENCES

[1] Bharti W. Gawali[1], Santosh Gaikwad[2], Pravin Yannawar[3], Suresh C.*Mehrotra*[4] "Marathi Isolated Word Recognition System using MFCC and DTW Features" *Proc. of Int. Conf. on Advances in Computer Science 2010.*

[2] Rabiner L. and Juang B.H., "Fundamentals of Speech Recognition". *New York:Prentice Hall Publishers,1993.*

[3] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuth, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques" *Journal of Computing, Volume 2, Issue 3, March 2010, ISSN 2151-9617.*

[4] Anjali Bala*, Abhijeet Kumar, Nidhika Birla, "VOICE COMMAND RECOGNITION SYSTEM BASED ON MFCC AND DTW" *International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7335-7342.*

[5] Kuldeep Kumar R. K. Aggarwal "Hindi Speech Recognition System Using HTK" *International Journal of Computing and Business Research ISSN (Online) : 2229-6166 Volume 2 Issue 2 May 2011.*

[6] Gopalkrishna Anumanchipalli, Rahul Chitturi , Sachin Joshi , Rohit Kumar, Satinder Pal Singh*, R.N.V. Sitaram*, S P Kishore, *"Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems".*