Automatic Speech Recognition System Using MFCC And DTW For Marathi Isolated Words

Kishori R. Ghule , Ratnadeep R. Deshmukh

MTech Student, Department of CS & IT, Dr. BAM University, Aurangabad, India. kishori.ghule@gmail.com, 2HOD, Department of CS & IT, Dr. BAM University, Aurangabad, Maharashtra, India

Abstract: Speech Recognition is the process of identifying words and phrase from spoken language and converts them in to machine readable format. This paper describes an approach of isolated words speech recognition by using Mel-Scale Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW). This ASR develops for large vocabulary. MFCC extract the features of spoken words speech signals. For database 100 isolated words are taken from Marathi language and recorded from 100 native speakers with 3 utterances. Dynamic Time Warping algorithm is used for the feature matching purpose. DTW algorithm measures similarity between two sequences, which may vary in speed or time.

Key Words: Feature Extraction, Mel frequency cepstral coefficients (MFCC), Speech Recognition, Dynamic Time Warping (DTW), Feature matching.

INTRODUCTION

Human interact with each other by using speech. They express thoughts, feelings, and ideas by speech. Speech recognition is the process to identify words or phrase from spoken language and convert into machine readable format. Very little work has been done for Indian languages compared to non Indian languages. This work is done for Marathi language. The database is collected from Marathi language. The 100 words are recorded from 100 speakers with 3 utterances of each word. First two utterances used for training purpose and third utterances used for testing purpose. The most popular spectral based parameter used in recognition approach is the Mel Frequency Cepstral Coefficients called MFCC. MFCCs are coefficients, which represent audio, based on perception of human auditory systems. The experiment performs in Matlab.

1 METHODOLOGY

1.1 Dataset Collection

1. Selection of the Speakers

- One hundred speakers were selected from all over Aurangabad district.
- Selected speakers whose native language is Marathi and also those speakers whose native language is not Marathi.
- Speaker's age is in between 20 to 40.

2. 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. While recording the sampling frequency was set to 16 KHz with Mono sound type. The speakers were asked to read each word and recorded sample was saved as .wav file. Each word is repeated three times so one speaker gives 300 speech samples. In this way recording taken from 100 speakers. So dataset has 30,000 speech samples.

3. Data Collection Statistics

We have target 100 speakers from Aurangabad district. The Speech samples are recorded 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 speech sample as a dataset. The 100 words are the Marathi names of medicinal plants which are growth in India. This data collected from different sources like internet, medical students etc. Some of these are given in Table 1.

TABLE 1
MEDICINAL PLANTS

Marathi name	Botanical name
चांगेरी	*Oxalis corniculata*
मंजिष्ठा	*Rubia cordifolia*
बेल	*Aegle marmelos*
कडुलिंब	*Azadirachta indica*
साजी पर्णी	*Desmodium gangeticum*
अशोका	*Saraka asoca*
गंध प्रसारनी	*Paederia foetida*
जांभूळ	*Syzygium cumini (Eugenia jambolana)*
हिंग	*Ferula northax (F. foetida)*
जिरे	*Jiraka - Cuminum cyminum*

This dataset divided into two groups training and testing. The first group training dataset contains first two utterances of 100 words and second group testing dataset having third utterance of 100 words recorded from 100 speakers. All speech signals are recorded under most similar setting condition such as the same length of recording time etc. In training, the Matlab program named as "train" extracts the features of 20,000 speech samples from training group and stored these features into "train_features.mat" file. For testing Matlab program "test" is executed. User can choose any speech sample for testing from test file. MFCC at the back end extracts the features of the chosen speech sample. Then "train_features.mat" file is called for feature matching. DTW first matches the features of the selected sample speech signal with "train_features.mat" by measuring the local distance. DTW then measures the

global distance and the part that matches with the chosen sampled speech is the result of the “test” program that shows the correct spoken word in the command window.

1.2 Feature Extraction by MFCC

Mel Frequency Cepstral Coefficients called MFCC is the most popular spectral based parameter used in recognition approach is the. Due to its advantage of less complexity in implementation of feature extraction algorithm, certain coefficients of MFCC corresponding to the Mel scale frequencies of speech Cepstrum are extracted from spoken word samples in database. [1]

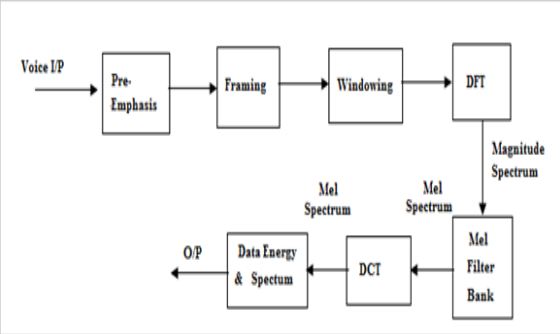

Fig -1: *Block Diagram of MFCC*

1. Preprocessing

To get the accuracy and efficiency of the extraction processes Speech signals are normally pre-processed before features extraction. The speech is first pre-emphasis with the first order high pass FIR filter to spectrally flatten the signal.

2. Framing and windowing

A Speech signal is assumed to remain stationary in periods of approximately 20ms. A discrete signal s[n] splits into number of frames in the time domain so each frame can be analyzed in the short time instead of analyzing the entire signal at once and truncating the signal with a window function w[n]. After dividing the signal into frames that contain nearly stationary signal blocks, Each time frame is then windowed with Hamming window to eliminate discontinuities at the edges [2], [3].

3. Fast Fourier transform

Fast Fourier Transformation (FFT) is calculated to extract frequency components of a signal in the time-domain. After the windowing FFT is calculated. FFT is used to speed up the processing. [4]

4. Mel filter bank processing

A subjective pitch is measured on a scale called the Mel scale for each tone of signal with an actual frequency f, measured in Hz because human ear perception of frequency contents of sounds for speech signal does not follow a linear scale. The Mel-scale used in this work is to map between linear frequencies scales of speech signal to the logarithmic scale for frequencies higher than 1 kHz. This makes the spectral frequency characteristics of signal closely corresponding to the human auditory perception. MFCCs use Mel-scale filter bank where the higher frequency filters have greater bandwidth than the lower frequency filters, but their temporal resolutions are the same.

5. Discrete Cosine Transform

Convert the log Mel spectrum from frequency to time domain, the result is called the Mel Frequency Cepstrum Coefficients (MFCC). The Mel-Cepstrum coefficients contain only real part. In a frame, there are 24 Mel-Cepstral coefficients, out of these only 13 coefficients have been selected for the recognition system.

1.3 Classification by DTW

Dynamic time warping algorithm measures the similarity between two sequences 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 used to find equivalently regions among the two time series or to finds the similarity between two times series [5]. Same user can give different utterances of same word which may differ in time. DTW resolves this problem by aligning the words properly and calculating the minimum distance between two words. 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 [6].

Suppose we have two time series Q and C, of length n and m respectively,

$Q = q1, q2,\ldots, qi,\ldots,qn$
$C = c1, c2,\ldots, cj,\ldots,cm$

So there is n-by-m matrix, To align two sequences using DTW the (ith, jth) element of the matrix contains the distance d (qi, cj) between the two points qi and cj is constructed, Then, the absolute distance between the values of two sequences is calculated using the Euclidean distance computation:

$d(qi, cj) = (qi - cj)2$

Each matrix element (i, j) corresponds to the alignment between the points qi and cj. Then, accumulated distance is measured by:

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

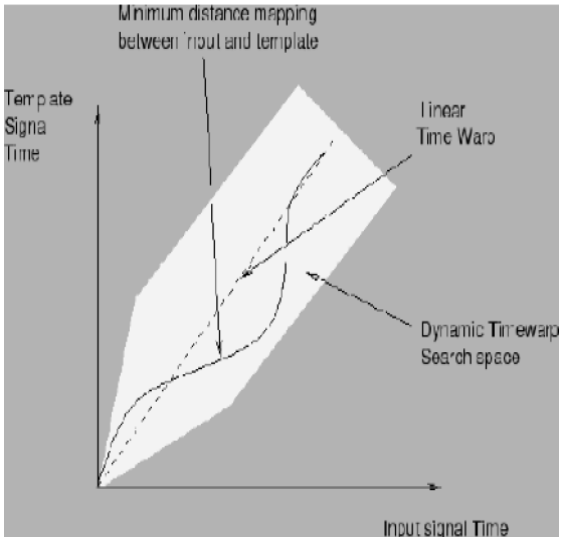

Fig -2: *Dynamic Time Warping*

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 (N2 V)

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

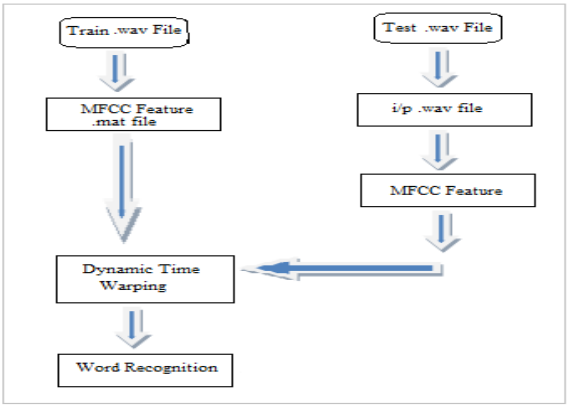

Fig -3: *Isolated Word Recognition System*

2 RESULT

Each speech sample in test file compared with all speech samples in train file. DTW calculates distance between words. From the experiment we obtained 89% (Out of 10,000 speech samples 8,900 could be classified correctly) recognition by using Dynamic Time Warping classification technique. Error rate is 11% (1,100 out of 10,000 could not classified correctly). The system gives better accuracy.

3 CONCLUSION

Develop a speech database and automatic speech recognition system of isolated words in Marathi language is the aim of this research. From this study we conclude that Dynamic Time warping is most powerful technique for classification. It gives good recognition result with a large database.

ACKNOWLEDGEMENT

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] Nidhi Desai, Prof.Kinnal Dhameliya, Prof.Vijayendra Desai3, "Feature Extraction and Classification Techniques for Speech Recognition: A Review", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com,ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 12, December 2013

[2] Goranka Zoric, "Automatic Lip Synchronization by Speech Signal Analysis", Master Thesis, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Oct-2005.

[3] Lahouti, F., Fazel, A.R., Safavi-Naeini, A.H., Khandani, A.K, "Single and Double Frame Coding of Speech LPC Parameters Using a Lattice-Based Quantization Scheme", IEEE Transaction on Audio, Speech and Language Processing, Vol. 14, Issue 5, pp. 1624-1632, Sept-2006.

[4] Namrata Dave "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition", G H Patel College of Engineering, Gujarat Technology University, India, International Journal For Advance Research In Engineering And Technology, Volume 1, Issue VI, July 2013

[5] Shivanker Dev Dhingra, Geeta Nijhawan, Poonam Pandit, "Isolated Speech RecognitionUsing MFCC And DTW", International Journal Of Advanced Research In Electrical, Electronics And Instrumentation Engineering, Vol. 2, Issue 8, August 2013

[6] Dynamic Time Warping-Wikipedia, the Free Encyclopedi