Odia Isolated Word Recognition using DTW

Anjan Kumar Sahu
Dept. of ECE
Centre for Advanced Post Graduate Studies
Rourlela, INDIA

Gyana Ranjan Mati
Dept. of ECE
Centre for Advanced Post Graduate Studies
Rourlela, INDIA

***Abstract*— Speech Recognition is the process of communication between human and computer, it has a wide area of applications in a security system, healthcare, military, telephony system, and equipment designed for handicapped. Speech is the vocalized form of communication, based upon the syntactic combination of lexica's and names that are drawn from very large (usually about 1,000 different words) vocabularies. Each spoken word is created out of the phonetic combination of a limited set of vowel and consonant speech sound units, so a proper algorithm is required for feature extraction and recognition process. For the feature extraction, we have used MFCC (Mel Frequency Cepstral Coefficients) and for recognition, we have used DTW (Dynamic Time Warping) all the implementation of Speech recognition have been done using MATLAB 2012b software.**

Keywords*— Speech Recognition, MFCC, DTW, FFT, Isolated word

I. INTRODUCTION

Speech is the ancient way to express ourselves. Speech recognition is the process of training the computer or a machine to identify spoken isolated word of a person using speech signal information. In the training phase a large number of data has to be stored for speech recognition process. In the second phase the model is used for classification. When a person speaks, the speech signal is captured. To identify the speech information acoustic analysis is carried out. First the feature is extracted of the data then from the feature extracted data it was given to a classifier which can classifies the isolated spoken word by matching it with the data base and the maximum match of utterance will recognized as the spoken word.

Basically speech recognition problem can focus on identifying the speech or speaker who uttered the speech. The proposed model aims to design an isolated word recognizer which can identifies the person.

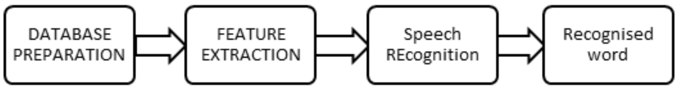


Fig.1 Proposed Word Recognition Model

II. LITERATURE REVIEW

Various methods have been proposed for the isolated word recognition over a decayed from which Hidden Markov Model (HMM) has been extensively used for large vocabulary data base for high reliability [10]. Artificial Neural Networks (ANN) is another classifier for speech recognition with accuracy which is acceptable [11], Support Vector Mechanism (SVM) have been used to classify speech pat-tern using linear and non-linear discrimination models [12]. For simple isolated word detection DTW and MFCC approach is enough and efficient [5]. However, if continuous speech detection with speaker discrimination is needed MFCC alone is not necessary for assuring the algorithm. Combination Various Classifiers required for high reliability, for simple word recognition and for small amount of database creation MFCC and DTW approach is the simpler approach then HMM, ANN and SVM [5].So in this Paper we have adopted MFCC and DTW as our proposed model.

III. METHODOLOGY

A. Recognition Module

In the process of isolated word recognition process we have to follow two approach. First process is the feature extraction model and second one is the feature matching model. For feature extraction model we have used MFCC and before going for feature extraction we have to calculate the energy, spectrogram and PSD (Power Spectral Density). In this process we have fist recorded the speech with 16KHZ sampling frequency then we do our pre-processing steps, in the pre-processing 1st we have calculated the energy and its energy spectrum is shown below fig.2.

The energy of an input speech is calculated for know-how much energy is present inside the signal, then we have to calculate the PSD, power is calculated to know the amount of power inside a speech signal and it can be shown in the below fig. 3 from that power spectrum we can know that which one is our desired signal and which one is our noise or the silence part after calculating the power we will go for the spectrogram analysis of the input speech signal using wideband spectrogram and narrowband spectrogram and it can be described in the below fig. 4.

Spectrogram is a visual representation of the spectrum of frequencies in a sound or other signal as they vary with time or some other variable. Spectrogram can be used to identify the spoken word phonetically and to analyze the various calls of human.

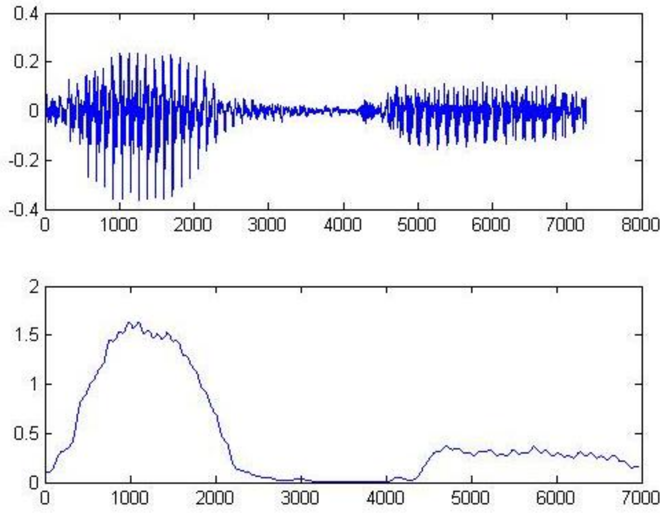


Fig.2 Energy of the given word

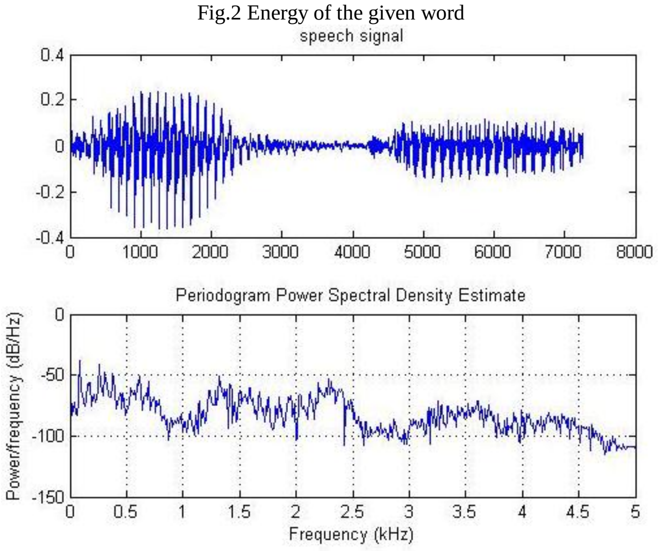


Fig. 3 Power Spectral Density of the Speech

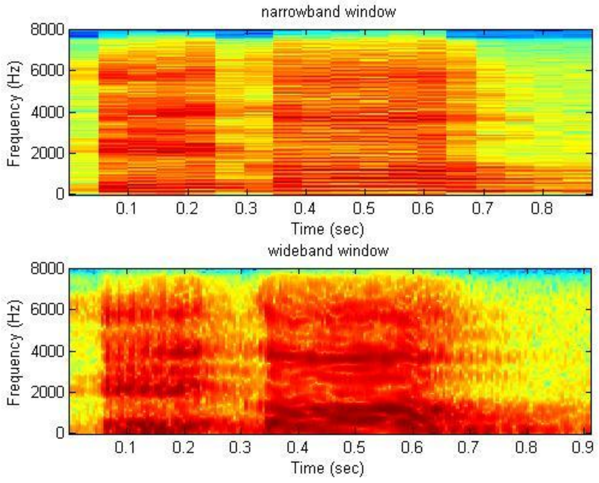
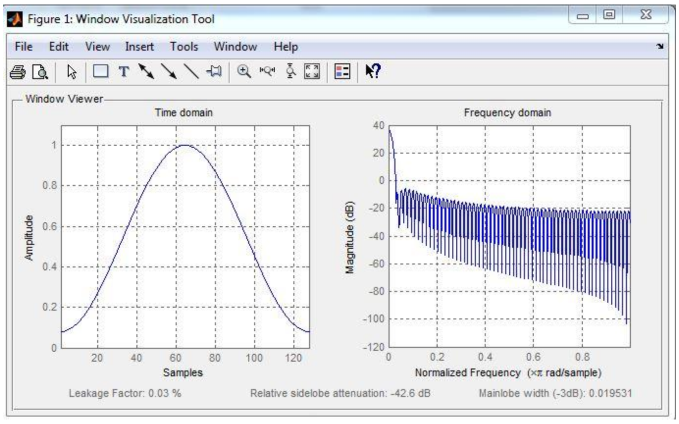


Fig.4 Spectrogram of the speech

B. *Feature Extraction*

After calculating the preprocessing steps we have to go for MFCC analysis of the spoken word and it can be described as, first the input voice is preprocessed it can be done by the help of three steps that are pre emphasis, normalization, and mean subtraction [1]. In the pre-emphasis a FIR high pass filter with the transfer function H (z) =1-0.98z-1 is used to flatten the signal spectrum. The high frequencies of the speech signal formed in the vocal tract are attenuated as the sound pass through the lips. Then normalization is done to reduce amplitude variation from speech samples for all the words. In normalization, all the samples of the signal are divided by the highest amplitude sample value in the signal. Mean subtraction is done to remove the dc offset introduced due to the microphone and some other effects introduced at the time of recording [4]. Then we go for framing and windowing and it can be done by taking the pre-emphasized signal is di-vided into short frame blocks and hamming window is applied to these frames. Hence, the speech signal is divided into frames and the assumption is that the signal is stationary for this small frame and features are calculated for each frame. The purpose of this window is to limit the time interval to be analyzed so that the properties of the waveform do not change appreciably.



Leakage Factor: 0.03 %
Relative sidelobe attenuation: -42.6 dB
Mainlobe width (-3dB): 0.019531

Fig.5 Hamming Window

Windowing also serves to remove the signal discontinuities at the beginning and end of each frame. Hamming window is used for this purpose since it provides smoother spectrum. Hamming window is given by Eq.

$$W(n)=0.54-0.46\cos[2\pi n/(N-1)] \qquad 0\leq n\leq N-1 \qquad (1)$$

Where, N is the number of samples in a single frame

Next we are going for FFT (First Fourier Transform), here we calculate the DFT of an input signal in an efficient manner and thus saving processing power and reducing computation time.

The FFT is characterized by the following equation:

$$X(k)=\sum x(j)W_N^{(j-1)(k-1)} \qquad (2)$$

Where x (j) is the jth sample, $W_N = e^{\frac{(2\pi i)}{N}}$ .This gives spectral coefficients of the windowed frame.

Then we go for MEL filtering which converts the frequency domain signal into its corresponding MEL domain. The process of obtaining Mel-Cepstral Coefficients involves the use of a Mel-scale filter bank. The Mel-scale is a logarithmic scale resembling the way that the human ear perceives sound.

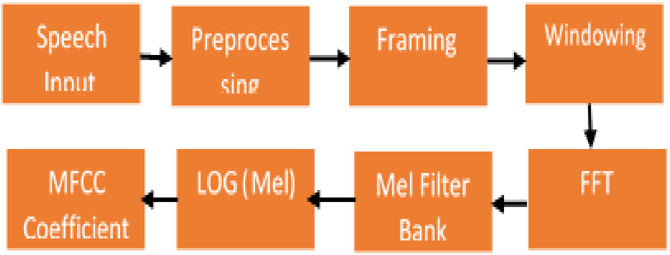


Fig.6 MFCC Feature Extraction Process

The filter bank is composed of 20 triangular filters that are not- uniformly placed in frequency such that these filters will have a linear frequency response at low frequency (up to 1 KHz) and logarithmic at high frequencies as shown in fig. 7 The Mel scale is represented by the following equation:

$$\text{Mel}(f) = 2595 * \log_{10}(1+f/700) \qquad (3)$$

Where f is the frequency.

The spectral coefficients of the frames are multi-plied by the filter gain and the result is obtained, the MEL filter output is shown in below figure 7.

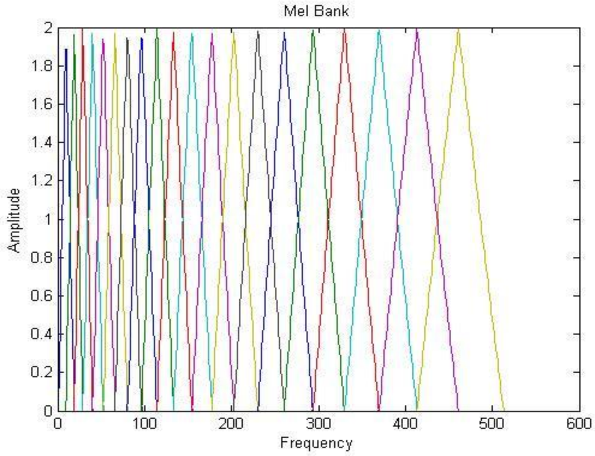


Fig. 7 Mel filter bank output

Then after MEL filtering we go for LOG and DCT which converts the MEL domain parameter into its corresponding frequency domain parameter. Which is given by

$$Mel^{-1}(m) = 700\ \exp\left(\left(\frac{m}{2595}\right) - 1\right) \qquad (4)$$

Where m is the MEL domain parameter and the output of LOG is shown in figure 10. And the DCT is calculated by

$$C_n = \sum_{k=1}^{K} (\log S_k)[n(k-0.5)\tfrac{\pi}{k}] \qquad (5)$$

Where n= 1, 2, 3, 4…..k (S_k= FFT coefficients)

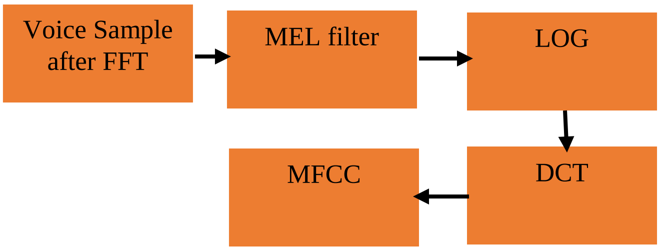


Fig. 9 steps in converting frequency domain to time domain

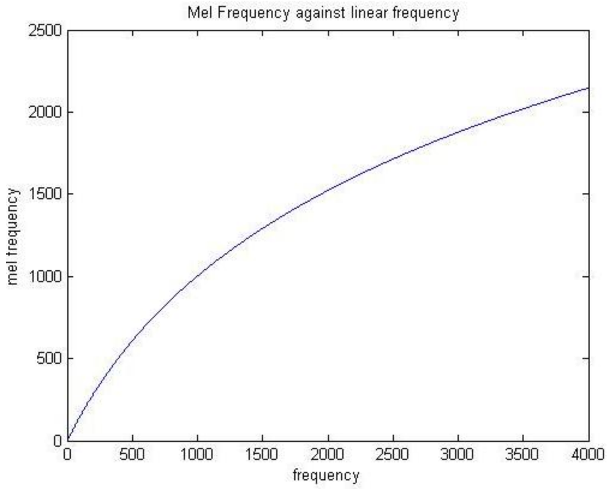


Fig.8 Mel Vs Frequency plot

The above figure shows the relationship between Mel and frequency and the process of converting of Mel to Frequency is shown in below figure 9

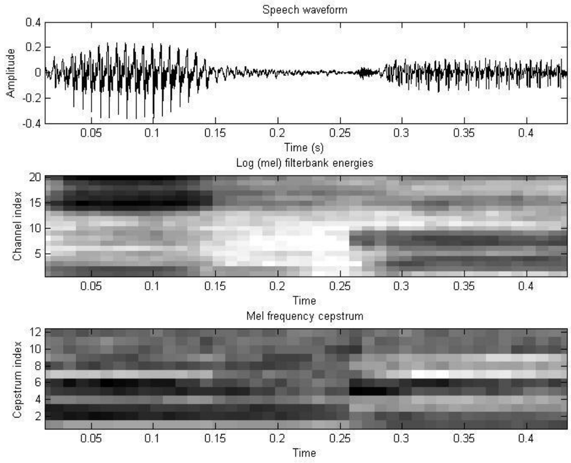


Fig. 9 LOG (Mel) output

C. Feature Matching

In this stage, the features of word calculated in pre-vious step are compared with the help of the data-base to calculating the exact spoken word. DTW algorithm is implemented to calculate the least distance between features of word utterance and reference templates [5]. Corresponding to least value among calculated score with each template, the word is detected. How can we find the optimum mapping path in DTW? To compute the smallest Euclidean distance between the paths the obvious choice is for-ward Distance Path, which can be summarized by simple three steps and the below are showing the optimum choice/ the 3steeps:

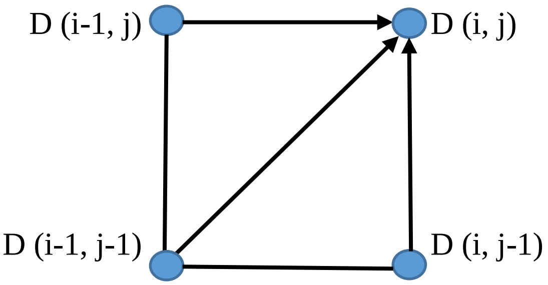


Figure 10. DTW distance mapping

1. Optimum value function: Define D(i, j) as the DTW distance between t(1:i)and r(1:j), with the mapping path starting from (1,1) to (i, j) .
2. Recursion:

$$D\,(i,j)=|t(i)-r(j)|+\min\left\{\begin{matrix} D(i-1,j) \\ D(i-1,j-1) \\ D(i,j-1) \end{matrix}\right\} \qquad (6)$$

With the initial condition D (1, 1) = $|t(1)-r(1)|$ (7)

3. Final answer : D(m, n)

In practice, we need to construct a matrix D of dimensions m×n first and fill in the value of D (1, 1) by using the initial condition. Then by using the recursive formula, we fill the whole matrix one element at a time, by following a column by column or row by row order. The final answer will be available as D (m, n), with a computational complexity of
O (mn) corresponds to the least distance is the word detected.

IV. RESULTS AND DISCUSSION

The distance while comparing similar words and different words are shown in table1. With similar words distance is below 150 with different words the distance is more than 300 i.e., 397.9128. Thus a threshold of 200 or less can filter a given word from set of saved templates. As DTW calculates possible alignment between two vector paths, the distance obtained when two same sequences compared should be 0.

Table 1 Pronunciation of English to Odia words

No. Pronounced in English	No. Pronounced in Odia
ONE	EKA
Two	DUI
THREE	TINI
FOUR	CHARI
FIVE	PANCHA

Table 2 Comparison Between Different Words

Word 1	Word 2	DTW Distance
EKA	DUI	397.913
EKA	TINI	447.869
EKA	CHARI	338.086
EKA	PANCHA	323.089
DUI	TINI	412.192
DUI	CHARI	377.377
DUI	PANCHA	522.532
TINI	CHARI	350.23
TINI	PANCHA	587.635
CHARI	PANCHA	430.281

Table 3 Comparison Between Same Words

Similar Words	DTW distance
EKA	130.695
DUI	112.619
TINI	125.9196
CHARI	120.129
PANCHA	101.722

V. CONCLUSION

With MFCC and DTW, isolated word detection system is generated in MATLAB 2012b environment. System is trained by saving templates of five separate words. Results showed that saving ten templates for each word in training phase gives good results compared with five templates. Efficiency in detecting isolated words is 100percent for two syllable words compared with one syllable word. From the results above, we can infer that DTW distance between identical words is less than 150 and between different words is more than 300. So setting the threshold of 200 we can easily filter the word uttered by the user from the other words whose templates are saved in the training phase. In the future scope we can use LPC instead of MFCC for a Comparing result between MFCC and LPC and instead of DTW we can use HMM for sentence recognition and converting the sentence into its corresponding English sentence.

REFERENCE

[1] Lawrence Rabiner and Biing-Hwang Juang, "Fundamentals of Speech Recognition", Englewood Cliffs, NJ: Prentice Hall, pages 333-352 and 434- 450, 1993.

[2] L.R.Rabiner, R.W.Schafer, "Digital Processing of Speech Signals", Prentice-Hall, Englewood, cliffs, NJ1978

[3] Urmila Shrawankar, Dr. Vilas Thakare "TECHNIQUES FOR FEATURE EXTRACTION IN SPEECH RECOGNITION SYSTEM: A COMPARATIVE STUDY".

[4] Koustav Chakraborty, Asmita Talele, Prof. Savitha Upadhya, "Voice Recognition Using MFCC Algorithm", International Journal of Innovative Research in Advanced Engineering (IJIRAE), ISSN: 2349-2163, Volume 1, Issue 10, (November 2014)

[5] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", JOURNAL OF COMPUTING, ISSN 2151-9617 VOLUME 2, ISSUE 3, MARCH 2010

[6] Kashyap Patel, R.K. Prasad, "Speech Recognition and Verification Using MFCC & VQ", International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue-7, May 2013.

[7] Vikas C. Raykar, S. R. Mahadeva, "Speaker localization using extraction source information in speech". IEEE Transactions on Audio and Speech Processing, vol.13, September 2005.

[8] Bishnu S. ATAL, and L. R. RABINER, "A Pattern Recognition Approach to Voiced—Unvoiced—Silence Classification with Applications to Speech Recognition", IEEE transaction on acoustics, speech, and signal processing, VOL. ASSP-24, NO. 3, JUNE 1976.

[9] H.K. Palo, Mihir Narayan Mohanty, "Classification of Emotional Speech of Children Using Probabilistic Neural Network", International Journal of Electrical and Computer Engineering (IJECE) Vol. 5, No. 2, April 2015, pp. 311~317 ISSN: 2088-8708.

[10] L. Rabiner, "A tutorial on Hidden Markov Model and selected applications in Speech Recognition", Proceedings of the IEEE, pp 257-286, vol. 77, No. 2, 1989

[11] Garima Vyas, Malay Kishore Dutta, "An Integrated Spoken Language Recognition System Using Support Vector Machines", CONFERENCE PAPER, DOI:10.1109/IC3.2014.6897156.

[12] B.P.Das, R. Parekh, "Recognition of Isolated Words using features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers", International Journal of Modern Engineering Research, pp. 854-858, vol. 2, No.3, June 2012.