Automatic Speech Recognition Technique for Bangla Words

*Md. Akkas Ali[1], Manwar Hossain[2] and Mohammad Nuruzzaman Bhuiyan[1]

[1]Lecturer, Department of CSE & IT,
University of Information Technology & Sciences (UITS),
Baridhara, Dhaka-1212, Bangladesh
[2]Software Engineer
akkas.buet@gmail.com, manwar.cuet@gmail.com, mdnuruzzaman2001@yahoo.com

Abstract

Automatic recognition of spoken words is one of the most challenging tasks in the field of speech recognition. The difficulty of this task is due to the acoustic similarity of many of the words and their syllabi. Accurate recognition requires the system to perform fine phonetic distinctions. This paper presents a technique for recognizing spoken words in Bangla. In this study we first derive feature from spoken words. This paper presents some technique for recognizing spoken words in Bangla. In this work we use MFCC, LPC, GMM and DTW.

Keywords: *Feature extraction, speech recognition, framing, overlapping, hamming window.*

1. Introduction

The technique for automatic speech recognition varies for working language. Every language has its own style of pronunciation, i.e. ‘Khabor’ and ‘Kobor’ are very close to each other in uttering. However the meanings are very far-apart. This pronunciation difference cannot be tolerated in most of the languages i.e. French because controlling the vibration of vocal cord and movement of lips differ from language to language [1]. However in Bengali based on locality, for the same word a notable tolerance limit can be seen. So, different approach is demanded for speech recognition in Bengali language.

The main task of this paper is to recognize Bengali words through speech recognition technique of our proposed model. Initially we analyze the set of speech and extract essential features based on signal processing concept. From these features we have done parametric representation, mathematical model, and signal flow diagram for stepping towards our desired result. We used appropriate technique to calculate accurate distance between feature and reference matrix [2]. Consequently we represent four speech recognition models and compare them showing recognition rate (%) and elapsed time (sec). Finally we observed that from one of our models we can get highest rate of perfection for a set of Bengali words.

2. Objectives

- An efficient feature extraction method will propose to recognize Bangla isolated speech.
- Measure the success rate of the proposed model.
- Create a Bangla isolated speech recognizer using that feature extraction method.

3. Methodology

The process of speech recognition system typically consists of two phases:

a) Training and

b) Recognition

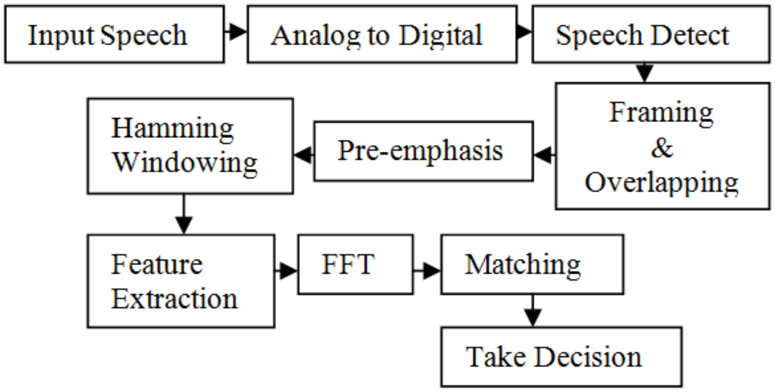


Figure 1. Models of Speech Recognition Process

In training phase the speaker has to provide sample of his voice so that the reference template model or database can be built. In training phase we record each word for 3 sec duration and 10 times. So for 100 words we record 1000 times. Then we extract the features of these 1000 words and save into a database [3].

The recognition phase ensures the inputted test voice is match with stored reference template model and recognition decision is made.

3.1. Input Speech

We can divide human’s speech into two types, voiced and unvoiced. Voiced sounds are produced by the vibration of vocal cords. On the other hand unvoiced sounds are not produced by the vibration of vocal cords.

3.2. Analog to Digital

As a computer is a digital system, so we cannot process an analog signal by a computer. So we need to digitize the input signal which takes via microphone.

3.3. Framing and Overlapping

When we analyze audio signal, we need to convert the signal into a set of frames, because audio signals are more or less stable within a short period of time, say 20 ms or 30 ms. if the frame duration is too big, we cannot catch the time-varying characteristics of the audio signal [4]. On the other hand, if the frame duration is too small, then we cannot extract valid acoustic features. In our project we take 20 ms or 160 sample point as a frame. Usually the overlap is 1/2 to 1/3 of the original frame [5]. The more overlap, the more computation is needed. Here we overlap 1/2 of original frame that is 80 overlapping frames.

3.4. Speech Detect

We record each word for 3 seconds. But total 3 seconds do not contains voice activity, they contains some silence or background noise. We need to suppress those silence parts and take only speech signal part. There are a lot of algorithms for speech detection and the scientists till now researching on it for effectively detect voice activity within a signal. Here we use an algorithm of voice activity detection [6]. This algorithm calculates the energy and number of zero crossing rate in a frame, compares these values with a threshold to determine if we have a possible voice activity or not. If either energy or zero crossing rates exceed the threshold, it continues analyzing frames and start buffering. If number of contiguous frames does not exceed buffer length (10 frames), we have a false alarm. If the number of contiguous frames exceeds buffer length, we have detected a word. An example is given in following figure, which shows how this algorithm detects a speech signal for 'পৃথিবী':

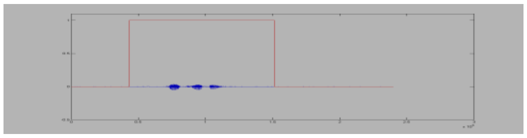


Figure 2. Spectrum of a Bangla Word 'পৃথিবী'

3.5. Pre-emphasis Filter

The filter with range of 1 to 0.97 is called pre-emphasis filter. It is a high pass filter, which removes the DC component of the signal [7]. DC component is the constant or zero frequency. Let us consider a signal $A + B\,\sin(2\pi f t + \theta)$; here A is the DC component of that signal. The speech after pre-emphasis sounds sharper with a smaller volume. The following figure shows the effect of pre-emphasis filter.

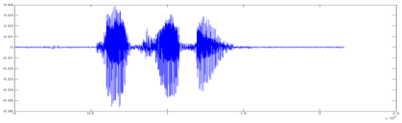


Figure 3. Before Pre-emphasis Spectrum of 'পৃথিবী'

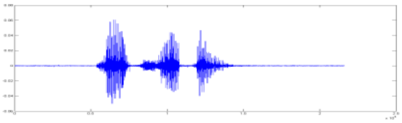


Figure 4. After Pre-emphasis Spectrum of 'পৃথিবী'

3.6. Hamming Window

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. It removes the ripples of the signal, which ripples found after Fourier Transform.

$$Ham\_Window = 0.54 - 0.46 * \cos\left(2 * \pi * \frac{0:160-1}{160}\right); \textit{here } 160 \textit{ is frame size}$$

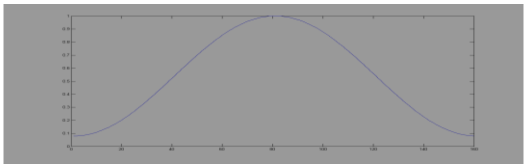


Figure 5. Hamming Window

The effect of applying hamming window on the pre-emphasis signal of 'পৃথিবী' shows bellow:

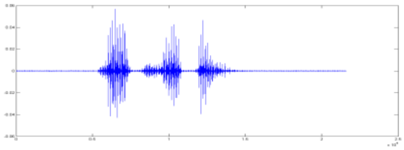


Figure 6. Spectrum of 'পৃথিবী' after Passing Hamming Window

3.7. Fast Fourier Transform

Fourier transform decompress a signal into it's constitute frequencies. Or simply it transforms a signal from time domain to frequency domain. Let us consider a 20 ms signal y, which consists of two frequencies 100 hz and 180 hz. $y = 0.7 * \sin(2 * \pi * 100 * t) + \sin(2 * \pi * 180 * t)$; After Fourier Transform we get the following figure (discarding imaginary values):

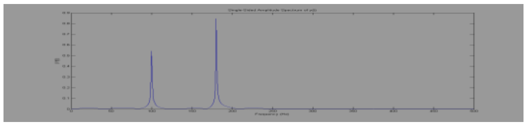


Figure 7. Single-sided Amplitude Spectrum of y

Our voice signal is not a periodic signal. It contains many frequencies. In different time interval we can get different frequencies. Example let us consider a 0.25 sec in which 5 types of frequencies remain constant in first 0.10 sec, 2 types of frequencies comes in next 0.05 sec and also 6 types of frequencies comes in next 0.10 sec [8]. In this case if we apply Fourier transform on the full signal at a time; then there is a chance of losing some frequencies. For that we need to divide the whole signal into some parts (frames) and Transform on each frames of the speech signal of 'পৃথিবী' we get the following Figure 8:

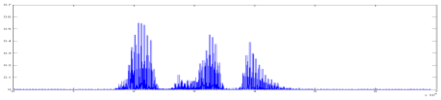


Figure 8. Short Term Fourier Transform on the Spectrum of 'পৃথিবী'

3.8. Feature Extraction

Feature extraction means converts the speech waveform to some of type of parametric representation. This parametric representation is then used for further analysis and processing.

The features of speech signal are amplitude of the signal, energy, intensity, velocity, acceleration, vibration rate, fundamental frequency etc. Feature extraction is process of obtaining different features such as power, pitch, and vocal tract configuration from the speech signals [9]. In this project for feature extraction we use MFCC, GMM and LPC model.

3.8.1. MFCC

The frequency range in FFT spectrum is very wide and voice signal does not follow the linear scale. The mel-filter bank is a nonlinear filter. The filter bank is constructed using 13 linearly-spaced filters (133.33Hz between center frequencies,) followed by 27 log-spaced filters (separated by a factor of 1.0711703 in frequency) [10]. Each filter is constructed by combining the amplitude of FFT. The 40 filters have this frequency response:

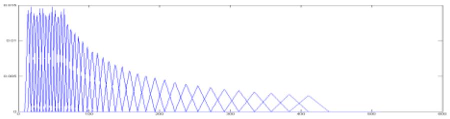


Figure 9. Mel-scale Filter Bank for 40 Filters

Each frame is converted to 12 MFCCs plus a normalized energy parameter (total 13 coefficients). The first and second derivatives, (13 delta (Δ´s) or velocity features and double delta (Δ Δ´s) or acceleration features) of MFCCs and energy are estimated, resulting in 39 numbers representing each frame.

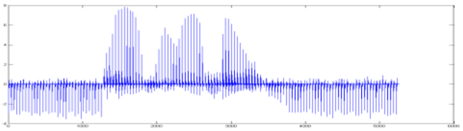


Figure 10. Coefficients Found for the Word 'পৃথিবী'

3.8.2. Gaussian Mixture Model

Gaussian Mixture model calculate some sets of mean and variant of signal frequencies by Expectation Minimization (EM) algorithm. Mean is the average, which calculating mean frequency of the power spectrum. The variance is used as a measure of how far a set of numbers are spread out from each other or simply describe how far the numbers lie from the mean.

3.8.3. Linear Predictive Coding

Linear Predictive coding (LPC) is defined as a digital method for encoding an analog signal in which a particular value is predicted by a linear function of the past values of the signal. An alternate explanation is that linear prediction filter attempt to predict future values of input signal based on past signal [11]. It represents a signal by an equation with a given number coefficients. The sound produced by the vibration of vocal cords is called voiced sound, this kind of sounds are quasi-periodic sound. Unvoiced sounds are not produced by the vibration of vocal cords, this kind of sounds are aperiodic sound. The quasi-periodic sound waveform consists of similar repeated patterns. For that during unvoiced and transient region of speech, the LPC model is less effective than for voiced region.

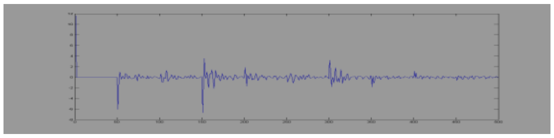


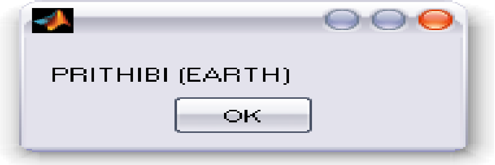
Figure 11. LPC Coefficients Found for the Word 'পৃথিবী'

3.9. Matching

Template matching is the simplest technique and has the highest accuracy when used properly, but it also suffers from the most limitations. There are a lot of techniques used for matching feature matrix and reference matrix [12]. In this project we use dynamic time warping, posterior probability function and Euclidian distance for matching or measuring distance between feature matrix and reference matrix.

3.10. Take Decision

We calculate the distance between feature matrix and all reference matrixes. Then we consider which reference matrix is close to the feature matrix. We consider the minimum distance for taking decision. For that the recognizer found minimum distance for input signal with পৃথিবী, and then it will show this message box:



PRITHIBI [EARTH]

Figure 12. Output Message

4. Results

We represent four speech recognition models. The implementation results of four models are discussed below:

4.1. Model 1

In this model we use Mel Frequency Cepstral Coefficient for extract feature and Dynamic Time Warping for matching.

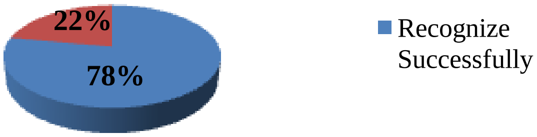


Figure 13. Recognition rate of model 1

4.2. Model 2

In this model we use Linear Predictive Coding and calculate linear predictive coefficients for extract feature and Dynamic Time Warping for matching.

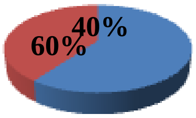


Figure 14. Recognition rate of model 2

4.3. Model 3

In this model we use Mel Frequency Cepstral Coefficient and Gaussian Mixture Model for extract feature and posterior probability function for matching.

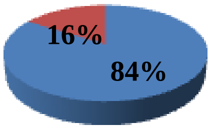


Figure 15. Recognition rate of model 3

4.4. Model 4

In this model we use Mel Frequency Cepstral Coefficient and compress it by Linear Predictive Coding for extract feature and Dynamic Time Warping for matching.

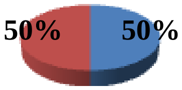


Figure 16. Recognition rate of model 4

4.5. Comparison between Four Models

If we compare four models together we found the following graph:

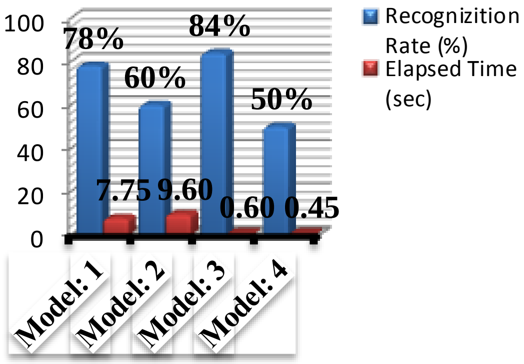


Figure 17. Graphical Comparison between Four Models

5. Conclusion

Speech recognition is a very challenging field of digital signal processing. Scientists achieved remarkable success in speech recognition of many languages. In English continuous speech can be recognized with accuracy rate more than 95%. Unfortunately in bangla, works on speech recognition is in the very preliminary stage. In this report we discussed how a spoken word can be recognized and proposed four models for bangla speech recognition system. Speech detection and recognition systems are extremely sensitive to different laptops, different microphones and different environment. Our speech recognizer can recognize highest 84% rate for one hundred bangla word in normal room environment. By ensuring the perfection (i.e. noise free) of the recorded signal it is possible to increase the accuracy of recognition. Hope our effort will help to take the research on speech recognition one step toward continuous speech recognition in bangla and encourage further development in this interesting field.

References

[1] Dudley, "The Vocoder", Bell Labs Record, vol. 17, **(1939)**, pp. 122-126.

[2] A. Hasanat, Md. R. Karim, Md. S. Rahman and Md. Z. Iqbal, "Recognition of Spoken Letters in Bangla", Proceedings of 5th International Conference on computer and Information Technology (ICCIT), Dhaka, Bangladesh, **(2002)**.

[3] http://en.wikipedia.org/wiki/Bengali_language.

[4] G. Muhammad, Y. A. Alotaibi and M. N. Huda, "Automatic Speech Recognition for Bangia Digits", Proceedings of 2009 12th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, **(2009)** December 21-23.

[5] A. K. M. M. Hoque, "Bengali Segmented Automated Speech Recognition", Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh, **(2006)** May.

[6] E. H. Bourouba, M. Bedda and R. Djemili, "Isolated Words Recognition System Based on Hybrid Approach DTW/GHMM", Informatics, vol. 30, **(2006)**, pp. 373–384, Algeria.

[7] M. P. Kesarkar and P. Rao, "Feature Extraction for Speech Recognition", Credit Seminar Report, Electronic Systems Group, EE. Dept., IIT Bombay, **(2003)** November.

[8] R. Jang, "AudioSignalProcessing", http://mirlab.org/jang/books/audioSignalProcessing/idex.asp.

[9] A. Klautau, "The MFCC", **(2005)** November 22.

[10] P. Lama and M. Namburu, "Speech Recognition with Dynamic Time Warping using MATLAB", CS 525, SPRING 2010 – PROJECT REPORT, **(2010)**.

[11] P. Senin, "Dynamic Time Warping Algorithm Review", Information and Computer Science Department, University of Hawaii at Manoa, Honolulu, USA, **(2008)** December.

[12] L. Muda, M. Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", journal of computing, vol. 2, Issue 3, **(2010)** March, ISSN: 2151-9617.

Authors

Md. Akkas Ali is working as a Lecturer at the Department of CSE & IT, University of Information Technology & Sciences (UITS), Baridhara, Dhaka-1212, Bangladesh. I completed my B.Sc Engg. in CSE from Chittagong University of Engineering and Technology (CUET), Chittagong-4349, Bangladesh. My M.Sc Engg. in CSE is running at Bangladesh University of Engineering and Technology (BUET), Dhaka-1000, Bangladesh. My research interest areas are the image processing, Computer Networks, Computer Networks and Data Security, Compiler, Theory of Computations, etc. My several papers accepted in International Journals.

Manwar Hossain is working as Software Engineer at NNS-Solution Ltd. Dhaka, Bangladesh. I completed my B.Sc Engg. in CSE from Chittagong University of Engineering and Technology (CUET), Chittagong-4349, Bangladesh.

Mohammad Nuruzzaman Bhuiyan is working as a Lecturer at the Department of CSE & IT, University of Information Technology & Sciences (UITS), Baridhara, Dhaka-1212, Bangladesh. I completed my B.Sc Engg. and M.Sc Engg. in CS from The University of Sheffield, Western Bank, Sheffield S10 2TN, UK.