**Final Project Report – Bengali Vowel Recognition**

| | |
|---|---|
| Name | Md Golam Tanvir Zim |
| Student No | 180715305 |
| Course | Real Time DSP (ECS732P) |

## Introduction

The aim of the project was to recognize Bengali vowels by Beaglebone Black board using the Bela platform. Bengali is the 8th most spoken language in the world. This is the national language of Bangladesh and widely spoken in West Bengal, Assam and Tripura in India. This project was attempted keeping in mind that there is an increasing demand for Bengali Speech Recognition across the region.

The most notable work done on Bengali Speech Recognition so far is Shruti II [1]. It's a SPHINX3 based computer application, which converts continuous Bengali speech to Unicode.

There are seven linguistic vowels in Bengali language. These are অ, আ, ই, উ, এ, ও, অ্যা. The recognition of vowels usually requires two stages, first one is a signal processing step, in it, the distinguishable features of a vowel are extracted. The second stage is about classifying a feature using different types of algorithm. This stage uses the extracted feature vector of a vowel to decide which class it belongs to. In this project, the first stage was realized and it was done using Mel Frequency Cepstrum Analysis.

There are a number of techniques used to extract features of a vowel which makes it distinguishable from others, these are Mel Frequency Cepstral Coefficients(MFCC), Linear Predictive Coding(LPC), Perceptual Linear Prediction (PLP) etc. In this project, Mel Frequency Cepstrum analysis was used to extract the feature in real time. As for the classification of recognition of the vowel from the extracted features, there are multiple techniques as well, i.e. Hidden Markov Model (HMM), Dynamic Time Warping (DTW), Vector Quantization (VQ), Artificial Neural Network (ANN) are used, which was not implemented in this project.

## Design Process

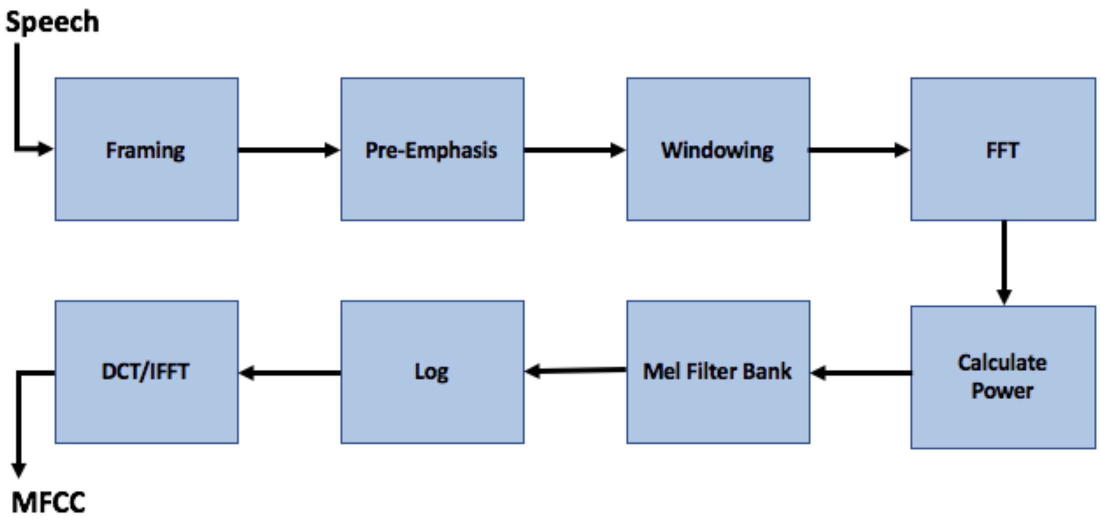
Speech can be modelled as modulated signal. In cepstrum based analysis, vocal tract shape is considered as a filter and its spectral peaks are considered as resonance and are commonly referred as formants[2]. Here speech signal is represented as,

$$s(t) = g(t) \otimes h(t)$$

Where the speech signal is a convolution product of excitation g(t) and vocal tract filter h(t).

The most fundamental block of a speech recognition system is extracting feature of the utterance. Features extraction in Automatic Speech Recognition is the computation of a sequence of feature vectors which provides a compact representation of the given speech signal [3]. In this project MFCC was calculated to create the features for each frame. The reason why Mel Frequency Cepstral analysis was used is this is computationally less expensive and easy to implement.

The standard procedure [4] for generating Mel Frequency Cepstrum Coefficients is as follows:



### Pre-emphasis

For automatic speech recognition, the sampled speech signal is passed through a low pass filter, to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing [5]. Usually, the following transfer function is is implemented to build the filter.

$$H(z) = 1 - az^{-1}$$

The difference equation is as follows:

$$y(n) = x(n) - a * x(n-1)$$

where 0.9<a<0.95

### Framing

Before performing Fourier Transform, the sampled input speech is segmented into small frames. The standard size of a frame is 20-40 milliseconds long which is statistically stationary, with 10 milliseconds hop/step size which is optional. If the file size can't be divided into an even number of frames, zero-padding is done after the end of file.

### Windowing

The segment of the input speech on which the Discrete Fourier Transform is performed is windowed as part of pre-processing of the speech. Fourier Transform is assuming the segmented discrete samples are periodic, in reality which is not, the end points are discontinuous resulting spectral leakage. Windowing is performed to get rid of that, it smooths the signal, by reducing the amplitude of the discontinuities. So, if the input signal is x(n) and window function is w(n), the windowed signal y(n) is,

$$y(n) = x(n) * w(n)$$

What essentially happens is each sample of input is multiplied with corresponding element of the window function. In this project, the hanning window was used. The hanning window function is as follows:

$$\mathrm{w(n)} = \ 0.5 * [1 - \cos\left(\frac{2 * \pi * n}{N}\right)]$$

Where N = Length of the window.

### Fast Fourier Transform (FFT)

The Discrete Fourier Transform of samples x(n) is as follows:

$$X(k) = \sum x(n) * \ e^{(-j*2*\pi*k*n/N)}$$

where k = 0,1,2,3 …n

FFT is an algorithm that rapidly computes the discrete Fourier transform (DFT) of a sampled signal. In this project, FFT was performed.

### Power of a framed speech

The Power spectral estimate of a speech frame is as follows:

$$\mathrm{P(k)} = \frac{1}{N}\sum_{k=0}^{n-1} |X(k)|^2$$

**Mel Frequency Spectrum and Mel Filter Bank**

MFCCs are one of the most popular feature extraction techniques used in speech. In Mel-Frequency Cepstral Coefficients (MFCC, a nonlinear frequency scale is used, which approximates the behaviour of the auditory system of human[6]. Additionally, these coefficients are robust and reliable to variations according to speakers and recording conditions.

To get how much energy is distributed in different frequency region, Mel Filter banks are used. Mel Filters are overlapping triangular windows. The length of each filter is based on Mel Scale. The relation between frequency (f) and Mel-frequency (M) is as follows:

$$M = 1125 * \ln{(1 + \frac{f}{700})}$$

The standard practice is to create 26 triangular filters. A triangular filter has lower and upper cut-off frequency, at which the value is 0 and at mid frequency, the value is 1. The next filter's lower frequency is same as the mid frequency of the first filter and so on. Therefore, if n-triangular filters are used, then n+2 points are needed to design all of them. In this project 26 triangular filter banks were used.

### Energy Distribution in Mel Filter Bank

After calculating the Mel Filter Banks, the power spectrum of framed each signal is multiplied with each filter bank and resulting coefficients are added up. So, in total 26 coefficients are generated which represents the energy distribution across the Mel frequency spectrum.

### Log and DCT

The logarithmic function is applied to 26 coefficients log Filter bank energies are obtained. The final step is Discrete Cosine Transform (DCT), which is performed over 26 coefficients. The purpose of DCT is to distinguish higher order and lower coefficients. Because the higher order represent the excitation signal, on the other hand, the lower order coefficients represent the vocal tract shape[7].

DCT II type was used in DCT operation. The DCT II is as follows:

$$X(k) = x(n) * \cos{(\frac{\pi}{N}(n + 0.5)k)}$$

From the resultant 26 DCT coefficients, 2nd to 13th coefficients are kept, as they contain necessary feature information and others are discarded. These are the desired MFCCs.

## Implementation

In this section, the details of implementation of the above-mentioned design will be described.

1) The Bela platform has three sampling frequency for audio input - 22.05kHz, 44.1kHz and 88.2kHz. In this project, 22.05kHz was used. It means, in 1 second, 22050 samples were taken as input.

The buffer or window size was limited to 512 samples, so the time duration of a frame = (512/22050) seconds = 23.22 milliseconds, which falls between the standard range 20~40 milliseconds. No overlapping was used.

2) The value of a in low pass filter was 0.95, hence the difference equation is,

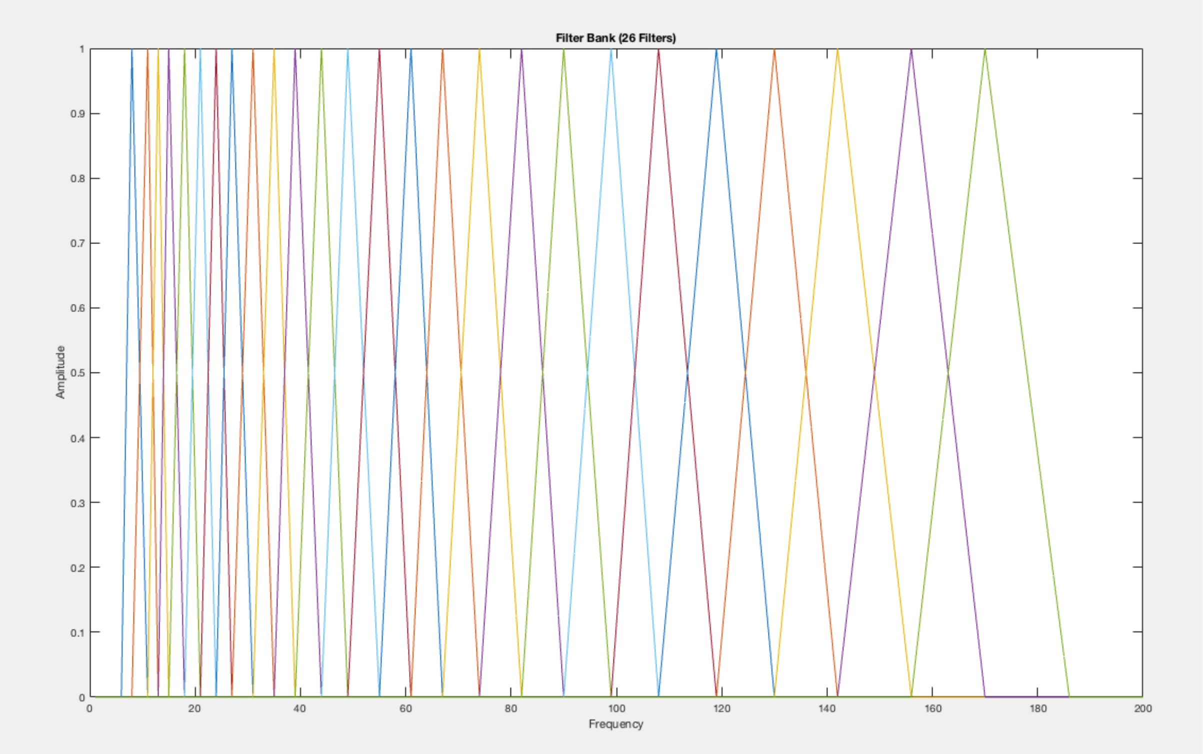
$$y(n) \;=\; x(n) - 0.95 * x(n)$$

3) Whenever 512 samples were taken, an auxilary thread `gFFTTask` was run to perform the FFT. This thread's task was to execute the function `process_fft_background()` which called the function `process_fft()` in which the FFT, power calculation, Filter Energy and DCT were performed.

4) The function FFT was performed using the function `ne10_fft_c2c_1d_float32_neon()`, which is a function of Ne10 Library. This function is a complex to complex FFT function. The complex FFT has two parts, one is real and another is imaginary. As the FFT is symmetric, half of the window size, i.e. 257 coefficients were used for power calculation of the signal. The power was calculated using the following formula.

$$P(k) = \frac{\sqrt[2]{Re(X(k))^2 + Im(X(k))^2}}{N}$$

5) There were 26 Mel Filter used in this project. The lowest and uppermost frequency were 300Hz and 8000Hz. Converting these values into Mel Scale, the values are 401.25 Mel and 2834.99 Mel. Between these two Mel points, 26 additional points were equally spaced. Then these 28 Mel points are converted back into normal frequency range and rounded to floor. The following Filterbank was obtained in Matlab.



6) Filter Energy was calculated for each filter by multiplying the Power of the signal with one each filter bank and the coefficients were summed up. Thus total of 26 coefficients were obtained. Then DCT was performed on these 26 coefficients.

7) The FFT and subsequent signal processing was done using an auxiliary thread. The purpose of this was to do that in real time. The FFT and subsequent signal processing was performed for a block of 512 samples, not each sample itself. Without the thread, the whole calculation couldn't be performed in (1/22050) seconds.

Most of the render functions did almost nothing except taking the audio sample, because only when a block of 512 samples were formed, then the FFT was performed, so other 511 render call were almost idle.

The threading solves it. The render function has got the highest priority and the gFFTask was of secondary priority. So, if the gFFTTask was not completed before the next render function is called, the OS will perform the rest of the task of gFFTTask after finishing the render task which is the higher priority.

## Evaluation

To create a class for each vowel, a model needs to be trained using lot of samples which couldn't be done. In terms of outcome, it was not possible to verify whether the signal processing algorithm was functioning properly.

The project ended up in generating 12 MFCC for each sample in real time, no underrun occurred.

## References

[1] Sandipan Mandal, Biswajit Das and Pabitra Mitra. Shruti-II: A vernacular speech recognition system in Bengali and an application for visually impaired community. *IEEE Students Technology Symposium (TechSym).* Kharagpur, 2010.

[2] Tonmoy Ghosh, Subir Saha and A.H.M Iftekharul Ferdous. Formant Analysis of Bangla Vowel for Automatic Speech Recognition. *Signal & Image Processing: An International Journal (SIPIJ) Vol.7, No.5.* August 2016.

[3] Namrata Dave. Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition. *International Journal for Advance Research In Engineering And Technology.* Volume 1, Issue VI, July 2013

[4] Mel Frequency Cepstral Coefficient (MFCC) tutorial. 2013. http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[5] Thiang and Suryo Wijoyo. Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot. *International Conference on Information and Electronics Engineering IPCSIT vol.6.* 2011, Singapore.

[6] Sayf A. Majeed, Hafizah Husain, Salina Abdul Samad, Tariq F. Idbeaa. Mel Frequency Cepstral Coefficients (Mfcc) Feature Extraction Enhancement In The Application Of Speech Recognition: A Comparison Study. *Journal of Theoretical and Applied Information Technology, Vol.79. No.1.* September 2015.

[7] P. Ehkan, F.F. Zakaria, M.N.M. Warip, Z. Sauli and M. Elshaikh. Hardware Implementation of MFCC-Based Feature Extraction for Speaker Recognition. Lecture note - *https://www.researchgate.net/publication/268508674*. November, 2015.

[8] J.W. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE, Volume: 81, Issue: 9.* Sep 1993.