Phoneme Based Bangla Digit Recognition

Tanvir Anjum

201606115

Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology

tanvir1167052@gmail.com

Mehedi Hasan Emon

201606113

Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology

emon00713@gmail.com

Ataher Sams

201606105

Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology

asnsamsniloy@gmail.com

Khyrun Nesa Neesa

201606124

Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology

khyrunneesa3@gmail.com

Abstract:

This research work aims at recognizing the Bangla digit from 0-9 by separating the phonemes of the digits. The ambiguity in phonemes in Bangla speech is more extreme and varied than that of English speech since Bangla stems from the "Indo-European language family". In this research, dataset is collected in manual fashion. Then, the digits are broken into their respective phonemes (both voiced and unvoiced part). Next, Mel frequency Cepstral Coefficients (MFCC) features of the segmented digits are calculated and trained using artificial neural network, which is used to recognize the unknown digit. The proposed system is implemented through MATLAB and accuracy of the system averages above 90%.

Keywords:

BANGLA SPEECH RECOGNITION, MFCC, ARTIFICIAL NEURAL NETWORK, BANGLA DIGITS, PHONEME .

Introduction:

Research in Automatic Speech recognition by machine has been done for almost four decades.[1] Though Bangla has approximately 260-300 million total speakers worldwide , speech recognition is not done as much as it should be. The phonemes of different language of the world is different, so a system for another language will not provide accurate result for Bangla language. So a system unique to Bangla language is designed in the paper, which can accurately predict the digit uttered which can be useful for implementing this system for further development in this field like predicting any isolated or connected word of Bangla language. It can have a huge impact for people with disabilities and can eliminate the language barrier between people. It can map human speech to text or commands or subtitling or indexing video recordings and streams, speech translation, language learning etc.

In this research study, an effort is made to develop such a system using phonemes which consists of both Voiced and unvoiced part from the segmented digits to extract significant features from MFCC and to train the feature vectors using artificial neural network to recognize the unknown digit.

Abbreviations and Acronyms

- MFCC - Mel frequency Cepstral Coefficients
- FIR - Finite Impulse Respose
- FFT - Fast Fourier Transform
- DCT - Discrete Cosine Transform
- IFFT - Inverese Fast Fourier Transform

Units

- Hertz (Hz) -it is the unit of frequency
- Watts (W) - it is the unit of power; the rate of change of energy per second
- Miliseconds (ms) - refers to 0.001 seconds

Methodology:

Our work began with collecting test data from different persons to build up a useful data set. We inserted the data by sorting and labeling those from 0 to 9 accordingly . Thus we have built up our data set which is consisted of data from around 100 different people (including both male and female vocal).

Now that we had our data set ready, we started to write MATLAB program to implement our project. At first we linked the data set to the m.file by writing MATLAB code, we have done this in such a way that the ten folders (labeled 0 to 9) and their corresponding data files path and location can be traced. We first resampled the data just to make sure the sampling rate remains the same throughout the data set. Followed by resampling, we have selected an audio channel from the dual audio channel to avoid interference. After that we have started framing the data set .We have used an FIR filter (coefficients [1 -0.95]) . This FIR filter processing has been done to pre-emphasize the data, ensuring system stability and noise attenuation. Each audio file in the data set has been time framed within a window of 25ms of which only 15 ms was set to overlap.

Now that we had divided each audio record in a 25ms window with 15ms overlap; we determined the power of each window. Then we observed the power spectrum of each window . We also had to know the CVC pattern for 0 to 9 bangla digit which are as follows:

Bangla Digits	Pronunciation	Phoneme Type
০ (শুণ্য)	Shunno	CVCV
১ (এক)	Ek	VC
২ (দুই)	Dui	CV
৩ (তিন)	Tin	CVC
৪ (চার)	Char	CVC
৫ (পাঁচ)	Pach	CVC
৬ (ছয়)	Choy	CV
৭ (সাত)	Sat	CVC
৮ (আট)	Aat	VC
৯ (নয়)	Noy	CV

Figure 1: ***CVC pattern of bangla digit 0-9***

We subtracted the first and last portion of the frame below 30% max power and considered the rest of the part as the voiced part . The duration of 56ms before and after the voiced part duration, we considered as unvoiced potion of the audio signal. Also 265ms around the highest peak has been taken for the voiced part. These three CVC portions of each audio signal has been taken separately to extract features. A short note is that, we have considered “Shunno” to have the CVC pattern.

After that we have started to extract features using MFCC.

Mel Frequency Cepstral Coefficient (MFCC)

In any automatic speech recognition the first step is to extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc.

The main point to understand about speech is that the sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope. This page will provide a short tutorial on MFCCs.

Mel Frequency Cepstral Coefficents (MFCCs) are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since[2].

Steps in MFCC Implementation

1. Frame the signal into short frames.
2. For each frame calculate the periodogram
3. estimate of the power spectrum.
4. Apply the mel filterbank to the power spectra, sum
5. the energy in each filter.
6. Take the logarithm of all filterbank energies.
7. Take the DCT of the log filterbank energies.
8. Keep DCT coefficients 2-13, discard the rest.

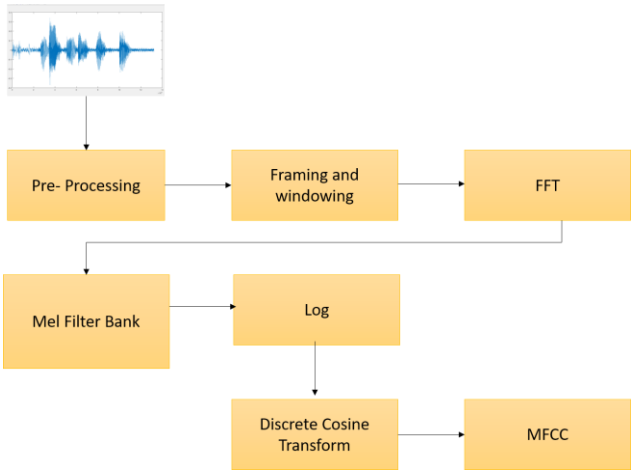

Figure 2: ***MFCC Flow chart***

Since an audio signal is constantly changing, so to simplify things we assume that on short time scales the audio signal doesn't vary much (when we say it doesn't vary, we mean statistically i.e. statistically stationary, obviously the samples are constantly changing on even short time scales). This is why we frame the signal into 20-40ms frames. If the frame is much shorter we don't have enough samples to get a reliable spectral estimate, if it is longer the signal changes too much throughout the frame.

What is the Mel scale?

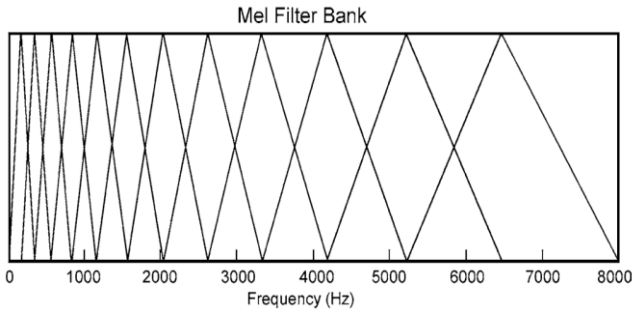

Figure 3: ***Mel filter bank [3]***

The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely what humans hear [4].

The formula for converting from frequency to Mel scale is:

$$M(f) = 1125 \ln(1 + f/700) \qquad (1)$$

To go from Mels back to frequency:

$$M^{-1}(m) = 700(\exp(m/1125) - 1) \qquad (2)$$

We considered only first 14 coefficients of MFCC to extract features. The first 14 Mel filter bank list is given below-

Table 1: ***Coefficients of MFCC[5]***

Filters	Passband Edges (Hz)
Filter 1	[133 267]
Filter 2	[200 333]
Filter 3	[267 400]
Filter 4	[333 467]
Filter 5	[400 533]
Filter 6	[467 600]
Filter 7	[533 667]
Filter 8	[667 800]
Filter 9	[533 667]
Filter 10	[733 867]
Filter 11	[800 933]
Filter 12	[867 999]
Filter 13	[933 1071]
Filter 14	[999 1147]

After using those fourteen Mel filter banks and frequency wrapping we have observed the following figures i.e for voiced part to differentiate features among different digits and also tried to analyze the reason behind the difference- i.e.Here x-axis is the time axis and each row from 1to 14 are different filters output-

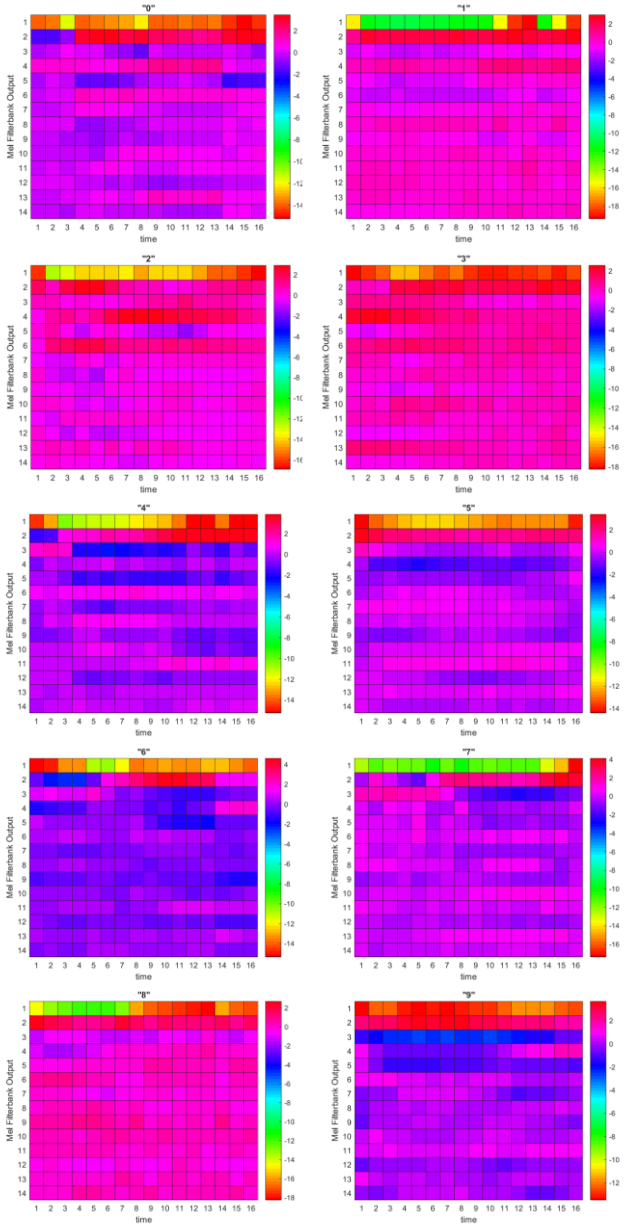

Figure 4: ***Heat map of Bangla digit 0 to 9***

From the heat map, it gives us the mel filter output spectrum vs time graph. 1-3 no are the unvoiced region, 4-13 are the voiced region and 14-16 are again unvoiced region. If looked closely, it can be seen that there are characteristic strips in filter no 3,5,9 and 12 for Bangla digit 4(char),5(pach) and 7(sat) at the beginning of the voiced part. The 6(choy) and 9(noy) digits' features also show similarity in different voiced strips. Again there are distinctive difference between two(dui) or three(tin) and four(char) or five(pach). We have extracted 1104 features(14*16) in total. Then we formed feature vectors

and inserted them into MATLAB builtin Artificial Neural Network (ANN).

MATLAB Artificial Neural Network

An artificial neural network, often just called a neural network, is a mathematical (or computational) model that is inspired by the structure and function of biological neural networks in the brain. An artificial neural network consists of a number of artificial neurons which are connected to each other via synaptic weights. Desired continuous mapping or a desired task is acquired in an artificial neural network by learning.[6]. Its behavior is defined by the way its individual elements are connected and by the strength, or weights, of those connections. These weights are automatically adjusted during training according to a specified learning rule until the neural network performs the desired task correctly. We have used MATLAB built in artificial neural network to break down the input features into layers of abstraction. It had been trained over many audio data to recognize patterns in speech .

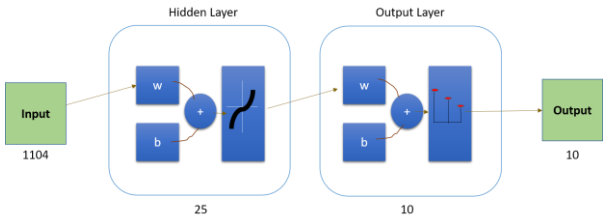

Figure 5*: Block Diagram of Artificial Neural Network*

The output of the neural networking is a function which is used to done resampling, audio channel selection, voiced unvoiced detection, features extraction, feature vector formation and vector input to the neural network once more. After following all the procedures, we get the desired output- the phoneme based speech recognition of Bangla digit from 0 to 9.

Result Analysis

When we train the dataset in the ANN, the data is analyzed there and similarities in features are trained, validated and tested accordingly. Looking at the outcome below, we find the error percentage in those parts to be less than 15%.

Results

	Samples	CE	%E
Training:	1905	2.14245e-0	4.14698e-0
Validation:	408	6.24737e-0	13.97058e-0
Testing:	408	6.23408e-0	14.46078e-0

Figure 6: *Output of analyzed data in the ANN*

The final outcome of our project is Bangla 0 to 9 phoneme based speech recognition and we have become pretty much successful to get the desired outcome.

	শুণ্য (0)	এক (1)	দুই (2)	তিন (3)	চার (4)
Test 01	99.96	99.70	99.16	99.03	99.95
Test 02	99.91	98.6	97.95	75.98	99.26
Test 03	99.99	99.8	99.95	85.85	93.31
Average	99.95	99.7	99.68	86.95	97.5

	পাঁচ (5)	ছয় (6)	সাত (7)	আট (8)	নয় (9)
Test 01	99.14	83.56	99.48	99.7	99.98
Test 02	97.87	86.78	83.87	87.8	89.80
Test 03	82.15	82.55	99.80	99.3	95.92
Average	93.05	84.29	94.38	95.60	95.23

Figure 7: *Recognition Accuracy for digits*

Pronunciation of “NOY” and “CHOY” are pretty close, so the accuracy for them sometimes detoriates. Again as the voiced part of “PACH”, “SAT”, “AAT” is same, unvoiced part of those digits helped to get the accurate result. It is clear from the final output chart that the accuracy of the system averages above 90% for almost all of them which is quite satisfactory.

Limitations

- If we have collected more data, the accuracy level will have also increased.
- Our dataset is comprised of audio data of people of a certain age. There are no audio data of children and old persons.So our designed system may not work accurately for old people and children.
- Also the system is designed maintaining pretty much ideal circumstances like – audio recording in almost noiseless situation, audio inputs which are the bangla digits were pronounced accurately as bangla grammar suggests,so almost no local pronunciation exists.In that case the system may not work accurately for person using local pronunciation.

Conclusion:

Neural networks behavior is defined by the way its individual elements are connected and by the strength, or weights, of those connections. So change in network architecture can bring in significant difference in output result. As in the paper we use MFCC more than once to find the features of the voiced and the unvoiced parts, we get a quite satisfactory result of above 90% accuracy. This work can be further implemented to recognize all different kinds of words of Bangla language.

ACKNOWLEDGMENT

A special thanks goes to our mentors- **Professor Dr. Shaikh Anowarul Fattah** and **Sadman Sakib Ahbab Jarif.**

References

[1] Rabiner, L., Juang, B. H., Yegnanarayana, B., 2010, *"Fundamentals of Speech Recognition"*, Second Edition, Pearson Education.

[2] S. Molau, M. Pitz, R.S. Uter, H. Ney, *"Computing Mel-Frequency Cepstral Coeffients on the power spectrum"* in Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on, Volume: 1

[3] Minh N. Do, *"An Automatic Speaker Recognition System"*, Audio Visual Communications Laboratory, Swiss Federal Institute of Technology, Lausanne, Switzerland, 2001.

[4] Navin K Manaswi, (2018) Speech Intelligence for Dummies. *Deep Learing with Applications Using Python*, Apress; 1st ed. edition

[5] Mathworks® MFCC(R2019b), Retrieved from https://www.mathworks.com/help/audio/ref/mfcc.html

[6] Kenji Suzuki, (2013), *Artificial Neural Networks-Architectures and Applications,* Janeza Trdine 9,1st ed. edition