Bangla Voice Based Text Independent Speaker Recognition Using Machine Learning and CNN: A Comparative Model Performance Analysis

Toiyob Hossain[1], Md.Jahidul Hoq Emon[2], Mir Jubaier Hassan[3], Abu Saleah[4], and Redwan Hossen[5]

[1]1706013@buet.ac.bd
[2]1706017@buet.ac.bd
[3]1706015@buet.ac.bd
[4]1706011@buet.ac.bd
[5]1706019@buet.ac.bd
[1,2,3,4,5]Department of EEE, Bangladesh University of Engineering and Technology

July 30, 2021

Abstract

Bangla Voice Based Text Independent Speaker Recognition is the process of identifying a person using his or her voice. Speaker recognition is a very difficult task because different speaker speaks different language and styles.
In this paper, the main goal was to develop a Bangla voice based attendance system and to compare with various models that can give a good accuracy. The experiment was performed using the database of voice of EEE 17A almost 5 minutes of recorded voice of reading an essay. As we did not have enough data for perfect CNN architectur, we had to augment the data to increase data size for better accuracy.
This project is based on feature extraction and neural networks. We have also analyzed three different features.They are MFCC, Spectrogram and Mel-spectrogram. We have also tested two different models: SVM and CNN.
Index Terms: Bangla Voice Based Text-independent Speaker Recognition, CNN, MFCC, SVM, Spectrogram, Mel-spectrogram.

1 Introduction

Identifying speakers using a specific language is an emerging area in speech processing. All device are giving priority to English speakers. However, no model can give a 100 percent accuracy for a speaker across all languages. Mystical nature of speech is responsible for this. The way a speaker talks can vary significantly from time to time and depends on lots of factors like language spoken, emotion, environment and nature of listener. This makes speaker recognition a very challenging task.
We have used "Augly" an open source library to augment the speaker's data. We have also extracted three audio features for this experiment. The features are: MFCC, Spectrogram, and Mel-spectrogram. All of these features used 256 points DFT and the same length windowing. Hanning window was used for better accuracy. Though there are lot of training models available,we have used two models: CNN and SVM. Convolutional Neural Networks and Support Vector Machine usually require a large amount of data for predicting speaker properly. However, from the confusion matrics of both model, we observed that MFCC with SVM model works better for this dataset. The description of database is shown below:

Items	Levels	Male	Female
Age	21-23	21	6

Total Speaker	Total audio files	Total Train Data	Total Test Data
27+1noise	2800	2240	569

2 Proposed Method

Although there are different techniques for speaker recognition, matching extracted features with prebuilt database is chosen in this

experiment.At first 3 features were chosen and then 2 models were trained and tested based on these features.

- **Feature Extraction :**

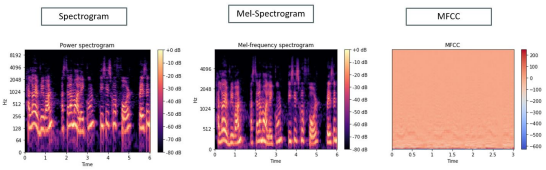
  MFCC: Mel Frequency Cepstral Co-efficients (MFCC) is an internal audio representation format which is easy to work on. This is similar to JPG format for images. MFCC stands for it is an acronym for Mel Frequency Cepstral Co-efficients which are the coefficients that collectively make up an MFC. MFC is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.
  Sound Spectrum : A sound spectrum is a representation of a sound – usually a short sample of a sound – in terms of the amount of vibration at each individual frequency. It is usually presented as a graph of power as a function of frequency.
  Linear Cosine Transform : They can also be called as sine and cosine tranforms which can be easily calculated using fourier transform. During our converion we will be needing both short time and fast time fourier tranform ar different stages.

  Power Spectrum : It is the result of fourier tranform which we get as a representation.Also known as periodogram.

  Mel Scale : Mel scale is a scale that relates the perceived frequency of a tone to the actual measured frequency. It scales the frequency in order to match more closely what the human ear can hear .
  A frequency measured in Hertz (f) can be converted to the Mel scale using the following formula :

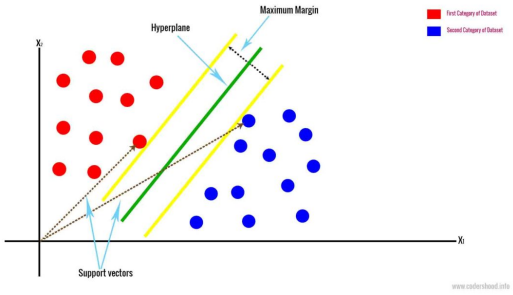
$$f_{mel} = 2595 \log_{10}(1 + \frac{f}{700})$$

  Mel-spectrogram : A mel-frequency spectrogram is related to the linear-frequency spectrogram, i.e., the short-time Fourier transform (STFT) magnitude. It is obtained by applying a nonlinear transform to the frequency axis of the STFT, inspired by measured responses from the human auditory system, and summarizes the frequency content with fewer dimensions. Using such an auditory frequency scale has the effect of emphasizing details in lower frequencies, which are critical to speech intelligibility, while de-emphasizing high frequency details, which are dominated by fricatives and other noise bursts and generally do not need to be modeled with high fidelity. Because of these properties, features derived from the mel scale have been used as an underlying representation for speech recognition for many decades.
  Visualization of these 3 features:



- **Chosen Models :**

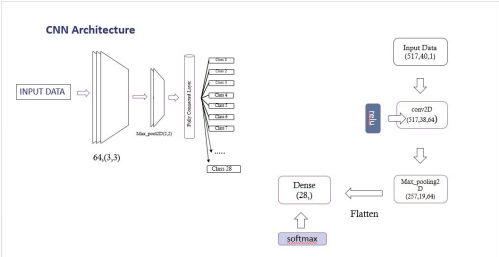
  Support Vector Machine: An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).



  SVM Kernel : In practice, SVM algorithm is implemented with kernel that transforms an input data space into the required form. SVM uses a technique called the kernel trick in which kernel takes a low dimensional input space and transforms it into a higher dimensional space. In simple words, kernel converts non-separable problems into separable problems by adding more dimensions to it. It makes SVM more powerful, flexible and accurate. The following are some of the types of kernels used by SVM.
  Linear Kernel can be used as a dot product between any two observations. The formula of linear kernel is as:
  K(x,xi)=sum(x∗xi)K(x,xi)=sum(x∗xi)

From the above formula, we can see that the product between two vectors say x and xi the sum of the multiplication of each pair of input values.We have used this kernel in our model.

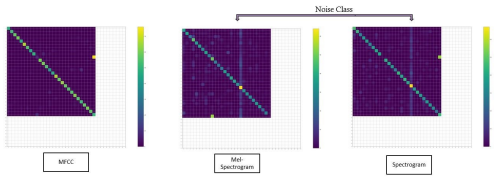
Convolutional Neural Network: Convolutional Neural Network (CNN/ConvNet) is a class of deep neural networks, most commonly applied to analyze visual imagery. Now when we think of a neural network we think about matrix multiplications but that is not the case with ConvNet. It uses a special technique called Convolution. Now in mathematics convolution is a mathematical operation on two functions that produces a third function that expresses how the shape of one is modified by the other.
Our proposed CNN works like this :



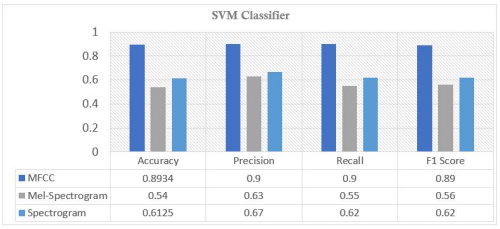
3 Results and Analysis

- **Training and Testing Description:** For training the model, a sample speech of 5 minutes has been collected from the audience. The speech was divided into 100 samples of 3seconds each. Among 100 samples of per person 20 has been used to test the model. Rest of the samples are used for training where per samples are augmented with some features like time stretching, percussion, pitch shift etc. and total 14 samples have been generated. So, total number of train data for per class is 1120. As there is total 28 classes, the train data set has total 31360 samples.

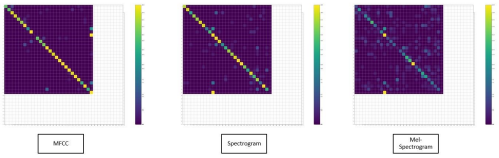
- **Using Support Vector Machine as a Model and MFCC, Mel-spectrogram, Spectrogram as Feature:** For different features the confusion matrices of SVM model has been shown in the following figure:



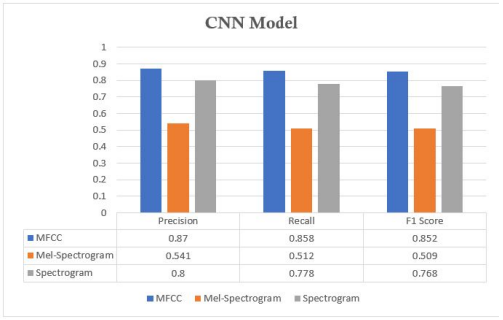
For different features the result of SVM model has been given below:



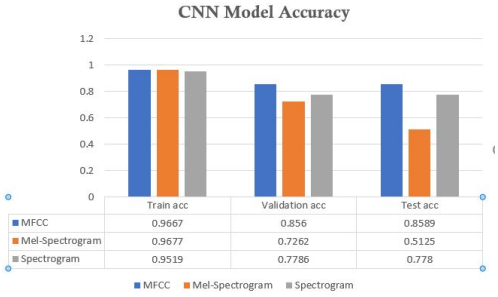
- **Using Convolutional Neural Network as a Model and MFCC, Mel-spectrogram, Spectrogram as Feature:** The confusion matrices of CNN as speaker recognition model for different features are shown below:



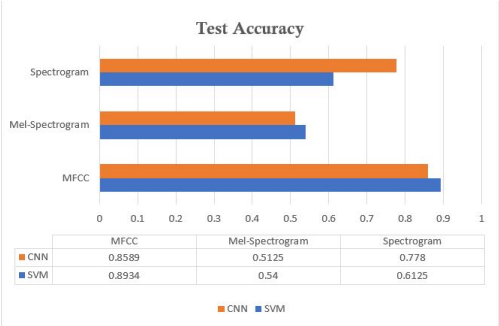
Precision, Recall and F1 score of the CNN model for various features are given below:



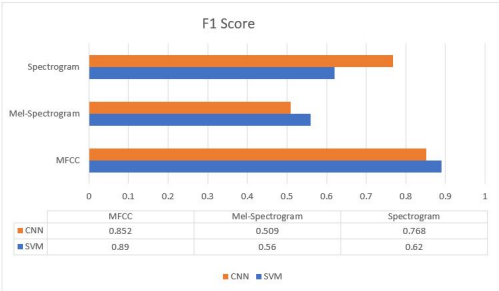
Then the accracy for different models have been predicted. The accuracy for CNN model for different features is given below:



- **Comparison:** Test accuracy of the both models for all three features has been shown in the following figure:



F1 score for both the models for the features is given below:



From the above figures,it can be said that among all the features MFCC works better in both CNN and SVM model. It has better test accuracy and F1 score compared to all other features. On the other hand, mel-spectrogram is not a good feature for speaker recognition. It has scored poorly in both of the models.

Between both of the models,SVM comparatively works better than CNN model. It scores higher than the CNN model for MFCC and mel-spectrogram. But for spectrogram CNN is better.

4 Conclusion

Bengali speech recognition, especially digits recognition can help in speech-based command for Internet-of-things devices. For example, an Intelligent traffic controlling system can get benefited from this.A bigger dataset covering all dialects of Bengali language from all age-groups can be created in the future.We had some limitations while implementing the project. We created a huge data set. But could not use it to full potential due to computational backdrop. We did not the time to evaluate CNN architecture. We hope that, in near future, we could run the whole dataset with solutions to computational problems. Also, we may use this dataset in ML, DL models that would work on Bangla voice.

5 References

1) Md. Afzal Hossan,Sheeraz Memon,Mark A Gregory,"A novel approach for MFCC feature extraction",2010 4th International Conference on Signal Processing and Communication Systems.
2) Chen Wang, Zhenjiang Miao and Xiao Meng,"Comparison of different implementations of mfcc",J. Computer Science Technology,vol. 16, no. 16, pp. 582-589, 2001.
3) A. V. Oppenheim, "Speech spectrograms using the fast Fourier transform", IEEE spectrum, vol. 8, no. 7, pp. 57-62, 1970.