Bangla Voice Command Recognition for Wheelchair Control

Mrinmoy Kundu
Department of EEE, BUET
Dhaka, Bangladesh
Email: mrinmoykundu14@gmail.com

Md. Jawad Ul Islam
Department of EEE, BUET
Dhaka, Bangladesh
Email: jawaduid@gmail.com

Nabila Tasfiha Rahman
Department of EEE, BUET
Dhaka, Bangladesh
Email: nabila.tasfiha@gmail.com

Md. Rafiqul Islam Rafi
Department of EEE, BUET
Dhaka, Bangladesh
Email: rafiqulrafi1999@gmail.com

Jahid Hasan Tushar
Department of EEE, BUET
Dhaka, Bangladesh
Email: tushar33.jht40@gmail.com

Ayan Biswas Pranta
Department of EEE, BUET
Dhaka, Bangladesh
Email: ayanbiswas199@gmail.com

***Abstract*—Despite being one of the most widely spoken languages of the world, no significant efforts have been made in Bangla speech recognition. Speech recognition, in particular keyword recognition, is a difficult task especially if the demand is to do so in noisy real-life conditions. With inspiration to control wheelchairs for handicapped persons with Bangla voice commands, we tried to develop a keyword recognition system in this project. Our convolutional neural network model can recognize five voice commands– 'shamne' (forward), 'pichone' (back), 'dane' (right), 'bame' (left) and 'thamo' (stop) and a class for unknown words. Among different feature extractors such as Mel spectrum, Bark spectrum and Bark spectrum with the modification of Teager Energy operator, Bark spectrum for mixed data sets outperformed the others with 93.8% accuracy.**

***Index Terms*—Convolutional Neural Network(CNN), Bark Spectrum, Teager energy operator**

I. Introduction

Voice has been the primary method of communication for humans and other animals from their birth. With rapid technological growth, the urge to use voice to interact with devices is only increasing. Researchers have been working on manipulating voice signals for easier and faster recognition. Tech giants like Google and Apple have their own voice recognition system. Google even associated their search engine and android platform with this facility. Keyword spotting has the potential to provide a convenient hands-free interface for digital devices. In fact, keyword spotting can assist disabled people. People unable to walk rely on wheelchairs for locomotion. Users have to use their hands and arms to rotate a rim connected to the rear wheels and move. Patients who don't have hands can't control a wheelchair. Integrating keyword spotting with their wheelchair can make their movement more convenient. Many researchers have proposed voice as the medium to control wheelchairs in English.[1]

Since our mother tongue is Bangla and most of our population aren't comfortable speaking English, a convenient Bangla keyword recognition system can assist disabled patients in our country more effectively. Our motivation for this project is to develop a CNN model that can detect necessary Bangla keywords to operate a wheelchair. Unfortunately this particular problem has no open source database. So we built a database collecting samples from our friends and family. This database contains 1647 .wav audio files of people saying 5 different words– "shamne"(forward), "pichone"(backward), "dane"(right), "bame"(left), "thamo"(stop) and 6624 WAVE audio files for unknown class. Each sample is 2 seconds long and contains a single word. The model tries to classify a two second audio as unknown or any 5 of these keywords. We developed a CNN model with 5 layers to classify an audio into the possible 6 outcomes and output the predicted label. For feature extraction, we have used bark spectrum, mel spectrum and teager energy operator (TEO) combined with bark spectrum and investigated their effect on accuracy.

Latency, accuracy and computational power are the key factors while designing a keyword spotting system. This system is designed to have low memory footprints and require low computational power, so it is compatible for a raspberry pi.

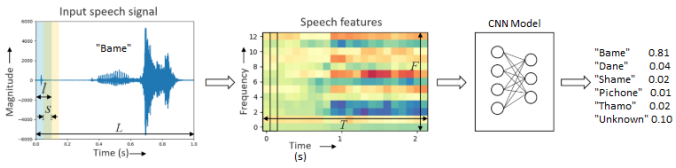

Fig. 1. Keyword spotting pipeline

II. Methodologies

A. Data Acquisition

We have performed command word detection for six classes in our project. We have collected 340 raw files of wav audio per command. Then the data is cropped or zero padded so every data is 2 seconds long. Every data is sampled at 44.1kHz. The data is further augmented 10 times. Augmentation process includes time shifting, pitch shifting, amplitude level changing, modifying tone, adding noise and stretching time.

1) Gender Ratio: Male and female voices are naturally different. A lot of characteristics differ between male and female voices. For better understanding of our model, we trained our model three different ways

- Trained model with only male dataset
- Trained model with only female dataset
- Trained model with both male and female dataset (mixed)

Command Key Word	Male (53%)		Female (47%)		Total	
	Raw	Aug	Raw	Aug	Raw	Aug
shamne	175	1750	154	1540	329	3290
pichone	185	1850	158	1580	343	3430
dane	168	1680	155	1550	323	3230
bame	172	1720	150	1500	322	3220
thamo	182	1820	148	1480	330	3300
Unknown	—				6624	

TABLE I
GENDER RATION OF THE DATASET

2) Unknown Class: The term unknown class refers to any word or sound apart from the above five command words.This belongs to the 'unknown' class in our project. We incorporated 6624 random unknown data in our model.

3) Source of dataset: The five command words are taken from us,our family members ,relatives and classmates.Then these audio files are augmented using MATLAB. Unknown data is taken from different bengali audio clips and then they are also augmented.

B. Preprocessing

1) Bark Spectrum: In this algorithm, the audio input is first buffered into frames of 198 samples. The frames overlap by 0.015s. Then the Hanning Window is applied to each frame, and the frame is converted to frequency-domain representation with 2048 points. Each frame of the frequency-domain representation passes through a bark filter bank. The spectral values output from the barkl filter bank (spectral centroid and spectrum flux) are summed, and then the channels are concatenated so each frame is transformed to a 50 element column vector.

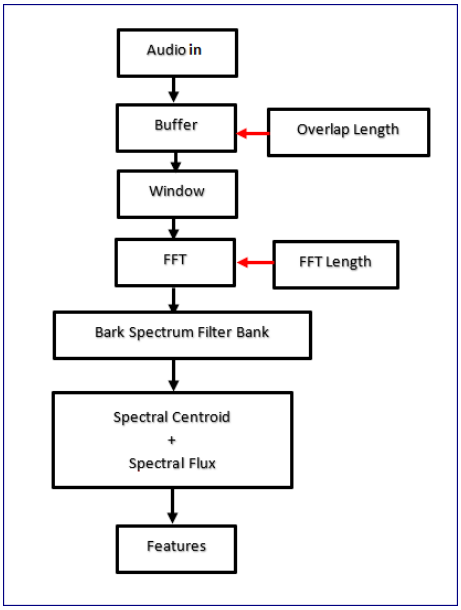

Fig. 2. Workflow for Bark feature extractor

2) Mel Spectrum: Likewise, by changing the spectrum filter banks for mel spectrum, Mel spectrogram feature extraction was performed.

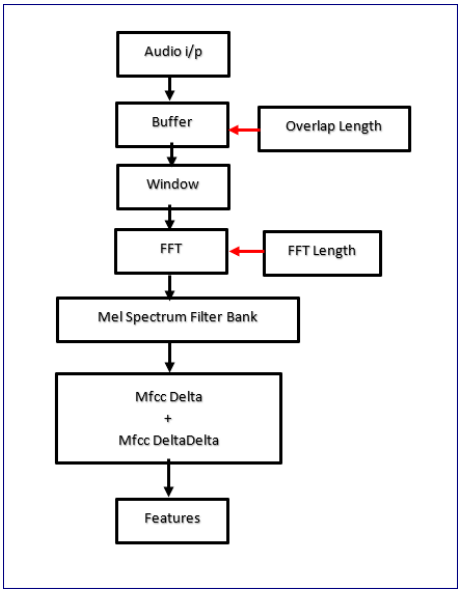

Fig. 3. Workflow for Mel feature extractor

3) TEO + Bark Spectrum: Here, before extracting the features using the Bark spectrum method,the discrete time domain signal was modified according to the Teager Energy algorithm. The formula is [2]

$$x[n] = x[n]^2 - x[n-1]x[n+1] \quad (1)$$

Apart from this modification, every step was followed like the Bark spectrogram method.

Description of the feature extraction parameters -

- Sampling Frequency, Fs = 44.1KHz
- Window Type: Hanning
- Duration of Windowing:2s
- Frame Length = 0.025s
- Overlapping = 0.015s
- FFT Length = 2048
- Number of bands or feature per frame = 50
- Number of frames = 198
- Size of spectrum coefficient matrix = 198x50

C. CNN model Architecture

After extracting the 198x50 2D feature spectrum matrices from each of the 1D audio vectors, the 2D data were passed into the CNN mode. Training, validation and test split for each set of dataset was 60:20:20 respectively.

The CNN model architecture consists of 5 convolutional layers with MaxPooling layers of stride-3 and padding-2 in-between. Number of channels for each convolutional layer were 12,24,48,48,48 respectively. Batch normalization and the ReLU activation layer were sandwiched between each convolutional and MaxPooling layers. With dropout probability 0.2, the last convolutional layer was pipe-lined with a fully connected neural network. Lastly, the 6x1 prediction vectors were found at the end with softmax activation. Weighted cross entropy loss function and Adam optimizer were used for training the network.

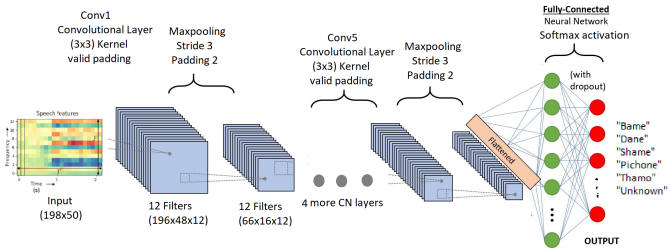

Fig. 4. CNN Architecture

D. Interface

We use a microphone for user input and feed data into our CNN model. The microphone updates the audio data every 100 milliseconds. Finally a mono channel wav data of 2 seconds is prepared for feature extraction and classification. To detect silence, we have neglected signals having average amplitude below 0.005(normalized). A voice controlled wheelchair must be responsive,as well as accurate. Since command refresh rate is 100 milliseconds, CNN detects a command multiple times. Using this to our advantage, we have developed a small 10 length memory system for FP detection. So it activates a command only after CNN detects it more than 5 times previously. This enables much smoother operation, with minimum latency of 0.5 seconds. To demonstrate our wheelchair, we developed a virtual environment with the MATLAB app developer. It takes voice input and displays corresponding motion. Initially our wheelchair can perform linear and rotary motion. To accomplish rotation, it rotates upto 90 degrees for every rotation command. However, any type of motion can be stopped using the termination command.

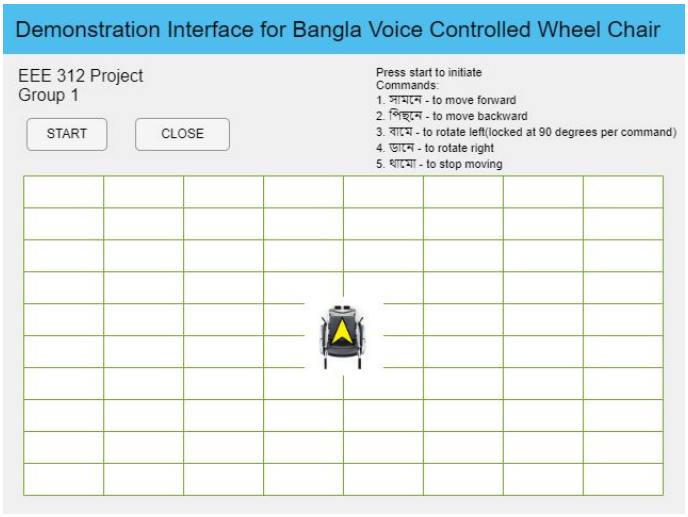

Fig. 5. Virtual interface

III. Result Analysis

As it has been described before, we have evaluated the model for three sets of dataset which were labeled in terms of gender. In this result analysis section, the results for the techniques namely Bark spectrum and TEO+Bark spectrum are shown.

A. Results for three datasets with Bark spectrum

1) Male Dataset: Firstly, performance table for male dataset shows that class-wise accuracy for the test dataset was quite high around 96% for all class, and overall accuracy was calculated to be 91.37%. But F-score for class 'Bame', 'Dane' and 'Shamne' was comparatively low.

Class	Bame	Dane	Pichone	Shamne	Thamo	Unknown
Accuracy	0.958	0.9645	0.9810	0.9682	0.9642	0.9909
Sensitivity	0.6926	0.9373	0.8293	0.7705	0.9555	0.9992
Specificity	0.9886	0.9677	0.9988	0.9918	0.9654	0.9838
Precision	0.8750	0.7730	0.9876	0.9184	0.7951	0.9811
F-score	0.7732	0.8472	0.9015	0.8380	0.8679	0.9901

TABLE II
MALE DATASET

2) Female Dataset: For female dataset, accuracies for individual classes are comparable to the previous male dataset with overall accuracy 91.22%. F-score was found to be the lowest for the 'Shamne' class.

Class	Bame	Dane	Pichone	Shamne	Thamo	Unknown
Accuracy	0.9652	0.9628	0.9768	0.9608	0.9604	0.9988
Sensitivity	0.7258	0.8866	0.7946	0.7326	0.9719	0.9992
Specificity	0.9916	0.9708	0.9947	0.9871	0.9590	0.9984
Precision	0.9045	0.7617	0.9368	0.8670	0.7527	0.9984
F-score	0.8054	0.8194	0.8599	0.7941	0.8484	0.9988

TABLE III
FEMALE DATASET

3) Mixed Dataset: For mixed dataset, F-score improved for all classes, this dataset found to be generating best results among the three with highest overall accuracy 93.80%.

Class	Bame	Dane	Pichone	Shamne	Thamo	Unknown
Accuracy	0.9760	0.9849	0.9847	0.9646	0.9808	0.9861
Sensitivity	0.8653	0.9828	0.9124	0.8383	0.9934	0.9826
Specificity	0.9934	0.9852	0.9963	0.9849	0.9788	0.9876
Precision	0.9536	0.9105	0.9753	0.8991	0.8830	0.9735
F-score	0.9073	0.9453	0.9428	0.8676	0.9350	0.9780

TABLE IV
MIXED DATASET

Fig. 6. Confusion matrix for mixed dataset

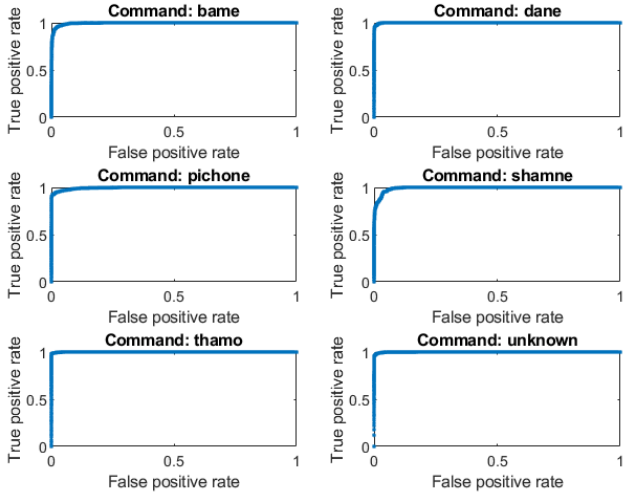

Fig. 7. ROC curve for mixed dataset

B. Result with TEO+Bark spectrum

For feature extraction technique TEO+Bark spectrum, we analyzed the results for only the previously best performed mixed dataset and overall accuracy for this case was found to be 89.12%, which is the lowest. Though the average F-score was comparatively good with 88.4%.

Class	Bame	Dane	Pichone	Shamne	Thamo
Accuracy	0.94	0.93	0.98	0.95	0.90
Sensitivity	0.91	0.85	0.94	0.82	0.91
Specificity	0.95	0.96	0.99	0.98	0.97
Precision	0.82	0.84	0.96	0.93	0.90
F-score	0.86	0.84	0.95	0.87	0.90

TABLE V
MIXED DATASET WITH TEO+BARK

C. Comparison among datasets and Feature extraction methods

For three datasets and three different kinds of feature extraction techniques – Mel, Bark, TEO+Bark, a comparison bar chart with a total four cases was generated. In terms of test accuracy, the mixed dataset with Bark feature extraction outperformed the individual male and female dataset. Though we hoped for better results for the teager energy operator, it didn't perform very well against the bark spectrum alone. The Mel spectrum technique also performed poorly for all four cases. Thus, the best result was found for the combination of mixed dataset with Bark feature extraction technique.

	Validation Accuracy(%)	Test Accuracy(%)	Overall Test F-score
Female Dataset (BarkSpec)	92.03	91.22	84.67
Female Dataset (MelSpec)	86.46	85.4	86
Male Dataset (BarkSpec)	91.33	91.37	86.5
Male Dataset (MelSpec)	82.35	79.1	69.5
Mixed Dataset (BarkSpec)	95.4	93.8	92.33
Mixed Dataset (MelSpec)	90.5	89.6	86.1
Mixed Dataset with TEO (BarkSpec)	76.25	89.12	88.4
Mixed Dataset with TEO (MelSpec)	89.6	88.4	87.3

TABLE VI
PERFORMANCE TABLE

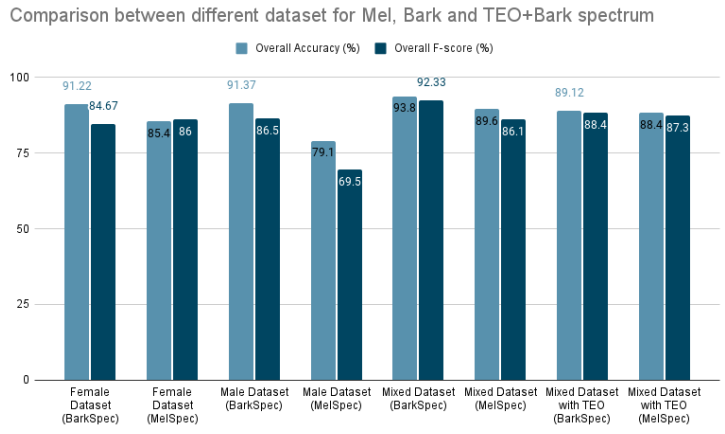

Fig. 8. Comparison between four cases

D. Comparison with related works

Our best result is compared with two other related works. Shakil Ahmed Sumon, et. al [3] created a CNN model for bangla short speech commands recognition with MFCC feature extraction. And Xuejiao Li, et. al. [4] used a similar feature extraction technique to detect 6 different english short commands, which are basically translated versions of ours. It is clear from this chart that our result is quite comparable with Xuejiao Li, et. al., with our test accuracy 93.8% and theirs 94.5%.

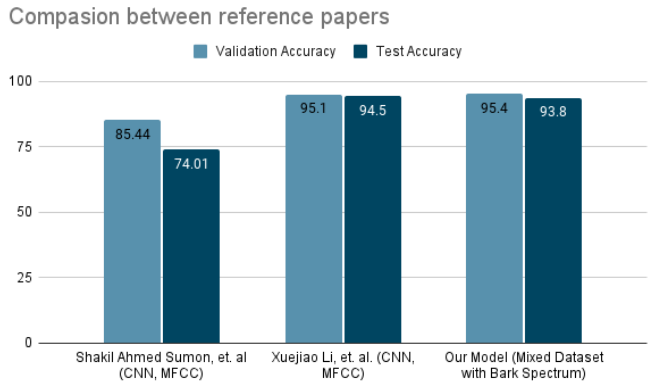

Fig. 9. Comparison with two works

IV. Future Prospectus

One limitation of our project is lack of data. With more versatile data, we believe the model can achieve better accuracy. We have used five commands for a voice controlled wheelchair in our project. We can add more commands to further ease the operation. We can add a sensor for bumpy roads so that the wheelchair can operate more smoothly. We can also upgrade our programme in such a way that it can distinguish between the command and other conversation which may include exact words of commands. If the user has a sore throat or similar medical condition which may worsen or temporarily change his/her voice, it can adapt its dataset according to that change and work accordingly. For better response time,we can optimize its algorithm.

V. Conclusion

We have built a bangla dataset for five commands and developed a CNN model in this project. We have demonstrated the performance using a Graphical User Interface. For both male and female datasets, the results were quite satisfactory. Our project showed best accuracy of 93.8% for mixed dataset using Bark feature extractor. One of the limitations of our project was insufficient data, which can be overcome in the future to produce more robust results. For a developing country like ours, this project will be beneficial for people from every sector as they can operate it using their own language- Bangla. In short, we can say that our project can be a saviour to aid people for an easier life.

Acknowledgment

We have collected 340 wav files per command to built the dataset. This would not be possible without the help of our family members, friends, classmates and relatives. We thank these wonderful people for their cooperation. We are also grateful to our respected teachers, Dr. Celia Shahnaz Madam and Shahed Ahmed Sir for their valuable guidance.

References

[1] Sumet Umchid, Pitchaya Limhaprasert, Sitthichai Chumsoongnern, Tanun Petthong, and Theera Leeudomwong. Voice controlled automatic wheelchair. In *2018 11th Biomedical Engineering International Conference (BMEiCON)*, pages 1–5, 2018.

[2] Md Tauhidul Islam, Celia Shahnaz, Wei-Ping Zhu, and M. Omair Ahmad. Modeling of teager energy operated perceptual wavelet packet coefficients with an erlang-2 pdf for real time enhancement of noisy speech. 02 2018.

[3] Shakil Ahmed Sumon, Joydip Chowdhury, Sujit Debnath, Nabeel Mohammed, and S. Momen. Bangla short speech commands recognition using convolutional neural networks. *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–6, 2018.

[4] Xuejiao Li xjli and Zixuan Zixuan. Speech command recognition with convolutional neural network. 2017.