সহকারী - A Bengali Virtual Assistant

Taslima Akbar
Institute of Information Technology
University of Dhaka
Dhaka Bangladesh

Naushin Nower
Institute of Information Technology
University of Dhaka
Dhaka Bangladesh

ABSTRACT

Bengali is the 7th most widely spoken and morphological rich language in the world. However, Bengali Natural Language Processing research is not that rich compared to the other less spoken languages. In addition, a large portion of the Bengali native speaker is illiterate and facing difficulty using the computer because of English commands. A Bengali virtual assistant can solve this problem where people with difficulty in English can easily operate a computer or smartphone by instructing their native language Bengali. In this paper, a Bengali virtual assistant named ’সহকারী’ is developed that utilizes the CNN model to recognize the Bengali command. The CNN uses a spectrogram to identify the type of the commands and executes the responses. The designed Bengali virtual assistant can provide almost 87% accuracy in command detection.

Keywords

Bengali, Virtual Assistant, CNN, Spectrogram

1. INTRODUCTION

Among the 265 million speakers in the world, Bengali ranks 7th biggest language. There are almost 100 spoken languages and Bengali holds the 7th position [1] among them. However, the research on Bengali speech recognition is insufficient compared to the other fewer population languages like German, French, etc [2]. Moreover, most of the people in Bangladesh live in the village and they are not techno-oriented, as a result, it is easier for them to give a Bengali voice command to operate mobile phone and computer. Voice command makes an easier to operate mobile/computer for disabled people, elder people, and illiterate people. Thus, a virtual assistant makes it comparatively easier to operate the smart phone/ computer by just giving a voice command without requiring any extra knowledge. Virtual assistance makes an environment where users can use a smart phone/computer just like they are communicating with another person.

Bengali is a morphologically, rich language with eleven vowels and thirty-nine consonants. Besides these, there are compound characters that combine consonants and/or vowels. As a result, there are a total of three-hundred characters set in Bengali. Although Bengali is one of the widely spoken languages and rich language, there is a lacking of a proper dataset in the Bengali language for research. In most cases, researchers collect their corpus for research. Besides this, the existing Bengali corpus is not well labelled. Thus lacking benchmark corpus, labelled data, and the morphological analysis makes research on the Bengali virtual assistant more challenging [3].

Huge researches have been done on the English virtual assistant. As a result, various English virtual assistants such as Google's Google Assistant, Microsoft's Cortana, Apple's Siri, Amazon's Alexa, and Bixby by Samsung are ready to accept voice commands in English [4]. On the other hand, very little researches have been done on Bengali language recognition, virtual assistant, and chatbot. According to Euromonitor International, 18% of the total population in Bangladesh can be benefited from English virtual assistants because they can speak and understand English [5]. Moreover, a large part of the Bengali-speaking people is illiterate and unable to communicate operate phone/ laptop using the text-based command. As a result, there is an urgent need for Bengali virtual assistant so that low literate people can be benefited directly from that. A virtual assistant can be used to set a reminder, play a song, know weather updates, PC restart, current time, search information in Google, update to-do list, and much more daily activity.

Recently, a few research initiatives have been shown in Bengali language recognition or language-based automation. A Bengali chatbot named 'Golpo' was proposed by T.D. Orin in 2017 which is mainly used for text conversions [6]. Besides this, Anirudha Paul et al. proposed a chatbot framework for close domain resource-poor languages including Bengali [7]. Both of them have only some limited functionalities with general text-based conversations on Bengali. Besides these, a Bengali virtual assistant named 'Adhtee' for the smart device is proposed in the [5]. This proposed virtual assistant can respond to some frequently used commands. The authors show that their virtual assistant can recognize the basic commands with almost 90% accuracy. However, the proposed desktop application has no learning ability and uses external API which is not free always.

In this paper, a Bengali virtual assistant named ‘সহকারী’ means Assistant is designed as a desktop application. It collects a set of frequently used commands with variations from various groups of people. From this collected command, a spectrogram is produced and used to train the CNN model. CNN learns the keyword from the images during the training phase and predicts the class using the learned keyword when the user gives the voice command. CNN then calculates the confidence score of the predicted class and executes the command.

The paper is organized as follows: Section II discusses the related work on Bengali speech recognition. The proposed application is demonstrated in section III. The functionalities of the application are shown in section IV. Finally, section V concludes the paper with a conclusion and limitation.

2. RELATED WORK

Initial research [8] applies basic techniques (MFCC, LPC, GMM, and DTW) to extract features from Bengali words. Speech recognition in Bengali is challenging because of the acoustic similarity of lots of words and their syllabi. As a result, it is needed to determine the fine phonetic distinctions among the words. The authors use MFCC+ DTW, LPC+DTW, MFCC+GMM and MFCC+LPC+DTW to recognize Bengali speech recognition. Among them, MFCC+GMM works best for Bengali speech recognition.

The paper [9] proposes two models for Bengali speech recognition. At first, they used CNN to recognize spoken Bengali words. Then, they utilized a recurrent neural network to predict the Bengali word. The output of these two deep learning techniques is merged using the Connectionist Temporal Classification (CTC). They claimed that, merging these techniques improves Bengali word recognition.

The paper [2] shows that CNN and LSTM can better recognize Bengali speech. They have applied different dimension reduction techniques with CNN and LSTM and show that well-known dimension reduction techniques (PCA, k-PCA, and T-distributed Stochastic Neighbour Embedding) have no beneficial effect on CNN and LSTM on Bengali speech recognition. CNN is also used in the short Bengali command speech recognition in the paper [10]. Authors use three different approaches to determine which feature extraction method works well with CNN for Bengali short command recognition. In the first approach, MFCC is extracted from the Bengali speech to train the CNN model. In the second approach, raw audio is used as input to the CNN model. And lastly, features learned from the English language are used as input for the CNN model. Among them, MFCC with CNN shows better accuracy compared with others. From the analysis, it can conclude that MFCC with CNN suits better with Bengali long speech recognition also. Another paper [6] designs a chatbot for the Bengali language. The designed chatbot can respond in real-time. To do this, at first, it creates a Bengali corpus and matches the input with that corpus. After that, it calculates the Jaccard Similarity function to determine how much similarity exists between the input and database. In the case of 50% of the word similarity, two sentences are considered as match, and responses are generated from the corpus. The main problem of the Bengali chatbot is the lacking of the proper dataset. Moreover, a designed chatbot is implemented by comparing only similarities without using any machine learning techniques.

Besides chatbots, some virtual assistants [11, 12] are developed for visually impaired people in Bengali. For example, M.R. Sultan et al. developed a virtual assistant named ‘Adrisya Sahayak’ which means invisible helper [11] for the visually impaired in the Bengali language. This virtual assistant uses Google's Web Speech API for speech recognition. It identifies the keyword from the speech and executes the corresponding service by responding. A similar type of work is proposed by M.M. Hoque et al. for visually impaired people in their paper [12]. The proposed virtual assistant ‘Always By Your Side (ABYS)’ uses Microsoft Speech Application Programming Interface (SAPI) for local speech recognition. The solution also uses the cloud-based Bing Speech API approach. Besides the above research, the Bengali virtual assistant named Adheetee is proposed in the paper [5]. The system can accept the Bengali command as voice or text. In the case of voice command, it converts the voice to text using the Google speech to text (STT) method. After that, the keyword is extracted from the text and detected the command category by comparing it with the keyword database. However, recognizing the variation of the same command is a very challenging task, to solve this problem, name entity recognition and cosine similarity are used. In the case of unknown commands, 'Adheetee' uses an external API call to resolve the issue. However, this approach mainly depends on the external API call and Google speech to text conversion approach which is not free always. Moreover, there is no learning mechanism to improve the performance of 'Adheetee'.

3. METHODOLOGY FOR THE PROPOSED APPLICATION -‘সহকারী’

The proposed system aims to develop an effective virtual assistant for Bengali voice command. The virtual assistant takes the command in the Bengali language and executes the command to produce action. To implement virtual assistance, it is needed to train the system so that it can recognize the command. As a result, it is needed a benchmark database for Bengali to train the system. However, the benchmark database is not available for the Bengali language. To solve this problem, various commands with a variation on the different categories are collected in the Bengali language. The audio command is then pre-processed and an image is generated from the audio. This generated image is used as input for CNN to prepare the model. After training the model is used for used for Bengali voice command detection.

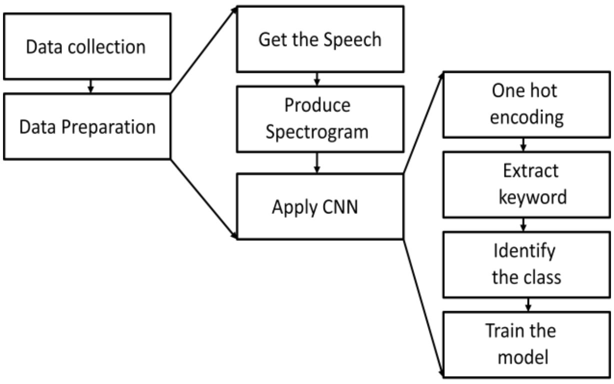


Fig. 1. Overall Process of the proposed ‘সহকারী’ Bengali Virtual Assistant

3.1 Data Collection

Initially, twelve basic commands frequently used commands are chosen. The topic of the command is given to others to get variations of that command. For example, the command is 'আজকের আবহাওয়া কেমন' means 'How is the weather toady'. Different variations of this command are possible. Such as 'Give me the weather update', 'Tell me today's temperature' and many more. The aim is to collect the variations of command because people can express command in their way. Twenty-four people of different ages are selected and asked them to give the twelve categories of commands. These data are collected virtually and stored in a Google drive. The size of the audio file is almost 650 kb and a total of 15 GB. For simplicity, it is assumed that collected voices are noise-free. In the Table 1, two variations of twelve commands are given.

Table 1: Variations of Commands

No.	Automated Task List	Command Variation (1)	Command Variation (2)
1	Fetching Data from Wikipedia	উইকিপিডিয়ায় সার্চ করো	উইকিপিডিয়া চালু করো
2	Open Google	গুগল ওপেন করো	গুগল চালু করো
3	Open YouTube	ইউটিউব খুলো	ইউটিউব চালু করো
4	File Explorer	ফাইল এক্সপ্লোরার খুলো	ফাইল এক্সপ্লোরার চালু করো
5	Computer Restart	কম্পিউটার রিস্টার্ট করো	পিসি রিস্টার্ট করো
6	Computer Shut Down	কম্পিউটার বন্ধ করো	কম্পিউটার শাটডাউন করো
7	Play Music	গান প্লে করো	মিউজিক ছাড়ো
8	Take a Screenshot	স্ক্রিনশট নাও	স্ক্রিনশট তুলো
9	Tell the time	কয়টা বাজে	টাইম বলো
10	Volume Decrease	ভলিউম কমাও	সাউন্ড কমাও
11	Volume Increase	ভলিউম বাড়াও	সাউন্ড বাড়াও
12	Weather Update	আজকের আবহাওয়া কেমন	বাহিরে ওয়েদার কি রকম

3.2 Audio Pre-processing

The recorded audio file is saved as .wav files and a spectrogram is generated from the audio files. A spectrogram represents signal strength over time by showing various frequencies present in a waveform. Non-stationary or nonlinear signal characteristics of a signal can be represented by the spectrogram. Thus, the spectrogram is a useful tool to analyze the signal with various frequency components and/or electrical and mechanical noise. To generate the spectrogram, the Short-time Fourier transform (STFT) of the audio data is used. Fourier transform is used to convert a signal from the time domain to the frequency domain. This is used to know how much the signal is varying. STFT is an extension where one takes small windows and convolves them with the signal and applies Discrete Fourier transform (DFT) within the convolved window. In DFT methods, the computational complexity is too long. It can be reduced through FFT or fast Fourier transform. So, FFT is nothing but the computation of discrete Fourier transforms in an algorithmic format, where the computational part can be reduced. After generating the spectrogram, the audio samples are saved as an image with its corresponding labels. From these spectrograms, the class of the command is determined. These saved image files are loaded when needed for comparing without generating spectrograms every time. From the spectrograms of the voice, different words can be distinguished which can be used to detect commands. A spectrogram is a detailed view of audio that represents time, frequency, and amplitude in one graph.

3.3 Build CNN Model

In this part, spectrograms are treated as an image and try to identify features from these images that would help to identify the class of the audio sample. Convolutional Neural Network (CNN) is used for classification. CNN would take images as input and learn spatial features of the images and predict class. TensorFlow is used for building CNN. At first, it loads the model and then reads the spectrogram images. These images are representations of our spoken words. Every spoken word corresponds to a spectrogram. It is expected that spectrograms of the word 'weather' sound would be similar across different speakers and genders. It is expected that despite the difference in volume, pitch, timbre, etc, the word 'Weather' spoken by anyone should have similarities with other weather sounds. The same is true for the other commands also. Thus, if there exists certain similarities between the same word sounds across all these variables, then CNN will catch them in the spectrogram.

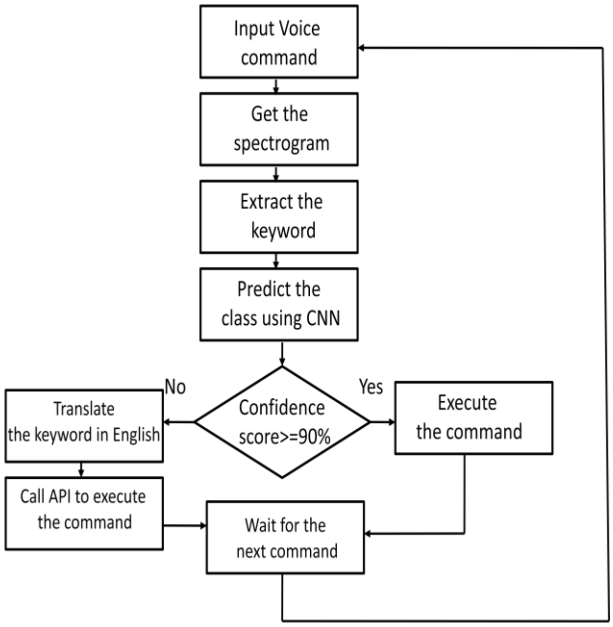

Input Voice command
Get the spectrogram
Extract the keyword
Predict the class using CNN
Confidence score>=90%
No
Yes
Translate the keyword in English
Execute the command
Call API to execute the command
Wait for the next command

Fig 2: Flowchart of the proposed সহকারী virtual assistant

3.4 Prepare the train and test dataset

To prepare the model, a command dataset is created with corresponding actions. From the spectrogram image, the categorical/ nominal data is extracted for further processing. Such as the keyword 'weather', 'shutdown', etc. However, some machine learning algorithms cannot operate on label data directly. The entire input data variables are needed to be numeric for most of the ML algorithm. As a result, it is needed to convert our categorical data into numerical form. To do this, a one-hot coding algorithm is used because data has no ordinal relationship between them. One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. Then, shuffle and split the data into training and testing sets. 60% of the total data is used for training the model and the rest 40% are used for testing purposes.

3.5 CNN architecture

To recognize different commands by their spectrograms, a CNN is used. The main attraction is that it can efficiently extract image features by learnable convolution operators. The output of the convolution operators is transmitted into a neural network for classification. The first network layer is the convolutional layer composed of two 3×3 kernels followed by a rectified linear unit (shortened to ReLu). The second layer consists of four 3×3×2 kernels. Then, a max-pooling operation is applied into the two convolutional layers + ReLu layer to reduce the data size by half. The next layer consists of eight 3×3×4 kernels with ReLu and ignores 40% of its output to avoid overfitting using a dropout method. Finally, a full-connected neural network is used to provide the output of the classification. For evaluating the model, cross-entropy loss function is used and for optimization, RMSProp optimizer is used.

4. AUTOMATING THE TASK

The microphone of the laptop is used to continuously record the audio and feed it into the CNN model for prediction. If the CNN model can classify the command with a high confidence score (>=90%), the system will execute that task.

4.1 The Designed Desktop App

In this section discusses how the designed app for Bengali command detection performs on different voice commands. At first, it will show the welcome window (as shown in Fig. 3) and welcome the user with a welcome message. Then it will take the commands. There are twelve commands that the desktop application can recognize. Users can give any commands from those twelve commands and the app will show the result. In the welcome window, the user can search for his/her desired topic. Fig. 4 shows the result of the command that ask to open the Wikipedia. Then, the user can search the topic in the Google and result is shown in Fig. 5. The user can ask to play a song by giving the command in Bengali and in the case of successful recognition the designed application will play a random song as shown in Fig 6. Suppose, the user wants to restart the PC by using the command in Bengali. In the case of successful recognition of the command, the PC will restart within 10 second and close all the running applications. The output of this command is shown in Fig. 7.

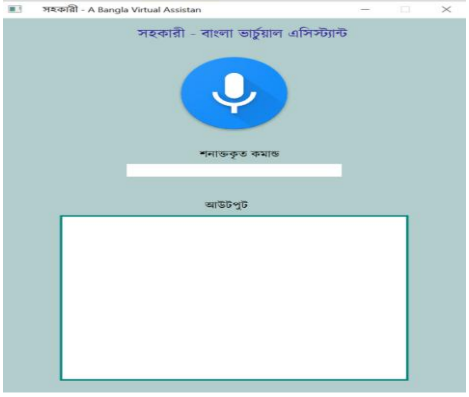
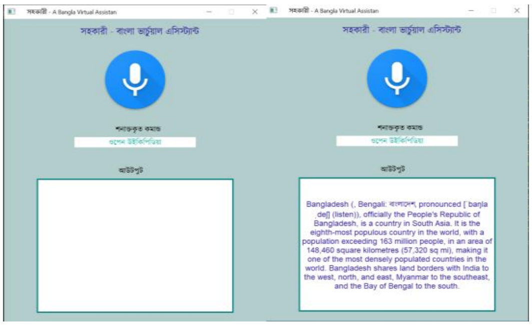
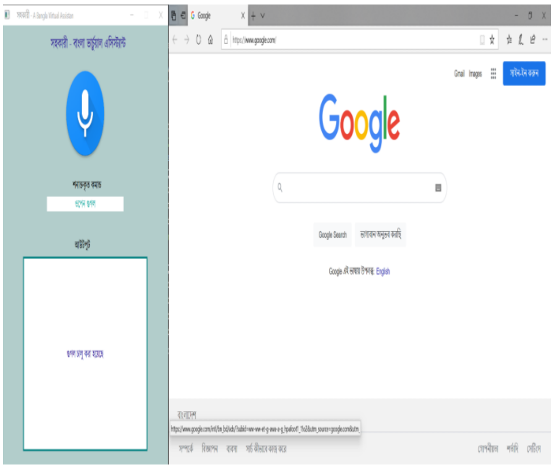

সহকারী - A Bangla Virtual Assistan
সহকারী - বাংলা ভার্চুয়াল এসিস্ট্যান্ট
শনাক্তকৃত কমান্ড
আউটপুট

Fig 3: Welcome window of সহকারী


সহকারী - A Bangla Virtual Assistan
সহকারী - বাংলা ভার্চুয়াল এসিস্ট্যান্ট
শনাক্তকৃত কমান্ড
ওপেন উইকিপিডিয়া
আউটপুট
সহকারী - A Bangla Virtual Assistan
সহকারী - বাংলা ভার্চুয়াল এসিস্ট্যান্ট
শনাক্তকৃত কমান্ড
ওপেন উইকিপিডিয়া
আউটপুট

Bangladesh (, Bengali: বাংলাদেশ, pronounced [ˈbaŋla ˌdeʃ] (listen)), officially the People's Republic of Bangladesh, is a country in South Asia. It is the eighth-most populous country in the world, with a population exceeding 163 million people, in an area of 148,460 square kilometres (57,320 sq mi), making it one of the most densely populated countries in the world. Bangladesh shares land borders with India to the west, north, and east, Myanmar to the southeast, and the Bay of Bengal to the south.

Fig 4: Output of the command Open Wikipedia



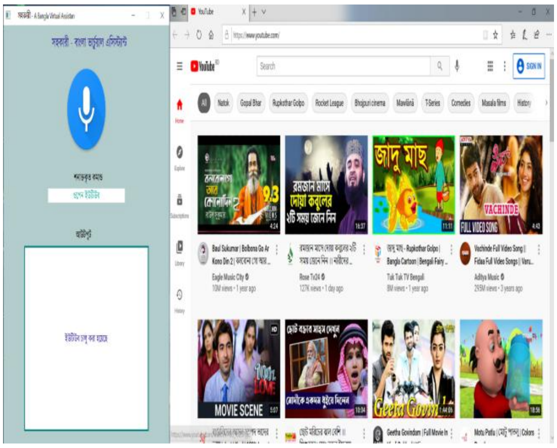
সহকারী - A Bangla Virtual Assistan

সহকারী - বাংলা ভার্চুয়াল এসিস্ট্যান্ট

শনাক্তকৃত কমান্ড: ওপেন গুগল

আউটপুট: গুগল চালু করা হয়েছে

Fig 5: Open Google using the Bengali virtual assistant



সহকারী - A Bangla Virtual Assistan

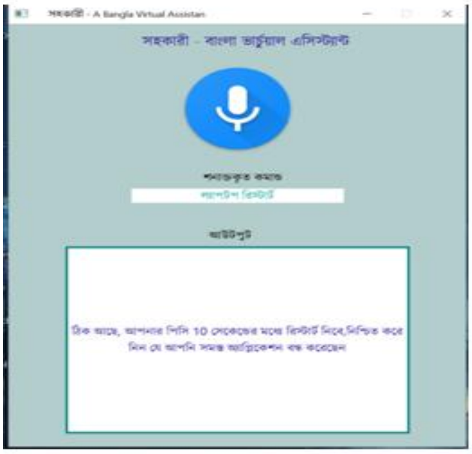
সহকারী - বাংলা ভার্চুয়াল এসিস্ট্যান্ট

শনাক্তকৃত কমান্ড: ওপেন ইউটিউব

আউটপুট: ইউটিউব চালু করা হয়েছে

Fig. 6: Open Youtube using the virtual assistant

The user can ask to play a song by giving the command in Bengali and in the case of successful recognition the designed application will play a random song. For the taking the screenshot, the user asks for screenshot in Bengali and application takes the screenshot. After taking the screenshot, it will create a folder with date and time and save the screenshot as shown in the Fig. 8.



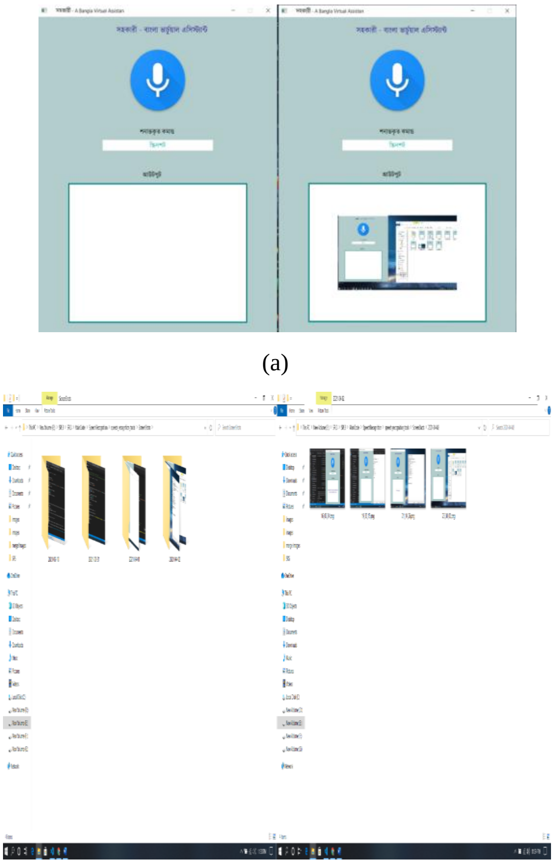
সহকারী - A Bangla Virtual Assistan

সহকারী - বাংলা ভার্চুয়াল এসিস্ট্যান্ট

শনাক্তকৃত কমান্ড: ল্যাপটপ রিস্টার্ট

আউটপুট: ঠিক আছে, আপনার পিসি 10 সেকেন্ডের মধ্যে রিস্টার্ট নিবে,নিশ্চিত করে নিন যে আপনি সমস্ত অ্যাপ্লিকেশন বন্ধ করেছেন

Fig 7: Output of the PC restarts command



(a)

(b)

Fig 8: (a) Responses of taking screenshot and (b) folder to save the screenshot

Using the application user can ask for the current time, increase/decrease the volume of the PC, and weather update of the current city. The output of the successful recognition of these commands is shown in the Fig. 9 and Fig. 10 respectively.

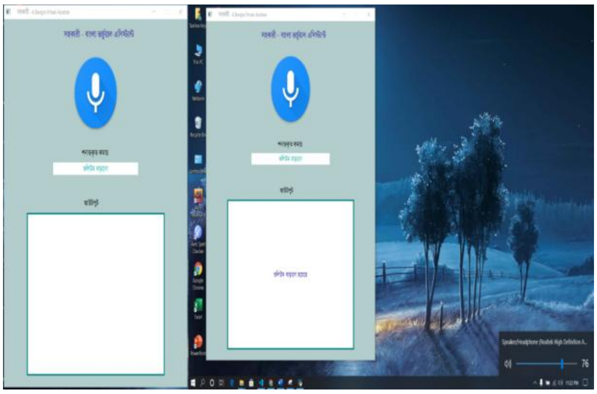


Fig. 9: Screenshot of the volume increase command

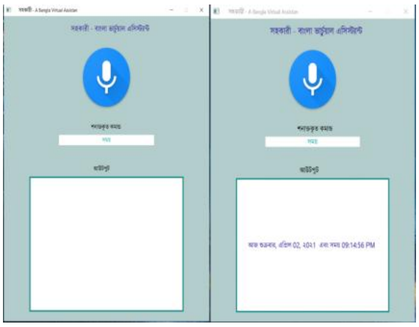
Finally the designed application named সহকারী’s training and testing accuracy is depicted in the Table 2. It is shown that during the training সহকারী can provide 94% accuracy while testing accuracy is almost 87%.

Table 2 Model Accuracy

	Training Accuracy	Testing Accuracy
‘সহকারী	94%	86.4322%

5. CONCLUSION

In this paper, a Bengali virtual assistant named ‘সহকারী’ is designed and implemented. The detail of the virtual assistant system is described here. The proposed system accepts the command in Bengali and executes the response for laptop or desktop computers. The system can accept the widely used commands from different gender and age’s people in Bengali and provide almost 87% accurate responses. The main advantage is that it does not depend on external API and has learning ability. However, the main problem is the lacking of a proper label dataset in Bengali. Only twelve frequently used commands is used and their variations to train the proposed system. More complex commands with their variation are needed to make the system robust. Moreover, our designed application works only on the computer, it is needed to develop a mobile app for Bengali virtual assistant because the smart phone is more widely used in the rural area than a computer.


সহকারী - A Bangla Virtual Assistant
সহকারী - বাংলা ভার্চুয়াল এসিস্ট্যান্ট
শনাক্তকৃত কমান্ড
আউটপুট

সময়

আজ শুক্রবার, এপ্রিল 02, ২০২১ এবং সময় 09:14:56 PM

Fig 10: Responses of the asking time using ‘সহকারী’

6. REFERENCES

[1] Visual Capitalist", [Online] Available at: https://www.visualcapitalist.com/100-most-spokenlanguages. Accessed on December 24, 2021.

[2] M.N. Sabab, M.A.R. Chowdhury, S.M.I. Nirjhor, J. Uddin, “Bangla speech recognition using 1D-CNN and LSTM with different dimension reduction techniques”, International Conference for Emerging Technologies in Computing pp. 158-169, Springer, Cham August 2020.

[3] O. Sen, M. Fuad, M.D.Islam, J. Rabbi, M.D. Hasan, M.Baz, M. Masud, M. Awal, A.A. Fime, M. Fuad, and T. Hasan, M.D. Iftee, “ Bangla Natural Language Processing: A Comprehensive Review of Classical, Machine Learning, and Deep Learning Based Methods” *arXiv preprint arXiv:2105.14875 2021*.

[4] S. Varma (2018, June 14). [Online] Siri vs Google Assistant vs Alexa: Which Is the Smartest Virtual Assistant in 2018? Available at: https://gadgets.ndtv.com/apps/features/siri-vs-google-assistant-vsalexa-which-is-the-smartest-virtual-assistant-in-2018-1867292. Accessed on January 5, 2022.

[5] S.M. Islam, M.F.A. Houya, S.M. Islam, S. Islam, and N. Hossain, “Adheetee: A Comprehensive Bangla Virtual Assistant” IEEE 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) pp. 1-6, May 2019.

[6] T.D. Orin, “Implementation of a Bangla Chatbot (Unpublished Bachelor Thesis)”, BRAC University, Dhaka, Bangladesh, 2017.

[7] A. Paul, A. Latif, A. A. Foysal, M. Rahman, “ Focused domain contextual AI chatbot framework for resource poor languages” Journal of Information and Telecommunication, pp. 1-22. 2018.

[8] M.A. Ali, M. Hossain, and M. N. Bhuiyan. "Automatic speech recognition technique for Bangla words." International Journal of Advanced Science and Technology, 50 (2013).

[9] J. Islam, M.Mubassira, M.R. Islam, A.K. Das, “A speech recognition system for Bengali language using recurrent neural network”, IEEE 4th international conference on computer and communication systems (ICCCS), pp. 73-76, February 2019.

[10] S.A. Sumon, J. Chowdhury, S. Debnath, N. Mohammed, S. Momen, “Bangla short speech commands recognition using convolutional neural networks” IEEE International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1-6, September, 2018.

[11] M.R. Sultan, M.M. Hoque, F.U. Heeya, I. Ahmed, M.R. Ferdouse, S.M.A. Mubin, “Adrisya Sahayak: A Bangla Virtual Assistant for Visually Impaired” IEEE 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 597-602, January 2021.

[12] M.R. Sultan, M.M. Hoque, (2019, December). “ABYS (Always By Your Side): A Virtual Assistant for Visually Impaired Persons”, IEEE 22nd International Conference on Computer and Information Technology (ICCIT), pp. 1-6, 2019.