2019 IEEE International Conference on Robotics, Automation, Artificial-intelligence and Internet-of-Things (RAAICON)
29 November - 1 December 2019, Dhaka, Bangladesh

A Digital Personal Assistant using Bangla Voice Command Recognition and Face Detection

Dipankar Gupta[1], Emam Hossain[2], Mohammad Shahadat Hossain[3], Karl Andersson[4] and Sazzad Hossain[5]

***Abstract*—Though speech recognition has been a common interest of researchers over the last couple of decades, but very few works have been done on Bangla voice recognition. In this research, we developed a digital personal assistant for handicapped people which recognizes continuous Bangla voice commands. We employed the cross-correlation technique which compares the energy of Bangla voice commands with pre-recorded reference signals. After recognizing a Bangla command, it executes a task specified by that command. Mouse cursor can also be controlled using the facial movement of a user. We validated our model in three different environments (noisy, moderate and noiseless) so that the model can act naturally. We also compared our proposed model with a combined model of MFCC & DTW, and another model which combines cross-correlation with LPC. Results indicate that the proposed model achieves a huge accuracy and smaller response time comparing to the other two techniques.**

***Index Terms*—bangla voice recognition, speech recognition, face detection, personal assistance, handicapped people**

I. Introduction

Day by day, computer is becoming an integral part of human life and can do almost everything a human asks it to do. To tell a computer what to do, human must provide it necessary commands by any communication means. Voice command has been the most convenient way to communicate with computers. Moreover, it is also helpful to the people who are not so comfortable in operating computers due to lack of knowledge as well as to people with disabilities.

A lot of researches have been done for speech recognition in the English language in the last few years. Although 265.0 million people around the world speak in Bangla, very little work has been done so far for Bangla speech recognition [1]. Moreover, digital personal assistants like Microsoft's Cortana, Apple's Siri, Google's Google Assistant, Amazon's Alexa, etc. don't take input in Bangla language.

The main aim of our research is to make an efficient hands free personal assistant for native Bangla speakers. Our proposed system will take their voice commands given in Bangla as input and then the operating system will perform specific action based on that command. Besides, the operator can move the mouse cursors using his face movements. In addition to being beneficial to disabled people, it can be useful for normal people as well to shorten the amount of time to give commands by eliminating the hassle of typing.

[1] Department of Computer Science and Engineering, Port City International University, Chattogram, Bangladesh, Email: dipucpi50@gmail.com
[2] Department of Computer Science and Engineering, University of Chittagong, Chattogram, Bangladesh, Email: ehfahad01@gmail.com
[3] Department of Computer Science and Engineering, University of Chittagong, Chattogram, Bangladesh, Email: hossain_ms@cu.ac.bd
[4] Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Skellefteå, Sweden, Email: karl.andersson@ltu.se
[5] Department of Computer Science and Engineering, University of Liberal Arts Bangladesh, Dhaka, Bangladesh, Email: sazzad.hossain@ulab.edu.bd

II. Related Works

Speech recognition has been a key interest for the researchers over the decades. Our literature work focuses on the research works on Bangla speech recognition and face detection which are published on renowned academic journals, like Elsevier, Springer, ACM Digital Library, IEEE Digital Library, etc. in the previous five years.

A. Speech Recognition

A lazy learning-based language identification from speech developed in 2019 by Mukherjee et al. [2]. MFCC is used to predict correct phonetic words.

In 2018, Nasib, Kabir, Ruhan Ahmed and Jia Uddin developed a model using an API named Sphinx4, which is written in java [3]. This model is unable to recognize any unknown voice.

A short Bangla speech commands recognition system is developed by Sumon, Chowdhury, Debnath, Mohammed and Momen based on three different CNN architectures [4]. A pre-trained model is developed using a dataset of English short speech commands, and later fine-tuned by re-training it on Bangla dataset.

Md. Masudur, Debopriya and Md. Mahbub developed an automatic speech recognition system for isolated Bangla word using SVM with Dynamic Time Warping(DTW) [5]. MFCC and DTW methods were used for feature matching and SVM with Radial Basis Function(RBF) was used for classification.

In 2016, Khalil and Rahman developed an ANN based approach for recognizing Bangla speech [6]. MFCC analysis is used here to elicit meaningful features for recognition. For MFCC computation, framing, pre-emphasis, windowing, FFT and DCT has been followed.

Nahid, Islam and Saiful developed a system for recognizing Bangla real number automatically using CMU Sphinx4 [7]. Their corpus contains Bangla real digit speech signal and its corresponding text. MFCC features are used in feature extraction.

In 2015, Md Yasin, Hossain and Hoque developed a speaker detection and Bangla word recognition system using semantic Modular Time Delay Neural Network(MTDNN) [8]. The noises are managed by clustering technique. MFCC features are used to recognize Bangla words and speaker detection.

978-1-7281-5852-5/19/$31.00 ©2019 IEEE

B. Face Detection

In 2019, Abiyev and Arslan proposed a CNN based system to move mouse cursor using user's head position and to send command using his eye-blink [9].

In 2016, Gyawal et al. developed a camera mouse system using Robust Camera Mouse for Disable People algorithm [10]. Haar like features and adaboosting algorithm are used for face detection.

In 2014, Epstein, Missimer and Betke developed a video-based mouse-replacement interface using kernels [11]. They used three different kernel-subset-tracker and compared their performance to the optical-flow tracker under two sessions. Each session contains five different experimental conditions. First and second sessions lasted 4.7 minutes and 6.9 minutes respectively.

In 2012, Krolak and Strumillo developed a system which can detect eye-blink for human-computer interaction [12]. Haar-like features and a cascade of boosted tree classifiers are used to detect face.

The reviewed literature indicates that all the previous systems for Bangla speech recognition were developed for a noise-free well-controlled speech data and none of the methods experimented with continuous speech or in noisy environments. In this research, we have tried to developed a digital personal assistant system with Bangla voice command recognition and face detection technique in noiseless(normal room), moderate(class room) and noisy(road) environments especially for the peoples having physical disabilities.

III. Overview of Cross-Correlation

This section presents a short overview of correlation for speech recognition and face detection. It discusses how correlation method finds the similarity between pass-phrase and the user's voice command for speech recognition.

Cross-correlation is a measure of the resemblance of two series as a function of the displacement of one relative to the other. This is also assumed as a sliding dot product. It is a popular technique for searching a long sign for a shorter and known characteristic. It has applications in pattern recognition, single particle analysis, etc. The term cross-correlation is employed for find out to the relationship between the sections of two absolute vectors X and Y. *xcorr* function of MATLAB is a cross-correlation function for succession for a random process which has auto-correlation.

Syntax for Correlation in MATLAB is derived as

$$z = xcorr(x, y)$$

where z returns the cross-correlation of two discrete-time arrangement, x(feature vector) and y(reference vector). Cross-correlation measures the oppressiveness amongst x, y and y's shifted vector(the reverse of vector y). If vector x and y have different lengths, then MATLAB's *zero* function reduces the length of x vector so that it has the similar length of y. This is done by eliminating the silenced portion in x.

The energy of the feature voice is calculated based on their linear transformation of log power apparition. According to [13], cross-correlation of *x(n)* and *y(n)*, can be mathematically expressed as:

$$r_{xy}(n) = \sum_{k=0}^{2.(N-1)} x(k).y(k-n) \tag{1}$$

The outcome of the cross-correlation function relies on the intensity of the recorded voice(y) and the recording length. The standardized cross-correlation is used to conduct cross-correlation regardless of scale and length. Each human voice command is then compared with each reference voice and illustrates how the command is matched with each of them. The recognition results are expressed in 0% to 100% interval [13] as expressed in equation (2).

$$\rho_{xy}(n) = \frac{\sum_{k=0}^{2.(N-1)} x(k)y(k-n)}{\sqrt{E_x.E_y}} \times 100 \tag{2}$$

where,

$$E_y = r_{yy}(0) = \sum_{k=0}^{N-1} y(k).y(k) = \sum_{k=0}^{N-1} |y(k)|^2 \tag{3}$$

$$E_x = r_{xx}(0) = \sum_{k=0}^{N-1} x(k).x(k) = \sum_{k=0}^{N-1} |x(k)|^2 \tag{4}$$

Here *x(n)* and *y(n)* are the feature signal and reference signal. r_{xy} is the value of cross-correlation between two signals which is measured based on time and energies of those signals. r_{xx} and r_{yy} are the energies of *x(n)* and *y(n)* respectively.

IV. Methodology

A. Problem Description

The primary step was to find out cross-correlation between the featured signal and the pre-recorded (reference) signals within the time domain. We recorded all the reference signals in 44100 KHz.

At first, the system prompts for voice command to be executed. If the cross-correlation between the input command and the reference signals is more than the threshold value, then the appropriate voice command associated with each reference voice will be executed. And if the value of cross-correlation is below the limit, the system will reply 'Not Found'. We have also integrated a method for face detection so that people can control mouse cursor using their face movement. To go to the cursor movement mode, user need to say the 12th Bangla voice command given in figure 2.

We implemented our work in MATLAB 2018 and tested the system in Windows 10 operating system. The Document window will be opened if the first command is said. The second command will open the picture window and similarly Video, Desktop, C drive, E Drive window will be opened

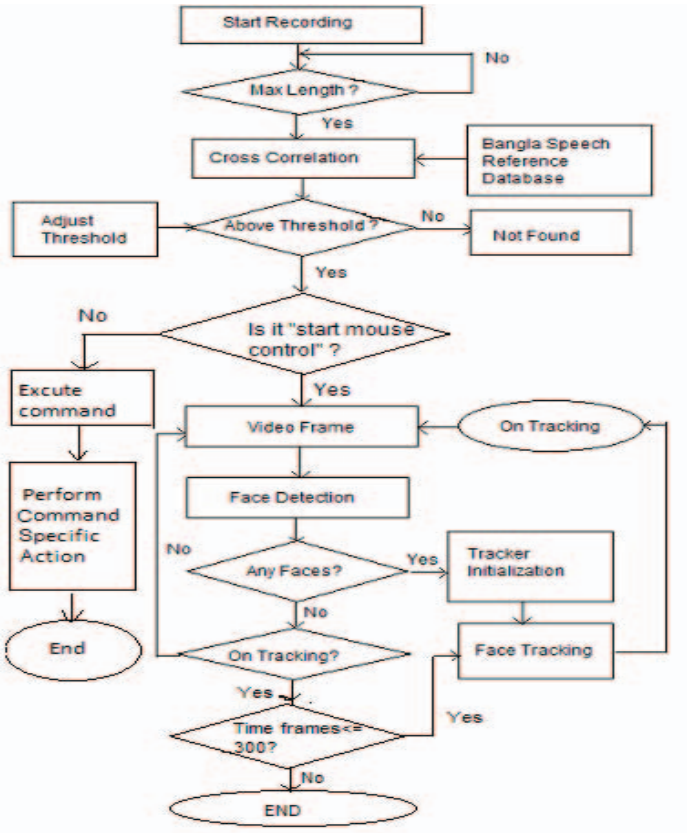

Fig. 1: System architecture of proposed digital personal assistant

for the 3rd, 4th, 5th and 6th commands respectively. Notepad application will be started if 7th command is recognized. A new folder will be created in the last known location if the user says the 8^{th}. For example, if a user opens C drive by using the 5th command, and then says the 8th command, a new folder will be created in the C drive. Here we set by-default folder name as ”new folder”. Similarly, user can delete a new folder by saying the 9th command. If the location has already a folder named ”new folder”, then another new folder will not be created. User can open Google Chrome web browser using 10th command. If the system recognizes the 11th command, then the Facebook's homepage will be opened on Google Chrome web browser. Users can turn on the mouse control system by saying the 12th command. After saying the 12th command, the webcam will be turned on and a live video frame will be shown on screen. Then the user needs to be stabilized for a moment so that a good picture of the user's face can be captured. Then it will automatically detect the user's face and start moving the mouse cursor by tracking the movement of the user's face. Once the user enters in the mouse control mode, it will be active for 20 seconds. After that, user will be shifted back to the normal execution mode.

B. Dataset Creation

The data set contains the speech signals which is mandatory for developing a speech recognition system. We recorded voice commands and added them to the database. These commands are given in figure 2.

১। ডকুমেন্ট খুলুন (Goto Documents)
২। পিকচার খুলুন (Goto Pictures)
৩। ভিডিও খুলুন (Goto Videos)
৪। ডেস্কটপ খুলুন (Goto Desktop)
৫। সি ড্রাইভে যান (Goto C-drive)
৬। ই ড্রাইভে যান (Goto E-drive)
৭। নোটপ্যাড খুলুন (Open Notepad)
৮। ফোল্ডার খুলুন (Create folder)
৯। ফোল্ডার মুছুন (Delete Folder)
১০। গুগুল ক্রোমে যান (Goto Google Chrome)
১১। ফেসবুকে যান (Goto Facebook)
১২। মাউস চালু (Start Mouse Control)

Fig. 2: Voice commands used in the proposed systems

The summary of the training dataset is given below:

- It contains 240 audio files of 12 different Bangla reference voices
- Number of speaker : 5 (from different parts of Bangladesh, age range: 20-25).
- 24 reference signals of each person, 2 for each command.
- Average words per sentence: 2-3.
- Total hours: 1 hour.

We have recorded these sentences which we generated in noiseless environment(normal room), moderate environment(class room) and noisy environment(road) using a microphone. We have taken speech data from different speakers from different parts of Bangladesh so that it would be more natural. We can download the reference dataset from GitHub [14].

C. Speech Recognition

In this research, we use the following algorithm for speech recognition technique:
Step 1: Record reference voice commands from the users
Step 2: Save those voice commands in vector form to create training data
Step 3: At the time of accessing the system, prompt for giving Bangla voice commands to be executed
Step 4: Record and save the spoken command in a different vector
Step 5: Cross Correlate the new vector with the previously saved vectors
Step 6: Find the maximum of $\rho(0)$ [13]
Step 7: If $\rho(0) > threshold$, then recognize the Bangla voice command. Otherwise, tell the user ”Sorry, didn't understand your command”

D. Face Detection

Mouse cursor movement using face detection scheme consists of three primary stages and occurs in the following order: detection, monitoring and tracking. We detect the human face using the Viola-Jones algorithm [15] which is implemented

on MATLAB's Computer Vision toolbox [16]. The system extracts features from the identified area once a face is detected. The extracted features are then monitored as the user moves his head around to regulate the mouse pointer. The movement of the user's head is translated into the movement of the mouse pointer on the computer. If the tracking features are lost at any point during the tracking phase, the face is detected again. The following subsections explain how to perform detection, monitoring, and pointer control.

1) Detection: This scheme detects the face using Histogram of Gradients (HOG) matching templates. At first, an image of user's face is captured and converted to a vector. And then it continuously captures still images from the live face movement of user and compares with the reference image. Then it compares the HOG feature of the reference images with continuously captured images and moves the mouse cursor accordingly.

The HOGs of an image is computed by dividing the image into 8×8 disjoint block of pixels. The orientation of each of these blocks is binned into 9 equal spaced bins between $-pi/2$ and $pi/2$. Hence, an image with height H, and width W, will have a HOG represented by a three-dimensional array with the dimensions of $H/8 \times W/8 \times 9$.

Cross-correlation is used to compare the HOGs of two images. After performing non-maxima suppression, the 8×8 block with the largest cross-correlation measure is considered to be the center of the detected face.

Once the face is detected, the initial features in the detected region are extracted using the minimum eigenvalue algorithm [17]

2) Monitoring: The feature points are then tracked using the Kanade-Lucas-Tomasi (KLT) algorithm [18]. To make the tracking robust, the threshold for the bidirectional error of the KLT is set to 3 pixels. To implement the feature extraction and the tracking algorithms, we used the MATLAB's *detectMinEigenFeatures* function [19] and *PointTracker* object [20] from the Computer Vision System Toolbox [16]. If the lighting conditions are not constant or if the head movements are too fast, a significant portion of the feature trackers is lost. Hence, when less than 10 feature points remain, the system notifies the user about the loss of tracking, and asks the user to stay still to re-detect the face of the user using the HOG of the template specified.

3) Pointer Control: To control the mouse pointer according to the head movements, we used a Linear Rate Control(LRC) approach [21]. In a Rate Control approach, the movement of the pointer depends only on the succeeding motions of the head and doesn't consider any absolute reference position/frame.

When the face has been detected and the tracker has been initialized, mouse pointer will be displayed to the center of the screen. In every frame, we compute the mean of the x-coordinates of all the feature points and the mean of the y-coordinates of all the feature points being tracked, and use the resulting means to demarcate the head-position. We apply Ax, a scaled version of this displacement to determine the next coordinates of the mouse pointer. Empirically, a value of 15 for A worked reasonably well. The whole source code of our research work can be downloaded from GitHub [22].

V. Results & Comparison

We validated our model in three different environments as we mentioned earlier. Following three subsections discuss about these environments. We considered a normal room environment as noiseless environment, class room as moderate environment and city road as noisy environment. We have taken each command 5 times from different users and tested our model using these 60 Bangla voice commands. Our system successfully recognizes 50 commands in noiseless and moderate environments and 45 commands in noisy environments. The accuracy of the model is 83%, 83% and 75% in noiseless, moderate and noisy environments respectively. On average, our proposed Bangla digital personal assistant remarkably recognizes 80.34% Bangla voice commands and responds within 2-3 seconds. Figure 3, 4 and 5 show the confusion matrix for noiseless, moderate and noisy environments respectively.

In Noiseless Environment
Expected Result

System Result	Time Of test	5	5	5	5	5	5	5	5	5	5	5	5	60
	Voice Command	ডকুমেন্ট খুলুন (Goto Document)	পিকচার খুলুন (Goto Picture)	ভিডিও খুলুন (Goto Video)	ডেস্কটপ খুলুন (Goto Desktop)	সি ড্রাইভে যান (Goto C drive)	ই ড্রাইভে যান (Goto E drive)	নোটপ্যাড খুলুন (Open Notepad)	ফোল্ডার খুলুন (Create Folder)	ফোল্ডার মুছুন (Erase Folder)	গুগল ক্রোমে যান (Goto Google Chrome)	ফেসবুক যান (Goto Facebook)	মাউস চালু (Start Mouse Control)	Total
	ডকুমেন্ট খুলুন (Goto Document)	5												5
	পিকচার খুলুন (Goto Picture)		5											5
	ভিডিও খুলুন (Goto Video)			5	1									5
	ডেস্কটপ খুলুন (Goto Desktop)				4									4
	সি ড্রাইভে যান (Goto C drive)					3	1	1						3
	ই ড্রাইভে যান (Goto E Drive)					2	4							4
	নোটপ্যাড খুলুন (Open Notepad)							4	2					4
	ফোল্ডার খুলুন (Create Folder)								3	2				3
	ফোল্ডার মুছুন (Erase Folder)									3				3
	গুগল ক্রোমে যান (Goto Google Chrome)										5			5
	ফেসবুক যান (Goto Facebook)											5	1	5
	মাউস চালু (Start Mouse Control)												4	4
	Total	5	5	5	4	3	4	4	3	3	5	5	4	50

Fig. 3: Confusion Matrix for noiseless environment

A. Face detection

Under constant lighting conditions and steady head movements, the system performs quite well, more than 10 tracking features are successfully track within 300 time frames. With regular tracking and display of the tracking features, the system achieves a frame rate of 9.08 fps. Without the display of the tracking features, the system achieves a frame rate of 10.25 fps. With no feedback at all, the system achieves a frame rate of 18.88 fps.

Over a set of 500 frames, we compared the user's true head position which is determined by the system in every $100th$ frame. On average, the Euclidean distance between the

In Moderate Environment
Expected Result

System Result \ Voice Command	ডকুমেন্ট খুলুন (Goto Document)	পিকচার খুলুন (Goto Picture)	ভিডিও খুলুন (Goto Video)	ডেস্কটপ খুলুন (Goto Desktop)	সি ড্রাইভে যান (Goto C drive)	ই ড্রাইভে যান (Goto E Drive)	নোটপ্যাড খুলুন (Open Notepad)	ফোল্ডার খুলুন (Create Folder)	ফোল্ডার মুছুন (Erase Folder)	গুগল ক্রোমে যান (Goto Google Chrome)	ফেসবুক যান (Goto Facebook)	মাউস চালু (Start Mouse Control)	Total
Time Of test	5	5	5	5	5	5	5	5	5	5	5	5	60
ডকুমেন্ট খুলুন (Goto Document)	5									1			5
পিকচার খুলুন (Goto Picture)		5											5
ভিডিও খুলুন (Goto Video)			4	2		2							4
ডেস্কটপ খুলুন (Goto Desktop)				2									2
সি ড্রাইভে যান (Goto C drive)					5					1			5
ই ড্রাইভে যান (Goto E Drive)			1			1							1
নোটপ্যাড খুলুন (Open Notepad)				1			5						5
ফোল্ডার খুলুন (Create Folder)						2		5					5
ফোল্ডার মুছুন (Erase Folder)									5				5
গুগল ক্রোমে যান (Goto Google Chrome)										3			3
ফেসবুক যান (Goto Facebook)											5		5
মাউস চালু (Start Mouse Control)												5	5
Total	5	5	4	2	5	1	5	5	5	3	5	5	50

Fig. 4: Confusion Matrix for moderate environment

In Noisy Environment
Expected Result

System Result \ Voice Command	ডকুমেন্ট খুলুন (Goto Document)	পিকচার খুলুন (Goto Picture)	ভিডিও খুলুন (Goto Video)	ডেস্কটপ খুলুন (Goto Desktop)	সি ড্রাইভে যান (Goto C drive)	ই ড্রাইভে যান (Goto E Drive)	নোটপ্যাড খুলুন (Open Notepad)	ফোল্ডার খুলুন (Create Folder)	ফোল্ডার মুছুন (Erase Folder)	গুগল ক্রোমে যান (Goto Google Chrome)	ফেসবুক যান (Goto Facebook)	মাউস চালু (Start Mouse Control)	Total
Time Of test	5	5	5	5	5	5	5	5	5	5	5	5	60
ডকুমেন্ট খুলুন (Goto Document)	5			1			1				1		5
পিকচার খুলুন (Goto Picture)		5											5
ভিডিও খুলুন (Goto Video)			5	1									5
ডেস্কটপ খুলুন (Goto Desktop)				3			1						3
সি ড্রাইভে যান (Goto C drive)					3								3
ই ড্রাইভে যান (Goto E Drive)					2	5					1		5
নোটপ্যাড খুলুন (Open Notepad)							1						1
ফোল্ডার খুলুন (Create Folder)								3					3
ফোল্ডার মুছুন (Erase Folder)							1	2	5	1			5
গুগল ক্রোমে যান (Goto Google Chrome)										4	1		4
ফেসবুক যান (Goto Facebook)											2	1	2
মাউস চালু (Start Mouse Control)							1					4	4
Total	5	5	5	3	3	5	1	3	5	4	2	4	45

Fig. 5: Confusion Matrix for noisy environment

true head-position and the mean head-position was 401 pixels which are 1.06% of the face template's diagonal length.

B. Comparison

Finally, we compared our proposed model with two hybrid techniques: i) combination of MFCC (Mel-Frequency Cepstrum Coefficients) and DTW (Dynamic Time Wrapping), and ii) combination of Cross-Correlation with LPC (Linear Predictive Coding).

1) Combination of MFCC and DTW: MFCC method is used for audio data retrieval from the user's voice commands. We could not use more than 65 voice samples to train the model as this model's response time exceeds 60s for larger training dataset. We used high-pass pre-emphasis filter which removes the DC component of the signal and Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. It removes the ripples of the signal which are found after Fourier Transform. DTW is used for comparing the voice command with reference signals. The accuracy of this approach is only 36.37% and also takes around 50 seconds to respond.

2) Combination of Cross-Correlation and LPC: LPC methods provide extremely accurate estimates of speech parameters, and does it extremely efficiently. We have used a high-pass hamming window filter for finding the maximum similarity between the references matrix and feature matrix using auto correlation. This model produces a quick response, but it has an accuracy of just 27.28%.

VI. Conclusion & Future Work

In this research, we proposed a digital personal assistant for Bangla native speakers so that they can use their commands to operate a computer system. This model should be very helpful to the physically handicapped people. We developed our system using the cross-correlation method, which finds the closely matched pre-recorded reference signal with the given Bangla voice commands. After successfully recognizing a command, the system performs that action specified in that command. We also embedded a mouse cursor movement feature so that the user can control the mouse using their facial movements. We compared the performance of our system with two other hybrid models. It is clear from the results that, our proposed method imposes its superiority in terms of accuracy and response time.

We have validated our system for 12 Bangla voice commands. In future, we will build a more sophisticated system so that it can recognize a huge number of Bangla voice commands and can perform those actions. We will also work on to reduce the response time within a second. Instead of face movement, mouse cursor control can be done using eyeball tracking. Moreover, we will also train our model in a wide variety of environments with a large number of training voices under uncertainty and will use the sophisticated existing models [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] to handle this phenomenon in a more natural way.

References

[1] Ethnologue, "Bengali: A language of bangladesh." https://www.ethnologue.com/language/ben, 2019. [Online; accessed 19-August-2019].

[2] H. Mukherjee, S. M. Obaidullah, K. Santosh, S. Phadikar, and K. Roy, "A lazy learning-based language identification from speech using mfcc-2 features," *International Journal of Machine Learning and Cybernetics*, pp. 1–14, 2019.

[3] A. U. Nasib, H. Kabir, R. Ahmed, and J. Uddin, "A real time speech to text conversion technique for bengali language," in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pp. 1–4, IEEE, 2018.

[4] S. A. Sumon, J. Chowdhury, S. Debnath, N. Mohammed, and S. Momen, "Bangla short speech commands recognition using convolutional neural networks," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–6, IEEE, 2018.

[5] M. M. Rahman, D. R. Dipta, and M. M. Hasan, "Dynamic time warping assisted svm classifier for bangla speech recognition," in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pp. 1–6, IEEE, 2018.

[6] K. Ahammad and M. M. Rahman, "Connected bangla speech recognition using artificial neural network," *International Journal of Computer Applications*, vol. 149, no. 9, pp. 38–41, 2016.

[7] M. M. H. Nahid, M. A. Islam, and M. S. Islam, "A noble approach for recognizing bangla real number automatically using cmu sphinx4," in *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pp. 844–849, IEEE, 2016.

[8] M. Y. A. Khan, S. M. Hossain, and M. M. Hoque, "Isolated bangla word recognition and speaker detection by semantic modular time delay neural network (mtdnn)," in *2015 18th International Conference on Computer and Information Technology (ICCIT)*, pp. 560–565, IEEE, 2015.

[9] R. H. Abiyev and M. Arslan, "Head mouse control system for people with disabilities," *Expert Systems*, vol. 0, no. 0, p. e12398. e12398 EXSY-Sep-18-364.R2.

[10] P. Gyawal, A. Alsadoon, P. W. C. Prasad, L. S. Hoe, and A. Elchouemi, "A novel robust camera mouse for disabled people (rcmdp)," in *2016 7th International Conference on Information and Communication Systems (ICICS)*, pp. 217–220, April 2016.

[11] S. Epstein, E. Missimer, and M. Betke, "Using kernels for a video-based mouse-replacement interface," *Personal and Ubiquitous Computing*, vol. 18, no. 1, pp. 47–60, 2014.

[12] A. Królak and P. Strumiłło, "Eye-blink detection system for human–computer interaction," *Universal Access in the Information Society*, vol. 11, no. 4, pp. 409–419, 2012.

[13] M. R. Alam, "Isolated speech recognition system based on cross-correlation technique," 11 2015.

[14] Github, "A digital personal assistance using bangla voice command and face detection for handicapped people." https://github.com/dipu528447/a-digital-personal-asistance-using-bangla-voice-command-and-face-detection/tree/master/auto-correlation/train3, 2019. [Online; accessed 24-October-2019].

[15] P. Viola, M. Jones, *et al.*, "Rapid object detection using a boosted cascade of simple features," *CVPR (1)*, vol. 1, no. 511-518, p. 3, 2001.

[16] Mathworks, "Computer Vision System Toolbox." http://www.mathworks.com/help/vision/index.html, 2019. [Online; accessed 19-August-2019].

[17] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pp. 593–600, IEEE, 1994.

[18] C. Tomasi and T. Kanade, *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

[19] Mathworks, "detectMinEigenFeatures MATLAB function," 2019.

[20] Mathworks, "vision.PointTracker MATLAB object." http://www.mathworks.com/help/vision/ref/vision.pointtracker-class.html, 2019. [Online; accessed 19-August-2019].

[21] R. Kjeldsen, "Improvements in vision-based pointer control," in *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pp. 189–196, ACM, 2006.

[22] Github, "A digital personal assistance using bangla voice command and face detection for handicapped people." https://github.com/dipu528447/a-digital-personal-asistance-using-bangla-voice-command-and-face-detection/tree/master/auto-correlation, 2019. [Online; accessed 24-October-2019].

[23] T. Uddin Ahmed, S. Hossain, M. S. Hossain, R. Ul Islam, and K. Andersson, "Facial expression recognition using convolutional neural network with data augmentation," in *Joint 2019 8th International Conference on Informatics, Electronics & Vision (ICIEV)*, 2019.

[24] R. R. Chowdhury, M. S. Hossain, R. Ul Islam, K. Andersson, and S. Hossain, "Bangla handwritten character recognition using convolutional neural network with data augmentation," in *Joint 2019 8th International Conference on Informatics, Electronics & Vision (ICIEV)*, 2019.

[25] M. Islam, M. S. Hossain, R. Ul Islam, K. Andersson, *et al.*, "Static hand gesture recognition using convolutional neural network with data augmentation," in *Joint 2019 8th International Conference on Informatics, Electronics & Vision (ICIEV)*, IEEE, 2019.

[26] R. Karim, K. Andersson, M. S. Hossain, M. J. Uddin, and M. P. Meah, "A belief rule based expert system to assess clinical bronchopneumonia suspicion," in *2016 Future Technologies Conference (FTC)*, pp. 655–660, IEEE, 2016.

[27] M. S. Hossain, K. Andersson, and S. Naznin, "A belief rule based expert system to diagnose measles under uncertainty," in *World Congress in Computer Science, Computer Engineering, and Applied Computing (WORLDCOMP'15): The 2015 International Conference on Health Informatics and Medical Systems 27/07/2015-30/07/2015*, pp. 17–23, CSREA Press, 2015.

[28] M. S. Hossain, S. Rahaman, R. Mustafa, and K. Andersson, "A belief rule-based expert system to assess suspicion of acute coronary syndrome (acs) under uncertainty," *Soft Computing*, vol. 22, no. 22, pp. 7571–7586, 2018.

[29] M. S. Hossain, S. Rahaman, A.-L. Kor, K. Andersson, and C. Pattinson, "A belief rule based expert system for datacenter pue prediction under uncertainty," *IEEE Transactions on Sustainable Computing*, vol. 2, no. 2, pp. 140–153, 2017.

[30] M. S. Hossain, I. B. Habib, and K. Andersson, "A belief rule based expert system to diagnose dengue fever under uncertainty," in *2017 Computing Conference*, pp. 179–186, IEEE, 2017.

[31] R. Ul Islam, K. Andersson, and M. S. Hossain, "A web based belief rule based expert system to predict flood," in *Proceedings of the 17th International conference on information integration and web-based applications & services*, p. 3, ACM, 2015.

[32] M. N. Jamil, M. S. Hossain, R. Ul Islam, and K. Andersson, "A belief rule based expert system for evaluating technological innovation capability of high-tech firms under uncertainty," in *Joint 2019 8th International Conference on Informatics, Electronics & Vision (ICIEV)*, IEEE, 2019.