Virtual Personal Assistant (VPA) for Bengali language with BanglaSTT (Whisper) with RL, BanglaBERT (BERT) and BanglaTTS (SileroTTS)

Sadidul Islam
Student, MSCSE
United Internation University
Dhaka, Bangladesh

Khandokar A. Mamun
Professor & Director - IRIIC
United Internation University
Dhaka, Bangladesh

***Abstract*—In the current technological world, virtual personal assistants (VPA) have become necessary companions, offering seamless voice-activated interaction with technologies, personalized assistance, and streamlined task management for enhanced productivity and convenience. In this paper, we propose a virtual personal assistant development with speech to text, chatbot functionality, and text to speech synthesis where speech to text is implemented as Reinforcement Learning model so that it can personalize the user specific voice characteristics. By leveraging chatbot capabilities, the assistant can engage in natural language processing, understand user intents, and provide relevant responses. The text-to-speech synthesis component ensures that the VPA can generate human-like speech output, enhancing the user experience. The developed VPA showcases the potential of combining multiple technologies to create a versatile and user-friendly virtual assistant.**

I. Introduction

Technology has made remarkable strides in recent years, revolutionizing the way we live, work, and connect with each other. We can make life easier by having a personal assistant installed on a personal computer or mobile phone. All the technologies require human interaction to operate, e.g., computer, mobile, IoT devices, etc. There are situations that humans can't provide proper interactions due to disabilities, lack of knowledge, and also for limited time as well. A virtual personal assistant (VPA) is built with Machine Learning technologies such as Natural Language Processing (NLP), Speech Recognition, which can help in these situations by taking a voice command from the human and completing the tasks automatically. For example, A virtual personal assistant (VPA) can be used to set a reminder, send email, know weather information, search information in Google, update to-do list, etc [1].

Numerous works for the Virtual Personal Assistant (VPA) have been done for the English language throughout recent years. Such as, Apple's Siri[11], Google's Voice Actions[12] and Google Now[13], Microsoft's Bing Voice Search[14], and Nuance's Dragon Go! [15] and Nina[16], and many startup efforts like Speaktoit[17], and many more.

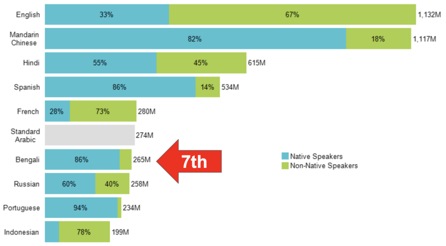


Figure 1. Top Language Speaker in the World

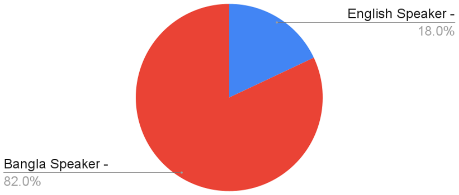


Figure 2. Top Language Speaker in Bangladesh

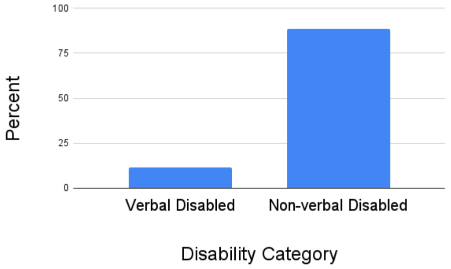


Figure 3. Statistics for Categories in Disability in Bangladesh

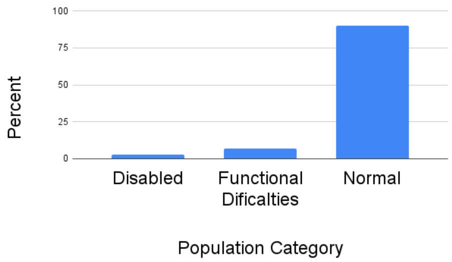


Figure 4. Population distribution in Bangladesh in regard to Disability

Bangla is the 7th most widely spoken language in the world, 272.7 million people (Figure 1) around the world speak in Bangla [18]. Euro-monitor International reports only 18% (Figure 2) of the total population in Bangladesh can understand and speak English[22]. In addition, to that 25.34% (literacy rate 74.66% according to the Population and Housing Census 2022) of the total population is illiterate who cannot even read or write Bangla language[20] According to Unicef Bangladesh, 2.8 percent (Figure 4) of the population and 1.7 per cent of children have at least one disability[19] in the country where Bengali is the primary language. Among the disabled people, 11.43 (0.32 percent in the whole population) percent has limitations in speaking, and other 88.57 percent can at least speak (Figure 3), but still they are away from technologies by other disabilities.

Although, very little work related to virtual personal assistant (VPA) had been done yet[1]. We can't find any production grade virtual personal assistant (VPA) in the Bengali language as of today.

In this research, we propose a system with Speech to Text, Chatbot, and Text to Speech combined to develop a virtual personal assistant (VPA) that can help disabled (verbal) people to interact with technologies, save time for busy people by doing their technological tasks in no time wasted with just a voice command. This virtual personal assistant will take voice commands, and based on that voice command it will do some actions like calling someone, interact with IoT devices, etc., and reply with the required information to answer the users' voice command. In the other virtual personal assistant (VPA) we found that it lacks of personalizing for a specific person, but the solution to this problem is more required when we consider a disabled person. Because they might not speak properly like other people, if we don't consider this problem, then it won't help them at all. So, we introduce Reinforcement Learning (RL) for Speech-to-Text model to personalize user-specific voice characteristics.

II. Review

Virtual personal assistant (VPA) has been researched and utilized by researchers over the decades. In our literature work, we focused on the latest works including research, production related information which are published in reliable sources. Like, IEEE Digital Library, ACM Digital Library, Springer etc. We have found a couple of research directly related to our works, and there are some partially related works. We have separated discussion of the researches in separate sections.

A. *Virtual Personal Assistant*

Adheetee: A Comprehensive Bangla Virtual Assistant is developed based on Natural Language Understanding (NLU) using Deep Learning (DL) models such as Recurrent Neural Network (RNN) for Bangla Language [1]. They have collected and created their own corpus then trained the model on that dataset.

In the paper "Large-Scale Personal Assistant Technology Deployment, specifically in the context of the Siri Experience" [8], the researchers provided a lot of information about development and deployment a production grade Siri, Siri is a virtual personal assistant developed by Apple. They have showed how they have implemented architecture, deployment architecture and how they have improved performance Natural Language Processing (NLP) and Natural Language Understanding (NLU) capabilities.

B. *Speech Recognition*

The paper State of Art Research in Bengali Speech Recognition [2] presents a comprehensive methodology for advancing the field of Bengali speech recognition. They have pointed out about challenges specific to Bengali speech recognition. The researchers proposed a hybrid approach that combines acoustic modeling, language modeling, and lexical modeling using Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) to improve accuracy and performance.

The Whisper has tremendous success in speech recognition in recent times by implementing an encoder-decoder transformer model with attention the capability of a multitask training format to reduce the complexity of the model. [3] For example, any fully featured speech recognition system can involve many components like, voice activity detection, speaker diarization, and inverse text normalization. Instead of using all of them separately, whisper uses one model to do all the tasks.

C. *Chatbot*

The paper "Doly: Bengali Chatbot for Bengali Education" [4] presents the methodology employed to develop an intelligent chatbot named Doly, designed specifically for Bengali language education.

The methodology emphasizes the importance of continually updating and expanding Doly's knowledge base to keep up with evolving educational content and trends.

BanglaBERT is a language model/dataset developed for Bengali (Bangla), so far the best open-sourced dataset and pre-trained model found online. [5] It serves as a pre-training and benchmarking tool, enabling effective natural language processing tasks in Bengali, addressing the need for improved language models in Bengali language.

LLaMA offers various foundation language models based on parameters ranging from 7B to 65B competing all the existing best performing LLMs. [7] To improve the performance of the transformer architecture, they have multiple improvements proposed by recent researches like PaLM, GPT3. For example, to improve training stability, they normalize the input of the transformer sub-layer instead of normalizing the output of the transformer.

D. Text to Speech

Tacotron 2 is one of the best performing Text to Speech models as of today, which consists of a recurrent sequence-to-sequence feature prediction network, inspired by Tacotron-style models, which generates mel-scale spectrograms from character embeddings and a modified WaveNet model functioning as a vocoder, converting these spectrograms into time-domain waveforms to produce synthesized speech. [6] This paper presents an innovative neural approach to speech synthesis by combining the strengths of previous methods, like WaveNet, Deep Voice.

III. Methodology

Our solution has three machine learning (ML) components, Speech to Text, Chatbot, and Text to Speech. Representing sounds or human voice in digital media involves converting them into analog signals. The main challenge lies in converting these analog signals into text format, which enables machines to process and work with the data effectively, and the Speech to Text solves this problem[2]. A Chatbot is a computer program or an artificial intelligence which conducts a conversation via auditory or textual methods[9]. Text-to-Speech (TTS) is a technology that converts written text into spoken words[10].

Figure 5 shows a user provides a voice input, then our system processes the voice data and provides text data from Speech-to-Text (BanglaSTT)[24] model. Speech to Text returns text data for the voice data, this data is the input of the chatbot (BanglaBERT)[5], chatbot takes the decision, like what to do, what to reply, or what to ask, these kind of questions. If the response is to take an action based on the output of the chatbot, then it takes the action, like call someone, interaction with IoT devices etc. Also, provide the voice feedback to the user using Text to Speech (BanglaTTS)[23]. There are plenty of solutions for other top languages, but we have developed for Bengali language.

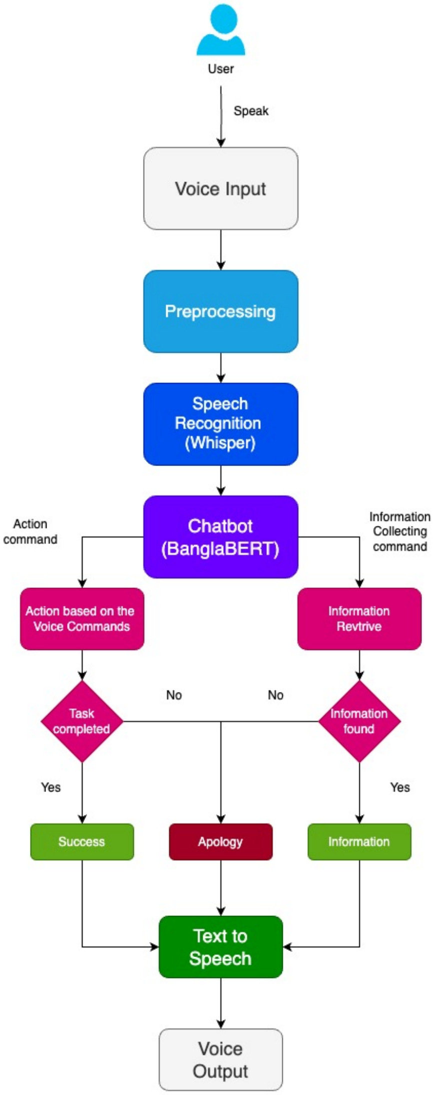


Figure 5. Flowchart of the system

Our deployment environment (Figure 6) is in the cloud, using API access a mobile/computer app can interact with the system. A basic model will be downloaded into the local device as well to make the communication with reduced latency. When a user provides some basic voice command, the local model will be sufficient to reply or interact with that. But when a complex or unknown reply is required, then it will communicate with the cloud version of the model. The models are self-improvable, for example, it will learn to understand a specific user in a better way by user's feedback and Reinforcement Learning (RL). Also, the local model is trained with the most

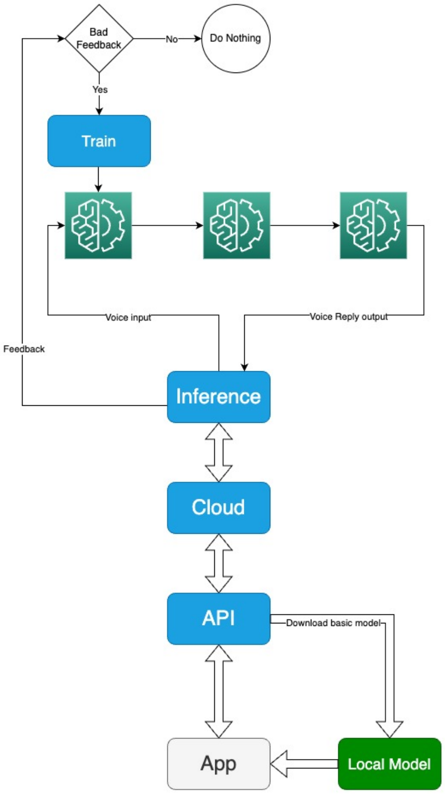


Figure 6. Deployment architecture of the system

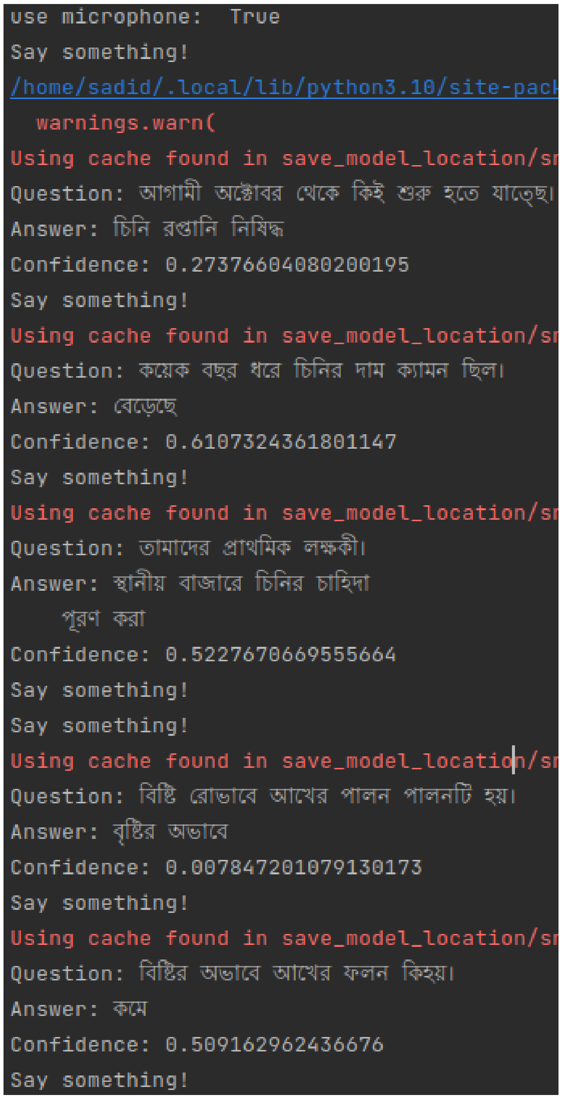
commonly used commands, so that it can respond quickly for the most common words for a specific user.

IV. Result

The implemented Virtual Personal Assistant combines state-of-the-art technologies for speech recognition (Whisper), chatbot functionality (BanglaBERT), and Bangla text-to-speech capability. The system effectively comprehends user speech input in the Bengali language and demonstrates the ability to accurately extract relevant answers from provided contexts using the BanglaBERT model. This extracted information is then vocalized in Bengali through text-to-speech synthesis.

The system's performance showcases its proficiency in understanding spoken Bengali input and subsequently utilizing BanglaBERT to locate and articulate correct responses. Although the system has not been extensively customized or fine-tuned to cater to specific user demands, the successful implementation serves as a proof-of-concept that customization is possible. This adaptability suggests potential for tailoring the system to fulfill a wide array of user requirements.

As evidence of functionality, a provided screenshot [Figure 7] displays a dialogue interaction, wherein the user speaks in Bengali and the system responds appropriately. This interaction underscores the system's real-time capabilities in processing user input, deciphering context, and generating coherent spoken Bengali output.



```
use microphone:  True
Say something!
/home/sadid/.local/lib/python3.10/site-pack
  warnings.warn(
Using cache found in save_model_location/sr
Question: আগামী অক্টোবর থেকে কিই শুরু হতে যাচ্ছে।
Answer: চিনি রপ্তানি নিষিদ্ধ
Confidence: 0.27376604080200195
Say something!
Using cache found in save_model_location/sr
Question: কয়েক বছর ধরে চিনির দাম ক্যামন ছিল।
Answer: বেড়েছে
Confidence: 0.6107324361801147
Say something!
Using cache found in save_model_location/sr
Question: তামাদের প্রাথমিক লক্ষকী।
Answer: স্থানীয় বাজারে চিনির চাহিদা
    পূরণ করা
Confidence: 0.5227670669555664
Say something!
Say something!
Using cache found in save_model_location/sr
Question: বিষ্টি রোভাবে আখের পালন পালনটি হয়।
Answer: বৃষ্টির অভাবে
Confidence: 0.007847201079130173
Say something!
Using cache found in save_model_location/sr
Question: বিষ্টির অভাবে আখের ফলন কিহয়।
Answer: কমে
Confidence: 0.509162962436676
Say something!
```

Figure 7. Screenshot—Output of the program

In summary, the Virtual Personal Assistant effectively integrates prevalent technologies to comprehend and respond to Bengali speech inputs, demonstrating its potential for customization to meet diverse user needs.

V. Conclusion

In this paper, we presented the development of a Virtual Personal Assistant (VPA) developed with state of the earth technologies to enhance user interactions with digital world e.g. IoT devices. VPAs have become a necessary companions by streamlining tasks and providing personalized assistance through voice-activated interactions. Our proposed VPA system integrates speech-to-text conversion, chatbot capabilities, and text-to-speech synthesis, collectively designed to offer a versatile and user-friendly virtual assistant experience.

We have proposed to implement a Reinforcement Learning model for speech-to-text conversion. So that the VPA can personalize its understanding of user-specific voice characteristics and serving individuals with diverse speech patterns, it can also adapt to those who are having disabilities. After getting the speech into a text form then it passes that text into the chatbot, the chatbot functionality enables the assistant to engage in natural language processing, parse user intents, and provide relevant responses from the given context. Furthermore, the text-to-speech synthesis enhances the user experience by generating human-like speech output.

The deployed VPA effectively handles Bengali speech input, utilizing the BanglaBERT model to accurately extract relevant information from provided contexts. Currently, the system is not implemented to specific user needs to do the tasks, but its successful implementation validates the potential for future customization and making it a production grade system.

Provided a screenshot7 to validate the system's functionality, illustrating a real-time dialogue interaction where the VPA comprehends and responds to Bengali speech input. Overall, our work represents the combination of multiple technologies in creating a versatile VPA, which holds promise for a wide area of applications and user scenarios, especially for users with varying speech capabilities and accent preferences.

References

[1] Syed Mohidul Islam, Most Fowziya Akther Houya, Syed Mynul Islam, Salekul Islam and Nahid Hossain, Adheetee: A Comprehensive Bangla Virtual Assistant, 2019.

[2] S.M. Saiful Islam Badhon, Md. Habibur Rahaman, Farea Rehnuma Rupon, State of art Research in Bengali Speech Recognition, 2020.

[3] Radford et al., Robust Speech Recognition via Large-Scale Weak Supervision, 2022.

[4] Md. Kowsher, Farhana Sharmin Tithi, M Ashraful Alam, Mohammad Nurul Huda, Mir Md Moheuddin, Md. Golam Rosul, Doly: Bengali Chatbot for Bengali Education, 2019.

[5] Bhattacharjee et al., BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla, 2022.

[6] Shen et al., NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS, 2018.

[7] Touvron et al., LLaMA: Open and Efficient Foundation Language Models, 2023.

[8] Jerome R. Bellegarda, "Large–Scale Personal Assistant Technology Deployment: the Siri Experience, 2013.

[9] Shafquat Hussain, Omid Ameri Sianaki, Nedal Ababneh, A Survey on Conversational Agents/Chatbots Classification and Design Techniques, 2019.

[10] Dosik Moon, Web-Based Text-to-Speech Technologies in Foreign Language Learning: Opportunities and Challenges, 2012.

[11] AppleSiri, http://www.apple.com/ios/siri/, 011.

[12] Google Mobile, http://www.google.com/mobile/voice- actions, 2008.

[13] Google Now, http://www.google.com/landing/now, 2012.

[14] Microsoft Tellme, http://www.microsoft.com/en-us/Tellme/consumers/default.aspx, 2008.

[15] Nuance Dragon Go!, http://www.nuance.com/products/dragon-go-in-action/index.htm, 2011.

[16] Nuance Nina, http://www.nuance.com/for-business/by-solution/customer-service-solutions/solutions-services/mobile-customer-service/nina/index.htm, 2012.

[17] Speaktoit Assistant, http://www.speaktoit.com/index.htm, 2012.

[18] Chad Emery, ”The 33 Most Spoken Languages in the World” Langoly https://www.asme.org/engineering-topics/articles/renewable-energy/catching-the-sun (accessed July 14, 2023).

[19] Moyukh Mahtab, Faria Selim, ”UNICEF concerned that more than half of children with disabilities in Bangladesh do not go to school” unicef Bangladesh https://www.unicef.org/bangladesh/en/press-releases/unicef-concerned-more-half-children-disabilities-bangladesh-do-not-go-school (accessed July 14, 2023).

[20] Mir Mohammad Jasim, ”Bangladesh's slow march towards 100% literacy” The Business Standard https://www.tbsnews.net/bangladesh/education/bangladeshs-slow-march-towards-100-literacy-492058 (accessed July 14, 2023).

[21] GovṫBangladesh, ”Report on National Survey on Persons with Disabilities (NSPD) 2021 (December 2022) [EN/BN]” reliefweb https://reliefweb.int/report/bangladesh/report-national-survey-persons-disabilities-nspd-2021-december-2022-enbn (accessed July 14, 2023).

[22] Pinon, Robert; Haydon, Jon (December 2010), The Benefits of the English Language for Individuals and Societies: Quantitative Indicators from Cameroon, Nigeria, Rwanda, Bangladesh and Pakistan, Euromonitor International Ltd

[23] Sifat Hossain, https://github.com/shhossain/BanglaTTS, 2023.

[24] Sifat Hossain, https://github.com/shhossain/BanglaSpeech2Text, 2023.