Automated Sinhala Voice Assistant to Manage Tasks using Natural Language Processing - ශ්‍රී voice

Abstract—Voice assistants are programs on digital devices that listen and respond to verbal commands. In this dynamic world, users can use these voice assistants to manage daily tasks, plan their day, get answers to problems, and for entertainment purposes. Most of the existing voice assistant applications functioned using the English language. Since Sinhala is the native language in Sri Lanka, it is not recognized internationally as well as within Sri Lanka for technical applications. As Sri Lankans are more inclined to use the Sinhala language, it is expected to develop this for the benefit of all the Sri Lankans despite their age and to use their native language through a technical application. Furthermore, the lack of English knowledge will lead to the decline of Information technology literacy. This project expects to take the Sinhala language forward to a standard where it is recognized locally and internationally. Thus, building a mobile application that supports Sinhala voice commands will solve the above-mentioned issues. It will facilitate the individuals to do daily activities efficiently and effortlessly within a less time. Machine Learning and Natural Language Processing are the main technologies used in this project. A computer vision-based algorithm from face detection technology is trained. This application is capable of answering questions, and following the instructions for daily tasks and also can be used for entertainment. Apart from the mentioned specialty, it also supports and responds to the Sinhala language that will be shown off on a display.

Keywords—Voice Assistance application, Sinhala Voice, Natural Language Processing, CNN Model, Neural Machine translation

I. INTRODUCTION

The technology sector is one of the most important industries in the world. In the modern world, people live with technology. Technology has transformed the business world of the twenty-first century. In today's modern world, technology has developed rapidly. All the technologies are designed and developed to make people's lifestyles and daily tasks easier and reduce the busyness of life. Because in this technology era, most people are looking to do something easier without wasting their valuable time and effort. Because of this, today's society is rapidly moving towards technology. The business, Information and Communication Technology (ICT), and fashion sectors rapidly use emerging technologies to automate tasks and events in their fields. When it comes to the way consumers communicate overall, modern technology has had a powerful influence.

Voice assistants have a long history that dates back over a century, which may seem surprising given that apps such as Siri were only released within the last decade. In this technological era, the Usage of voice commands is highly increased.

Because of their busy lives, people are looking for easy options to automate their tasks instead of doing them manually. A voice assistant is a virtual assistant that employs speech recognition, language understanding algorithms, and vocal synthesis to listen for and respond to particular spoken commands, providing relevant information or doing specified actions as desired by the user. By listening for precise keywords and filtering out background noise, voice assistants may deliver useful information to respond to the user's explicit orders, referred to as intents [4]. Virtual assistants, becoming an increasingly common feature of many consumer electronics devices, can respond to user commands, provide information, and assist users in controlling other connected electronics. There are approximately 110 million virtual personal assistant users in the United States, and the software is widely popular on smartphones and smart speakers. Virtual assistants have become an integral part of the smart device business, influencing how users engage with their gadgets. Businesses are increasingly searching for larger and better uses of "smart" technology as the sector evolves and technology progresses. Technologically adept consumers may now speak with their linked homes and automobiles like they do with their smartphones [5].

Sri Lankans also use voice assistant applications daily to automate their tasks. The majority of people use voice assistant applications that function using English as their primary language. The English language is widely used in the business and IT sectors in Sri Lanka. However, some fields in Sri Lanka do not use the English language frequently. Since Sinhala is a regional language in Sri Lanka, it is not recognized internationally or within the country for technical purposes.

Vendors produce the majority of applications in English. To work with these apps, Sri Lankans need to have English knowledge. Nowadays, teenagers and younger generations are learning the English language. Nevertheless, most of the elderly have not learned English as a language and have not experienced technology.

Since Sinhala is a local language in Sri Lanka, Sri Lankans are more inclined to use the Sinhala language. Therefore, making a voice assistant application function using the Sinhala language will help all users use this voice assistant application. Furthermore, this application will have an emotion recognition feature to understand the user and uplift the user's mental health or freedom by providing suggestions, a unique concept in Voice assistant applications. Since this application will have users' sensitive data, the application has enhanced security methods from facial unlock and a very fast and effective voice-to-text and text-to-voice algorithm. This application is expected to develop across all Sri Lankans without any age difference. Thus, by building a mobile application that supports Sinhala voice commands, this project aims to take the Sinhala language forward to a standard where it is recognized locally and internationally. Humans use the internet to save time and work efficiently.

Furthermore, the internet has become universal, and it helps many with socializing. To save time and do things efficiently, people try to do their important work using the internet because it has become universal and delivers services quickly. By building an application that works with the Sinhala language, you get the ability to use modern technology for those who are poor in English [3].

II. RELATED WORKS

A Natural Language Question and Answer System is a grammatical framework that converts English questions and assertions into predicate logic. The user feedback is interpreted and translated into KIF (Knowledge Exchange Format) and transmitted to a logical individual using a series of reasoning first-order axioms to translate and return an answer. They also built a specific English grammar to incorporate this system, based on recent research in head-based grammar phrase structure. The program follows the Theory of English Problem Constructions, Ginsburg, and Sag (future).

Triplet Loss Based Cosine Similarity Metric Learning for Text-independent Speaker Recognition [9]is a research in a text-independent speaker identification mission, the deep network-based embedding of speakers is increasingly popular. Unlike the generically trained I extractor, a DNN speaker embedded extractor in the closed classification scenario typically is equipped with SoftMax. The question we discussed in the paper is to pick a DNN solution for built-in speaker testing. with several choices. One is to use a basic heuristic speaker similarity (e.g. cosine metric) to calculate. [9]

Knowledge Enhanced Hybrid Neural Network for Text Matching [10] is a long text that poses a major challenge because of its complex structures, for neural network-based text approaches. We propose a hybrid Neural Network (KEHNN) which builds on previous knowledge to recognize useful information and filter noise out in long text and performs with multiple perspectives to tackle the challenge.

According to research done in India, they have created a voice personal assistant desktop application. It is designed to make it easy for a person to handle a number of services. Some of the services available are mail exchange, google search, music player, open a camera, and many more.[6] As tools and technology, they use artificial intelligence, natural language processing, voice recognition and machine learning.[6] As tools and technology, they use artificial intelligence, natural language processing, voice recognition, and machine learning. Accordingly, the latest AI programming algorithm is used to give good output to the user. The purpose is to learn from the data input and get a good output to the user. Designed to handle voice and text using natural language processing. They also say that it is set up to reset the response to the user's response. Machine learning is the process of continuously learning new words and phrases. That means the machine learning model learns on its own. Accordingly, for my component, they use machine learning, artificial intelligence, and natural language. Also, since all these things are made for the English voice assistant, hope to develop new algorithms and models.[10]

III. RESEARCH METHODOLOGY

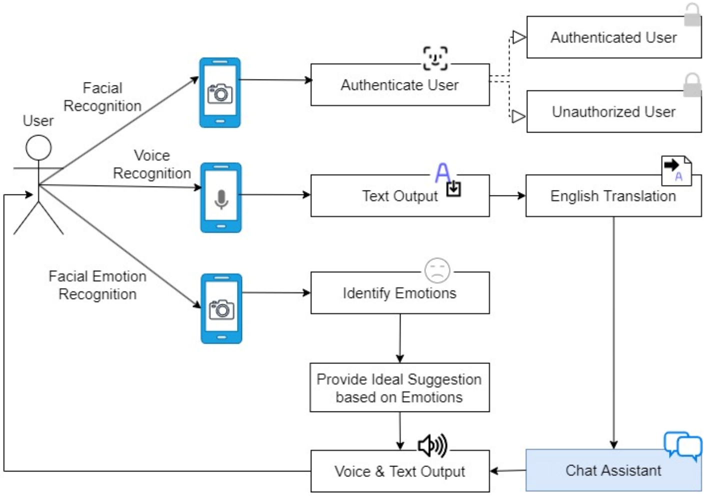

Figure 1: System Overview Diagram

A. *User Authentication using facial Recognition*

This research part is to authenticate user login using facial recognition. In here user's face is stored by the trained model which identifies that face and gives access to enter the application through the login.

There are various face recognition algorithms with varying degrees of accuracy, and some of them beat human performance in ideal situations. However, the majority of these algorithms are proprietary and were pre-trained on a massive non-public dataset. A few are open-source and freely accessible. The models were first created using face recognition data training utilizing open cv, Media pipe, and TensorFlow. Face identification using Haar cascades is a machine learning-based technique that involves training a cascade function using a collection of input data.

Face recognition, picture preprocessing - detecting face landmarks and positioning, producing face embeddings, and classification are all crucial processes in this methodology. Face detection is the initial stage in the face recognition process. Face detection is a well-studied area in the field of computer vision. As a consequence of decades of study, there are currently several machine learning algorithms that are suitable for this purpose. This project seeks to test the Face Net system for face recognition, and a pretrained Face Net model is implemented. To achieve a more accurate output the model is finetuned using the local dataset for this project. CNN's have obtained advanced results in image classification and object identification in recent years. A state-of-the-art CNN cascade is employed for a face identification job in this stage because of its runtime performance. The cascade is made up of six CNNs: three for binary classification (face and non-face) and three for bounding box calibration.

Although doing the facial alignment and front realization is a typical technique in facial recognition, the classifier employed in this study is solely for frontally aligned faces. Both the classifier training and face recognition procedures include picture preprocessing techniques such as photo capture/selection, grey scaling, face identification, image cropping, and image upload. The fundamental picture pre-processing step is shown in the following code sample.

The picture is converted to grayscale and the histogram is equalized during the image preparation step. The picture is subsequently sent to the classifier, which gives the discovered face's dimensions and position. After that, the picture is cropped to the recognized face and shrunk to 160x160 pixels.

B. *Sinhala Voice to text conversion and provide accurate English phrases*

In this component of research, mainly the entire process is based on the voice input that is given by a user. Other than recognizing the voice of the user, the noise removal process is also happening here. The second part is to provide accurate English phrases to the recognized Sinhala text. To fulfill the above-mentioned requirements, a few python libraries, Neural machine translation, Attention mechanism and google cloud technologies were used.

Considering the voice recognition process, one of the most accurate systems for the purpose is using the Google Cloud speech recognition method. The Google Cloud Speech recognition method Gives tips to improve the accuracy of the transcription of uncommon and domain-specific words or phrases. Users can use classes to translate spoken numbers into things like addresses, years, currencies, and more. Also, it chooses from a variety of trained models for voice control, call transcription and video transcription that are targeted toward achieving domain-specific quality standards. Google speech recognition can be used in many languages also without the need for extra noise cancellation, speech-to-text can handle noisy audio from many contexts. Firstly, the Google cloud voice recognition needs to connect to the python code by the google cloud specified python library. The text output of the speech command will be output with the accuracy rate of it. The text is received to the front end through an API. So that the user can make sure what he said was identified correctly.

In the English translation process, basically, Neural Machine Translation (NMT) is the main method. For this part, the Attention mechanism is used which is a complex NMT mechanism in NLP. By basing on the attention algorithm, BERT (Bidirectional Encoder Representations from Transformers) will be implemented to accomplish this methodology. BERT is a deep network architecture that is used in NLP. The output of the Neural Machine Translation (NMT) will be an English-translated text of the users' initial voice command [1].

$$argmax_y P(y|x) \rightarrow `argmax_y P(y|x)P(y) \qquad (1)$$

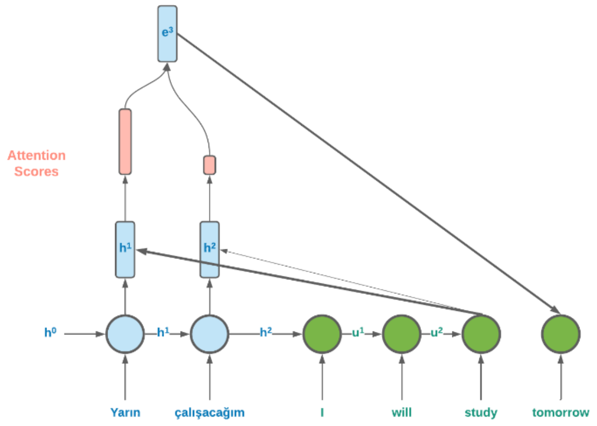

Figure 2: RNN Encoder and Decoder with Attention

C. *Providing accurate voice & outputs according to the user's commands*

Sinhala voice output, as well as word display, was done using python. He also used libraries such as TensorFlow, NumPy, karas, sklearn for this purpose. Providing accurate voice & outputs according to the user's commands, also this part is powered by Machine Learning and Soft Computing (MLSC) and Natural Language Processing (NLP). In speech recognition, there is a cycle of events that occurs between a vocal utterance and the response to that utterance from the computer. This set of events are referred to as a voice dialog circle. To ensure that the voice assistant understands a command, a sophisticated component of human speech must also be broken down. Natural language processing analyzes user intent, context, unique speech patterns, slang, and accents, among other aspects of human language, to assist the voice assistant in understanding the command. TensorFlow, Matplotlib, and Pytorch are technologies (python libraries) used in Deep Learning applications. This allows the user to get the output voice clearly. TTS engine collects the speech from the detected, and the TTS engine realizes the scenario. The Joiner Algorithm is used to combine TTS engines to function together. The system leverages the TTS engine to recognize the speech and turn the voice to text. The TTS engine is deployed by the system to modify the input text into a voice for communication options between the deaf person and the typical person. The system also employs the TTS engine to display the text as a dynamic object. By using the Sinhala dataset and providing better output, the dataset has been trained by the machine learning model and obtained from the .h5 file format. The text-to-Speech process will catch it using both intents (patterns) and Utter or Action (Responses) techniques connected by stories. Those stories are written in markdown format. Neural machine translation is a recently proposed technology for automatic translation (NMT). Neural machine translation seeks to construct a single neural network that can be collaboratively modified to optimum translation performance, in contrast to standard statistical machine translation. The neural machine translation (NMT) is used to generate meaningful voice output. Based on the command given by the user, it is optimized to give him the most suitable answer at that time.

This part will provide what the user asks for in Sinhala as soon as possible. Also, since it has been done with a machine learning model, the accuracy has shown a maximum value.

D. Facial Expression detection and suggestion generating

Facial expression recognition is essential for nonverbal communication between humans and the production. Using this facial expression detection and identification the solution which is the Automated Sinhala voice assistant identifies the expressions from the face without telling the application. Because most of the time people are not going to express their ideas to a digital device. In the context of research, Uplift users' moods by providing suggestions mainly based on the facial expressions of the user. The objective is to obtain a fast baseline to determine if the CNN architecture performs better when simply the image's raw pixels are used for training. Or if it is preferable to provide CNN with further information (such as face landmarks or HOG features). The outcomes demonstrate that the additional data improves CNN's performance.

To train this model, Fer2013 which is a popular publicly available dataset has been used. This dataset contains 35,887 images of facial expressions grouped in seven categories: Angry, Disgust, Sad, Surprise and Neutral. The dataset is difficult since the represented faces vary considerably in terms of human age, facial posture, and other aspects (Fig. 1), mimicking true settings. There are 28,709 training samples, 3,589 validation samples, and 3,589 test samples in the training, validation, and test sets, respectively. Basic expression labels are given for all examples. Each image is grayscale and has a 48-by-48-pixel resolution.

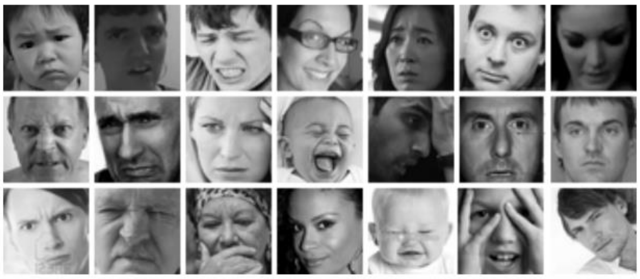

Figure 3: Example images from FER2013 dataset

In order to emphasize the methodological distinctions between these efforts, can be split each technique into three components: (i) preprocessing, (ii) CNN architecture, and (iii) CNN training and inference. 1) Preprocessing: Preprocessing comprises actions that are performed only once on each picture. This often comprises face detection, face registration to adjust for position differences, and lighting correction methods. Face detection (FD) is described as the process of determining where human faces reside inside an image. FD is required for either face recognition or expression recognition. The extraction of facial features is necessary for face expression recognition. If the face is not detected, additional redundant information will be present when the picture is fed into a network or feature extraction is performed, which will have a negative impact on the recognition result; hence, face detection is crucial. The Viola-Jones method was chosen for FD algorithm of choice. Using AdaBoost and Haar face identification technology, this system treats Haar characteristics as a poor classifier. Then, this technique combines numerous weak classifiers into a single robust classifier. The strongest classifiers are combined in sequence to build a cascade classifier, which is the highest possible score. Face detection may utilize cascaded classifiers. Next, AdaBoost algorithm and Haar characteristics are introduced.

Haar characteristics

Haar features are mostly utilized to extract face characteristics. There are four types of Haar characteristics that may be efficiently integrated into feature templates. In the feature template, there are just two types of rectangles: white and black. The template covers a certain region of the picture while extracting features, and then calculates the sum of the pixel values of two distinct types of rectangles. The template covers a certain region of the picture while extracting features, and then calculates the sum of the pixel values of two distinct types of rectangles. It is highly compatible with Haar characteristics given that the face features include bright information and there is a light-dark connection in local locations. Reason why the feature template must compute the sum of pixels for each region in the picture, resulting in a huge number of repetitive computations of region pixel values. To remedy this issue, the notion of an integral graph is introduced. According to the following formula, the integral graph is the total of the pixels in the rectangle area produced from the image's beginning point to each point. [2]

$$SAT(x,y) = \sum_{x_i \le x, y_i \le y}^{n} I\,(x_i, y_i) \qquad (2)$$

Using this Hara characteristic, the face will identify the algorithm. Based on features, emotion can be identified using the CNN model. After identifying the emotional expressions, suggestions can be generated to uplift users' mental health. These suggestions are generated according to the users' characteristics and users' emotions. The application will be able to prompt ideal suggestions to the user to increase their mental freedom.

IV. RESULTS AND DISCUSSION

The focus of the research was to simulate a friendly voice assistant application that functioned using the Sinhala language via uplifting mental freedom of the Sri Lankan people using natural language processing concepts. Since foreign people usually use voice assistant applications in their native language, there is no option for Sri Lankans who use Sinhala voice commands. Any person can use this application without fluent spoken English knowledge. The environment provided by the proposed system would allow all users to identify their real-time feelings without human involvement. Furthermore, the system will take most of the suggestions to uplift their real-time feelings.

On the other hand, the proposed system would be beneficial for identifying the users given Sinhala voice commands without interfering with background noises. The application will remove the background noises and unwanted words and get relevant input for the system. Basically, users would be accessed the application through the security layer. The voice assistant application mainly contains users' personal details. So that the application security increased using the facial recognition mechanism. Users can give a voice command to the application to fulfill their needs. The application will process according to the users' given command, and the results will finally be presented to the users by providing the display prompt and with Sinhala voice output. This mechanism led to finding suitable output according to the user's needs. During the development of each research component, the developers met a variety of obstacles, for which they advised and implemented many solutions. Significant issues were the inaccuracy of speech detection libraries and text matching algorithms used to discover keywords, the reduction of background noise, and the search for a solid dataset.

While the facial expression identification section was developed, the facial expressions mismatching issue was raised, and while developing the application's security layer, various assumptions were considered. Which is OpenCV and performing it with MATLAB, Facial recognition TensorFlow model, and MediaPipe method. Here the Haar Cascades had issues identifying the face area correctly, and it took considerable time to identify the face. Those issues were fixed using the MediaPipe method, and the application's security and performance have increased.

Furthermore, by doing an internet search, many issues have been raised while providing Sinhala texts related to the search results. During implementations, modifications were made to a few components. A trained CNN model is incorporated into the facial expression detection component.

V. CONCLUSION

ශ්‍රී voice mobile application is software that can automate the conventional use of the application by sending commands to the application process utilizing contemporary natural language processing techniques and deep learning applications. The proposed system mainly works with human voice and facial expressions. The application will be able to identify the real-time feeling from the user's face. The technology captures and translates human language into text-based inputs that the machine can interpret. ශ්‍රී The voice mobile application is designed with Python and the Flutter front-end framework. This is the research paper on the customized mobile application of ශ්‍රී voice for automating daily tasks of the users while uplifting the users' mental health by providing ideal suggestions. This research paper comprises five chapters, with the majority of the discussion focusing on the Back-End development process of the voice mobile application. Overall, all system objectives have been accomplished, and the intended system is operational. After a quick review of the evaluations, it has been determined that the ශ්‍රී voice mobile application project has been a complete success to date.

REFERENCES

[1] Y. Perera, N. Jayalath, S. Tissera, O. Bandara, and S. Thelijjagoda, "Intelligent mobile assistant for hearing impairers to interact with the society in Sinhala language," *Int. Conf. Software, Knowl. Information, Ind. Manag. Appl. Ski.*, vol. 2017-Decem, 2018, doi: 10.1109/SKIMA.2017.8294116.

[2] J. H. Kim, B. G. Kim, P. P. Roy, and D. M. Jeong, "Efficient facial expression recognition algorithm based on hierarchical deep neural network structure," *IEEE Access*, vol. 7, no. c, pp. 41273–41285, 2019, doi: 10.1109/ACCESS.2019.2907327.

[3] J. H. Kim, B. G. Kim, P. P. Roy, and D. M. Jeong, "Efficient facial expression recognition algorithm based on hierarchical deep neural network structure," *IEEE Access*, vol. 7, no. c, pp. 41273–41285, 2019, doi: 10.1109/ACCESS.2019.2907327.

[4] M. S. Amarasekara, K. M. N. S. Bandara, B. V. A. I. Vithana, D. H. De Silva, and A. Jayakody, "Real-time interactive voice communication - For a mute person in Sinhala (RTIVC)," *Proc. 8th Int. Conf. Comput. Sci. Educ. ICCSE 2013*, no. Iccse, pp. 671–675, 2013, doi: 10.1109/ICCSE.2013.6553993.

[5] M. Tkalčič, M. Elahi, N. Maleki, F. Ricci, M. Pesek, and M. Marolt, "Prediction of music pairwise preferences from facial expressions," *Int. Conf. Intell. User Interfaces, Proc. IUI*, vol. Part F1476, pp. 150–159, 2019, doi: 10.1145/3301275.3302266.

[6] M. S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan, "Machine learning methods for fully automatic recognition of facial expressions and facial actions," *Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern.*, vol. 1, pp. 592–597, 2004, doi: 10.1109/icsmc.2004.1398364.

[7] J. Rani and K. Garg, "Emotion Detection Using Facial Expressions-A Review," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 4, no. 4, p. 2277, 2014, [Online]. Available: www.ijarcsse.com

[8] S. Tivatansakul, M. Ohkura, S. Puangpontip, and T. Achalakul, "Emotional healthcare system: Emotion

detection by facial expressions using Japanese database," *2014 6th Comput. Sci. Electron. Eng. Conf. CEEC 2014 - Conf. Proc.*, pp. 41–46, 2014, doi: 10.1109/CEEC.2014.6958552.

[9] M. Asad, S. O. Gilani, and M. Jamil, "Emotion Detection through Facial Feature Recognition," *Int. J. Multimed. Ubiquitous Eng.*, vol. 12, no. 11, pp. 21–30, 2017, doi: 10.14257/ijmue.2017.12.11.03.

[10] H. Wang and H. Zhang, "Movie genre preference prediction using machine learning for customer-based information," *2018 IEEE 8th Annu. Comput. Commun. Work. Conf. CCWC 2018*, vol. 2018-Janua, pp. 110–116, 2018, doi: 10.1109/CCWC.2018.8301647.

[11] D. D. S. Rajapakshe, K. N. B. Kudawithana, U. L. N. P. Uswatte, N. A. B. D. Nishshanka, A. V. S. Piyawardana and K. N. Pulasinghe, "Sinhala Conversational Interface for Appointment Management and Medical Advice," IEEE, 26 February 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9357155.

[12] F. Nasirian, M. Ahmadian, and O. K. D. Lee, "AI-based voice assistant systems: Evaluating from the interaction and trust perspectives," *AMCIS 2017 - Am. Conf. Inf. Syst. A Tradit. Innov.*, vol. 2017-August, no. May, 2017.

[13] M. S. Amarasekara, K. M. N. S. Bandara, B. V. A. I. Vithana, D. H. De Silva, and A. Jayakody, "Real-time interactive voice communication - For a mute person in Sinhala (RTIVC)," *Proc. 8th Int. Conf. Comput. Sci. Educ. ICCSE 2013*, no. Iccse, pp. 671–675, 2013, doi: 10.1109/ICCSE.2013.6553993.

[15] A. Poushneh, "Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors," *J. Retail. Consum. Serv.*, vol. 58, p. 102283, 2021, doi: 10.1016/j.jretconser.2020.102283.

[16] T. Dinushika, L. Kavmini, P. Abeyawardhana, U. Thayasivam and S. Jayasena, "Speech Command Classification System for Sinhala Language based on Automatic Speech Recognition," IEEE, 19 March 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9037648.

[17] P. M. Dias and K. Jayakody, "Virtual Assistant in Native Language," IEEE, 22 June 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9452751.

[18] Y. Zhang and D. Xie, "The Application Of Face Recognition Technology," vol. 631, no. Sdmc 2021, pp. 242–247, 2018, doi: 10.2991/icaset-18.2018.4.

[19] M. I. P. Nasution, N. Nurbaiti, N. Nurlaila, T. I. F. Rahma, and K. Kamilah, "Face Recognition Login Authentication for Digital Payment Solution at COVID-19 Pandemic," 2020 3rd Int. Conf. Comput. Informatics Eng. IC2IE 2020, pp. 48–51, 2020, doi: 10.1109/IC2IE50715.2020.9274654

[20] D. D. S. Rajapakshe, K. N. B. Kudawithana, U. L. N. P. Uswatte, N. A. B. D. Nishshanka, A. V. S. Piyawardana and K. N. Pulasinghe, "Sinhala Conversational Interface for Appointment Management and Medical Advice," 2020 2nd International Conference on Advancements in Computing (ICAC), 2020, pp. 85-90, doi: 10.1109/ICAC51239.2020.9357155.

[21] M. Khan, S. Chakraborty, R. Astya, and S. Khepra, "Face Detection and Recognition Using OpenCV," *Proc. - 2019 Int. Conf. Comput. Commun. Intell. Syst. ICCCIS 2019*, vol. 2019-Janua, pp. 116–119, 2019, doi: 10.1109/ICCCIS48478.2019.8974493.