

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

Project ID - 2022-200

Final Thesis

B.Sc. (Hons) Degree in Information Technology Specializing in  
Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology  
Sri Lanka

September 2022

**Automated Sinhala Voice Assistant to Manage Tasks using  
Natural Language Processing - ශ්‍රී Voice  
Project ID - 2022-200**

Supervisor - Ms. Dinuka Wijendra

Co-Supervisor - Ms. Jenny Krishara

B.Sc. (Hons) Degree in Information Technology Specializing in  
Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

September 2022

## **DECLARATION**

We declare that this is my own work, and this dissertation does not incorporate without acknowledgment any material previously submitted for a degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgments made in the text. Also, we hereby grant to Sri Lanka Institute of Information Technology, the nonexclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or another medium we retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:

The above candidate has carried out research for the bachelor's degree Dissertation under my supervision.

Signature of the supervisor:

Date:

## 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 entertainment purposes. Most of the existing voice assistant applications are 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 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 for 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. 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 Assistant Application, Sinhala Voice, Convolutional Neural Network (CNN), Natural Language Processing, Emotion Recognition, Neural Machine Translation

## **ACKNOWLEDGEMENT**

We desire to take advantage of the progression of this part in order to offer my heartfelt appreciation to everyone who has assisted me along this trip from the very beginning. At the beginning, I'd want to express my gratitude to the Sri Lanka Institute of Information Technology (SLIIT) for instituting this requirement and allowing me the chance to talk about some of my more original course material. In addition, We desire to thank each professor and lecturer who provided guidance and assistance throughout this research project.

It was a honor to have a supervisor who helped us get back on track when we took the wrong turn. Therefore, Ms. Dinuka Wijendra, who kindly consented to supervise this study throughout the year and gave ideas to increase the value of the final result, has my sincerest thanks. We desire to take this opportunity to convey my gratitude to her in the most heartfelt way. We would like to thank Ms. Jenny Krishara, our co-supervisor, for her commitment to oversee this project throughout the year.

Finally, we desire to express my appreciation to my team members, my friends, and my family members for their encouragement and support.

Moreover, we'd like to express my thankfulness to everyone else who has helped me along the journey but whose names aren't included here.

## TABLE OF CONTENTS

DECLARATION .....	i
ABSTRACT.....	ii
Acknowledgement.....	iii
TABLE OF CONTENTS .....	iv
LIST OF FIGURES .....	vii
LIST OF TABLES .....	viii
LIST OF EQUATIONS .....	ix
LIST OF ABBREVIATIONS .....	x
1 INTRODUCTION .....	xi
1.1 Background .....	xi
1.2 Literature Review.....	1
1.3 Research Gap .....	7
1.4 Research Problem .....	9
1.5 Research Objectives.....	11
1.5.1 Main Objective .....	11
1.5.2 Specific Objectives .....	12
2 METHODOLOGY .....	16
2.1 System Overview .....	16
2.2 Research Design.....	17
2.3 Data Gathering .....	18

2.3.1	Face recognition Data Collection .....	18
2.3.1	Chat Assistant Data Collection .....	20
2.3.1	Facial Expressions Recognition Data Collection .....	20
2.3.2	Custom Made Sri Lankan Database by me .....	21
2.4	Voice recognition and English Phrase Generation.....	22
2.4.1	Google Cloud Speech to Text recognition.....	22
2.4.2	Neural machine translation (NMT) .....	24
2.4.3	Text-to-Text Transfer Transformer .....	25
2.4.4	Attention mechanism.....	28
2.5	Improve Security using Facial recognition .....	29
2.5.1	Open CV .....	30
2.5.2	Media pipe .....	31
2.5.3	TensorFlow.....	32
2.5.4	Fine Tuning.....	33
2.5.5	Face Net .....	33
2.6	Chat Assistant .....	35
2.6.1	Neural Machine Translation Model .....	35
2.6.2	Attention Mechanism .....	36
2.6.3	Text-To-Speech API .....	37
2.7	Facial Emotion Recognition.....	39
2.7.1	OpenCV & Keras approach.....	39
2.7.2	TensorFlow and Keras Approach .....	41
2.7.3	Convolutional Neural Network Approach.....	42
2.7.4	Facial Expression Recognition (FER) Using Transfer Learning in Deep CNNs and Training the model .....	44

2.7.5	Fine Tunning.....	47
2.7.6	Provide ideal suggestions according to the real time situation.....	48
2.8	Technologies Used.....	49
2.8.1	Low Fidelity Prototyping – Figma, Adobe XD.....	49
2.9	Commercialization aspects of the Product.....	51
2.10	TESTING & IMPLEMENTATION .....	53
2.10.1	Testing.....	53
2.11	Implementation .....	54
2.11.1	Agile Methodology.....	54
3	RESULTS & DISCUSSION .....	55
3.1	Results.....	55
3.2	Research Findings.....	61
3.3	Discussion .....	63
4	Summary of the Student Contribution.....	64
5	CONCLUSION .....	68
6	REFERENCES .....	69
7	Appendix .....	77
7.1	Gantt Chart.....	77
7.2	Similarity Index.....	78



## LIST OF FIGURES

Figure 1: Problems identified by the Survey.....	9
Figure 2: System Overview Diagram.....	16
Figure 3 - Face data set .....	19
Figure 4 : Example images from FER2013 dataset .....	21
Figure 5 : Different training objective .....	26
Figure 6 : A flow chart of unsupervised objectives configuration.....	27
Figure 7 : Comparing different fine-tuning methods .....	28
Figure 8 : Weights are assigned to input words at each step of the translation .....	29
Figure 9 : Face recognition process .....	30
Figure 10 : Scanning via MediaPipe Technique .....	32
Figure 11: Face Net process.....	34
Figure 13 : Attention mechanism.....	36
Figure 14 : Summary of the 3 Phases of the Facial Emotion Recognition Model.....	42
Figure 15: Overall flow of the Learning based deep CNN model .....	45
Figure 16 : Illustration of Facial Expression Recognition Model.....	46
Figure 18 – Wireframe Design 1.....	49
Figure 17 – Wireframe Design 2.....	49
Figure 19 – Wireframe Design 3.....	50
Figure 20 – Wireframe Design 4.....	50
Figure 23 : Commercialization.....	52
Figure 24 : Gantt Chart .....	77
Figure 25 : Similarity Index .....	78

## LIST OF TABLES

Table 1: Abbreviation .....	x
Table 2 : Research Gap .....	8
Table 3 : Results.....	60

## LIST OF EQUATIONS

(1).....	24
(2).....	24

## LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
NMT	Neural Machine Translation
FER	Facial Expression Recognition
NLP	Natural Language Processing
AI	Artificial Intelligence
T5	Text-to-Text Transfer Transformer
DTP	Directional Ternary Patterns
ML	Machine Learning
TL	Transfer Learning
NLP	Natural Language Processing
DCNN	Deep Convolutional Neural Network
NMT	Neural Machine Translation
NN	Neural Network
BERT	Bidirectional Encoder Representations from Transformers
RNN	Recurrent Neural Network

Table 1: Abbreviation

# **1 INTRODUCTION**

## **1.1 Background**

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 [8]. 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 [9].

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].

## **1.2 Literature Review**

Using natural language processing and machine learning concepts, the automated Sinhala voice assistant is a mobile application that automates daily tasks so that users do not have to perform them themselves, and it will help to get answers to questions in the Sinhala language. Therefore, the whole study can be divided into 4 sections. To develop a comprehensive voice assistant application, you need other sections also. This study consists of a facial recognition system, which is utilized to improve the security of the application. And there is a chat assistant which enables you to interact with Sinhala language and do the commands and get outputs accordingly. Finally, the application consists of a facial emotion analyzing feature. This special feature enables us to identify users' real-time feelings and provide ideal suggestions according to the real-time situation of their lives. During the literature review, the research team discovered similar studies that had been conducted in previous years.

### **Facial recognition**

I concentrated on the same functional and research areas in this review of the literature, But there is only a small amount of research in Sri Lanka regarding my function. An intelligent conversational user interface (ICUI) has gained much attention recently since it can be used in many settings, such as intelligent personal assistant systems, customer care systems, and information retrieval systems. A few famous systems include Apple Siri, Microsoft Cortana, and Amazon Alexa. Those are a few successful ICUs. is backed by a lot of computational linguistic research in the English language. [3] Sinhala is a low resource language, and the Sinhala speaking community is a minority in a global context. In the contemporary Sinhala alphabet with 60 letters, the Sinhala language contains 18 vowels, two half vowels, and 40 consonants. [6]. Over the last three decades, hand-held devices such as mobile phones, tabs have developed with more portable technology and higher quality interconnection mechanisms. [3]. Humans use the internet to save time and work efficiently. Also, the internet has become universal, and it helps many with socializing. SCI-AMMA is a mobile application that provides a human computer interaction facility as a solution for making an appointment with a doctor in hospitals and provides efficient services for patients. This type of application is currently unavailable in Sri Lanka. In SCI- AMMA the rest API is used to create a connection between the microservice and



the mobile app. To make appointments in Sinhala, this system has implemented the Sinhala keyword dictionary. This system was developed using python webhook API with the Flask REST API framework and 2.3.0 Dart package is used for converting the Sinhala voice to text output.[3]

Z. B. Lahaw et al. described a face recognition method. The proposed investigation employs linear discriminant analysis, independent component analysis, principal component analysis, and support vector machine techniques. The project makes use of the AT&T database. This collection features 400 photographs of the faces of forty subjects, including ten images of each subject wearing sunglasses captured at different times and in different stances and settings. The dimensions of these grayscale photos are 112 by 92 pixels. The authors attained a 96% recognition rate using a hybrid method based on the Discrete Wavelet Transform (DWT), principal component analysis (PCA), or linear discriminant analysis (LDA) for dimension reduction, and support vector machine for face classification. There is study on mobile phone face recognition cases. Smartphones have evolved into the most practical and pervasive interface for Internet services. They are outfitted with a number of sensors that can considerably enhance the cognitive capacity of the user. Face recognition is one of the most rapidly expanding mobile applications due to its vast array of possible uses. However, facial recognition is a computationally demanding operation that exceeds the capabilities of even contemporary mobile smartphones. In this study, the authors describe a cloud-computing offloading-based facial recognition method that is both effective and efficient. Before delegating recognition to the cloud, the architecture performs early image processing on mobile devices to decrease network traffic and conserve battery power. Preliminary results indicate an overall response time reduction of over 230% and an energy savings of over 200% on the mobile device.

A face recognition program is a piece of software that verifies and identifies a person using a video or image from a source. The work on facial recognition that predates the 1950s and 1960s in psychology can be linked to technical literature. Among the early discoveries are experiments on Darwin's facial expression perceptions. With Intel's OpenCV platform, facial recognition can be performed rapidly and accurately. The preferred facial features are derived from a face and a picture database. It is commonly compared to biometrics such as fingerprints and eye surveillance systems, and is employed in security systems, such as thumb recognition systems. Common recognition techniques included the key element analysis utilizing Fisher face

algorithms, the Markov model, multilinear subspace learning utilizing tensor representations, and the dynamically driven dynamic reference matching, among others. The Intel open-source computer-View library makes programming simple. This delivers advanced features including facial detection, face tracking, and facial recognition, as well as a variety of ready-to-use artificial intelligence technologies (AI). It has the benefit of being a cross-platform framework; it supports Windows and Mac OS X, in addition to Mac OS and Mac OS X.

This paper discusses an application for the automatic detection and tracking of faces in video streams from public or commercial security cameras. It is helpful in a variety of settings, including exhibitions, retail malls, and public parts of buildings, to determine where people are looking. Using the open-source platforms Arduino and OpenCV, a face recognition and tracking system prototype was built to communicate with webcams. AdaBoost and Haar-Like facial characteristics are utilized by the system. This technology can be used for security purposes to detect, monitor, and record a visitor's face. OpenCV is used to create a program capable of facial detection and web camera monitoring.

Face detection is the most researched topic in computer science's vision field. It is a computer technology that recognizes human faces in digital photos and is utilized in a range of applications [1]. This field's research is developing in numerous scientific disciplines, including psychology. Face recognition is one of the most discussed technologies. Localization of human faces is considered the first and most fundamental step in face detection research. For example, in home video surveillance etc. Face localization is the extraction of facial characteristics using a pattern recognition algorithm. Both MATLAB and Open CV may be utilized for the development of such prototypes and systems. The authors of this publication conducted their research using Open CV. In this paper, the reasons for utilizing an open CV are explained in greater detail.

There is research into interactive voice communication in real time for mute individuals. Speech impairments should not be an excuse to forgo mobile communication. With the advent of technology, various assistive gadgets and equipment for people with varying abilities have become available. [7]. They created this application with a particular attention on that point. They recognized Sinhala Unicode characters and intelligently predicted words. Using text-to-speech synthesis, the author ensures that once the user enters a message, it is converted into a Sinhala audio clip. They have employed a text-to-speech (TTS) engine for this purpose.

### **Speech Command Classification System for Sinhala Language based on Automatic Speech Recognition [4]**

This system, a domain-specific speech command classification system, is capable of understanding spoken Sinhala and its intent as well as of identifying automatic speech. This system can also be used effectively for commercial purposes in the value-added applications such as Sinhala speaking dialogue systems. This system is equipped with an automatic speech recognition engine that can convert continuously spoken Sinhala words into their textual representation as well as a text classifier and new database for this purpose, corpus of Sinhala speakers in the financial sector. The Sinhala speech command classification system outperformed the most recent prior speech to intentional categorization systems created for the Sinhala language, with an accuracy of 89.7% in predicting the meaning of a statement. The word error rate was 12.04 percent, and the sentence error rate was 21.5 percent. The experiment provided appropriate and specific knowledge of speech to motivation taxonomy for researchers with limited resource language comprehension.

### **Intelligent mobile assistant for hearing impairers to interact with the society in Sinhala language [5]**

This was a cross-platform mobile development project for the Sinhala language known as "SANWADHA" that focused on hearing impaired people and used instant messaging (IM) chat. This system's primary input method is a text in the Sinhala language, which it transforms to Sinhala sign language and displays in GIF format to the hearing-impaired user. People without hearing issues can work with hearing impaired people using this system. The project's ultimate objective was to connect with the deaf community in Sri Lanka and use communication to empower hearing-impaired people.

## **Sinhala Speech to Sinhala Unicode Text Conversion for Disaster Relief Facilitation in Sri Lanka [6]**

People often speak in their native tongues when they are under stress or in a panic situation, and during disasters in Sri Lanka, the majority of call assistant services are provided in the local language. Speech recognition, also referred to as automatic speech recognition or computer speech recognition, is a process that turns the acoustic speech signals captured by a microphone or a telephone into a list of words. This method takes some time. Computers are equipped with the ability to understand the voice commands of users and perform the required tasks by the process of speech recognition. And due to the wide range of industries using speech recognition software, this technology is expanding quickly. The objective of the study was to create and refine a method for transcribing spoken Sinhala into documents using the Sinhala Unicode character set. The words were recognized using the hidden Markov model. The software used was able to achieve 65% accuracy in a quiet environment and 54% in a noisy one.

## **Sinhala Conversational Interface for Appointment Management and Medical Advice [7]**

SCI-AMMA is an intelligent conversational application for seeking medical advice and advice from doctors. The primary components are the Speech Recognition unit, Query Processing unit, Dialog Management unit, Voice Synthesizer unit, and User Information Management unit for the purpose of responding to user requests and having an engaging conversation. Responses are generated from the Dialogue Management Unit after the query processing unit has determined the user's intents, and they are then delivered to the user via a voice synthesizer. This project has been successfully implemented using state of the art technology stack including Flutter, Python, Protégé and Firebase.

## **Voice Assistant [1]**

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. They make use of machine learning, voice recognition, natural language processing, and artificial intelligence as tools and technologies. They employ artificial intelligence, natural language processing, speech recognition, and machine learning as tools and technologies. 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.

## **Voice Assistant Application for the Serbian Language [2]**

Most voice assistant applications are made in English, and some are designed to be used in their native language. They have created a voice assistant that works in the Serbian language. They say the result is very good and they hope to increase its efficiency in the future. As mentioned, c ++ has been used to write the native part.

According to a previously released research paper, Artificial Intelligence (AI) technologies are one of the newest and most difficult technologies growing at a breakneck pace. Although these technologies appear to be widely adopted, some individuals do not want to use them. For many years, technology adoption has been examined, and there are numerous generic models defining it in the literature.

However, it appears that developing more specialized models for emerging technologies based on their characteristics is important. To validate their model, they used a voice assistant system (VAS) as an example and validated a theory-based model using data from a field survey.

Their results substantiate. AI assists in providing the proper output to the user, as well as determining the correct response to provide in response to the user's instruction. Voice assistants or any type of voice-enabled artificial intelligence are no longer exclusive to science fiction films, according to research from California. Voice technology is presently integrated into a variety of products, such as mobile apps for smartphones and residential voice assistants. Moreover, voice assistants are increasingly integrated into our everyday lives.

#### **Sinhala Conversational Interface for Appointment Management and Medical Advice [4]**

According to Sinhala research, they did propose an intelligent conversational user interface for assisting Sinhala-speaking users in scheduling doctor appointments and obtaining health advice. To manage users' requests and preserve a meaningful dialogue, this Interface Sinhala Conversational for Appointment Planning and Medicine Guidance consists of a Speech Recognition unit, a Query Processing unit, a Dialog Management unit, a Voice Synthesizer unit, and a User Information Management unit. The (SCI-AMMA) acquires users' speech utterances and recognizes the language content of those utterances for subsequent processing. The language content is further processed by the query processing unit in order to ascertain the user's intent. A response is generated from the Dialogue Management Unit to satisfy the users' intent. The user will receive this reaction via a voice synthesizer. Their proposed system is implemented successfully using a cutting-edge technology stack that includes Flutter, Python, Mentees, and Firebase. The effectiveness of the method is demonstrated through a series of sample system potential situations.

## **Automatic Speech Recognition-based Sinhala Language Speech Command Classification System [6]**

The capacity of Communicative Artificial Intelligence to turn conventional computers into human-like computers is revolutionizing the global landscape. Exploiting the speaker's intent is one of the most crucial aspects of conversational Artificial Intelligence. The lack of language resources is a significant barrier that hinders the effectiveness of understanding the speaker's intent. To address this issue, we develop a domain-specific voice command classification system for the low-resource language Sinhala. Automatic Speech Recognition and Natural Language Understanding are used to provide intent identification for Sinhala speech. The proposed technique is applicable to applications with added value, such as Sinhala speech dialog systems. The system consists of an Automatic Speech Recognition engine that translates continuous authentic Sinhala human speech to text and a text classifier that understands the user's intent accurately. In addition, we provide a new dataset, 4.15 hours of Sinhala banking speech corpus, for this task. Our new Sinhala voice command classification system accurately identifies the intent of utterances with 89.7% accuracy. It outperforms the most sophisticated Sinhala language direct speech-to-intent classification algorithms. Moreover, the engine for Automatic Speech Recognition has a 12.04% Word Error Rate and a 21.5% Sentence Error Rate. In addition, our experiments provide useful categorization insights for speech-to-intent to researchers with limited resources for spoken language understanding.

## **Consistent facial features and emotions throughout civilizations. Journal of Personal and Social Psychology [13]**

Since the turn of the 20th era, Ekman et al. have identified seven fundamental emotions, regardless of the society in which a person develops (anger, feared, happy, sad, contempt, disgust, and surprise). Sajid et al. demonstrated the relevance of facial

asymmetry as an indication of age estimation in a recent investigation of the facial recognition technology (FERET) dataset. Their findings indicate that right-sided facial asymmetry is preferable than left-sided facial asymmetry. Facial detection still has a significant difficulty with face pose appearance [13]. Ratyal et al. gave the answer to the variable look of people. They have utilized a three-dimensional posture invariant method with subject specific descriptions. Recent advances in facial expression identification have led to advancements in neuroscience and cognitive science that encourage the expansion of facial expression research. In addition, advancements in computer vision and machine learning make emotion recognition considerably more precise and accessible to the general public [13].

### **Facial Expression Recognition Utilizing the TensorFlow Architecture [14]**

Early facial expression detection relies mostly on face recognition techniques to categorize and identify facial expressions. SVM, LBP, and Gabor are typically used to organize and identify facial expressions based on Haar, Adaboost, and neural network properties. Using a neural network, Kobayashi et al. [21] were able to categorize and recognize fundamental facial expressions. Using an SVM classifier based on LBP features, Caifeng Shan et al. achieved facial emotion recognition [18]. Using an SVM classifier based on ICA and Gabor features to classify facial expressions, Ioan Buciu et al. proved that a facial expression identification system that combines Gabor wavelet and SVM may achieve a higher detection rate [20]. Using RPCA and AdaBoost, Xia Mao et al. [17] demonstrated strong facial expression recognition. Early facial expression identification algorithms lacked analysis of face expression's distinctive characteristics.



### 1.3 Research Gap

English is the primary language of most voice assistant applications currently available, due to the fact that English is a global language. Additionally, there are no effective voice assistant programs that support Sinhala. As a result, the authors' proposed system is entirely original. Additionally, voice assistant applications for face recognition models have not been used in previous research projects. Here, voice assistant users could gain access to the application using a face recognition technique. And existing voice assistant applications didn't have this kind of face recognition login. The aforementioned researchers can only provide limited answers despite sharing a primary objective with the Sri Voice application. The found research projects are not able to answer many questions and instead instruct the user to repeat the phrase quickly or ignore it entirely. Additionally, it does not support sentences that are lengthy and complex.

Some of the voice assistant programs available today don't respond to user requests with the right information. Additionally, in contrast to several studies, this component has been created so that the user can finish part of his task with the least amount of understanding possible. Nevertheless, even existing English voice assistants do not provide ideal suggestions according to the users' emotions. Existing popular English voice assistant applications such as Apple Siri, Google Assistant, and Samsung Bixby suggest some standard suggestions. Facial emotion analyzing algorithm, preference gathering, and emotional level identification algorithm identify about user accurately. Combining these algorithms provides ideal Sinhala voice suggestions to improve users' mental freedom. This compelling feature will help users be calm, enjoy and get some release from stress. The following table compares existing research with the proposed system.

The results presented here demonstrate that it is feasible to go beyond conventional voice assistant apps and get a novel experience. Also, the user interface has been created by doing UX research so that outputs may be shown on the screen in a user-friendly manner.

Research	Speech to text	Sinhala Language	Facial expression detection	Text to speech	Answering To a wide range of questions	Authenticate the user and give access	Use local dataset
[5]	✓	✓	✗	✓	✗	✗	✓
[1]	✓	✓	✗	✓	✗	✗	✓
[2]	✓	✓	✗	✗	✗	✗	✓
[3]	✓	✓	✗	✗	✗	✗	✓
[4]	✓	✓	✗	✓	✗	✗	✓
[6]	✓	✓	✗	✓	✗	✗	✗
[7]	✗	✓	✗	✗	✗	✗	✓
🗣️ Voice App	✓	✓	✓	✓	✓	✓	✓

Table 2 : Research Gap

## 1.4 Research Problem

People are increasingly turning to technical applications in the modern world. They are able to complete their daily tasks with ease as a result of those factors. Although it is already well-liked among young people, some Sri Lankans chose not to use technology due to a lack of technological literacy. The following figure clearly shows most people suffer from these issues.

5. If you not use voice assistant applications, why is that?

29 responses

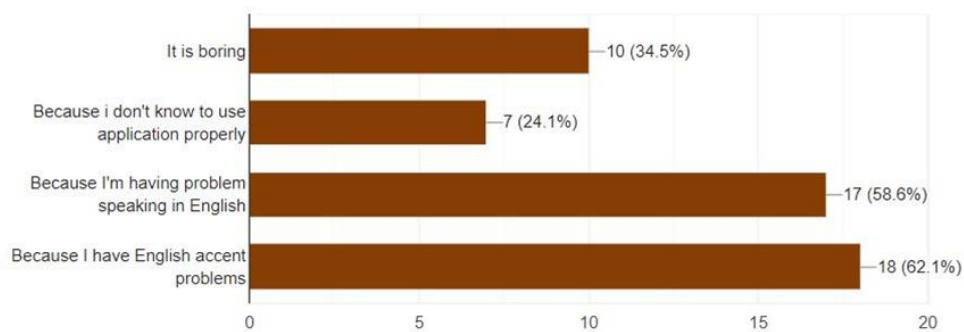


Figure 1: Problems identified by the Survey

This incident happens due to low technological literacy and the fact that the majority of technical jargon used in apps is in English. The English language is the foundation on which many of the tools and applications are created. Many Sri Lankans are reluctant to use such programs due to a lack of English proficiency and a fear of the English language.

Furthermore, because they are unfamiliar with the use of such applications, they might not be able to facilitate their actions with new technology and might even be excluded from society. There is also a possibility that their mental and health status will be

greatly reduced due to this existing epidemic. Additionally, there is a tendency for the status of Sinhala as the mother tongue to gradually deteriorate. Furthermore, there isn't a single application that functions correctly in Sinhala, and a few of the ones don't have the features they promise. In some current Sinhala applications, the display does not show the response, but it also has an accurate voice output.

Also, the command given by the user in some applications will not be answered. Additionally, research demonstrates that the user is unable of providing a brief and unambiguous response. However, the existing voice assistant cannot identify users' emotions and feelings. As a personal digital assistant, it should have a real-time feature to recognize users' emotions and feelings. That feature also lacks existing digital voice assistants. It is beneficial to have someone trustworthy to share feelings and seek relief or make suggestions to free their minds. Most people constantly lack trustworthy individuals in their hearts. Individuals may have difficulty reaching their loved ones when needed in various circumstances. This is a different type of issue. It is beneficial to have a trusted digital Sinhala assistant who can act as a friend in people's real-world situations.

Native languages like Sinhala are not given priority by major IT sector suppliers due to the low customer base and market share. Most people don't prioritize learning Sinhala as their first language. Sinhala is a language that is only commonly spoken in Sri Lanka, as is well known. This will result in Sinhala being underestimated and underused in technical situations. Due to their low educational levels, lack of fluency in English, and lack of technical illiteracy, some Sri Lankans do not use voice assistant software.

## **1.5 Research Objectives**

### **1.5.1 Main Objective**

As a solution to the aforementioned issue, a Sinhala voice assistant mobile application powered by Machine Learning and Soft Computing (MLSC) and Natural Language Processing (NLP) has been proposed to aid users with daily duties, problem-solving, and amusement. The proposed system focuses mainly on making the users' daily activities easier through Sinhala voice commands and making the lifestyle enjoyable and more arranged. With the help of this application, users can have an assistant like a real person and this application is working in their native language. Therefore, it's more familiar to the users rather than other voice assistant applications. The main four objectives that are related to the main four research components are the followings.

1. Identify the accurate face of the user and authenticate and provide easy access to the application while improving security.
2. Improve the capability of providing the most relevant responses to the users' voice commands by converting users' initial voice commands into text commands and converting them to the English language.
3. Providing accurate voice output and text, according to the user's command.
4. Improves the user's mental freedom by identifying their real-time emotions and emotional level through their facial expressions; based on the emotional level and the status of the emotion, and their personal preferences, the application provides the ideal suggestions. Because of the way things are in the user's life right now, these ideas will be used to improve and increase mental freedom.

### **1.5.2 Specific Objectives**

The followings are the sub-objectives that are related to the aforementioned main objectives that are related to the research project. To achieve the study's primary objectives, the particular sub-objectives must be met.

- Determine whether the object is a face or not.

To construct a fully functional system, each scenario must be followed. To enhance facial recognition via the app, the face must first be identified. First, the algorithm determines whether or not a face is present. If a true occurs, the program automates the subsequent activities. If not, the algorithm will prompt the user to correctly position the face.

- Faces are extracted from the background.

As stated previously, following face detection, the majority of images contain a background. As a result of this background algorithm, both performance and distraction may suffer. Second, for face landmark analysis, the background will be eliminated. so that the algorithm can assess the facial landmarks without being distracted.

- Capturing a face image by executing an algorithm.

In order to authenticate the user, a facial picture must be stored. In order to improve performance and precision, a high number of photos will be taken in a matter of seconds.

- Verify the user's identity and provide access.

Following their capture, photographs of the users will be saved in a protected database. When a user attempts to access the program, biometrics will be evaluated. For this method, scanning the face is required. During the scan, all face landmarks will be compared to previously captured images. Moreover, the system will determine the similarity index by comparing face landmarks. If the face landmarks become comparable and the similarity index is met, the algorithm will permit access to the application.

- Speech recognition of users' speech commands

Speech-to-text technology from Google can handle noisy audio from a variety of contexts and can be used for speech recognition in many languages without the need for additional noise cancellation. As a result, the noise cancellation component operates automatically. Initially, The Python code must connect to the Google cloud voice recognition using the Python library that Google cloud has specified. The text is delivered to the front end via an API. By that, the user can confirm that what he said was correctly identified.

- Converting the Sinhala text into English

For this English translation process, a Text-to-Text Transfer Transformer (T5) model will be used with the Attention mechanism which is a complex Neural Machine Translation (NMT) mechanism in Natural Language Processing (NLP). The Python programming language will be used in conjunction with React native to create this component.

- From the English phrase received from the user's command, the first thing to do is to search for the pattern that belongs to the command using the dataset.
- After finding the corresponding pattern, the corresponding response will be displayed.

- Also, the English language received from the user's command is prepared by a neural machine translation model (NMT), and the output from it is read in Sinhala using a Text-to-speech API.

- Identify if there is a face in the captured image.

Initially, to achieve the main task of detecting emotion, there should be a face in the captured image. So, the algorithm first checks whether the face is there or not. If there is a face in the image, then it automatically continues the process. Otherwise, it displays an error message.

- Remove the background and crop and get only the face.

To perform with good speed and better accuracy, the algorithm needs to meet specific conditions. To work accurately, the algorithm should remove the background content so that the algorithm speed will be increased, and it will not be distracted by the background. Then, by cropping the image, it can only focus on facial landmarks.

- Identify users' facial expressions.

Using facial landmarks, the algorithm will be able to identify the facial expressions. Based on well-trained data, an algorithm will analyze the inputted image and then identify the facial expressions accordingly.

- Identify users' emotions according to their expressions.

Based on the identified facial expressions, the emotions will be categorized according to the well-trained data. Then the algorithm will identify users' real-time emotions.

- Predict the emotion level in percentage to the corresponding emotional class.

There are emotional classes. So, the algorithm will predict what emotion class by checking the facial landmarks. Then, it will analyze the emotion level as a percentage based on the facial landmarks and the class you chose.



- Provide ideal suggestions according to the users' emotions and preferences.

Based on the identified emotion and the users' preferences, suggestions will be provided. These suggestions will be varied according to each user. To provide suggestions, the application will check the users' preferences. Based on preferences and the real-time emotion and emotion level, ideal suggestions will be provided with a user-friendly interface.

## 2 METHODOLOGY

### 2.1 System Overview

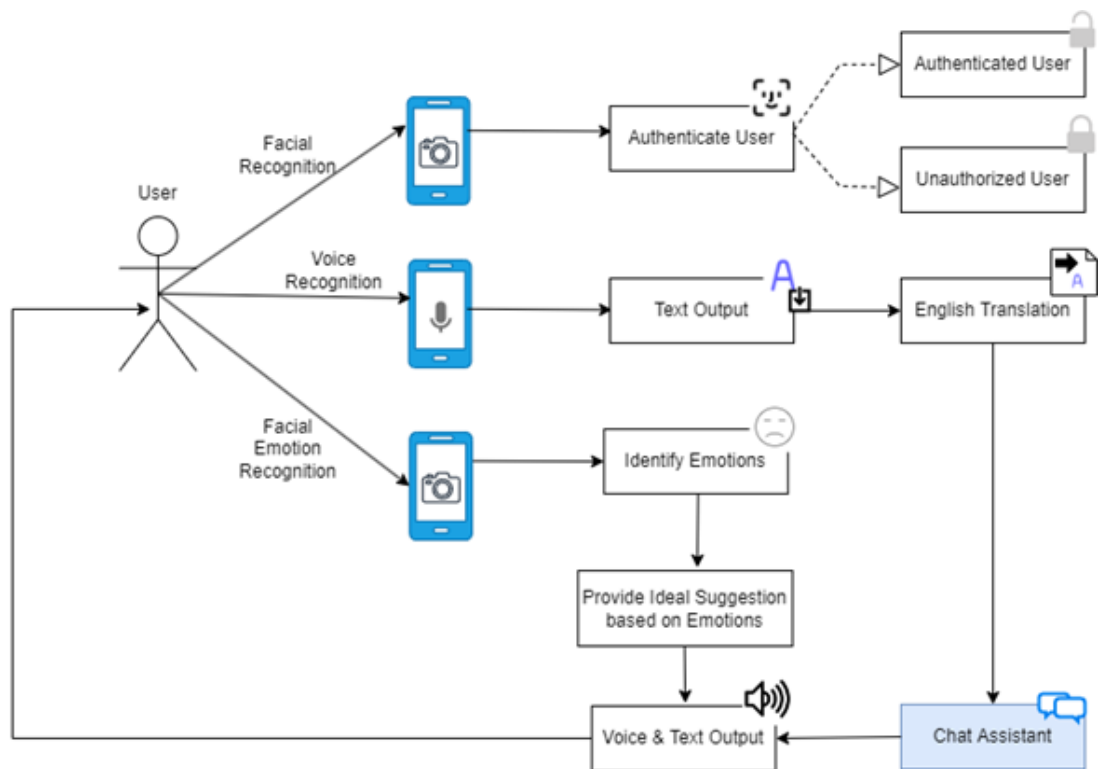


Figure 2: System Overview Diagram

## **2.2 Research Design**

Main research consists of four components. The combination of these four Sri voice applications functions using Sinhala language and also Sinhala voice commands and text outputs to be a reliable digital personal assistant in a real-time situation for the users. These situations will vary from user to user. The application works in these situations and helps people get relief from their lives with its emotion recognition system.

The primary research component of the Sri Voice Mobile Application uses Sinhala Language, Sinhala voice commands, and Sinhala text outputs to function as a dependable digital personal assistant for users in real-time situations. Here, he may have a unique experience compared to other apps. Additionally, they may get responses to their queries by voice command. Additionally, if the user wants to access other programs, they may do so via voice commands. These scenarios will differ amongst users. With the emotion detection technology, the application finds the realtime feelings on the face and provides ideal suggestions to each user based on the situation.

Before the research, the basic idea and the specific content of each part were needed. When focusing on the research scenario, the application's security layer must first be developed. Then to work with voice commands, there should be a voice to text section. Using this section, the application will convert inputted Sinhala voice to text while reducing the background noise. Then the chat assistant section can interact with the user to give commands and ask questions and the chat assistant will provide answers accordingly.

When improving users' feelings using facial expressions, initially a face should be detected on the captured image. From this CNN model, identify the users' facial expressions and detect the emotion from their faces. The level of emotion is expressed as a percentage. Initially, registration of the users, their hobbies, and their interests will

be inputted into the system. Finally based on the emotion status and the level, hobbies and interests, the ideal suggestions are provided to user.

### **2.3 Data Gathering**

For this research, different datasets have been used. Each and every component of the study has four distinct components.

1. Voice recognition data collection
2. Face recognition data collection
3. Chat assistant data collection
4. Facial expressions recognition data collection

The author used the internet to gather the data needed to create the learning model. The collected data was found from Hugging Face. Hugging Face is a community and data science platform that gives users access to tools for building, training, and using machine learning (ML) models that are based on OS tools and code. It serves as a gathering place for a sizable community of data scientists, researchers, and machine learning (ML) engineers to come together, share ideas, support one another, and take part in open-source projects. Overall, the data that was gathered was in text form. In its current form, collected data cannot be used. Therefore, a variety of techniques were employed to arrange the data in a way that would improve the efficacy and precision of the final product.

Therefore, the input and output were tokenized during preprocessing using the Auto Tokenizer API. In that section, the maximum length can be specified. The other parameters can also be set up from here. The dataset is then divided into mini badges and converted into a data load, which can then be passed to the T5 model.

#### **2.3.1 Face recognition Data Collection**

Before beginning this project, we need a large amount of face data to detect face attributes and train the model. Because of this, the research team tried to gather relevant data from Internet sources. It is the most convenient method. When searching

for faces, look for the most applicable data sets. Because the accuracy of the research effort depends on these data sets. The team begins by obtaining the Celebrity Faces Dataset and training the model. However, when testing the model with local individuals, it encounters various issues. Consequently, the research team had to collect local data subsequently. The team has created a successful local dataset by recording facial data from a variety of angles and at different times. Finally, the team has collected approximately one thousand faces, each with a resolution of 160 by 160 pixels. The research team also trained this dataset.

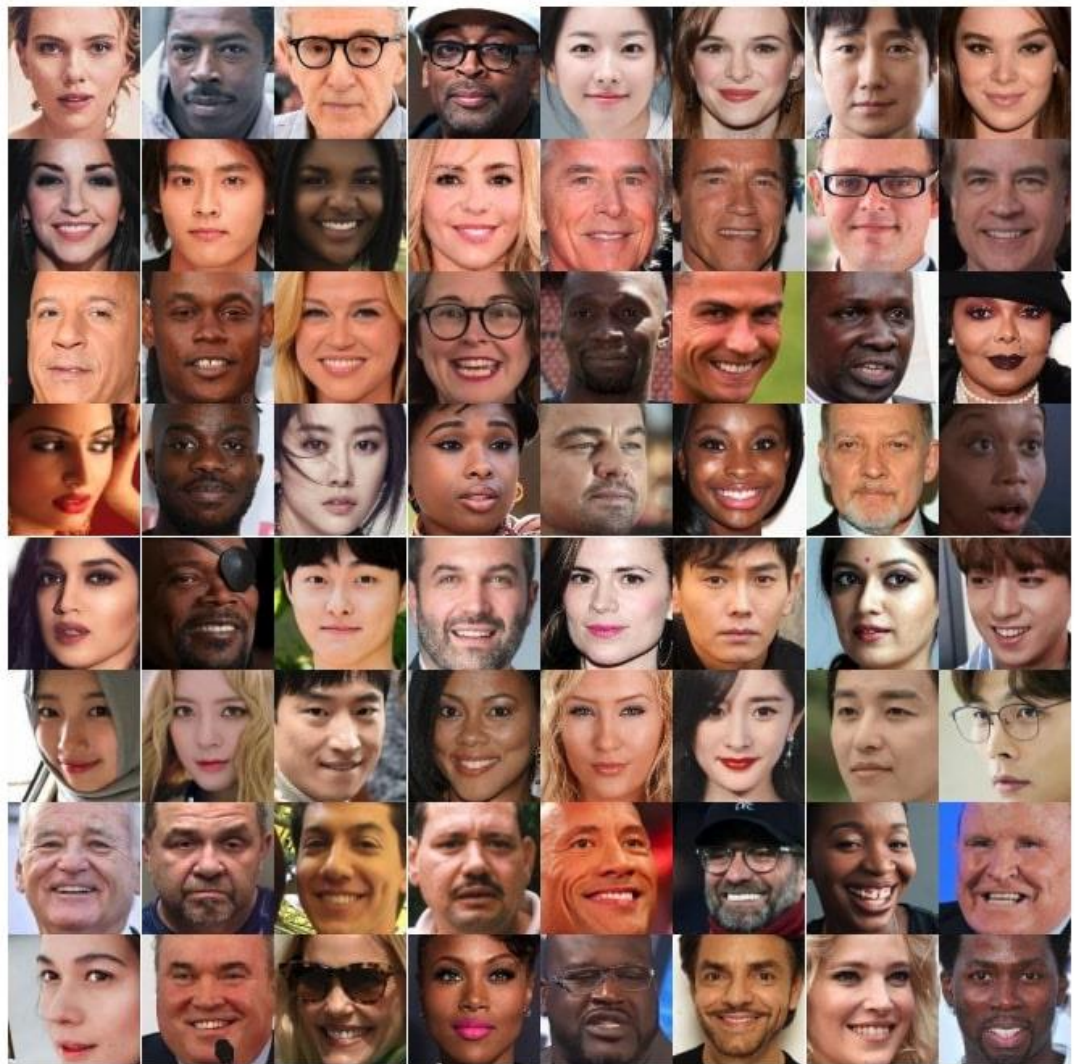


Figure 3 - Face data set

### **2.3.1 Chat Assistant Data Collection**

I used the internet to find a dataset suitable for this research, but it was difficult to find a dataset related to my component. Then used an intents base dataset to create a dataset suitable for a voice assistant. The aim is the goal the chatbot user has in mind while inputting a question or comment (called an utterance) inside the conversation flow. A chatbot entity is a modification or lower-level explanation of the intent. In other words, purpose represents the intended objective or action of the chatbot user. The fundamental role of an entity is to extract speech data in order to enhance intent prediction.

### **2.3.1 Facial Expressions Recognition Data Collection**

The special reason for using it is that, unlike the others, here according to the user's command, the relevant answer is shown on the display, and it needs to be read again in Sinhala. Further, the important thing to be considered while setting this is to maintain the accuracy at a good level. Therefore, the number of grant tags was taken as 10. And there are 950-1000 words here. There are usually 20-30 patterns in the tags here. And also are between 10-20 responses in one tag.

To develop this emotion recognition huge number of images had been collected. To get better performance from the model and the accurately trained model there need to be vast number of images on the training dataset. To increase the model's performance, dependability, and accuracy, two datasets were utilized.

- 1) FER2013
- 2) Custom Made Sri Lankan Database by me.

#### **FER2013**

This dataset is popular database among the industry and the researchers. This dataset consists of grayscale cropped images of the faces which are 48 x 48 pixels. The images of the faces have been automatically aligned such that each face takes up roughly the

same amount of the frame in each image. This dataset is also categorized according to the emotions. This consist of 7 emotion classes. Them are

- 0 = Angry
- 1 = Disgust
- 2 = Fear
- 3 = Happy
- 4 = Sad
- 5 = Surprise
- 6 = Neutral

There are a total of 28,709 images in the training dataset, with 3,589 available for use in the public test set.



Figure 4 : Example images from FER2013 dataset

### 2.3.2 Custom Made Sri Lankan Database by me

Custom Made Sri Lankan Database by me. While developing the emotion recognition algorithm, the issue has been raised when testing. So, this FER2013 dataset is a foreign dataset. It does not consist with any local people (Sri Lankan people) images in training set and also testing set. When compare with a foreign person there are different

characteristics on Sri Lankan people's faces. So, to avoid the accuracy related problems the custom created database has been used by me. To implement this dataset, I took different people's face images according to the emotion classes. The images of the faces are well categorized. This dataset includes main emotions (Happy, Sad, Angry, Neutral). This consists of 240 images. These images are cropped and there are some backgrounds also can be found. And also, that dataset is a colored database it is not like FER2013. Because FER2013 database is a grayscale image

## **2.4 Voice recognition and English Phrase Generation**

### **2.4.1 Google Cloud Speech to Text recognition**

One of the most accurate systems for speech recognition is Google Cloud Speech to Text. The Google Cloud Speech recognition method offers recommendations on how to improve the accuracy of the transcription of uncommon and domain-specific words or phrases. Users have access to classes that allow them to translate spoken numbers into a wide range of objects, such as addresses, years, currencies, and more. Furthermore, it makes a choice from a variety of trained voice control, call transcription, and video transcription models that are meant to meet 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. Therefore, the noise cancellation part is happening automatically. Initially, The Python code must connect to the Google cloud voice recognition using the Python library that Google cloud has specified. The accuracy rate of the speech command will be output along with the text output. The text is delivered to the front end via an API. By that, the user can confirm that what he said was correctly identified.

Since the API is the key component of the solution, it must first comprehend how it can be used, the service and what constraints or requirements it places on the other components. Three common usage scenarios are covered in the documentation:



transcription of short files, transcription of long files, and transcription of audio streaming input. In order to instantly recognize a user's speech, third scenario is the most interesting one.

The following features of the audio stream are advised by the documentation to get the best voice recognition results:

- 16 kHz sampling
- 16-bit signed sample format
- lossless compression
- single channel
- 100 milliseconds length of the audio chunk in each request in the stream

Additionally, any pre-processing like noise reduction, gain control, or resampling should be avoided.

Initially, the Google cloud should be activated using a google account. Then the user has to enable the Speech-to-text recognition API from it. Then the user has to create a service account to get the service. After that, since the language used in this component is python, the python code was implemented for the speech to text recognition process using the client service key that has been provided by the created client service account in the PyCharm IDE. The main python library used for this is google-cloud-speech and it can be installed using **pip install google-cloud-speech** command.

By using potent neural network models, Google's Speech-to-Text API enables users to convert audio files to text. Real-time transcription of audio content is possible, as is transcription from stored files, which may be more relevant for social science research. More than 125 languages are currently recognized by the API. The Accuracy is found more than 87% and most of the times it has the ability of recognizing the text 100% correctly.

### 2.4.2 Neural machine translation (NMT)

Considering the problem that is going to be solved with Neural Machine Translation, the machine translation task can be viewed as a model for the conditional probability  $P(y|x)$ , where  $x$  and  $y$  stand for the input and output sentences, respectively. Finding the best sentence in a different language,  $Y$ , for a given sentence  $X$  in language  $X$  is essentially what aiming to accomplish. This objective can be formalized as in Eq. In other words, output sentence  $y$  that maximizes the conditional probability  $P(y|x)$ . This goal can be rewritten as  $P(x|y) P(y)$  using Bayes' Rule.

$$\operatorname{argmax}_y P(y|x) \rightarrow \operatorname{argmax}_y P(y|x)P(y) \quad (1)$$

Here, the translation model  $P(x|y)$  represents how words and phrases are translated, and the target language model  $P(y)$  represents the model of  $Y$ . The latter probability can be viewed as awarding the output sentence a fluency score, for example,  $P(\text{"I am playing"}) > P(\text{"I playing am"})$ . To build a language model  $P(y)$ , there are several techniques that can be used. The answer to the more challenging question is how to build the  $P(x|y)$  translation model. A translation corpus with many sentence pairs is available for use. The translation model can be discovered using statistical machine translation techniques. However, it is a subtle issue. Even simple sentences can be challenging to translate into another language because language is inherently ambiguous and complex. Consider the following example in Sinhala “මම හෙට සෙල්ලම් කරන්නම්” which translates to “I will play tomorrow”. In Sinhala language ‘සෙල්ලම් කරන්නම්’ is referred to ‘play’ in English which is one word in English but two words in Sinhala. As you can see, it is even challenging to determine how these two straightforward sentences relate to one another. Think about the consequences if more sophisticated words and phrases were added to the mix.

In the Neural Machine Translation (NMT) process, a Text-to-Text Transfer Transformer (T5) model will be implemented by using the attention mechanism. The

Attention is a sophisticated NMT mechanism in NLP that can also be described as a text strategy. The Text-to-Text Transfer Transformer (T5) is a text-to-text conversion model which is mostly used in neural machine translation (NMT). The T5 model can be pretrained or rebuild according to the project's requirements. In addition, it also can be finetuned using a custom dataset. One of the most important advantages of the Text-to-Text Transfer Transformer model is there are many existing resources that could be found to finetune the model. This is the primary method used in the English translation process. The Attention mechanism, a sophisticated NMT mechanism in NLP, is utilized for this section. The output of the Neural Machine Translation (NMT) will be an English-translated text of the users' initial voice command.

The application will be able to provide more accurate responses in the chat assistant component part of the application with the aid of the English translation process. By doing this, users will benefit from getting the most relevant responses to their inquiries.

### **2.4.3 Text-to-Text Transfer Transformer**

Text-to-Text Transfer Transformer, also known as T5, is a Transformer-based architecture that employs this method. Every task, such as translation, question-answering, and classification, is viewed as a process of feeding the model text as input and training it to produce some target text. This allows for the use of the same model, loss function, hyperparameters, etc. across our diverse set of tasks. Comparing BERT to the changes is as follows:

- adding a causal decoder to the bidirectional architecture.
- replacing the fill-in-the-blank cloze task with a mix of alternative pre-training tasks.

Transfer learning enables the model to learn from an infinite amount of data and apply the trained model to downstream tasks even though downstream tasks and pre-trained tasks may differ. Due to the rarity and value of labeled data compared to the abundance of unlabeled data, it is particularly helpful in natural language processing (NLP) [11].

T5 is a model that enables text input while also producing text as output. The model is trained using a wide range of unlabeled data thanks to this adaptable design, and all tasks have the same goal, training method, and decoding procedure. A prefix text is required in order to inform the model of the task that needs to be resolved. A prefix for a machine translation task might be "translate from Sinhala to English:" The expected result for the regression task (STS-B) is a similarity score between 1 and 5. A text is produced by the model that corresponds to a number between 1 and 5. Training configurations are as following [12]

- Vocabulary: Text can be encoded using Sentence Piece (Kudo and Richardson, 2018), with a sub word size limit of 32k for Romanian, English, German, and French.
- Learning Rate: 0.01, exponential decay for the final 104 steps.
- Learning Rate Schedule:  $1/\sqrt{\max(n, k)}$ , while current training iteration and  $k$  is the number of warm-up steps ( $k$  is  $10^4$  in all of the experiments).

Similar to BERT, Masked Language Modeling (MLM) is used as an unsupervised training objective. In order to enable the model to learn from noisy data, 15% of the tokens are dropped [12].

Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party apple week .	(original text)
Deshuffling	party me for your to . last fun you inviting week Thank	(original text)
I.i.d. noise, mask tokens	Thank you <M> <M> me to your party <M> week .	(original text)
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>

Figure 5 : Different training objective

BERT substitutes a single token (i.e., "MASK") for 90% of the tokens and a random token for 10% of the tokens, whereas T5 substitutes a string of corrupted tokens (e.g., "X", "Y", and "Z") for the tokens in a given order. Additionally, a single span token will be used in place of several tokens for consecutive words. Various corruption rates

in place of the standard rate (15%) to determine the effects could be used. The performance of the model was found to be largely unaffected by the corruption rate. In order to determine how well the model performed, researchers evaluated the average span of a corrupted span token. As an illustration, the corrupted token rate is 15% while the total token and total corrupted span are 500 and 25, respectively. The total amount of tokens that are masked is 75 ( $500 \times 15\%$ ), and the typical span is 3 ( $75/25$ ). Its minimal impact on the model's performance is the conclusion.

In addition to the training goal, researchers assessed the effectiveness of training methods.

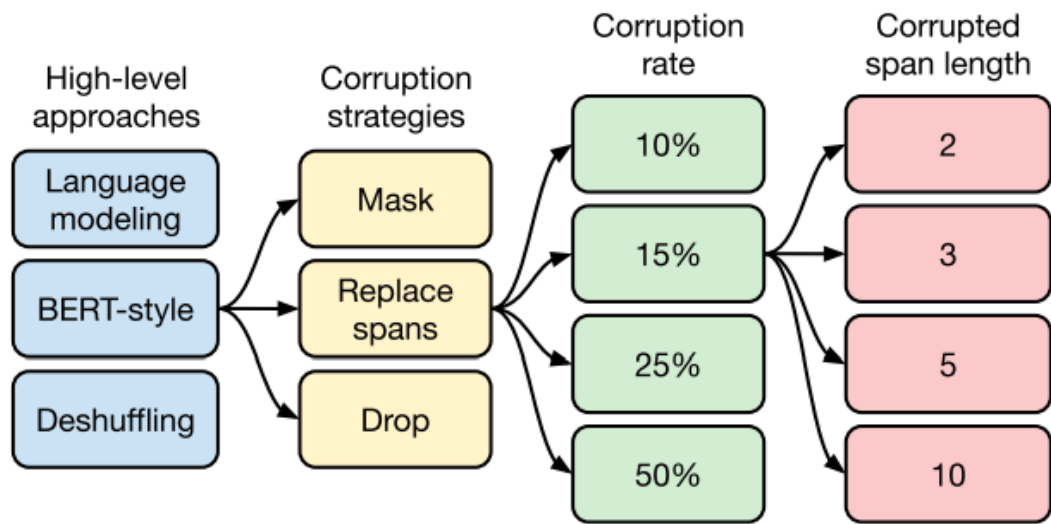


Figure 6 : A flow chart of unsupervised objectives configuration

In addition to the training goal, researches assessed the effectiveness of training methods. Finetuning is one of the most famous training methods. For fine-tuning the model for upcoming tasks, there are three methods. The first method is fine-tuning all pre-trained layers when training downstream tasks. The first approach entails fine-tuning all previously learned layers before training new ones. The second method only updates the adapter layers when training downstream tasks, leaving the pre-trained layer frozen. The third method is gradual unfreezing [13]. Over time, the pre-trained layers will become unfrozen. Howard and Ruder introduce gradual unfreezing in 2018.

Fine-tuning method	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ All parameters	<b>83.28</b>	<b>19.24</b>	<b>80.88</b>	<b>71.36</b>	<b>26.98</b>	<b>39.82</b>	<b>27.65</b>
Adapter layers, $d = 32$	80.52	15.08	79.32	60.40	13.84	17.88	15.54
Adapter layers, $d = 128$	81.51	16.62	79.47	63.03	19.83	27.50	22.63
Adapter layers, $d = 512$	81.54	17.78	79.18	64.30	23.45	33.98	25.81
Adapter layers, $d = 2048$	81.51	16.62	79.47	63.03	19.83	27.50	22.63
Gradual unfreezing	82.50	18.95	79.17	<b>70.79</b>	26.71	39.02	26.93

Figure 7 : Comparing different fine-tuning methods

#### 2.4.4 Attention mechanism

This idea of directing your attention is the foundation for Deep Learning's attention mechanism, which emphasizes particular aspects of the data when processing it. The model experiences a bottleneck when a fixed-length vector is used. While back-propagation can partially pass gradients backward, there isn't much room for the model to pass information from the encoder to the decoder. Bahdanau et al. use the attention mechanism to get around this so that the decoder can look at the source tokens that are pertinent while creating the subsequent token [14]. Across a wide range of sequence-to-sequence applications, the attention mechanism significantly boosts model performance. However, people did not stop there, and stronger strategies have been put forth.

Consider the following translation task as an example to demonstrate how attention mechanism functions. For example, the French translation of "How was your day" is "Comment se passe ta journée." The network's Attention component will map the important and pertinent words from the input sentence and give them higher weights for each word in the output sentence, making the output more predictable [15].

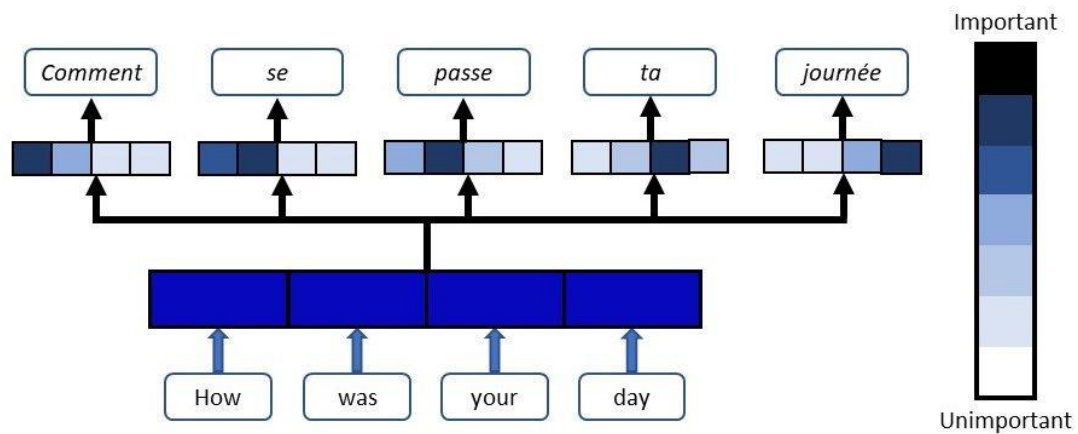


Figure 8 : Weights are assigned to input words at each step of the translation

## 2.5 Improve Security using Facial recognition

Face detection is an AI-based technology that uses facial recognition to identify faces in digital images. It can be used for various applications such as security, law enforcement, and personal safety. Face detection has progressed from rudimentary computer vision techniques to more sophisticated technologies such as artificial neural networks. It is now a vital step in many key applications, such as face tracking and facial recognition. Face detection is a process utilized for analyzing an image or video to identify parts of it that should be focused on, such as age, gender, and emotions. It uses facial expressions to generate a faceprint. Face detection algorithms are used to find human faces in large images, which often include other non-face objects such as buildings, vehicles, and landscapes. The goal of this training is to improve the algorithms' ability to detect faces in large data sets. This is done by training them on thousands of images. To get a face detection in a more accurate way the team had to tried with different methods. They are,

1. Face recognition using open cv
2. Using media pipe
3. Using TensorFlow

#### 4. Using Face Net

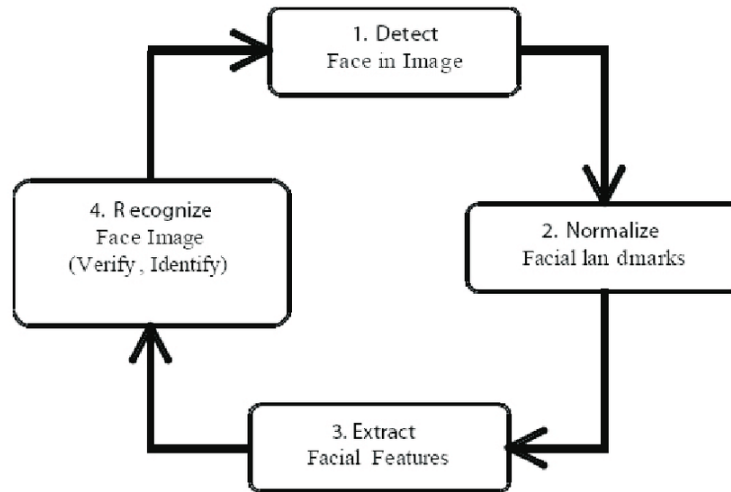


Figure 9 : Face recognition process

##### 2.5.1 Open CV

OpenCV is an open source, multi-platform software library for creating real-time computer vision applications. Face and object detection are only a few of the many imaging and video analysis tasks it tackles. OpenCV, short for "Open Source Computer Vision Library," is a set of APIs first developed by Intel and now maintained by Willow Garage for use in real-time computer vision applications. In accordance with the BSD open-source license, its use is without cost. This library works on a wide variety of operating systems. The main emphasis is on processing images in real time. This C interface allows OpenCV to run on devices like digital signal processors, which are native to the C programming language. Wrappers for languages like C#, Python, Ruby, and Java have been developed to facilitate wider use. However, beginning with version 2.0, OpenCV has supported both the C interface and the C++ interface. This new interface attempts to reduce the number of lines of code required to provide vision capabilities as well as common programming issues, such as memory leaks, that might arise when using OpenCV in C, as represented. Currently, the majority of OpenCV's



breakthroughs and algorithms are implemented through the C++ interface. Unfortunately, it is significantly more difficult to build wrappers in other languages for C++ code than for C code; hence, many of the more current OpenCV 2.0 features are missing from the wrappers for other languages. When compared to other models in terms of accuracy, open cv has a low level. Therefore, the team had to test their accuracy using alternative approaches.

### **2.5.2 Media pipe**

In June of 2019, Google's open-source Media Pipe was introduced for the first time. It integrates computer vision and machine learning skills to ease our lives. Media Pipe is a framework for developing multimodal, cross-platform (Android, iOS, web, edge devices), applied machine learning (ML) pipelines (video, audio, or any time series data). Media pipe also enables the insertion of machine learning technology into demos and apps running on a wide variety of hardware platforms.

Media pipe provides extremely effective and precise face detection models for detecting several faces from photos. It also gives detections for six face landmarks. First, the media pipe must be installed, then a media pipe face detection model must be initialized, and media pipe drawing utilities will be used to conveniently draw points and rectangles on an image. After that, researchers must convert the image to its original size by multiplying its x values by its width and its y values by its height. Using Media Pipe, obtain face detection and facial landmarks drawing.

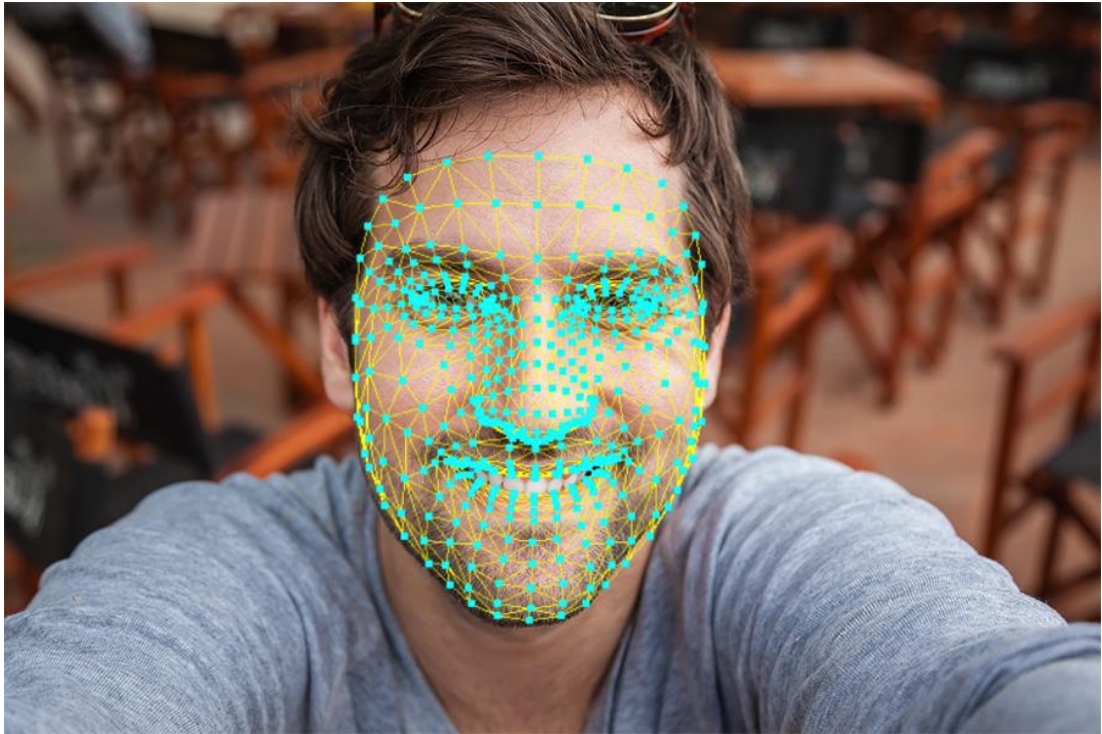


Figure 10 : Scanning via MediaPipe Technique

### 2.5.3 TensorFlow

TensorFlow is a general-purpose framework for machine learning. TensorFlow may be used anywhere, from training massive models across clusters in the cloud to running models locally on embedded systems such as your phone or IoT devices. Developing face recognition and detection models with TensorFlow may involve work, but it is ultimately worthwhile. TensorFlow is the most popular Deep Learning framework, and it includes pre-trained models that make picture classification simple. CNN is used to classify photos.

In most instances, to develop a model, it is sufficient for the categorization of the photos to provide a similar positive image. Through a process known as anchoring or Transfer Learning, the picture is then trained and retrained. TensorFlow enables the creation of dataflow graphs by developers. Each node in the graph symbolizes a mathematical process, and each edge between nodes is a tensor, which is a

multidimensional data array. TensorFlow applications can execute on nearly any handy target, including a local PC, a cloud cluster, iOS and Android devices, CPUs and GPUs. Using Google's cloud, you may run TensorFlow on Google's unique TensorFlow Processing Unit (TPU) silicon for additional acceleration. Therefore, the researchers tried to get the face recognition with this machine learning framework.

#### **2.5.4 Fine Tuning**

Transfer learning can be applied or utilized through the process of fine-tuning. Fine-tuning is a procedure that takes a model that has already been trained for a certain task and then tunes or modifies it so that it can execute a second similar task. If the new goal is similar to the previous one, we can use a pre-existing artificial neural network and benefit from the model's training without having to build it from scratch. Multiple iterations of trial-and-error procedures are often necessary when starting from scratch to build a model. Depending on the data we use to train our model, building and validating it can be a mammoth undertaking in and of itself.

This is what makes the method of fine-tuning so appealing. If we can discover a trained model that already does one task effectively, and that task is remotely similar to ours, then we can apply whatever the model has learned to our specific work. After modifying the structure of the old model, we need to freeze the layers from the original model in our new model. In this research component also, researchers had to make a local data set and later it finetuned to the face net model to get the more accurate output.

#### **2.5.5 Face Net**

As a test for other deep learning methods, Face Net uses deep convolutional networks to optimize its embedding rather than intermediary bottleneck layers. One-time learning describes this type of approach. This technique has the potential to produce the first model using a minimal set of face images, and it may be utilized with subsequent models without retraining. Directly in Euclidean space, where similarities make up the distance between facial models, Face Net trains the face. Once we have a

better understanding of how different face models are related, we can use the feature vectors that are connected to Face Net to do face recognition and classification with easy.

Face Net utilizes triplets by matching face to face using an online novel triplet mining method during training. Obviously, this trio consists of a variety of anchor images, with each image comprising both positive and negative images. Face Net has batch layers as inputs and a deep architecture consisting of a deep CNN followed by L2 normalization, the result of which is face embedding. Face Net has batch layers as inputs. During the training procedure, Face Net also pursued the triplet loss. Using the rest net backbone of the CNN architecture, the local dataset was optimized.

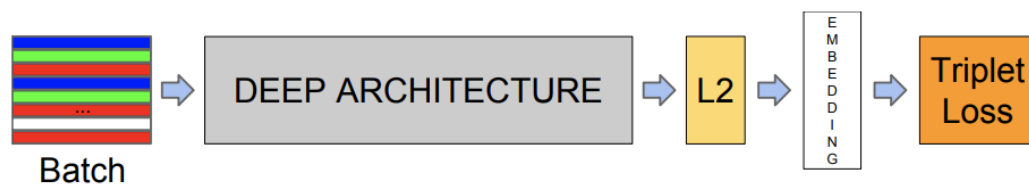


Figure 11: Face Net process

## 2.6 Chat Assistant

### 2.6.1 Neural Machine Translation Model

By including document-level context information, neural machine translation (NMT) may be enhanced. To capture the context in an organized and dynamic way, present a hierarchical attention model. The model is included into the original NMT architecture as an additional level of abstraction, dependent on the NMT model's prior concealed states. Experiments demonstrate that hierarchical attention considerably increases the BLEU score over a strong NMT baseline using context-aware approaches that are state-of-the-art, and that both the encoder and decoder benefit from the context in complementary ways.[12]

Here, I describe the issue I am attempting to address using NMT. The Machine Translation problem may be seen as modeling the conditional probability  $P(y|x)$ , where  $x$  and  $y$  represent the input phrase and the output sentence, respectively. For a given statement  $x$  in language  $X$ , we are effectively seeking the optimal sentence  $y$  in another language  $Y$ . We may explicitly express this objective using the equation: 1, i.e., we seek the output phrase  $y$  that maximizes the conditional probability  $P(y|x)$ . This aim may be rewritten using Bayes' Rule as  $P(x|y)P(y)$ .

$$\operatorname{argmax}_y P(y|x) \rightarrow \operatorname{argmax}_y P(y|x)P(y) \quad (2)$$

Here,  $P(x|y)$  models the translation model, i.e. how words and phrases are translated, whereas  $P(y)$  models the model of the destination language,  $Y$ .  $P(\text{"ආයුබෝවන්! මම ඔබට කෙසේද සහය වන්නේ?"}) > P(\text{"වන්නේ? සහය මම ඔබට ආයුබෝවන්! කෙසේද"})$ . There are techniques available for creating a language model  $P(y)$ . The much more challenging challenge is how to build the translation model  $P(x|y)$ . We may use a translation corpus containing many sentence pairings.

## 2.6.2 Attention Mechanism

The attention mechanism technique was used to further improve what is done by the above neural machine transaction. Accordingly, to increase the performance of the encoder-decoder paradigm for machine translation, the attention mechanism was devised. By combining all of the encoded input vectors with weights, with the most relevant vectors given the biggest weights, the attention mechanism was designed to allow the decoder to employ the most crucial elements of the input sequence in a flexible way.[13]

Bahdanau et al. (2014) established the attention technique to overcome the bottleneck issue that emerges with the usage of a fixed length encoding vector, where the decoder

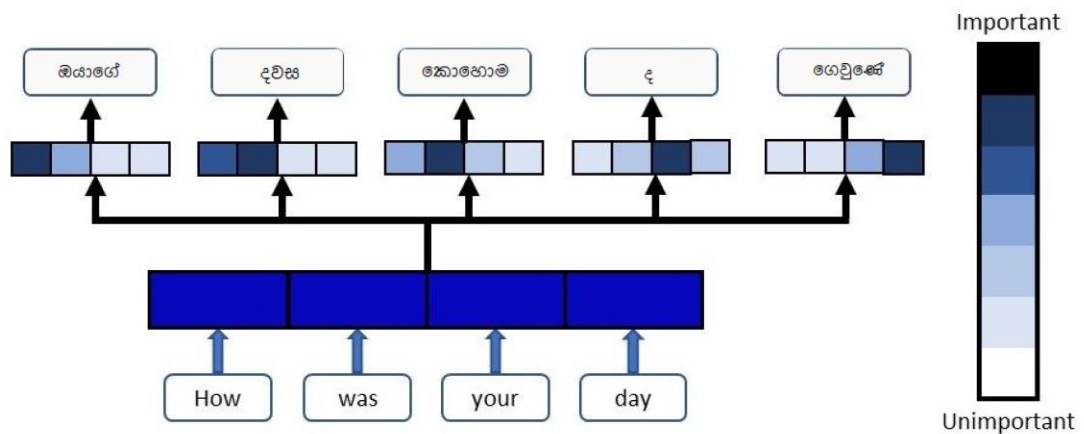


Figure 12 : Attention mechanism

would have restricted access to the input's information. It is considered that this is especially problematic for long and/or intricate sequences, whose representations would be needed to have the same dimensions as shorter or simpler sequences.

Take the following translation task as an illustration of how the focus the mechanism works. For instance, the Sinhala translation of "How was your day?" is "ඔයාගේ දවස

කොහොම ද ගෙවුණේ" The Attention component of the network will map the significant and relevant terms from the input phrase and assign them larger weights for each word in the output sentence, so making the output more predictable.

### **2.6.3 Text-To-Speech API**

The output obtained after the neural machine translation is given through the google cloud text-to-speech API. The reason for using this google text-to-speech API is because it can get a very good output. google cloud speech-to-text is a cloud-based speech-to-text transcription service powered by Google's AI-powered API. With Cloud Speech-to-Text, customers may transcribe their material with correct captions, improve the customer experience with voice commands, and get customer interaction analytics. The Cloud Voice-to-Text API enables users to modify speech recognition such that domain-specific phrases and rare words may be transcribed using hints. The program is capable of converting spoken numbers into precise locations, currencies, and years, among other things. The user may choose from a variety of trained models, including video, phone call, command, and search, or the default model. The speech-to-text API employs machine learning that has been taught to detect individual audio files from a given source, hence enhancing transcription outcomes. Google Speech-to-text may interpret audio streamed straight from a user's microphone or from a pre-recorded audio file and provide real-time transcription results. The Google Speech-to-Text API supports more than eighty different languages.

The voice synthesis technology developed by DeepMind is extraordinarily sophisticated and realistic. The majority of voice synthesizers, like Apple's Siri, use concatenative synthesis, in which individual phonemes are recorded and then assembled into words and sentences. [14]

## Prerequisites

- Create a Google Account
- Go to the Google cloud console and create project
- Billing enable for the created Google project

These are the main credentials to use Google Text To Speech API.

WaveNet, in comparison to earlier text-to-speech algorithms, generates more human-sounding speech. It incorporates human-like emphasis and intonation on syllables, phonemes, and words into standard speech. In compared to other text-to-speech frameworks, WaveNet provides voice sounds that are preferred by consumers. Unlike most other text-to-speech systems, the WaveNet model creates audio waveforms directly from text. A neural network that has been trained with a huge corpus of speech samples is used in the model. In the course of its training, the network learns the basic structure of speech, including the relationships between different tones and the shape of an actual speech waveform. When fed text, the trained WaveNet model can produce natural-sounding speech waveforms from scratch at up to 24,000 samples per second, with seamless transitions between phonemes.



## **2.7 Facial Emotion Recognition**

When researching about the field different studies and testing and results has found. To implement a good and highly accurate emotion analyzing model according to the given real-time face many approaches has been tried

1. OpenCV & Keras approach
2. TensorFlow & Keras approach
3. Convolutional Neutral Network approach

### **2.7.1 OpenCV & Keras approach**

To implement the development Python and the python library called NumPy has been used, A face has a number of traits that play a significant of emotions. The emotion recognition system contains of three phases:

- Face detection
- Feature extraction
- Emotion classification

Initially, the faces have been identified in FER2013 database. Dataset has two folders. Training set and Testing set. Each image structure represents the growth of an emotional expression, beginning with a neutral face and closing with a particular emotion. So, goal is to extract two photos that meet this requirement from each sequence.

#### **2.7.1.1 Extracting Faces**

The classifier will function successfully if the photos contain just faces, thus the images were processed to detect faces, cropped and placed in a particular folder [7]. OpenCV offers four pre-trained classifiers; therefore, it is desirable to recognize as many faces as feasible.

### **2.7.1.2 Training & Classification**

The dataset has been categorized and is prepared for recognition but train the classifier to distinguish specific emotions is must. In the dataset came with 2 split which is training set and testing set. Train the classifier to recognize labels using the training set. To be predicted, then utilized the classification set to evaluate the classifier's performance. Choose at random and train on 80% of the data while classifying the remaining 20%; repeat 15 times. After training the fisher face classifier, the emotions might be predicted.

### **2.7.1.3 In Real Time**

Although it is more difficult than with static images, It is still possible to identify emotions in real time because the camera is recording a video rather than just a single frame. In this instance, the Facial Landmarks approach [22] was utilized, which is more successful and robust than the sequence of static photos used with the Fisher face classifier [23], but required extra features and modules.

#### **2.7.1.3.1 Testing Landmark Detector**

Initially, front facing camera need to operate to record the real-time video. Every frame of the camera footage will have face recognized, and then those discovered faces will undergo further processing. The image is then processed by going from color to grayscale, improving the An adaptive histogram equalization is in contrast.

#### **2.7.1.3.2 Extracting features from the faces**

In the stage of feature extraction, the previously identified faces in the image undergo additional processing to identify the eye, eyebrow, nose, cheek, and lips. Initial estimations of the Y coordinates of the eyes were made using the horizontal projection. The regions surrounding the y coordinates were then analyzed to determine the precise locations of the features. Then, a technique for identifying corner points was used to get the requisite corner points from the feature areas.

## **Problem Raised**

This model has been trained successfully. But when testing the model's accuracy, it was not good. Because the emotion recognition model, which is developed using mainly OpenCV, unable to detect emotions correctly. Most of the time it returns wrong emotion class. To fix this issues number of epochs has been increased and it took 24 hours to train. After training again, it also failing. Because it was unable to analyze the emotion class from the real-time image. Because of this issue stared to develop new model using TensorFlow and Keras.

### **2.7.2 TensorFlow and Keras Approach**

To implement model using TensorFlow, some of libraries has been used. Them are Keras, OpenCV, matplotlib. A model for recognizing facial expressions in TensorFlow that is built on Inception-v3.

#### **Image Processing**

Image preprocessing is a crucial step for enhancing the effectiveness of image categorization. Cropping and converting images are two steps in getting a target picture ready for use. Format conversion for images. Using the Python image processing module PIL to convert between several picture formats. Image clipping. To increase the correctness of picture classification and decrease the influence of non-target info on image classification, the target image might be trimmed during image preprocessing.

The Inception-v3 [27] model relies heavily on the Inception structure's original network design concepts, with the most important methods including massive size filter convolution decomposition, an additional classifier, and shrunken feature maps. Inception-v3 [27] is a model that is trained using the ImageNet datasets.

### 2.7.3 Convolutional Neural Network Approach

Convolutional neural networks (CNNs) are a subset of neural networks specializing in image identification and classification, among other applications. In general, convolutional neural networks are comprised of numerous layers of tiny neurons that analyze the input image's receptive fields. The majority of a convolutional neural network consists of three-layer types: convolutional layers, max-pooling layers, and fully connected layers. The other two types of layers are responsible for feature extraction, the introduction of non-linearity into the network, and feature dimension reduction, with the latter being the least essential. The fully connected layer is tasked with categorizing the input based on the characteristics retrieved by the other layers.

When developing a model to recognize to emotions from the face there need to be a flow in Figure.2.

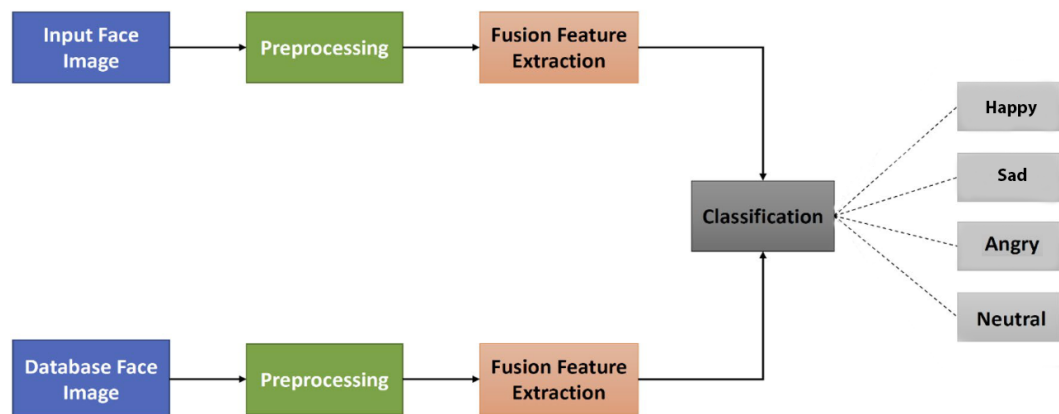


Figure 13 : Summary of the 3 Phases of the Facial Emotion Recognition Model

Initially, A simple CNN template consisting of numerous building components that are straightforward to comprehend and link with the suggested CNN model. As shown in Figure 4, Three different layer types make up a basic CNN: input, hidden, and output. Before reaching the output layer, the data travels through several hidden layers after entering the CNN through the input layer. The prediction of the network is mirrored in the output layer. In terms of loss and error, the network output is compared

to the actual labels. The network's hidden layers serve as the fundamental building blocks for data transformation. The four sub-functions of each layer are layer function, pooling, normalization, and activation. The following layers make up a convolutional neural network's architecture [24].

- Convolution Layer
- ReLu Layer
- Pooling Layer
- Fully Connected Layer
- Softmax
- Batch normalization

#### **2.7.3.1 Convolution Layer**

The basic objective of convolution in a Convolutional Neural Network is to extract features from the input image. Convolution pre-serves the spatial connection between pixels by learning picture characteristics (i.e., filtering using small squares of input data). CNN uses the phrases 'filter,' 'kernel,' and 'feature detector. The 'Convolved Feature', 'Activation Map,' or 'Feature Map' is the matrix produced by dragging the filter across the image and computing the dot product.

#### **2.7.3.2 ReLu Layer**

In this layer, each negative value in the cleaned image is removed and replaced with zero. This is done to prevent the sum of the values from becoming zero. Otherwise, the output is zero.

#### **2.7.3.3 Pooling Layer**

The goal of a pooling layer is to reduce the spatial dimension of the corrected feature map, resulting in a more condensed feature extraction. This layer of pooling produces a pooled featured map. The two most prevalent algorithms for pooling are maximal pooling and mean pooling. In Max Pooling techniques, the maximum parameter is eliminated at each step, while the remainder is diminished.

#### **2.7.3.4 Fully Connected Layer**

Fully Connected Layer transforms a two-dimensional pooled feature map into a one-dimensional vector, or feature vector. The outcome is a "flattened" map of combined features. This feature vector serves as a typical Fully linked layer for categorization.

#### **2.7.3.5 Softmax**

Softmax is accomplished by a neural network layer immediately preceding the output layer. The Softmax output layer must have the same number of nodes as the output layer.

#### **2.7.3.6 Batch Normalization**

Batch normalizer accelerates the training procedure by introducing a transition that maintains the mean activation close to 0 and the standard activation deviation close to 1.

### **2.7.4 Facial Expression Recognition (FER) Using Transfer Learning in Deep CNNs and Training the model**

When implementing the Facial Emotion Recognition (FER) model, The first layer of the CNN collects fundamental visual characteristics such as edges and corners. The subsequent layer recognizes more complicated elements, such as textures and forms, and the highest layer use the same method to learn more complex patterns. In the bottom layers of a Deep Convolutional Neural Network (DCNN), FER tasks are similar to other image-based processes, such as classification, because all images share the same essential features.

Because it is so time-consuming to train a DCNN model from scratch, the TL method can be used to fine-tune a DCNN model that has already been trained on an extra task for emotion identification. As such, FER works best with a deep convolutional neural network (DCNN) model (e.g., VGG-16) that has already been trained on a big picture classification dataset (e.g., Face Net).

To train this model huge computation power has been used. To train the model initially Google Collab has been used. But the free subscription was not enough to train the huge, weighed model. Because of that, to train this model using 25 epochs took 24hours with configuration of Nvidia GeForce 1660Ti 8GB. The subsequent subsections outline TL ideas for FER and the suggested FER technique with the necessary examples in depth.

Figure 3 depicts the overall design of a TL-based DCNN model for FER, in which the convolutional base is a piece of a pre-trained DCNN without its individual classifier and the classifier atop the basis is the newly added layers for FER. The DCNN has 2 steps: Replace the existing classifier with a modern one and refine the model. The new classifier portion of densely interconnected layers.

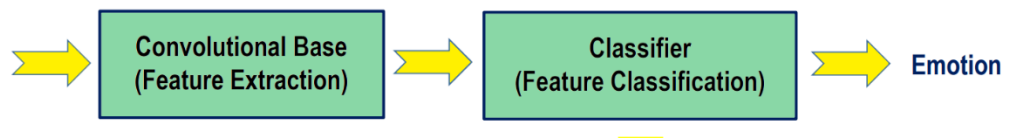


Figure 14: Overall flow of the Learning based deep CNN model

The suggested FER system is depicted in Figure 4 utilizing the famous pre-trained DCNN model VGG-16. The supplied VGG-16 model is trained to identify 1000 picture objects using the ImageNet dataset. Specifically, the deep layers of the previously trained model are redefined and tuned to improve emotion recognition (fine-tuning). To fine-tune the model two dataset has been used. Which are FER2013 contains 28,709 images and custom created dataset by including local Sri Lankan people images by me. This consists with over 500images. Modifications are made to the pre-trained model for emotion identification by redefining the thick layers, followed by fine-tuning with emotion data. In defining the architecture, the pre-trained model's final dense layers are replaced with new dense layers to classify face photos into one of four emotion classes (i.e., Happy, Sad, Neutral, Angry) The component and three additional classes no longer have a scenario. A dense layer is a normally, fully connected, linear layer of a Neural Network (NN) that receives a dimension as input

and produces the dimension vector necessary. Therefore, the output layer consists only 4 neurons. Fine-tuning is conducted on the structural design consisting of the convolution base of the model that has been pretrained and an additional dense layer. A cleaned, preprocessed (i.e., scaling, cropping, and other activities) emotion dataset is utilized for fine-tuning training. In testing, a cropped images is provided at the system's input, and the emotion with the highest output probability is considered.

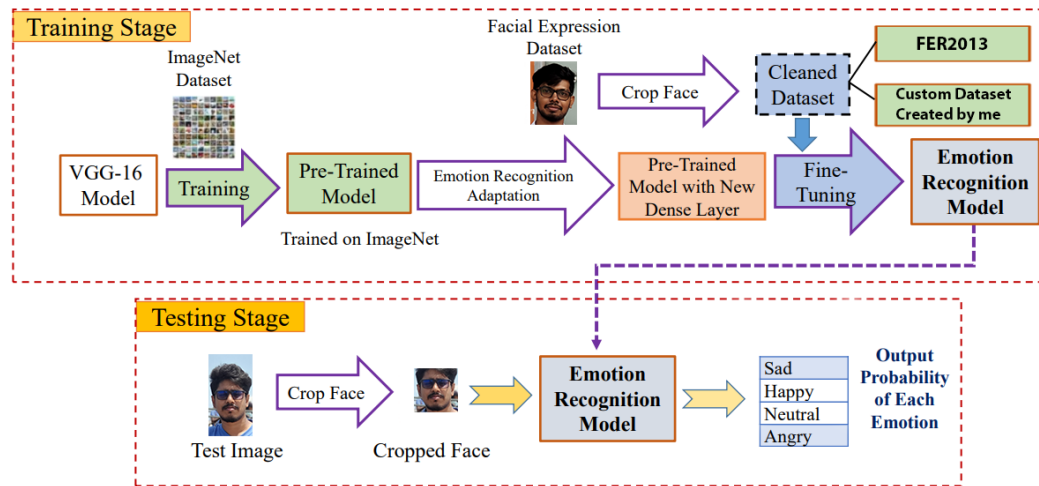


Figure 15 : Illustration of Facial Expression Recognition Model

During the testing, the model test image has been inputted and the face is identified as being there or not in the image. When inputting the image, there will be a background in the image. So, to get better performance from the FER model initially, the background is identified and cropped. Then the emotions will be analyzed by a cropped image according to the trained data. This algorithm has been trained for two datasets in order to achieve the main goal of the component. After analyzing the image and the expressions from the image, the probability value will be returned for each class. Then the max value will be chosen as the correct output. It depends on the value class that can be identified.



### 2.7.5 Fine Tunning

Due to the portability of their internal deep representations, pretrained CNNs are advantageous for a range of visual recognition applications. In order to fine-tune fresh workloads, CNNs can serve either as feature extractors or weight initializers depending on the application. Instead of utilizing the pre-trained network as a feature extractor, which may miss some of the discriminative information of the new dataset, performance is typically enhanced by fine-tuning the pre-trained network. As the network progresses, it extracts features that correlate closer to the particular visual aspects of classes included in the initial dataset, whereas earlier layers extract general attributes like edges or blobs. Common fine-tuning approaches include copying the whole pre-trained model exclude for the last layer (which is unique to the source classification job) and replacing it with a new layer whose number of neurons matches to the number of classes in the new target domain. Due to the fact that early layers extract features that are relevant to a variety of image recognition tasks, fine-tuning only a portion of the network, typically the last or last few layers, enables the network to adapt to the specifics of the target domain, which results in a performance boost for a variety of classification problems. This is accomplished by the fact that early layers extract features that are relevant to image recognition tasks.

Several fine-tuning scenarios are considered based on the target dataset size and the relationship between the target and source domains in order to discover the most efficient solution and prevent over-fitting. In each of these cases, the number of transferred layers that are allowed to frozen away once the new job's error has been propagated back through the network is altered. In this study, experiment with five different cases [25]:

- all - the weights of all layers are adjusted with each iteration.
- When using "skip first," the weights in the first convolutional layer (conv1) are not updated from their original values.
- the first two convolutional layers (conv1 and conv2) are skipped, but their weights are kept.

- just last 3 - only the weights of the final three completely linked layers (fc6, fc7, and fc8) are altered [25].

### **2.7.6 Provide ideal suggestions according to the real time situation**

The main objective for developing the component for this study is override features, which are found in the voice assistants on the market. Many integrations have been done to provide ideal suggestions based on the users' real-time situations. To provide personalized suggestions according to the users' preferences, during the registration, users' hobbies and interests are stored in the database. This is done as an initial step to the ideal suggestion, providing part. Then the application depends on the emotions of the users. To work with the FER model, the Restful API has been used. This suggestion-providing module interacts with the FER model. When given an image to analyze, the FER model uses CNN, which was talked about and explained in the section before this one. Finally, the FER model returns the emotion probability values. Whether it is happy, sad, angry, or neutral, based on the outputted emotion class and the combination of previously added hobbies and interests and the real-time situation, suggestions are provided by a rule-based mechanism to the user accordingly. The suggestion will be different when considering the emotional level as a percentage. These suggestions will be sent to the front end via the Restful API. The user is able to see all the suggestions with a nice-looking user interface. Based on their preferences and emotions in a real-time situation, users are free to choose the suggestions. Using these suggestions, they will be able to fix or relax their minds.

For instance, if a user is unhappy, the application will identify the situation from the FER model, which is powered by CNN. Then, with the combination of hobbies and interests, if a user likes to listen to relaxing music, the application suggests "Spotify වලින් mind relaxing music on අහමුද?", "YouTube වලින් mind relaxing music on අහමුද" The suggestions will vary according to the real-time situation and the emotion level of the user(angry-50-80%).

## 2.8 Technologies Used

When developing the solution for the application, to design prototypes and low fidelity prototypes Figma and Adobe XD has been used. To develop client application Popular JavaScript framework called React Native has been used. Server-side development has been done using Python with Flask framework. AWS S3 Bucket has been used as a database for the application

### 2.8.1 Low Fidelity Prototyping – Figma, Adobe XD

Before implementing the application, there should be a specific theme for the application. There should be a user flow and the application should produce a good user experience for the user with minimalistic design. User research techniques have been followed to design the UI of the application. To develop the UI and the user flow, popular tools like Figma and Adobe XD have been used.

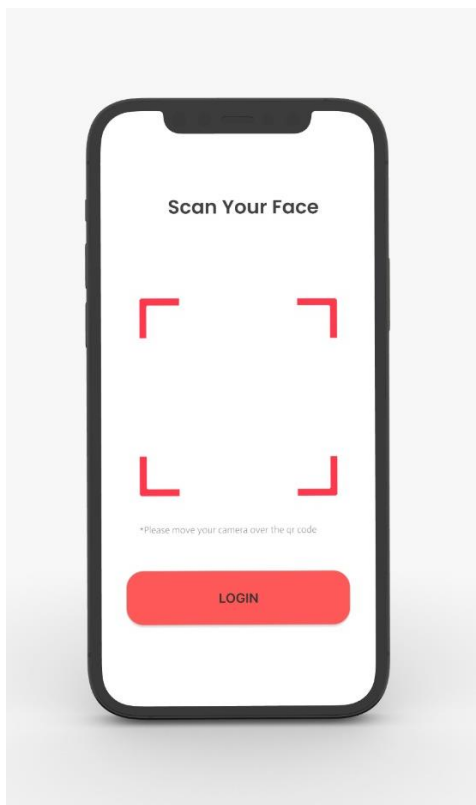


Figure 16 – Wireframe Design 1

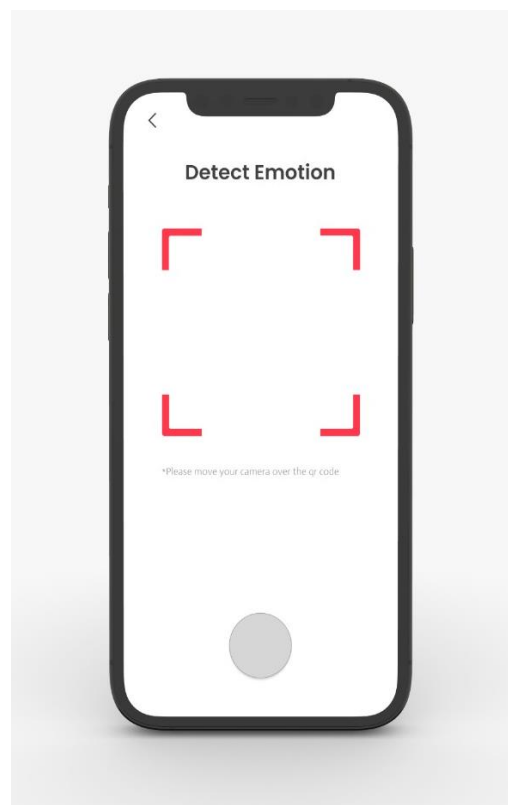


Figure 17 – Wireframe Design 2



Figure 18 – Wireframe Design 3



Figure 19 – Wireframe Design 4

## **2.9 Commercialization aspects of the Product**

As the users are unaware of the "Sinhala Voice Assistant" mobile application, which was developed to automate daily operations using the Sinhala language, many users are unaware of the emotion identification procedures and how they positively impact their real-time emotions or sensations. The research team has various commercialization ideas for the product.

“In essence, a web application will be created to provide consumers with a preview of the Sri Voice mobile application. This online application will allow users to discover the apps' functionalities. How-to guides for the application Each new feature and version release is made available via the online application. And this online application was designed with a very straightforward and attractive interface. All information will be made available on the official web application. The web application will receive traffic from Facebook, Instagram, and even WhatsApp groups. The download link for the Sri Voice Application will be provided on the official website. Therefore, visitors to the online application will be able to download and utilize the program.

The “Sri Voice” Voice Assistant application features an intuitive and clean layout. Therefore, any anyone having a smart phone will be able to comprehend the application flow. If necessary, the instruction is also available within the official online application. in addition to the online application There are already an plenty of social media platforms available on the Internet. Facebook Pages, Instagram Pages, and Twitter accounts are being built in order to market the application. Using these unique media outlets, attractive adverts and brochures are created to promote the Sri Voice application among the public



Figure 20 : Commercialization

## **2.10 TESTING & IMPLEMENTATION**

### **2.10.1 Testing**

This stage involves testing each unit or element of the Sri Voice system. The main goal of this phase is to determine whether the system is functioning in accordance with client requirements or to address any bugs. This phase includes various testing levels that make it easier to evaluate the behavior and functionality of the software. Individual components will be tested during unit testing. The integrated components will also be tested, and it will be seen if data flows from one model to another during the integration testing process. Therefore, the entire system will be tested in the system testing phase to see if it complies with the requirements. Finally, acceptance testing, which is primarily carried out by users or clients, will evaluate the final system. Debugging is the process of finding and fixing specific implementation bugs or errors in a program or system. The testing phase of Sri Voice application is done according to the components. In this component, the accuracy of the created Sinhala text and the accuracy of the translated English phrase was the main testing parts.

Other than that, the front end was tested to see if the functionalities are working according to the plan. After the entire system was tested the system will be going to the evaluation (maintenance) phase.

## **2.11 Implementation**

### **2.11.1 Agile Methodology**

Agile Methodology When developing the solution for the application Agile software development methodology has been used. The Agile methodology is a method of project management that involves segmenting the whole process into distinct phases. It is important that constant communication with stakeholders take place, as well as continuous improvement at each level. After getting started on the task at hand, teams go through iterations of the planning, executing, and evaluating processes. This is an ongoing engagement. To develop the solution in this study, initially plan the steps and that need to flow. Design and develop a flow for the solution. After developing the flow steps has been divided in to sprints. All the time tried to complete the work of the sprint on time. To manage this project, project management tool called Trello has been used. In there briefly can get overall understanding. The tasks can be labeled in the Trello. While doing the task, simple it can move to doing section. If it is done, simply card can be moved to done section and to test. After the sprint, the next sprint can be started by planning. Then it goes through the design phase, implementation phase, testing and integration phase, and maintenance phase. This cycle happens repeatedly.



### **3 RESULTS & DISCUSSION**

#### **3.1 Results**

The Sri Voice mobile application has been designed to be used by a person of any knowledge level, providing a new experience to Sri Lankans. This section explains the findings and conclusions of the Sri Voice virtual assistant application.

This application has been developed to allow users to automate their daily tasks using Sinhala voice commands and to increase their mental freedom. The mobile application is implemented for the Android platform using React Native, Python with Flask and an AWS S3 database. The research has used machine learning concepts and algorithms, image processing algorithms, and natural language processing techniques in the implementation. There is no sophisticated voice assistance option for Sinhala speaking Sri Lankans, but many foreigners use voice assistant software in their native language. This program is available to all users, regardless of English proficiency.

Various assumptions were considered during the development of the application's security layer, including the MediaPipe method, the TensorFlow model for facial recognition, and OpenCV for MATLAB implementation. In this instance, the Haar-Cascades had difficulty accurately detecting the facial region, and it took a considerable amount of time. MediaPipe was used to fix these issues, thereby enhancing the application's speed and security.

In addition, various concerns have been raised when displaying Sinhala texts linked with search results while doing an online search. Minor adjustments were made to various components during implementation. A CNN model has been trained in the face expression detection component.

For the development of the chat assistant, they decided to use distilBERT. But there it is most suitable for the dataset of a chatbot. Therefore, it was decided to use an intentbased dataset instead of one. because it allows us to use this data as required by

the voice assistant application. In the end, the expected results were achieved. And its precision could be improved over time.

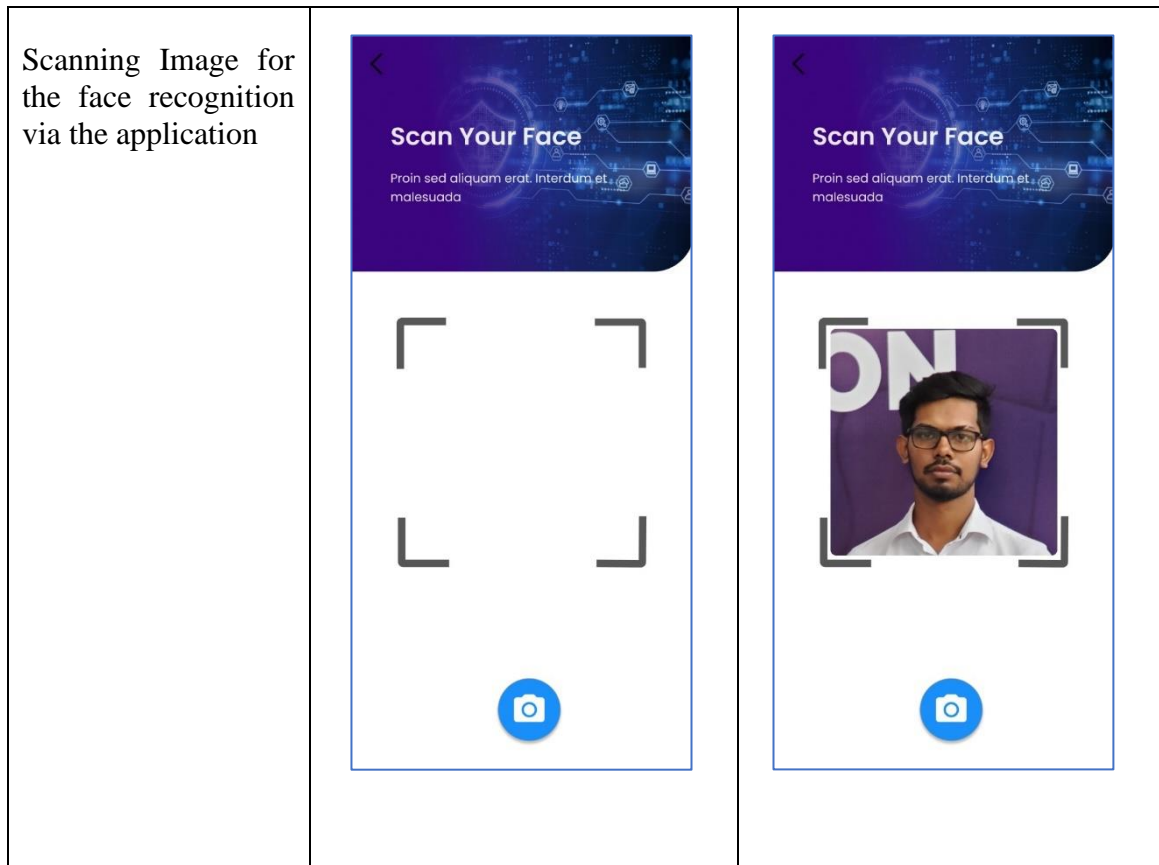
In addition, the suggested system would be useful for distinguishing between users when they are given Sinhala voice commands without being affected by background noise. The application will get pertinent input for the system and filter out background noise.

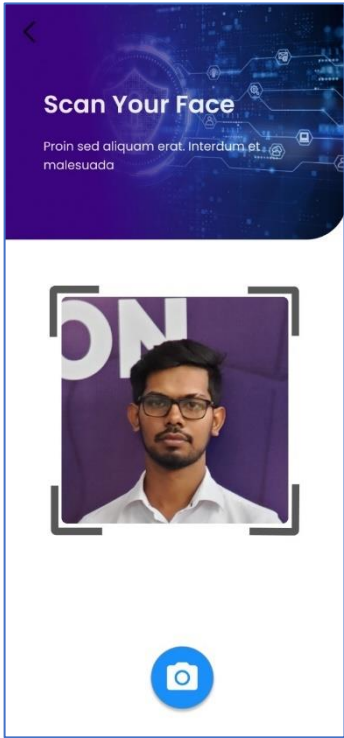
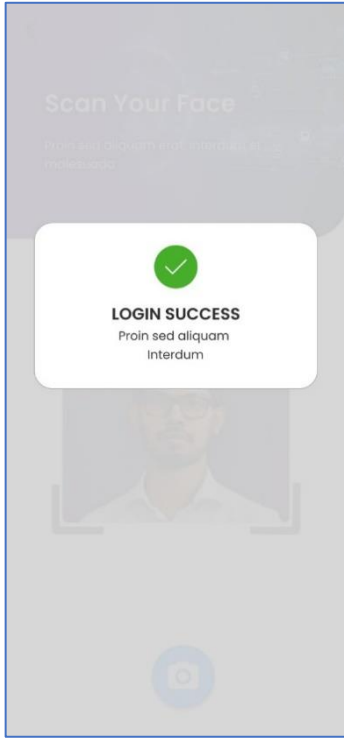


In essence, the security layer would be the user's gateway to the application. Users' personal information is primarily gathered in the voice assistant application. In order to increase application security, facial recognition technology was used. Users can use voice commands to get the application to meet 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. According to the needs of the user, this mechanism produced the best output.



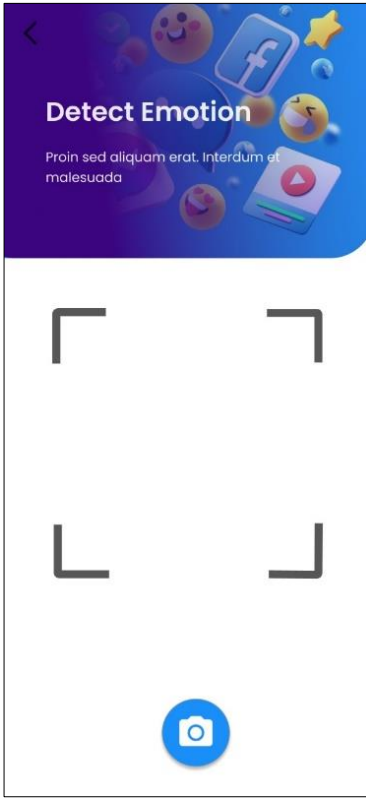
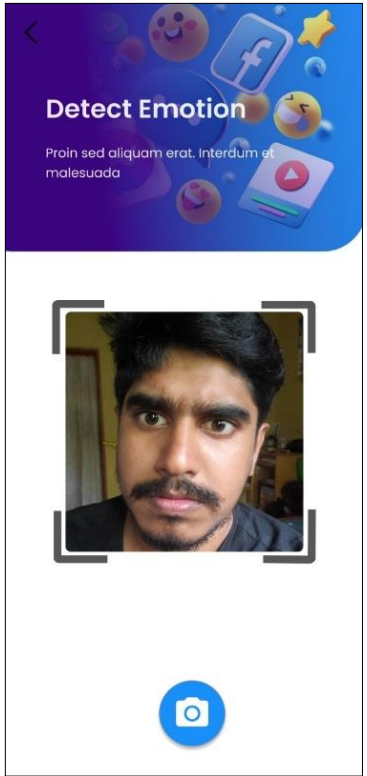
The researchers ran into a variety of issues as they implemented each research component, and they came up with and put into practice a number of solutions to fix them. The main problems were mismatching facial expressions, reducing background noise, finding a good dataset, and using less accurate voice detection libraries and text matching algorithms to find keywords.

Using the FER model, which was developed using CNN, the FER2013 dataset contains 28,709 facial expressions of the user that have been inputted and tested. Then, after performing the fine tuning of the Face-Net using cosine distance, the model's performance has been increased. After training the model, testing images from the FER2013 dataset have been used. A locally created dataset has been input since the main purpose is to analyze the expressions on the users' faces using a testing dataset. By inputting images to the model, four emotion classes have been identified, which are: happy, sad, angry, and neutral. By using the analyzed expressions from the inputted image, the maximum probability has been chosen to predict the emotion class. This scenario is again done by using real-time images with a colored background.

The goal of the research study is to simulate a helpful mobile voice assistant application that helps in various situations in life. The main goal has been accomplished through the combination of all the individual components. Users will be able to automate their tasks and raise their emotional level based on the real-time situation in their lives by using the "Voice" virtual assistant application.



<p>Authenticate the user face image by analyzing the face data and successful login</p>		
<p>according to user Sinhala voice command, the application gives replies in text and voice output.</p>		

<p>Also, the question about the date will be answered.</p>		
<p>Image Capturing to detect emotions via application</p>		

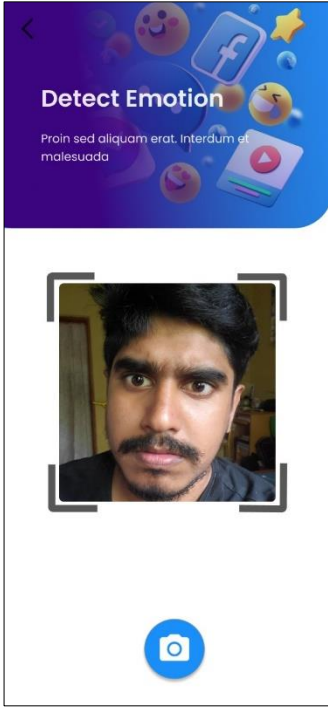

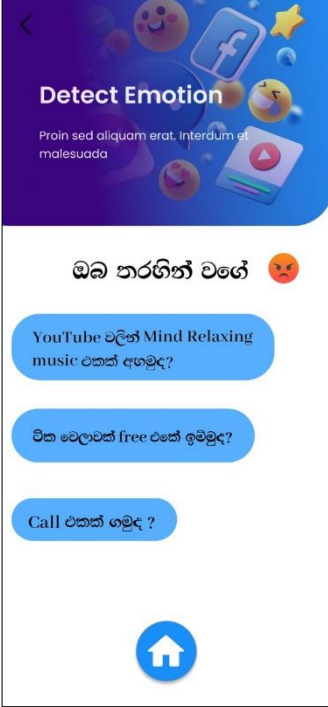
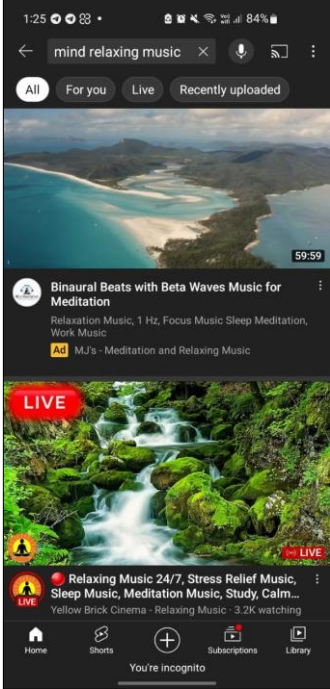
<p>Identify the emotions from the captured image by analyzing facial expressions</p>		
<p>Provide the Sinhala suggestions based on the identified emotion class, level and the preference</p>		

Table 3 : Results

### 3.2 Research Findings

According to the chat assistant, the main thing that is done here is to provide the correct answer that matches the command received from the user. There, an intent-based dataset was created to do it. The intent is the purpose the chatbot user has in mind while entering a query or remark (known as an utterance) inside the discussion flow.

Intent indicates the desired goal or activity of the chatbot user. The basic function of an entity is to obtain the data from the speech in order to improve the intent prediction. There are "Tags", "Patterns", and "Responses" in this section. It is trained using a model of machine learning. In this manner, the most suitable answer is determined based on the user's instruction. Here, even for the same question, the pertinent response might be provided in a variety of ways. Typically, between 20 and 30 patterns are contained in a tag, which consists of between 10 primary tags. The inclusion of so many patterns is intended to facilitate the user's response to any instruction. Additionally, many replies are incorporated in a tag. Because it provides the same response in several formats. It's also performed via an algorithm for machine learning.

The English language obtained from the user's instruction is processed by a neural machine translation (NMT) model, and its result is read in Sinhala using a Text-to-speech API. Using the Intents base dataset, the pattern associated with the user command is discovered. This is taught using a model of machine learning. It may respond appropriately to the user's demand. Here, two operations simultaneously use the English term and involve the user. The first step is to present the dataset's response to the inquiry. The neural machine translation model then adapts the English sentence to the Sinhala language. Then, Google Cloud Text to speech API reads the text.

During the application's investigation, several new discoveries were made. When it comes to the Facial Emotion Recognition (FER) model, the dataset will affect both model performance and accuracy. This occurs because of the facial look of foreign people. In comparison with Sri Lankans, their skin color and tone are distinct. The hair, facial hair, and eyebrows of foreigners differ from those of Sri Lankans. Due to this issue, the FER model does not always accurately recognize the face expression.

Therefore, it is trained using a dataset of Sri Lankan people's facial expressions. During this investigation, this issue was detected, and it has been resolved by the use of cosine distance to the FER model's fine tuning.

The main purpose of the facial recognition section was to identify the accurate face of the user and authenticate and provide easy access to the application while improving security. First, the team needs to develop a machine learning algorithm to train the face model to accurately capture the face data of the users. Therefore, here the team used the Face Net model to get a more accurate output, and again, the team used a local dataset and fine-tuned the data set using the rest net backbone of the CNN model. After attempting to use the OpenCV, media pipe, and TensorFlow models, the team concludes that the Face Net model is the best way to go. Then fine tuning to this model and completing the research. The above-mentioned attributes taken from the trained model.



### **3.3 Discussion**

This report is a mobile application for users to manage their daily responsibilities and enhance their entertainment. In the modern world, people are too busy and most of them are mentally down. because of this busy lifestyle. Our team intended for our application to provide some form of relaxation to people who are mentally stressed. This application has a Sinhala voice command feature with an emotion suggestion feature, and this feature gets the users' emotions and helps them relax. It also improves biometrics with facial recognition and an enhanced Sinhala voice command system and chat assistant. When logging in to the app, it detects the user's face and allows them to continue to the application. Here the team used international and local datasets so the users could get the fastest and most accurate login to the system. And this app has features that collect users' hobbies and interests, like things. And therefore, the team used facial recognition for this, so the users can add their favorites without any fear. Using the chat assistant features, users will be able to ask questions and see the response via the app.

## 4 SUMMARY OF THE STUDENT CONTRIBUTION

Personal	Functionality	Description
Senarathne K.H.I.R. IT19152592	Analyze users' emotions via facial expressions and based on emotion level and preferences, provide ideal suggestions using Sinhala language to uplift their mental freedom.	<p>Implement a FER model using OpenCV and Keras.</p> <p>Implement a FER model using TensorFlow and OpenCV.</p> <p>Implement a FER model using the CNN model.</p> <p>Identify emotion level as a percentage and predict of occurrence.</p> <p>Fine tune the implemented FER model to improve performance and accuracy of classification.</p> <p>Create a dataset of facial expressions from Sri Lankans.</p> <p>Fine tuning the FER model using Cosine distance</p>

<p>Nirash J.M.I IT19159768</p>	<p>Providing accurate voice &amp; outputs according to the user's commands.</p>	<p>Identify users' preferences and emotional level in each situation.</p> <p>Develop mechanism using rule-based technique to provide Ideal suggestions to the user.</p> <p>To interact with the React Native front-end application, develop RESTful APIs using python with Flask framework.</p> <p>Creating the intents-based dataset.</p> <p>Creating a model to provide the best outputs to the user's command.</p> <p>Using a neural machine translation existing model to accurately prepare the relevant words before performing the test to speech process.</p> <p>To perform the Google text- to-speech process, prepare the google development profile and enable the text- to-speech</p>
------------------------------------	---	---

<p>Herath H.M.C.P. IT19216010</p>	<p>Identify accurate face of the user and authenticate and provide easy access to the application while improve the security</p>	<p>API.</p> <p>To open the required app according to the users' request, the application needs to open the apps in an effective manner. To achieve this, the Flask RESTful API has been used to integrate with the React Native frontend. Using this API, necessary information is being passed accordingly.</p> <p>Gather users face images to create the local face dataset.</p> <p>Create a model to train the face data set.</p> <p>Fine-tuned the local data set to the rest net model.</p> <p>Trained models to increase the accuracy of the dataset.</p> <p>To connect front end with the backend of create and used the flask RESTful API.</p>
---------------------------------------	--	--

<p>Bandara V.D. IT19808994</p>	<p>Voice – to – text conversion of user’s initial voice command and providing accurate English phrases to the created Sinhala Text</p>	<p>Implementing the Dataset for T5 model</p> <p>Creating the T5 model using the Attention mechanism</p> <p>Training T5 model</p> <p>Training the T5 to increase the accuracy of the dataset.</p> <p>Using a neural machine translation existing model to accurately prepare the relevant words before performing the speech - to-text process.</p> <p>To perform the Google speech-to-text process, prepare the google development profile and enable the speech-to-text API.</p> <p>Implementing Google speech-to-text API using Python libraries.</p>
------------------------------------	--	---

## 5 CONCLUSION

voice mobile application is software that can automate the traditional usage of the application by providing the commands to the application process using modern 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 system extracts and converts human language into system understandable text-based inputs. The voice mobile application is developed using Python as the programming language alongside Flutter as the frontend framework. This is the research paper on the customized mobile application of the voice for automating daily tasks of the users while uplifting the users' mental health by providing ideal suggestions. This research paper consists of 5 main chapters, and mostly the Back-End development process of the voice mobile application is discussed throughout the paper. Overall, all the system's objectives are successfully achieved, and the proposed system is up and running. After briefly analyzing the evaluations done, the conclusion determines that the overall project of the voice mobile application has been 100% successful up to now.

## 6 REFERENCES

- [1] 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>. [Accessed 19 January 2022]
- [2] 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," IEEE, 19 February 2018. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8294116>. [Accessed 25 December 2022].
- [3] N. Prasangini and H. Nagahamulla, "Sinhala Speech to Sinhala Unicode Text Conversion for Disaster Relief Facilitation in Sri Lanka," IEEE, 28 November 2019. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8913360>. [Accessed 21 January 2022].
- [4] 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>. [Accessed 17 January 2022].
- [5] A. Shehan, Lanka Sinhala Personal Assistant, Ayesh Shehan, 2020.
- [6] P. M. Dias and K. Jayakody, "Virtual Assistant in Native Language," IEEE, 22 June 2021. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9452751>. [Accessed 23 January 2022].

- [7] M. Amarasekara, K. Bandara, B. Vithana, D. D. Silva and A. Jayakody, "Realtime interactive voice communication - For a mute person in Sinhala (RTIVC)," IEEE, 15 July 2013. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/6553993>. [Accessed 22 February 2022]
- [8] Statista. 2022. *Number of voice assistants in use worldwide 2019-2024* / Statista.[online]Available at: <https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/>
- [9] Statista. 2022. *Number of voice assistants in use worldwide 2019-2024* / Statista.[online]Available at: <https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/>
- [10] H. M. Kabir, F. Ahmed, "Face recognition with directional ternary pattern (DTP)," In 2012 International Conference on Graphic and Image Processing, 2013
- [11] Th. Prasartvit, B. Kaewkamnerdpong, T. Achalakul, "Dimensional reduction based on artificial bee colony for classification problem," In Bio-Inspired Computing and Applications, pp. 168-175, 2012.
- [12] S.Beniwal, J. Arora, "Classification and feature selection techniques in data mining," international journal of engineering research & technology (ijert) vol. 1, no. 6, pp. 1-6, 2012.
- [13] P.Ekman and W.V. Friesen, "Constants across cultures in the face and emotion..pdf." p. 124,129, 1980. [Online]. Available: [http://www.communicationcache.com/uploads/1/0/8/8/10887248/constants\\_across\\_cultures\\_in\\_the\\_face\\_and\\_emotion.pdf](http://www.communicationcache.com/uploads/1/0/8/8/10887248/constants_across_cultures_in_the_face_and_emotion.pdf)
- [14] H. Min, X. Yanxia, W. Xiaohua, H. Zhong, and Z. Hong, "Facial expression



- recognition based on AWCLBP,” vol. 01005, pp. 1279–1284, 2013, [Online]. Available: [https://www.itm-conferences.org/articles/itmconf/pdf/2017/04/itmconf\\_ita2017\\_01005.pdf](https://www.itm-conferences.org/articles/itmconf/pdf/2017/04/itmconf_ita2017_01005.pdf)
- [15] G. Erkan and D. R. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," in *Journal of Artificial Intelligence Research*, Vol 22, 2004
- [16] TensorFlow, “TensorFlow,” *TensorFlow*, 2019. <https://www.tensorflow.org/>
- [17] Xia Mao, Yu-Li Xue, Zheng Li, Kang Huang, ShanWei Lv: Robust facial expression recognition based on RPCA and AdaBoost. *WIAMIS 2009*: 113-116
- [18] I. Buciu and C. Kotropoulos, “ICA AND GABOR REPRESENTATION FOR FACIAL EXPRESSION RECOGNITION and I . Pitas Department of Informatics , Aristotle University ofThessaloniki,” *Human-Computer Interact.*, pp. 8–11, 2003.
- [19] 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.
- [20] Ioan Buciu, Constantine Kotropoulos, Ioannis Pitas: ICA and Gabor representation for facial expression recognition. *ICIP (2) 2003*: 855-858
- [21] 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.
- [22] 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 F147615, pp. 150–159, 2019, doi: 10.1145/3301275.3302266.

- [23] 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.
- [24] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, "Deep learning approaches for facial emotion recognition: A case study on FER-2013," *Smart Innov. Syst. Technol.*, vol. 85, pp. 1–16, 2018, doi: 10.1007/978-3-319-66790-4\_1.
- [25] E. Cetinic, T. Lipic, and S. Grgic, "Fine-tuning Convolutional Neural Networks for fine art classification," *Expert Syst. Appl.*, vol. 114, pp. 107–118, 2018, doi: 10.1016/j.eswa.2018.07.026.
- [26] 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.
- [27] X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," *2017 2nd Int. Conf. Image, Vis. Comput. ICIVC 2017*, pp. 783–787, 2017, doi: 10.1109/ICIVC.2017.7984661.
- [28] H. Min, X. Yanxia, W. Xiaohua, H. Zhong, and Z. Hong, "Facial expression recognition based on AWCLBP," vol. 01005, pp. 1279–1284, 2013, [Online]. Available: [https://www.itm-conferences.org/articles/itmconf/pdf/2017/04/itmconf\\_ita2017\\_01005.pdf](https://www.itm-conferences.org/articles/itmconf/pdf/2017/04/itmconf_ita2017_01005.pdf)
- [29] TensorFlow, "TensorFlow," *TensorFlow*, 2019. <https://www.tensorflow.org/>
- [30] V. Tatan, "Understanding CNN (Convolutional Neural Network)," *Medium*, Dec. 23, 2019. <https://towardsdatascience.com/understanding-cnn-convolutional-neural-network-69fd626ee7d4>
- [31] Xia Mao, Yu-Li Xue, Zheng Li, Kang Huang, ShanWei Lv: Robust facial expression recognition based on RPCA and AdaBoost. *WIAMIS 2009*: 113-116

- [32] I. Buciu and C. Kotropoulos, "ICA AND GABOR REPRESENTATION FOR FACIAL EXPRESSION RECOGNITION and I . Pitas Department of Informatics , Aristotle University of Thessaloniki," *Human-Computer Interact.*, pp. 8–11, 2003.
- [33] 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.
- [34] Ioan Buciu, Constantine Kotropoulos, Ioannis Pitas: ICA and Gabor representation for facial expression recognition. *ICIP (2) 2003*: 855-858
- [35] Kobayashi H, Hara F: The recognition of basic facial expression by neutral network. [C], *International Joint Conference on Neural Network*. 1991:460-466.
- [36] O. Çeliktutan, S. Ulukaya, and B. Sankur, "A comparative study of face landmarking techniques," *Eurasip J. Image Video Process.*, vol. 2013, no. 1, pp. 1–27, 2013, doi: 10.1186/1687-5281-2013-13.
- [37] M. Lin, C. B. R. Diller, Ming, N. Forsgren, Y. Huang and J. F. Nunamaker, Jr, "Segmenting Lecture Videos by Topic: From Manual to Automated Methods," in *Eleventh Americas Conference on Information Systems*, Omaha, NE, 2005.
- [38] H. J. Zhang and S. W. Smoliar, "Developing power tools for video indexing and retrieval," in *SPIE'94 Storage and Retrieval for Video Databases*, San Jose, CA, USA, 1994.
- [39] M. Chau, M. Lin, J. F. Nunamaker and H. Chen, "Segmentation of Lecture Videos Based on Text: A Method Combining Multiple Linguistic Features," in *Proceedings of the 37th Hawaii International Conference on System Sciences*, Hawaii, 2004.
- [40] X. Che, H. Yang and C. Meinel, "Lecture Video Segmentation by Automatically Analyzing the Synchronized Slides," in *MM'13*, Barcelona, Spain, 2013.

- [41] J. Allan, J. Carbonel, G. Doddington, J. Yamron and Y. Yang, "Topic detection and tracking pilot study: Final report," in Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.
- [42] R. Shah, D. Shah and L. Kurup, "Automatic Question Generation for Intelligent Tutoring Systems," in 2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA), Mumbai, 2017.
- [43] Y.-T. Huang, Y.-M. Tseng, Y. S. Sun and M. C. Chen, "TEDQuiz: Automatic Quiz Generation for TED Talks Video Clips to Assess Listening Comprehension," in 2014 IEEE 14th International Conference on Advanced Learning Technologies, Athens, 2014.
- [44] G. Erkan and D. R. Radev, "LexRank: Graph-based Lexical Centrality as Saliency in Text Summarization," in Journal of Artificial Intelligence Research, Vol 22, 2004.
- [45] M. Heilman and N. A. Smith, "Question Generation via Overgenerating Transformations and Ranking," in Technical report, Language Technologies Institute, Carnegie Mellon University Technical Report, 2009.
- [46] M. Heilman and N. A. Smith, "GoodQuestion! Statistical Ranking for Question Generation," in HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Log Angeles, California, 2010.
- [47] A. Krishna, P. Bhowmick, K. Ghosh, A. Sahu and S. Roy, "Automatic Generation and Insertion of Assessment Items in Online Video Courses," in IUI Companion '15 Proceedings of the 20th International Conference on Intelligent User Interfaces Companion, Atlanta, Georgia, 2015.
- [48] M. Heilman, "Automatic Factual Question Generation from Text," Doctoral thesis, Carnegie Mellon University, Pittsburgh, 2011.

- [49] M. Tasic and M. Cubric, "SeMCQ – Protégé Plugin for Automatic OntologyDriven Multiple Choice Question Tests Generation," in Procs of the 11th International Protege Conference, 2009.
- [50] M. Al-Yahya, "OntoQue: A Question Generation Engine for Educational Assesment Based on Domain Ontologies," in 2011 IEEE 11th International Conference on Advanced Learning Technologies, Athens, 2011.
- [51] A. Papasalouros , K. Kanaris and K. Kotis, "Automatic Generation Of Multiple Choice Questions From Domain Ontologies," in IADIS International Conference e-Learning 2008, Amsterdam, 2008.
- [52] M. Maher, H. Lipford and V. Singh, "Flipped classroom strategies using online videos," UNC Charlotte, North Carolina, 2013.
- [53] M. J. Rubin, "The effectiveness of live-coding to teach introductory programming," in Proceeding of the 44th ACM technical symposium on Computer science education - SIGCSE '13, Denver, Colorado, USA, 2013.
- [54] "Importance and Effectiveness of E-learning," 1 December 2015. [Online]. Available: <https://higheredrevolution.com/importance-and-effectiveness-of-e-learning-9513046ed46c>. [Accessed 6 March 2019]
- [55] Z. Woolfitt, "The effective use of video in higher education," 2015. [Online]. Available: <https://www.inholland.nl/media/10230/the-effective-use-of-video-in-higher-education-woolfitt-october-2015.pdf>. [Accessed 6 March 2019]
- [55] 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.
- [56] I. William, D. R. Ignatius Moses Setiadi, E. H. Rachmawanto, H. A. Santoso, and C. A. Sari, "Face Recognition using FaceNet (Survey, Performance Test,

and Comparison),” *Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019*, 2019, doi: 10.1109/ICIC47613.2019.8985786.

- [57] S. Saleem, J. Shiney, B. Priestly Shan, and V. Kumar Mishra, “Face recognition using facial features,” *Mater. Today Proc.*, no. xxxx, 2022, doi: 10.1016/j.matpr.2021.07.402.
- [58] "Speech to Text in Python with Deep Learning", Analytics Vidhya, 2022. [Online]. Available: <https://www.analyticsvidhya.com/blog/2021/09/okgoogle-speech-to-text-in-python-with-deep-learning-in-2-minutes/>.
- [59] G. Tang, R. Sennrich, and J. Nivre, “An Analysis of Attention Mechanisms: The Case of Word Sense Disambiguation in Neural Machine Translation,” *WMT 2018 - 3rd Conf. Mach. Transl. Proc. Conf.*, vol. 1, pp. 26–35, 2018, doi: 10.18653/v1/w18-6304.

## 7 APPENDIX

### 7.1 Gantt Chart



Figure 21 : Gantt Chart

## 7.2 Similarity Index

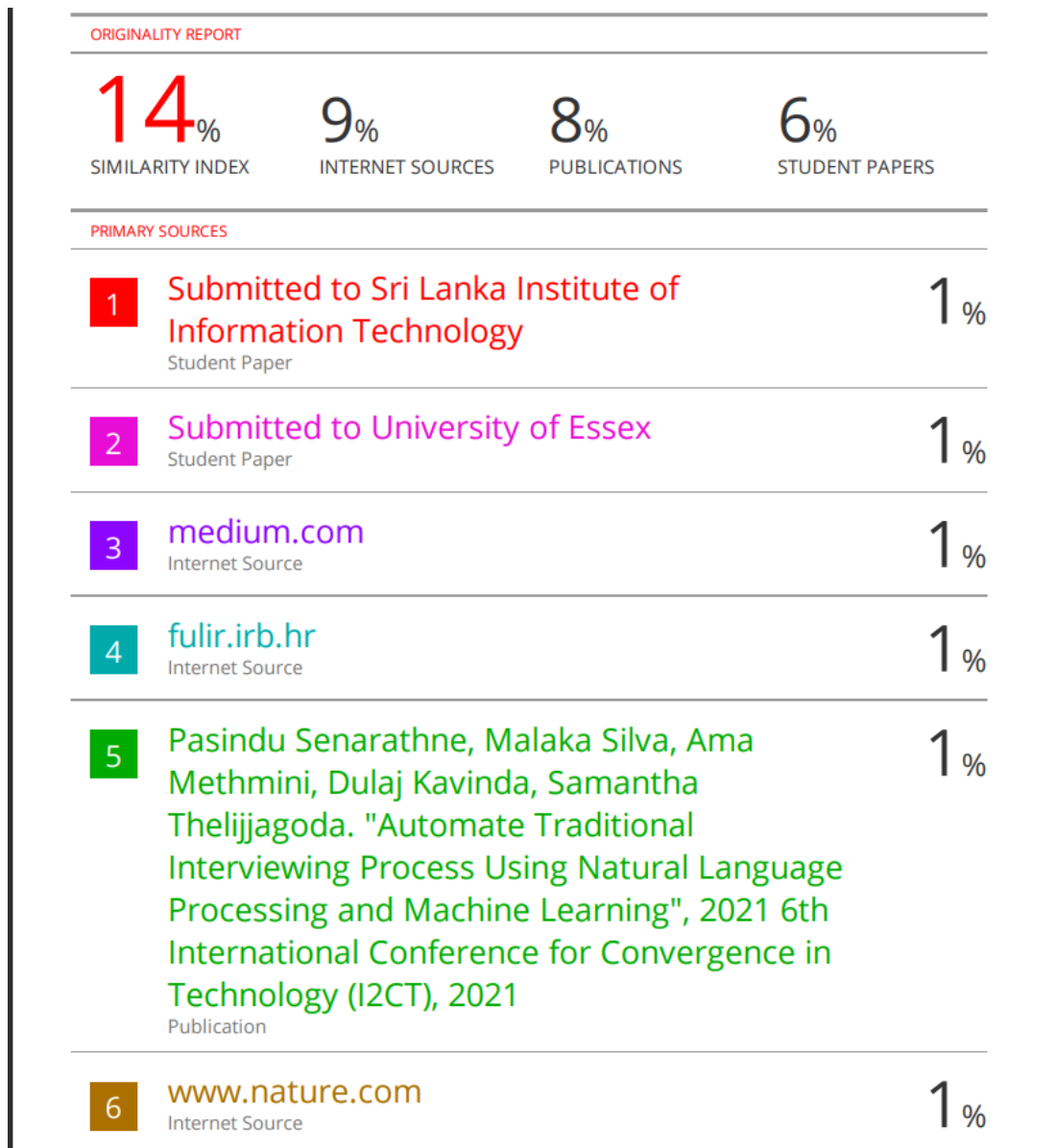


Figure 22 : Similarity Index